Characterizing Variation of Protein Complexes and Functional Modules on a Temporal Scale and across Individuals

Dissertation zur Erlangung des

naturwissenschaftlichen Doktorgrades

der Julius-Maximilians-Universität Würzburg

vorgelegt von

NATALIE ROMANOV

(Geburtsort: Wien, 1990)

Würzburg, 2018

Eingereicht am: 27.Juni.2018

Mitglieder der Promotionskommission:

Vorsitzender: ……..

Gutachter: Prof. Dr. Thomas Dandekar

Gutachter: Prof. Dr. Peer Bork

Tag des Promotionskolloquiums: ……..…

Doktorurkunde ausgehändigt am: …….

Name, Vorname: **Romanov, Natalie**

Straße:

PLZ und Ort:

Tel.

E-Mail:

Eidesstattliche Erklärungen nach §4 Abs. 3 Satz 3, 5, 8 der Promotionsordnung der Fakultät für Biologie

Affidavit

I hereby declare that my thesis entitled: „ **Characterizing variation of protein complexes and functional modules on a temporal scale and across individuals**" is the result of my own work.

I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore I verify that the thesis has not been submitted as part of another examination process neither in identical nor in similar form.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation: „**Characterizing variation of protein complexes and functional modules on a temporal scale and across individuals**", eigenständig, d. h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen, als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Würzburg, den 26.06.2018

Signature PhD-student

Abstract

A fundamental question in current biology concerns the translational mechanisms leading from genetic variability to phenotypes. Technologies have evolved to the extent that they can efficiently and economically determine an individual's genomic composition, while at the same time big data on clinical profiles and diagnostics have substantially accumulated. Genome-wide association studies linking genomic loci to certain traits, however, remain limited in their capacity to explain the cellular mechanisms that underlie the given association. For most associations, gene expression has been blamed; yet given that transcript and protein abundance oftentimes do not correlate, that finding does not necessarily decrypt the underlying mechanism. Thus, the integration of further information is crucial to establish a model that could prove more accurate in predicting genotypic effects on the human organism.

In this work we describe the so-called **proteotype** as a feature of the cell that could provide a substantial link between genotype and phenotype. Rather than looking at the proteome as a set of independent molecules, we demonstrate a consistent **modular architecture** of the proteome that is driven by molecular cooperativity. Functional modules, especially protein complexes, can be further interrogated for differences between individuals and tackled as imprints of **genetic and environmental variability**. We also show that subtle stoichiometric changes of protein modules could have broader effects on the cellular system, such as the transport of specific molecular cargos.

The presented work also delineates to what extent **temporal events and processes** influence the **stoichiometry** of protein complexes and functional modules. The re-wiring of the glycolytic pathway for example is illustrated as a potential cause for an increased Warburg effect during the **ageing** of the human bone marrow. On top of analyzing protein abundances we also interrogate proteome dynamics in terms of **stability** and **solubility** transitions during the short temporal progression of the **cell cycle**. One of our main observations in the thesis encompass the delineation of protein complexes into respective

sub-complexes according to distinct stability patterns during the cell cycle. This has never been demonstrated before, and is functionally relevant for our understanding of the dis- and assembly of large protein modules.

The insights presented in this work imply that the proteome is more than the sum of its parts, and primarily driven by variability in entire protein ensembles and their cooperative nature. Analyzing protein complexes and functional modules as molecular reflections of genetic and environmental variations could indeed prove to be a stepping stone in closing the gap between genotype and phenotype and customizing clinical treatments in the future.

Zusammenfassung

Eine fundamentale Frage in der heutigen biologischen Forschung ist durch welche Mechanismen eine gebenene genetische Variation sich in einem Phänotyp äußert. Etliche Technologien können heutzutage effizient und ökonomisch die genomische Komposition eines Individuals mit beispielloser Genaugikeit aufschlüsseln. Gleichzeitig gibt es wesentliche Erfolge und Bemühungen, große Datenmengen von Patienten zu sammeln, sowohl klinische Profile, als auch Diagnosen. Es gibt bereits mehrere genomweite Assoziationsstudien, die auf spezifische genomische Loci hinweisen, die womöglich einem bestimmenten phänotypischen Merkmalen zugrunde liegen. Obwohl für die meisten genetischen Assoziationen, eine veränderte Genexpression oftmals als Ursache diskutiert wird, ist dies wahrscheinlich nur ein Teil des zugrundeliegenden Mechanismus. Wir können dies annehmen, da RNA-Transkripte nicht unbedingt mit ihrem Protein-Produkt korrelieren aufgrund von post-transkriptioneller und translationeller Regulation. Um dementsprechend ein Modell zu etablieren, das die genotypischen Effekte auf den human Organismus akkurat vorhersagen kann, ist eine Integration von mehreren zellulären Informationsschichten notwendig.

In der folgenden Arbeit beschreiben wir den sogenannten **Proteotyp** als ein zelluläres Merkmal, das eine substanzielle Verknüpfung zwischen dem Genotyp und dem Phänotyp eines Individuums schaffen könnte. Statt das Proteom als ein Set unabhängiger Moleküle zu betrachten, zeigen wir eine konsistent **moduläre Architektur** des Proteoms auf, das durch die molekulare Kooperativität zustande kommt. Funktionelle Module, v.a. Proteinkomplexe, können weiters auf Unterschiede zwischen Individuen untersucht werden, sowie deren Variabilität aufgrund **genetischer oder umweltbedingter Ursachen**. Wir demonstrieren u.a. auch, dass leichte stöchiometrische Veränderungen in solchen Modulen zu weitläufigen Effekten im zellulären Haushalt führen können, z.B. im Transport von spezifischen Molekülen.

Die vorgestellte Arbeit beschreibt allerdings auch inwieweit **temporäre Ereignisse und Prozesse** die **Stöchiometrie** von Proteinkomplexen und funktionellen Modulen beeinflussen. Wir zeigen z.B. auf, dass eine Veränderung in der glycolytischen Enzym-Stöchiometrie die

Ursache für den Warburgeffekt in **gealterten Zellen** des humanen Knochenmarks darstellen könnte. Neben der Analyse von Protein-Abundanzen untersucht die vorliegende Arbeit Proteomdynamik auch in Hinblick auf **Stabilitäts- und Löslichkeitsveränderungen** von Proteine in kürzeren Zeitabläufen wie den **Zellzyklus**. Wir können dabei feststellen, dass Untereinheiten von größeren Proteinkomplexen verschiedene Stabilitätsmuster aufweisen. Dies ist durchaus eine neue Erkennis, die weittragende Folgen für unser Verständnis des Ab- und Aufbauprozesses von Proteinkomplexen haben könnte.
Die Einblicke, die aus dieser Arbeit gewonnen werden können, implizieren in jedem Falle, dass das Proteom mehr als die Summe der Einzelteile darstellt, und hauptsächlich durch die Variabilität von gesamten Proteinensembls und deren Kooperativität bestimmt wird. Proteinkomplexe und funktionelle Module sollten daher als molekulare Reflektionen von genetisch- und umweltbedingter Variation betrachtet werden. Solch ein Perspektivenwechsel könnte damit die Möglichkeit bieten eine mechanistische Verknüpfung von Genotyp und Phänotyp zu gewährleisten, und ein Fundament für zukünftige individuell angepasste klinische Behandlungen darstellen.

Acknowledgements

In undertaking my PhD project I was privileged to work with, and enjoy the support of a number of colleagues and friends. I would like to take this opportunity to gratefully acknowledge their contributions to this work.

First, and foremost, my heartfelt thanks to my supervisor, **Peer Bork**, who has given me the opportunity to work on many interesting projects at EMBL. Fortunately for me, he did not push me into a strict working framework, but let me manage my own projects and time, giving me a sense of scientific independence. Though such independence would usually spark some more explorative months that would not necessarily yield productive results, Peer would help me to push onwards and to 'get things done'. After this work I hope that this is indeed an attitude that I will maintain for the rest of my life and career.

From all people I had the pleasure to work with I would like to particularly highlight my TAC committee member **Martin Beck** for his mentorship in the proteomics field. Indeed, the work on protein complexes would not have been possible without his critical input, and his extensive knowledge on the structural confounders and complex assemblies. He also provided the grounds for the fruitful collaboration with **Mikhail Savitski** on the thermal stability of proteins, which has been nothing but joy to work on. The regular discussions on preliminary results with Misha, Martin, **Isabelle, Amparo** and **Frank** were indeed productive and inspiring sessions that I consider a happy collaborative experience.

Speaking of collaborations, I would also like to acknowledge members of another collaboration that I have been heavily involved in, and which has undoubtedly given me a lot of valuable lessons – the SyStemAge project. My thanks to **Anne-Claude Gavin**, **Anthony Ho**, and **Marco Hennrich**. I would also like to acknowledge members of the ongoing collaboration with the Pepperkok Group: **Rainer Pepperkok** himself, as well as **Georg Galea** and **Juan Jung**, who were very supportive and would try to put my bioinformatics hypotheses into real biological practice.

A key figure during my PhD that deserves special mention is of course **Alessandro Ori**, who at the very least can be described as the scientific and energetic driver of my PhD. It is his enthusiasm that got me started to think about investing more time in the exploration

of the issues and ideas presented in this work, and I thank him for always being up-beat with what I was doing.
I am grateful to my TAC committee member **Wolfgang Huber** for always having a spare minute or two for my scientific questions, and of course, my last TAC committee member and university advisor **Thomas Dandekar** deserves a mention for his advisory role and friendly attitude that would always lighten my day. Also **Michael Kuhn**, our group's staff scientist, is reserved for special thanks, as he considerably helped me with accomplishing my tasks and being one of the friendliest colleagues I have ever had.
I am grateful to the enjoyable environment provided by my colleagues from the **Bork Group** in general, particularly at our daily lunch breaks and subsequent coffee sessions with passionate discussions about the political and scientific landscape. Thank you for the daily doses of humor, inspiration and the great deal of scientific support as well.
Admittedly, this endeavor would not have been possible without my family and friends. My warm thanks to all!

Contents

List of Figures

S2.8 Leveraging module normalization for discovery of differentially expressed proteins in context of modules (related to Figure 2.5)

S3.1 Checking data quality with known cell cycle markers (related to Figure 3.2)

S3.2 Analysis of cell cycle- dependent stability effects on organelle-specific proteins (related to Figure 3.3)

S3.3 Overview on complex co-stability and co-abundance (related to Figure 3.7)

S3.4 Differential stability pattern of moonlighting subunits of complexes (related to Figure 3.7)

S3.5 Detailed line plots for all measured subunits of the NPC-complex (related to Figure 3.7)

S3.6 Detailed melting patterns of spindle-associated proteins (related to Figure 3.8)

S3.7 Selection of stabilized proteins, GO-analysis for stabilized set of proteins (related to Figure 3.9)

S3.8 Stabilization of sumoylated proteins, and leak-over (related to Figure 3.9)

S3.9 Detailed solubility track of proteins (related to Figure 3.10)

S3.10 Functions and modifications of proteins with prominent solubility transition during cell cycle (related to Figure 3.11)

S4.1 Labeling efficiency, protein numbers and outlier analysis (related to Figure 4.2)

S4.2 Overview of the mRNA expression levels, functional classes, as well as cellular compartments of all quantified proteins in comparison with the total human proteome (related to Figure 4.2)

S4.3 Comparison of the proteomes of the six investigated cell populations (related to Figure 4.2)

S4.4 Characterization of the number and relative abundance of the core and specific proteome in the individual cell populations (related to Figure 4.2)

S4.5 The stoichiometry of proteins of the glycolytic pathway are maintained across donors in the different cell populations (related to Figure 4.2)

S4.6 Overview on gender effects in ageing, and transcript expression changes upon ageing.

S4.7 Age-affected pathways in the individual cell populations (related to Figure 4.3)

S4.8 Single-cell analysis reveals lineage- and age-dependent increase of glycolytic enzymes (related to Figure 4.5)

S4.9 Overview on the computational data processing (related to Methods)

List of Tables

S4.1	Overview of the age, gender, available cell populations and purity of the samples used in the study.
S4.2	List of all proteins identified in our study
S4.3	List of all proteins quantified by label-free quantification
S4.4	List of pathway changes across different cell populations in the bone marrow
S4.5	List of all proteins quantified by TMT
S4.6	List of numbers of proteins relevant to the study
S4.7	List of pathways that contain proteins that are significantly altered upon ageing

List of Publications

Publications from Doctoral Study:

- **Natalie Romanov***, Alessandro Ori, Michael Kuhn, Martin Beck and Peer Bork (2018). Systematic meta-analysis of individual proteotypes reveals sex- and diet-specific functional modules. *In preparation.*

- Marco L. Hennrich*, **Natalie Romanov***, Patrick Horn*, Samira Jaeger, […], Peer Bork, Anne-Claude Gavin and Anthony Ho (2018). Cell-specific proteome analyses of human bone marrow reveal molecular mechanisms of age-dependent functional decline. *Nature Communications.*

- Isabelle Becher*, Amparo Andres-Pons*, **Natalie Romanov***, Frank Stein*, Maike Schramm, Florence Baudin, Dominic Helm, Nils Kurzawa, André Mateus, Marie-Therese Mackmull, Athanasios Typas, Christoph W. Müller, Peer Bork, Martin Beck and Mikhail M. Savitski (2018). Pervasive protein thermal stability variation during the cell cycle. *Cell.*

- Panagiotis Kastritis, Francis O'Reilly, Thomas Bock, Yuanyue Li, Matt Z. Rogon, Katarzyna Buczak, **Natalie Romanov**, Matthew J Betts, Khanh Huy Bui, Wim J Hagen, Marco L Hennrich, Marie Therese Mackmull, Juri Rappsilber, Robert B Russell, Peer Bork, Martin Beck, Anne-Claude Gavin (2017). Capturing protein communities by structural proteomics in a thermophilic eukaryote. *Molecular Systems Biology*, 13(7):936. doi: 10.15252/msb.20167412.

- Chunguang Liang, Dominik Schaack, Mugdha Srivastava, Shishir K. Gupta, Edita Sarukhanyan, Anne Giese, Martin Pagels, **Natalie Romanov**, Jan Pané-Farré, Stephan Fuchs and Thomas Dandekar (2016), A Staphylococcus aureus Proteome Overview: Shared and Specific Proteins and Protein Complexes from

Representative Strains of All Three Clades. Proteomes, 4(1). pii: E8. doi: 10.3390/proteomes4010008.

- Yvonne Heinze, Martin Bens, Enrico Calzia, Susanne Holtze, Oleksandr Dahovnik, Arne Sahm, Joanna M. Kirkpatrick, Karol Szafranski, **Natalie Romanov**, Kerstin Holzer, Stephan Singer, Maria Ermolaeva, Matthias Platzer, Thomas Hildebrandt and Alessandro Ori (2018). Species comparison of liver proteomes reveals links to naked mole-rat longevity and human aging. BMC Biology, 2018. *In press.*

Previous Publications & Other:

- Kristofer Bodvard, Ken Peeters, Friederike Roger, **Natalie Romanov**, Aeid Igbaria, Niek Welkenhuysen, Gael Palais, Wolfgang Reiter, Michel B. Toledano, Mikael Käll & Mikael Molin (2017). Light-sensing via hydrogen peroxide and a peroxiredoxin. *Nature Communications*, 8:14791. doi: 10.1038/ncomms14791.

- **Natalie Romanov***, David Maria Hollenstein, Marion Janschitz, Gustav Ammerer, Dorothea Anrather, Wolfgang Reiter (2017). Identifying protein kinase-specific effectors of the osmostress response in yeast. *Science Signaling*, 10(469). pii: eaag2435. doi: 10.1126/scisignal.aag2435.

Glossary

AGE	Advanced Glycation Endproduct	**MAF**	Minor allele frequency
AHA	Azidohomoalanine	**MOAC**	Metal oxide affinity chromatography
AUROC/AUC	Area under receiver operating characteristic	**MON**	Monocytes/monocyte progenitors
CETSA	Cellular thermal shift assay	**MS**	Mass spectrometry
CID	Collision-induced dissociation	**MSC**	Mesenchymal stromal cells
CNV	Copy number variation	**OMIM**	Online Mendelian Inheritance in Man
COSMIC	Catalogue of Somatic Mutations in Cancer	**PC**	Principal component
CPTAC	Clinical Proteomic tumor Analysis Consortium	**PCA**	Principal component analysis
DDA	Data-dependent acquisition	**PDB**	Protein data bank
DIA	Data-independent acquisition	**pQTL**	Protein Quantitative Trait Loci
ELM	Eukaryotic linear motif	**PSM**	Peptide spectrum match
eQTL	Expression Quantitative Trait Loci	**PTM**	Post-translational modification
ERP	Erythrocytes/erythrocyte progenitors	**QTL**	Quantitative Trait Loci
FDR	False discovery rate	**RNA-seq**	RNA sequencing

GRA	Granulocyte/granulocyte progenitors	**ROC**	Receiver operating characteristic
GWAS	Genome-wide association studies	**SILAC**	Stable isotope labeling amino acid in culture
HapMap	International haplotype map project	**SLiM**	Short linear motif
HCD	High collision dissociation	**SNP**	Single nucleotide point mutation
HPC/ HSC	Hematopoietic progenitor stem cells	**TCA**	Tricarboxylic acid
IMAC	Immobilized metal affinity chromatography	**TCGA**	The Cancer Genome Atlas
iTRAQ	Isobaric tags for relative and absolute quantitation	**TF**	Transcription factor
LC	Liquid chromatography	**TMT**	Tandem mass tag
LYM	Lymphocyte/lymphocyte progenitors	**TPP**	Thermal protein profiling

Chapter 1

Introduction

The cell is a sophisticated machinery, characterized by a number of intricate subsystems that coordinate and execute vital tasks, such as cell differentiation, cell growth and cell division, to name only a few [Alberts et al., 2008]. The sheer dimension of that coordination is nothing less but extraordinary. It suffices, for example, to know that an adult human loses around 50 to 70 billion cells every day due to programmed cell death [Alberts et al., 2008] and that the organism compensates that loss by creating new cells. Cell division, or mitosis, provides the mechanistic means to achieve that; it essentially involves a faithful duplication of the genetic material, chromosomal separation into precisely equal shares, and a division of the cytoplasm, organelles and the membrane [Carter et al., 2014]. Though not apparent from the summary, that fundamental process requires an army of tiny molecular machines [Satir et al., 2008], i.e. proteins. They are needed for duplicating and packaging the DNA-material, as well as for the kinetochore assembly to ensure the correct timing to pull the chromosomes apart [Santaguida et al., 2009], just to name a few critical elements during the cell cycle. Proteins are undoubtedly the molecules that make the cell actually work: They perform a vast array of functions, from catalyzing reactions, and DNA replication to signaling and transporting molecules from one cellular location to another. The fact that those tiny molecular machines routinely and faithfully execute their functions billions of times in the human body is truly a fascinating feat of nature. And it proves vital to understand how these systems operate to leverage their power in the future (design of artificial molecular machines [Strong, 2004; Lu et al., 2018]), and cure diseases of the human body.

Dissecting biological systems in the cell reveals molecular cooperativity as a fundamental principle underlying practically all mechanisms involved [Whitford et al., 2005]. The phenomenon becomes apparent with hemoglobin as an example: Once oxygen binds to one of the four binding sites of hemoglobin, the affinity of the three remaining sites increases, hence it becomes more likely that the hemoglobin molecule will have all of its binding sites occupied with oxygen (cooperative binding, Whitford et al., 2005). While cooperativity

equally manifests itself at the DNA, and lipid level, it is challenging to examine the effects of cooperativity across molecular layers (a) and in the entirety of the system (b). The case of (a) is probably best exemplified with the paradigm of Linus Pauling [Pauling, 1949], again in context of the aforementioned hemoglobin (Hb) molecule: A single mutation in the Hb gene - while not affecting the immediate functionality of its product itself - leads to a flawed assembly of hemoglobin molecules into tetrameric structures. Hence, it directly interferes and ruptures the required cooperativity of that molecular machinery in order to successfully bind oxygen molecules. It becomes even more difficult to estimate what other elements of the cellular system are then immediately affected by that missing cooperativity (b), and how molecules respond to avoid collateral damage.

The most prominent model providing a link between the genetic composition of an individual (genotype) to phenotypic variability, is guided by the principles of the so-called central dogma of molecular biology [Crick, 1970]. The flow of genetic information assumes that a mutation in a gene-coding DNA region further transcribes into a flawed transcript, and subsequently translates into a possibly functional or non-functional protein structure (**Figure 1.1**). While the dogma certainly provides a framework for understanding the information flow, it does inherently treat gene products on a singular basis, and not in a system-wide manner. Thus, if a flawed protein structure were to emerge from a mutated gene, how is the proteomics system modulated to possibly buffer it? What re-arrangements are necessary to achieve that? The large-scale effects on the protein landscape that is characterized by high levels of connectivity and molecular cooperativity remain largely unexplored.

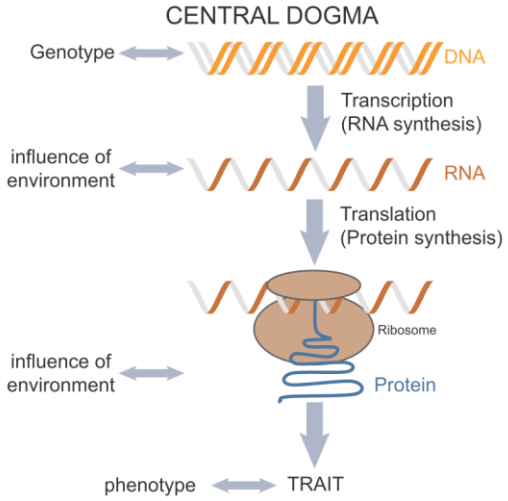


Figure 1.1. Illustration of the Central Dogma. The genotype of an individual is defined by the composition of the respective DNA molecules, which gets transcribed into RNA, and subsequently translated into a gene product, which could be a protein. In concert with other proteins the gene product elicits a certain functionality and a phenotypic trait. At every step of the central dogma, feedback from the environmental setting adds variation to the process.

Investigating systematic changes in the protein landscape and characterizing the proteome not only requires the accurate identification of individual proteins, but measuring their abundances as well. We assume that by studying those protein abundances and possible co-variation in abundances, it is possible to infer molecular cooperativity and to hence observe disruptions in protein organizations due to genotypic variability.
In the following sections, we will present current measurement methods to approximate protein abundances, and further delve into the concept of the proteome modularity and possible features that can be extracted from it.

1.1 Proteomics State of the Art and why Transcriptomics is not enough

1.1.1. The Transcriptome as a Proxy for the Proteome?

Protein concentration levels in general are determined by the levels of their coding mRNAs, by translation rates and by protein turnover [Beck et al., 2011; Schwanhäusser et al., 2013]. To what extent does each process, however, contribute to the eventual protein levels? Would a weak correlation of protein and mRNA levels indicate poor quality of data or rather imply post-transcriptional processes fine-tuning of proteins? This particular research question has already been investigated for decades, yet still there is no simple consensus on how the cell determines final protein concentrations, and whether the mechanisms in place are applied at random [Liu et al., 2016; Edfors et al., 2016] In the following section of the introductory chapter, we will elaborate on protein-mRNA correlations and their usefulness (a), as well as tackle the question on how (measured) protein levels are determined (b), and finally whether post-transcriptional regulation is relevant (c). These are important questions to understand how the proteotype is established in the first place.
A deviation from a perfect regression line could be due to noise (measurement errors), but also caused by post-transcriptional mechanisms, or intrinsic mRNA variation. At this point it is also vital to distinguish between two types of correlations that can be performed: 1) the same protein in different conditions is monitored against its respective transcript, answering the question on how much does transcript change affect protein level changes (**Figure 1.2A**) [Liu et al., 2016]. The actual R^2 across studies is not comparable since it is highly dependent on the conditions used. (2) Different proteins in the same condition are compared against their respective transcript levels (**Figure 1.2B)** thereby giving an estimation on how much transcript levels determine absolute protein levels. Various studies have been conducted on identifying the actual impact, ranging from R^2=0.4-0.9 (Pearson and Spearman, whereby Spearman is independent of linearity of relationship). One particular study by Willhelm et al. (2014) stated that protein-mRNA ratios are highly

conserved/constant across tissues and thus could be used to predict protein levels from mere mRNA. In fact, as seen from **Figure 1.2C**, they reach a correlation of up to 0.91 by applying this concept. Another publication by Fortelny et al. (2017) confirmed that predictions of protein levels indeed coincide with protein-mRNA ratios, which they incidentally label as "translation rates". Probably the ratio is defined by both a protein's half-life and its turnover-rate. The authors of this work performed and additional analysis, changing the translation rates to the median transcript levels across all tissues (the mRNA is rendered

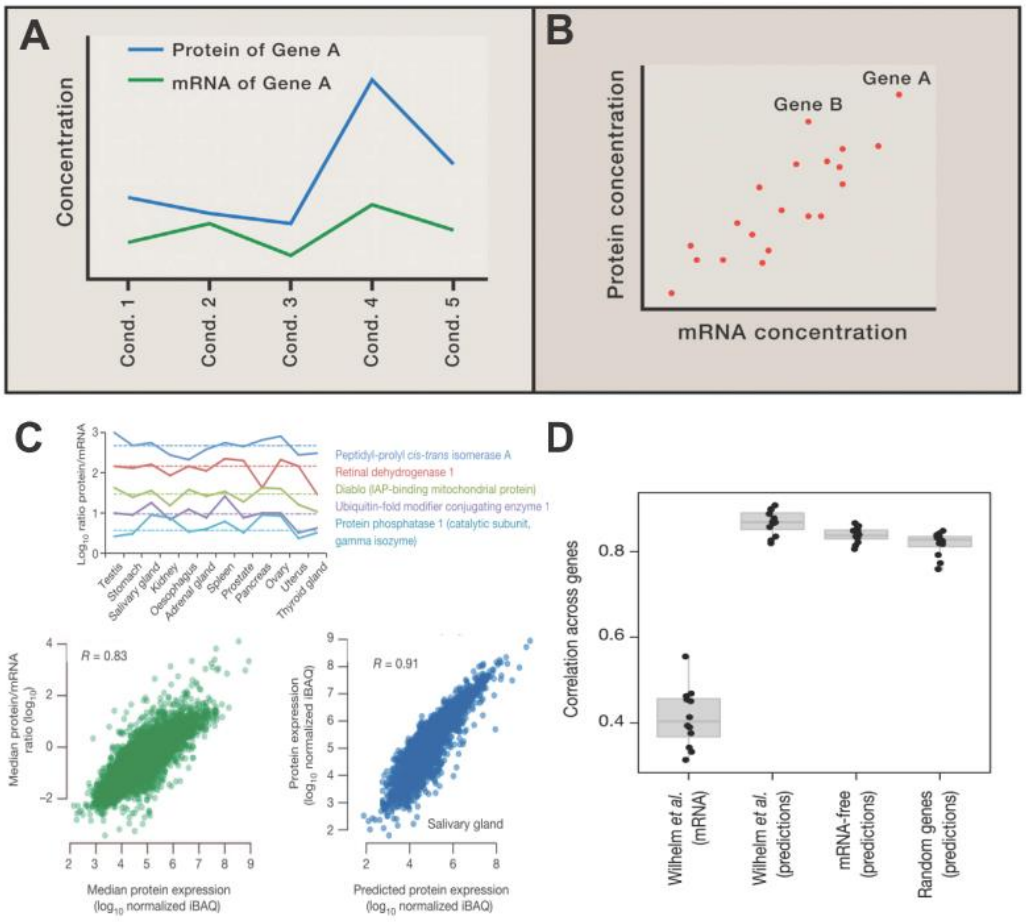


Figure 1.2. Overview on Protein-Transcript Correlation Aspects. (a) Correlation of one protein with its respective transcript across different conditions. (b) Given a condition, measured protein concentrations are correlated against respective mRNA levels. Both (a) and (b) were adapted from Liu et al., 2016. (c) Conservation of protein/mRNA ratio across 12 organs (upper panel). Lower left panel shows the correlation of median translation rates of transcripts across all tissues, with protein abundance. The lower right panel shows the correlation between protein expression (y-axis) and protein levels as predicted from protein/mRNA ratios (x-axis). Graphs adapted from Willhelm et al., 2014. (d) Box-plots illustrating correlations from 12 tissues between mRNA and protein levels (1st boxplot), between predicted and observed protein levels (2nd boxplot, sourced from Willhelm et al., 2014), between mRNA-free predictions and protein levels (3rd boxplot), and between predictions based on random transcripts and protein levels (4th boxplot). The graph has been adapted from Fortelny et al., 2017

independent of the tissue). The resulting protein-RNA correlations were almost as good as for the tissue-specific mRNA. Even randomizing protein-transcript pairs resulted in high correlations (**Figure 1.2D**). It can be concluded therefore that beyond the expression of a given transcript, other transcriptional and post-translational factors need to be considered to give a more accurate estimation about the protein levels in the cell [Willhelm et al. reply, 2017]. As indicated in the beginning of this chapter, the central dogma is affected by several processes. Levels of mRNA, for instance, are governed by their stability and localization [Trcek et al., 2011]; latter being crucial in neurons and developmental processes for translational purposes. Translation, on the other hand, depends largely on mRNA structure, mRNA-binding proteins, miRNAs, codon usage, (charged) tRNA pool, etc. [Svennigsen et al., 2017]. Protein turnover is then affected by proteasome availability/localization, autophagy, stoichiometry of protein complexes, folding/chaperones, etc. In fact, if the chaperone mechanisms fails to control solubility of certain proteins, the formation of aggregates may leads to a dramatic increase of half-lives as proteins get stabilized [Liberek et al., 2008]. Furthermore, the actual ex- and import of the protein, and therefore its decoupling from the mRNA, could influence the eventual correlation that we measure.

When analyzing protein-mRNA correlations, there are also severe technical issues to consider: Quantifying proteins tends to be difficult due to (i) post-translational modifications changing visible peptide levels, (ii) incomplete proteins/protein fragments, (iii) splice isoforms, and (iv) physico-chemical biases, with e.g. membrane proteins not being detected. On the transcriptomics side, there are also technical issues involving splice isoforms, stability of transcripts and ratios of nascent (immature) to mature mRNAs.

Finally, an important aspect in understanding the relationship between protein abundances and transcript levels is kinetics [Liu et al., 2016] : The actual mRNA might be gone by the time of protein measurement due to a lag-phase or signal delay, RNA being relatively unstable while proteins usually being stable (**Figure 1.3A**). Such kinetics become especially apparent in single-cell analysis: For a protein to remain stably expressed, it is believed so far that transcription works in so-called bursts (transcription bursts) over time, which are not apparent in bulk measurements (**Figure 1.3B**). A similar situation applies to cell-cycle dependent genes, where we also observe a delay between RNA and protein bursts (**Figure 1.3C**). Conclusively, the higher the measurement resolution (sampling density), the lower the protein-RNA correlation is bound to be, because of the reduced spatial and temporal averaging.

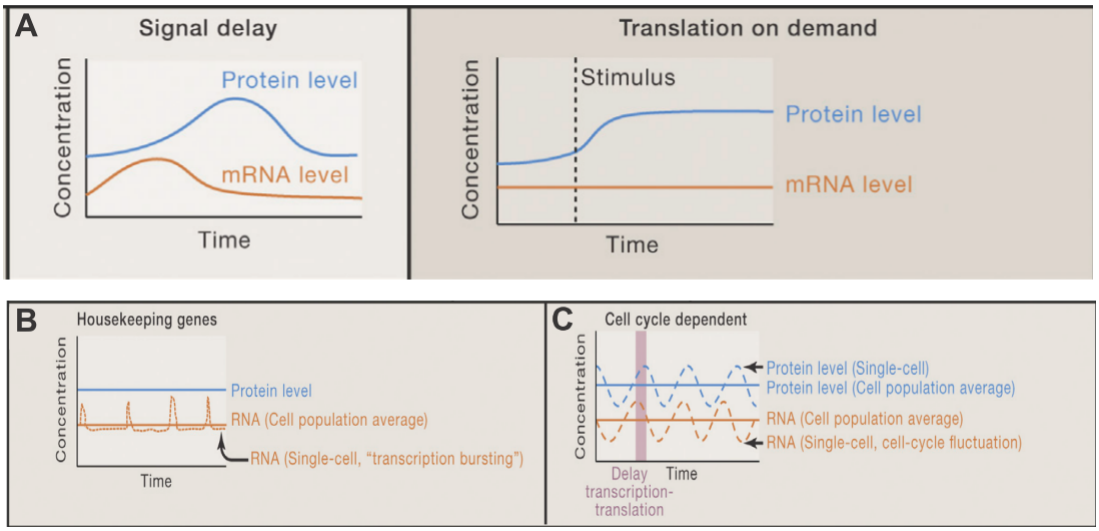


Figure 1.3. Dynamics of Protein-RNA relationship. (a) The left-hand panel illustrates the delayed signal upon transcription and translation of elevated RNA levels. 'Translation on demand' mechanism ensures that protein levels increase due to tweaking of translation rates despite constant RNA levels (right-hand panel). (b) At the single-cell level transcriptional bursting signals become apparent that ensure constant protein levels over time. A bulk measurement of RNA levels would be constant, however. (c) Processes like the cell cycle do also show protein- and RNA-levels lagging due to delays in transcription and translation. The graph has been adapted from Liu et al., 2016.

Many studies have investigated to what degree transcripts contribute to protein levels [Willhelm et al., 2017; Fortelny et al., 2017; McManus et al., 2015]. One study by Jovanovic et al. (2015) investigated LPS-stimulated dendritic cells, and collected comprehensive data on protein levels, protein level changes upon stimulation using a pulse-chase-SILAC set-up [Selbach et al., 2008], turnover of proteins and mRNA expression. This data was then used to construct a model that would consider all possible contributions of biological information. It essentially described a steady-state (before treatment), as well as dynamic changes (fold-changes) distinguishing relative levels versus absolute molecule counts, respectively (**Figure 1.4**). The authors found that before treatment around 70% of protein abundances can be attributed to RNA levels, 21% to translation and 10% to degradation. After treatment, however, the relative fold-changes are almost completely determined by RNA changes. Interestingly, the absolute number of molecules gives a different picture with around 40% transcript contribution. It is further argued by the authors that for very highly abundant proteins even small changes in the translation rate would have a huge impact on the total number of protein molecules created. At this point it also needs to be highlighted that the protein concentration of the cell remains more or less constant unless the cell is actively growing (e.g. G1-phase in the cell cycle). Hence, in order to increase the quantity of one specific protein, the quantity of another protein needs to be reduced.

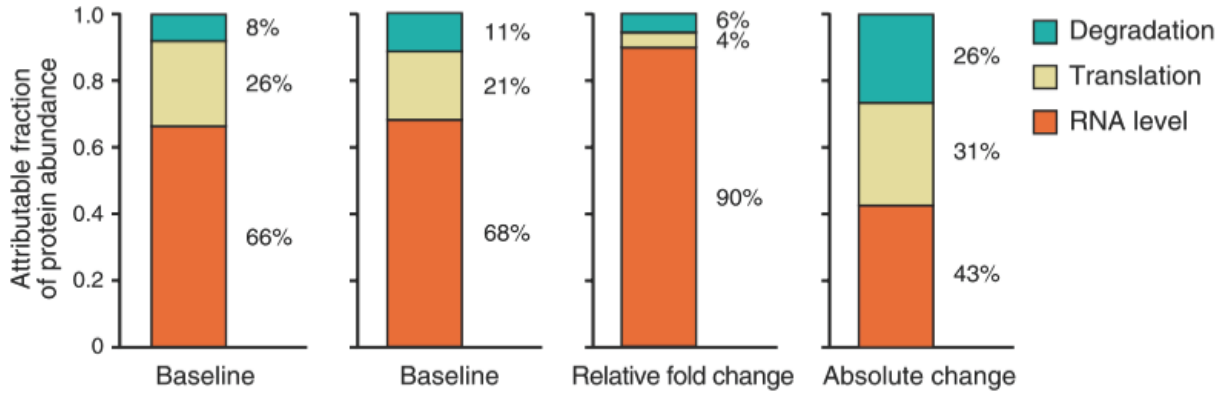


Figure 1.4. Contribution of mRNA abundance, translation, and degradation rates to protein levels. *(f.l.t.r.)* First and second bar plots illustrate contributions to protein levels before LPS induction, the two bar plots on the right show the contribution to protein abundance change between LPS-induced and vehicle-treated. The third and fourth bar plot differ by quantifying contribution to the relative fold-changes (3rd bar plot), and the absolute changes (4th bar plot) after LPS stimulation. The graph has been adapted from Jovanovic et al., 2015.

Conclusively, it can be established that mRNA indeed can serve as a reasonable proxy of protein abundances, are easy to measure in a genome-wide manner and usually are less noisy than the measurement of its proteomic counter-part. Hence, is it enough to look at the transcripts only? It is clear that for certain predictable changes, such as developmental programs of organisms, transcripts give a very accurate picture of the ultimate amount of proteins produced. In a system that is not as clear-cut as developmental programs or the cell cycle, post-transcriptional mechanisms come into play as well, such as in signaling mechanisms [Hinnebusch et al., 2011; Beck et al., 2011]. For such systems, notably, it has been reported that a mechanism of 'translation on demand' [Beyer et al., 2004] is maintained, which allows for more cautious use of energy resources. The final mechanism in case of unwanted mRNA fluctuation involves degrading excess proteins. That protein level buffering becomes especially apparent in protein complexes with cells maintaining specific stoichiometries [Dephoure et al., 2014; Stingele et al., 2012]. This has been shown to be the case for copy number variations [Goncalves et al., 2017] in general, and more specifically in cases of trisomy-21 [Liu et al., 2017].

We therefore conclude in this section that transcription is one of the major mechanisms by which cells control the abundance of proteins; yet there are also other processes that need to be taken into account with their respective importance varying from protein to protein. Using mRNA as a proxy for protein levels is therefore not a reliable approach in general. It can thus be argued that the proteome and the transcriptome do not contain the same biological information [Liu et al., 2016].

1.1.2. Mass-spectrometry technology - its strengths and caveats

The experimental method of choice for both identifying and quantifying proteins in a large-scale manner is mass-spectrometry (MS) [Aebersold and Mann, 2003; Domon and Aebersold, 2010]. Briefly, protein molecules get extracted from samples and cleaved by specific proteases into peptides. These peptides then get subjected to an analysis using a high-performance liquid chromatography (HPLC) coupled to an MS instrument (LC-MS) (**Figure 1.5**). The chromatography allows the physical separation of the peptides in time due to their interaction with the package material and the stationary phase in the column. As other solvents are pumped through the column, different peptides reach the end of the chromatography where a UV-detector records the absorbency at different wavelengths of the spectrum (260nm = DNA, 280nm = protein (Trp, Tyr, Phe), 595 = Coomassie blue dye). The area under the UV-trace is directly proportional to the amount of the given

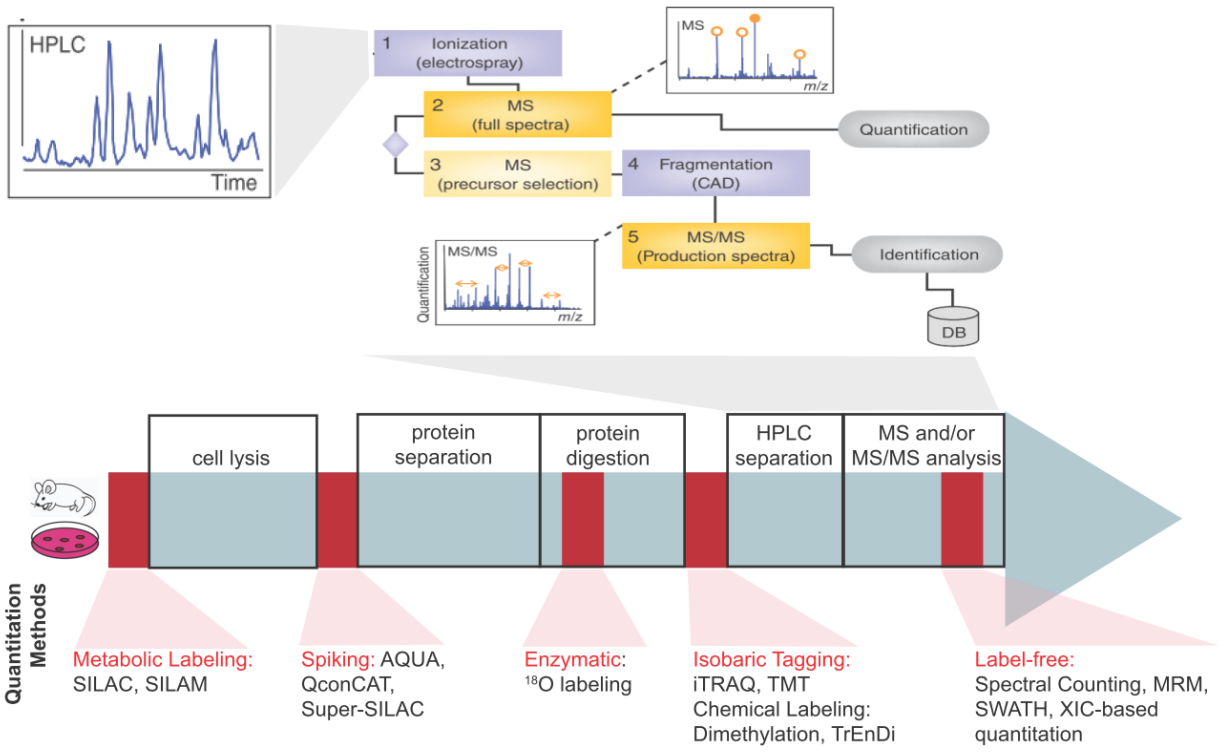


Figure 1.5. Workflow of a proteomic experiment, with exemplified quantification methods. The first steps in the workflow involve cell lysis, protein separation, and digestion to obtain peptides. These peptides are then separated by an HPLC prior to mass spectrometry analysis. More specifically, the peptides get ionized; MS1 spectra are acquired (full spectra), and top precursor ions are further fragmented and MS2-spectra (production spectra) are obtained. The combination of both MS1 and MS2 spectra is required for identification and quantification. In the lower part of the graph, different methods to acquire more precise quantification information, are outlined. The upper part of the graph has been adapted from Domon & Aebersold, 2010.

material eluting from the column, and therefore constitutes a critical part of the subsequent integration with actual MS-data. An electrospray then injects the resulting peptides into the mass spectrometer which records mass-to-charge ratios (m/z) of eluting peptides at a given time point (MS1) [Canas et al., 2006; Yates et al., 2009]. From such a full-scan spectrum the precursor ions with the highest intensity are selected (top12) and further fragmented using collision-induced dissociation (CID) or high-energy C-trap dissociation (HCD), depending on the instrument. For each precursor ion, a further fragmentation spectrum (MS2) is recorded, gathering the information on so-called b- and y-ions. Ultimately, both the MS1- and MS2-spectra are used to identify the peptide by interrogating a database of protein sequences. In order to quantify protein abundances, several methods have been developed (**Figure 1.5**), and it is important to understand their advantages, as well as their caveats to reliably draw conclusions on results obtained with them. **Table 1.1** gives a brief overview. The most common and straightforward way of quantifying protein abundance relies on extracting the precursor ion chromatogram at the MS1 level [Zhu et al., 2010], for a given m/z-value. The approach is genuinely independent of any chemical modifications of the peptides or multiplexing and can theoretically be obtained for an unlimited number of peptides and samples. However, the quantification method suffers from poor reproducibility due to the stochastic ion selection per MS run, and technical issues including sample handling, digestion efficiency, injection, shifts in retention times, instantaneous matrix effects, etc. [Nesvizhskii, 2007; Worboys et al., 2014]. This particular methodology will be used in **Chapter 4** as an orthogonal quantification method, but needs to be carefully interpreted due to its inherent caveats.

Table 1.1: Overview on MS-quantification methods

Method	Advantages	Disadvantages	Applications
Label-free (area under the curve)	○ no labeling required	○ poor reproducibility ○ requires specialized software (MQ has label-free quantification option)	biomarker discovery possible
SILAC	○ reduces technical variability ○ labeling incorporated in vivo	○ cells must grow on special media (difficult with entire living organisms, i.e. mice) ○ expensive ○ side effects?	Cell-culture based study
iTRAQ/TMT	○ no special growth conditions required ○ use on primary tissue possible ○ higher multiplex than SILAC	○ noisier than SILAC ○ expensive ○ interference with enrichment protocols	primary cells/tissue-based studies

Synthetic peptides	○ guaranteed observation precise assessment of concentrations	○ analyte must be known peptide preparation needed	Targeted proteomics
Spectral Counting	○ no labeling required	○ poor reproducibility	

Isotope labeling approaches, on the other hand, partially address the issue of poor reproducibility. The basic principle was introduced with stable isotope labeling of amino acids in culture (SILAC, Ong et al., 2002), where cultures to be compared are grown in light (C-12) and heavy (C-13) media. After mixing cells in a 1:1 ratio, proteins are processed together to guarantee low technical variability, and the relative protein abundance between the light and heavy condition can reliably be quantified in the mass spectrometer. This quantification method has found a lot of applications for monitoring protein and PTM expression [Battle et al., 2014; Romanov et al., 2017], protein interactions [Kleiner et al., 2017], as well as synthesis and degradation rates (pulse SILAC, Selbach et al., 2008). However, it remains restricted to lower organism, such as yeast and bacteria, than can synthesize their own amino acids from basic precursors and relies on the area under the curve- calculations from the extracted ion chromatograms. The introduction of higher multiplex-labeling, such as iTRAQ (isobaric tags for relative and absolute quantification), and TMT (tandem-mass tag, McAlister et al., 2012; Rauniyar and Yates, 2014), certainly represented a paradigm shift as it was circumventing latter mode of quantification. The essential idea involves chemical reagents that have stable isotopes at various positions, while their total mass is the same. The reagents typically consist of a reporter group, a balance group and a peptide reactive groups that interacts with primary amines (hence the N-terminal ends of peptides, e.g.). After sample processing and protein cleavage, peptides get labeled with those reagents, and analyzed in the MS. During precursor fragmentation, the balance group is bound to fall off, leaving the reporter ion as a single fragment with a very distinct mass (in a 4-plex system, 4 reporter ions with 114-117 m/z are created). Thus, the different variations of the reporter ion appear in a fragmentation spectrum of one single peptide, and reflect the actual abundances of the peptide across the different conditions tested. The TMT-MS currently supports ten-multiplexed systems allowing the coupling of several conditions in one single MS-run. The power and usefulness of this technology will be further demonstrated in **Chapter 3** and **4**.

In any case every quantification method remains biased towards abundant peptides given the inherent manner of ion selection in the mass spectrometer. These so-called data-dependent acquisition methods (DDA) allow covering 5.000 proteins in one single run (50.000 peptides, corresponding to 50.000 western blots an hour) in a reasonable amount of time spanning 6-8 hours from sample processing to acquiring the digital file [Lemeer and

Heck, 2009]. Recently developed technologies, such as SWATH-MS, aims at complementing such traditional methods that remain hampered in their quantification capability (data-independent acquisition, DIA) [Gillet et al., 2012]. In a typical SWATH set-up, the mass spectrometer scans a range of peptide precursors from 400-1200 m/z and fragments all precursors in 25Da isolation windows. The same isolation window is fragmented again at each cycle during chromatographic separation, thereby allowing for a time-resolved recording of fragment ions of all peptide precursors. While the obtained ion maps are difficult to de-convolute (OpenSWATH), the technology provides the first step towards accurate quantification of a wide range of proteins in a reasonable time frame.

1.1.3. Proteomic Findings Reaching Saturation Level

The current state of the art in proteomics research indicates that around 14,200 proteins have been detected, with more than 1 million distinct peptides from 133 million peptide spectrum matches (PSMs) (according to the latest survey on Human Peptide Atlas [Desiere et al., 2006]). While a more thorough investigation of post-translational modified peptides, as well as isoforms could certainly fill the remaining void in peptide detection, it is clear that in terms of actually mapping the human proteome, the discovery of new peptide/protein species are close to saturation level. **Figure 1.6A** demonstrates the tremendous increase in fragmentation spectra generated from the human proteome using classical data-dependent acquiring techniques (DDA), as recovered from the EBI databases [Cook et al., 2016].

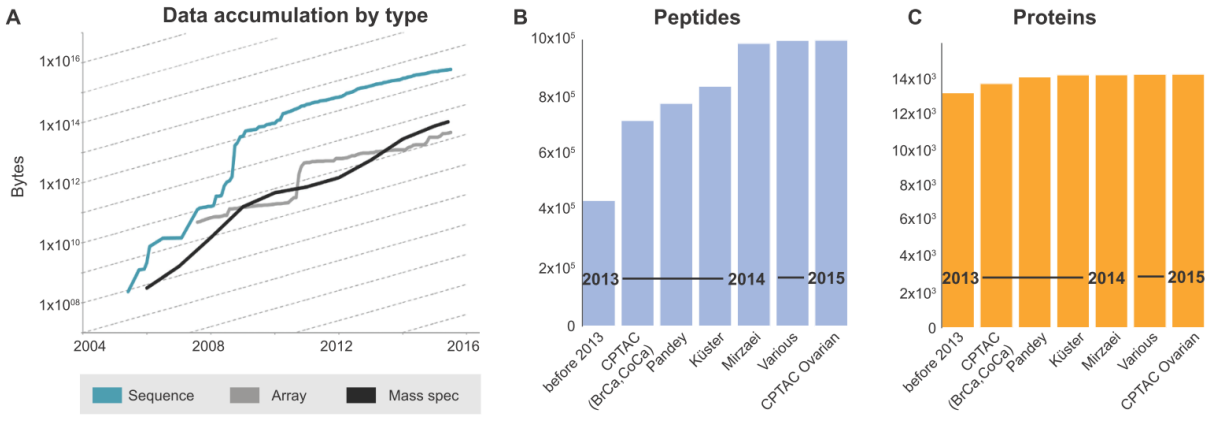


Figure 1.6. Saturation of the Human Proteome. (a) Overall size of data of different types (sequencing, array, mass spectrometry) accumulated over time in EMBL-EBI. (b) Number of identified peptides across large-scale datasets. (c) Number of identified proteins across large-scale datasets. Data for (b) and (c) have been sourced from the Human Peptide Atlas.

When further dissecting peptide and protein discoveries according to publications in **Figure 1.6B/C**, the plateauing becomes noticeable with both peptides and proteins (assuming that the FDR has been correctly estimated). It is therefore paramount to look at the acquired data from a different perspective. In context of the above described gap between genotype and phenotype, it makes sense, for example, to investigate the notion and relevance of protein cooperativity and the modular architecture of the proteome.

1.2 The Proteome in Context - the Proteotype

1.2.1. The modular architecture of the Proteome

For the most part proteomics research treats proteins as largely independent molecules, the same way it is usually done for transcripts [Liu et al., 2016]. Yet it is clear that this is not reflective of the mode of operation of those molecules. Rather, proteins act a social molecules, assembling and acting in modules and networks, which in turn is complex and rich in biological information. In fact, a pioneering observation made by Linus Pauling in 1949 already indicated the effect of protein re-organization leading to a clinical phenotype [Linus Pauling et al., 1949]. The established paradigm was exemplified in sickle cell anemia, showing that one mutation in a specific locus in the genome leads to a structural change in hemoglobin Hb, causing the phenotype. The defined mutation did indeed result in abundance changes of the specific gene product. Most importantly, however, it changed the way hemoglobin gets organized into tetrameric complex, which is the primary trigger for the emergence of the disease (**Figure 1.7A**). While the paradigm establishing the link between genetic variants and altered protein structure and function is still valid, it remains surprisingly ignored in the *omics* field. Hence, is it possible to extend Linus Pauling's principle to large-scale mass spectrometry data?
One particular attempt to do that involved establishing a model of the so-called proteotype, which basically describes the acute state of the proteome in a cell, all its components, their organization and inter-connections with each other [Aebersold et al., 2016] (**Figure 1.7B**). The given model has a number of properties that should be considered: (1) the proteotype, hence its composition and organization, is the result of complex processes and multiple layers of regulation that remain poorly understood (i.e. transcriptional & translational control, RNA interference, micro-mRNA modulation, phosphorylation, ubiquitination, etc.). The cell, however, perfectly integrates all these layers of information, interpreting each control level and generating a result entity. Incidentally, the resulting buffering of transcriptional variability renders the proteotype more stable than the analogous transcriptome. (2) The proteotype reflects the composite response of the cell to environmental perturbations. (3) The proteotype determines the biochemical state of a cell

and is therefore expected to define phenotypes, as postulated by Beadle & Tatum [Beadle and Tatum, 1941], as well as Linus Pauling. (4) Rather than mapping individual functions to gene products, analyzing the proteotype primarily centers on the question: How does the system react in case the function of a gene is impaired? (5) The proteotype can be represented at different levels of resolution which can be precisely measured if need be [Aebersold et al., 2016].
This informative entity that is fundamental to the translation of genotypic variability to the phenotype could be influenced by environmental cues, epigenetics or genetic information; however, the enormous task of fitting the proteotype into the genotype-to-phenotype model remains restricted to the identification and quantifications of individual proteins. Clearly, the task in the future encompasses the understanding of how the different proteins are connected and whether the level of connectivity is changed under certain conditions, e.g. due to genetic variation, copy number variation, environmental changes, etc. Using the proteotype as a model, several questions can be tackled to fully integrate it into our understanding of the cellular phenotype:

1) To what extent are cellular proteins organized in macromolecular structures?
2) To what extent is the genotype determining the proteotype?
3) Does the proteotype classify groups (biomarker)?
4) Does proteome context reveal information that is not apparent from conventional quantitative proteome measurements?
5) Do changes in the modular organization of the proteotype determine function and phenotype?

Ultimately, these fundamental questions are major incentives for the presented work here as well.

1.2.2. Optimizing the data matrix- technical aspects to be considered

An underlying necessity to answer the above questions involves data representation and defining technical parameters of the data matrix that needs to be interrogated [Röst et al., 2015]. Such a data matrix is typically derived from perturbed states of a biological specimen (columns usually corresponding to the samples), and contains quantification values of the measured variable (rows corresponding to the proteins). This data matrix would support correlative analyses, machine learning, etc. High reproducibility of MS-measurements as well as depth of identification and quantification are the foundation to such matrices; even quantifying a low number of proteins across several sample, significant biological results could be acquired [Röst et al., 2015]. What are the optical dimensions of the matrix though?

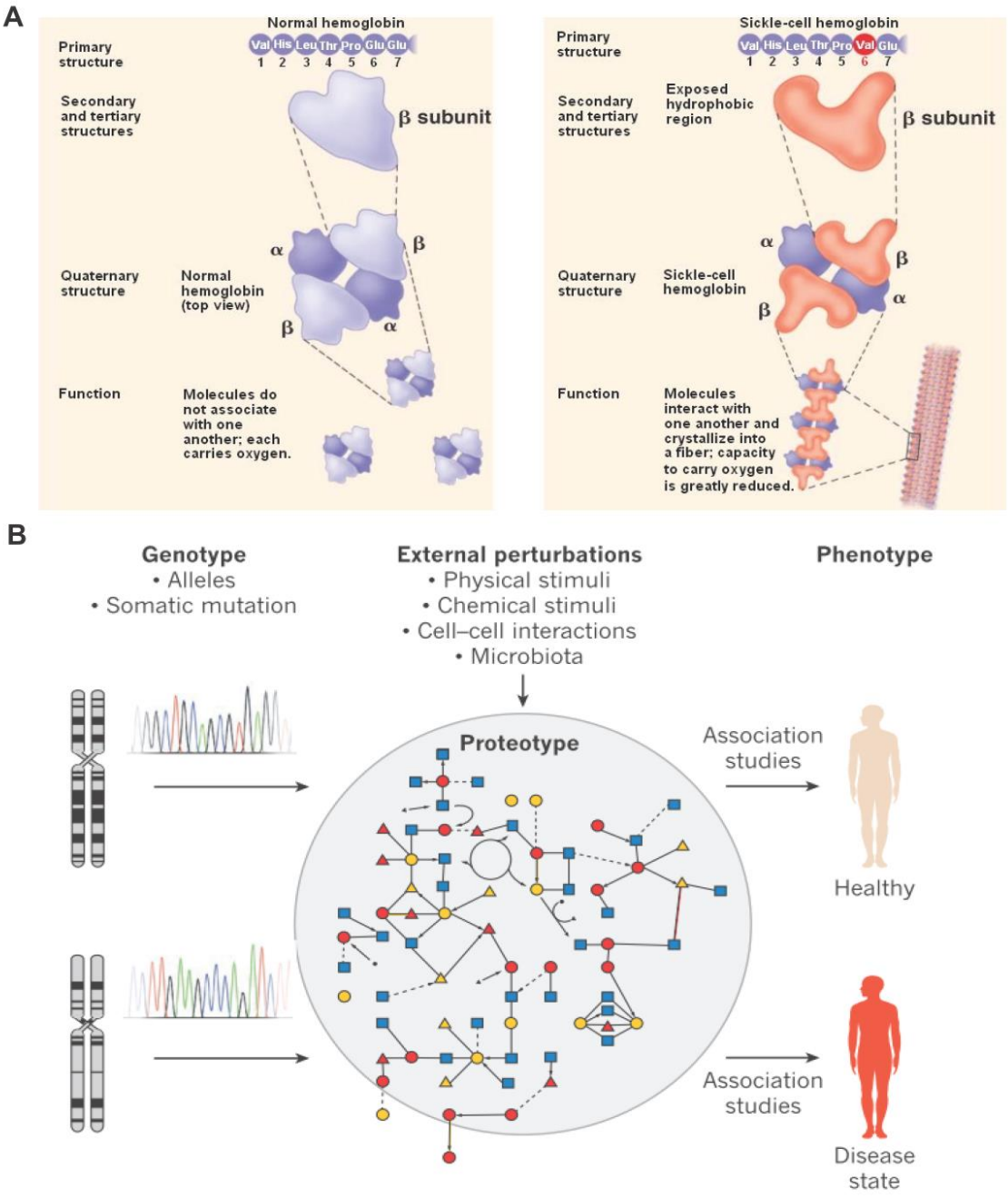


Figure 1.7. Extension of Linus Pauling's paradigm to the conceptualization of the proteotype. (a) Schematic illustration of Linus Pauling's paradigm, exemplified on sickle cell anemia. On the left-hand side, hemoglobin does not carry any mutations, and forms a functional quarternary structure. On the right-hand side, a missense mutation leads to an aberration of the tetrameric hemoglobin complex, ultimately resulting in fiber structures that give the blood cells the particular sickle-shape. Graph adapted from http://bio1151b.nicerweb.net/Locked/media/ch05/hemoglobin-sickle.html. (b) Schematic illustration of the proteotype in connection with potential phenotypes of an individual. The proteotype is shown as a network of proteins (colored shapes). The proteotype is influenced by both genotype (somatic mutations, alleles…) and external perturbations, such as physical or chemical stimuli, cell-cell interactions or the microbiota. The phenotype is then determined by the respective proteotype of an individual which can be established by association studies. To discover associations in a reliable manner, protein abundances need to be quantified across a large number of patients with qualitative mass-spectrometry techniques. Graph adapted from Aebersold et al. (2016).

Should we spend more effort increasing the depth of the sample or rather the sample number? Furthermore, what is the most informative subset of proteins in the matrix? Do all proteins in the data matrix have the same biological information? Can we infer information about proteotype organization from such a data matrix alone? If given the possibility of applying questions from GWAS studies [Lappalainen et al., 2013] to a data matrix derived from proteome measurements, it could indeed render this line of research powerful for (1) biomarker exploration, (2) GWAS association studies, (3) identification of clusters of proteins behaving similarly across conditions, (4) network inference, (5) population-based molecular biology, and (6) support of mechanistic models. Most studies on biomarker exploration, for example, have a very limited sample cohort, as well as a limited number of proteins; clinical specimen are unique, irreplaceable and small in number as well [Drucker et al., 2013]. It would thus be of high relevance to understand how biological significance is impacted by matrix dimensions.
An unpublished investigation on this matter by Ting Huang, Olga Vitek et al. gives some insight by determining factors that maximize prediction power of biomarkers, given a matrix of 70 proteins quantified in 200 subjects. Their preliminary study essentially found that the addition of more proteins reduces the predictive accuracy of markers since unfavorable noise behavior might dilute the marker signal (or predictive proteins). More subjects, on the other hand, increased the accuracy and sensitivity. It is therefore suggested to optimize for the number of available samples, and high reproducibility of a low number of proteins.

1.2.3. Proteotype Variation and what it depends on

The next question regards the information content of each protein: Arguably some proteins are more determined by environmental cues, by ageing, or genetic determinants. Studies so far have not quantified the influence of each of these determinants on the proteotype but provided a patchy network of dependencies, and relative contributions. There is one study, however, that should be highlighted as it does attempt to mathematically de-convolute the sources of variability of protein abundances based on plasma proteome maps of a human twin cohort [Liu et al., 2015]. By defining general linear-mixed effects models with genetic contributions, common environment of the twins, their individual environments, longitudinal effects (age), and non-biological effects, they assessed for a total of 342 plasma-derived proteins how their overall variability can be partitioned into different fractions according of those effects. For around 100 proteins they could observe a strong genetic impact on abundance variability (>20% of explained variance); yet other proteins were primarily influenced by the environment in their abundance (**Figure 1.8**).
To what extent is the knowledge of the heritability landscape useful? The authors observe that over time the genetic control on protein abundance actually decreases, hence the

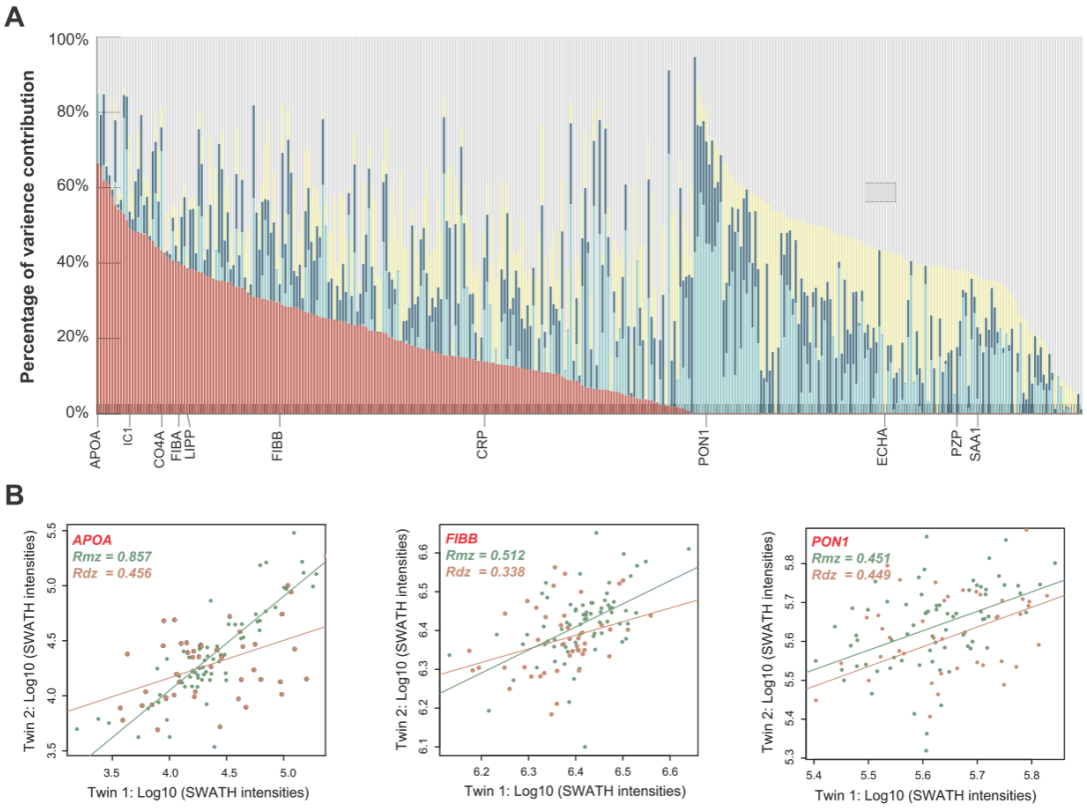


Figure 1.8. Heritability landscape of 342 human plasma proteins. (a) Histogram showing contribution of biological components to protein abundance variation (red: heritability; light blue: common environment; dark blue: individual environment; yellow: longitudinal effects; gray: unexplained fraction). (b) Examples of clinically assayed proteins with heritability are shown. Graph has been adapted from Liu et al., 2015.

longitudinal vector (age) changes the contribution of the genetic component to protein abundances. Such an observation can be crucial for our understanding and validation of biomarkers. More specifically, around 302,423 papers have been published with the keyword "protein biomarker" (257,568 papers with the keyword "protein biomarkers"), yet only a few markers make it through clinical validation [Drucker et al., 2013]. Liu et al. (2015) demonstrate that postulated marker proteins might have a very strong longitudinal component, meaning that these proteins are highly variable between people of different ages and environmental conditions (**Figure 1.9A**). It is hypothesized that these marker proteins could have been postulated as such due to under-sampling and not taking into account the longitudinal component. In contrast to that, FDA-cleared or approved plasma proteins tend to be more stable over time and exhibit stronger genetic components (**Figure 1.9B**). The identification of biomarkers should thus be correlated with high genetic heritability, decreased longitudinal effects and low variability in the cohort. Mapping sources of variability to protein abundances is therefore of high relevance, and should be taken into account for e.g. biomarker research.

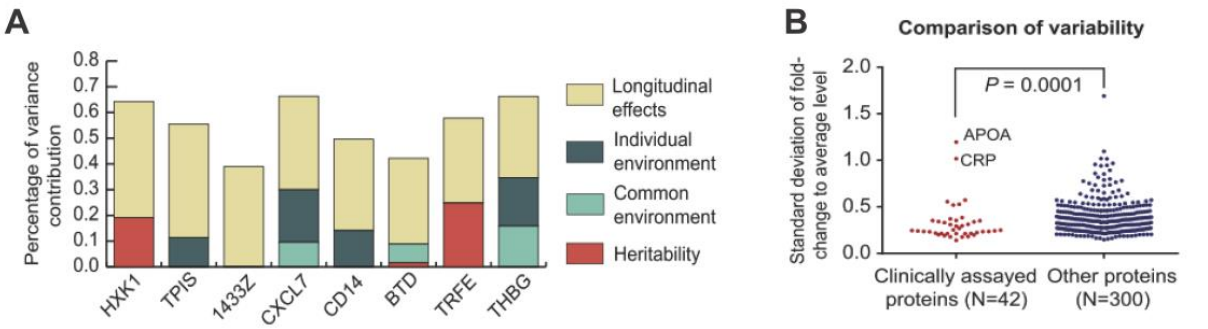


Figure 1.9. Biomarker analysis in twin proteomic dataset. (a) Contribution of different variance sources to abundance levels of reported protein biomarker candidates. (b) The variability of clinically assayed proteins tends to generally lower than the quantitative variability of other plasma proteins. Graph has been adapted from Liu et al., 2015.

1.2.4. Relevance of Proteotype for Personalized Medicine

Latter section is important in context of personalized medicine which has evolved as a major concept in recent years, and which essentially tries to optimize medical treatment for patients according to their genomic features [Ashley, 2016]. Since the costs and time efforts for sequencing an entire human genome dramatically decreased in recent years, the exploration of a patient's genetic mutations has fueled the design of drugs that would target a specific molecular alteration. Despite a number of successful tailored treatments (Gleevec-secondary drug combination in leukemia patients, Hochhaus et al., 2017), understanding the relationship between a patient's genetics and the ultimate medical treatment remains challenging. Biomarkers are being currently tested not only at the DNA and RNA level, but also at the protein level as well; yet, our ability to stratify patients into clinical target groups based on protein biomarkers is limited [Ashley, 2016]. One particularly striking aspect concerns the stratification of male and female organisms based on proteome data: Various studies have reported protein abundance variation due to the sex of an organism, but remain restricted to chromosome X/Y-specific protein expression rather than the systemic differences in the proteotypic pattern that latter could entail [Wu et al., 2013; Kukurba et al., 2016]. Exploring gender differences of the proteome is pivotal for our eventual understanding of human clinical phenotypes that often are sexually dimorphic. Apart from obvious anatomic differences, several studies have pointed towards diseases being more prevalent or severe with one or the other sex, such as autoimmune disorders [Whitacre, 2001], cancer susceptibility [Naugler et al., 2007], cardiovascular [Mendelsohn and Karas, 2005], and psychiatric diseases [Pigott, 1999; Hankin and Abramson, 2001]. Although the GWAS (genome-wide association studies) community is catching up on elucidating the role of the X-chromosome in the heritability of such phenotypes [Chang et al., 2014; Tukianen et al., 2014], sex-specific genetics needs to be further interrogated at the

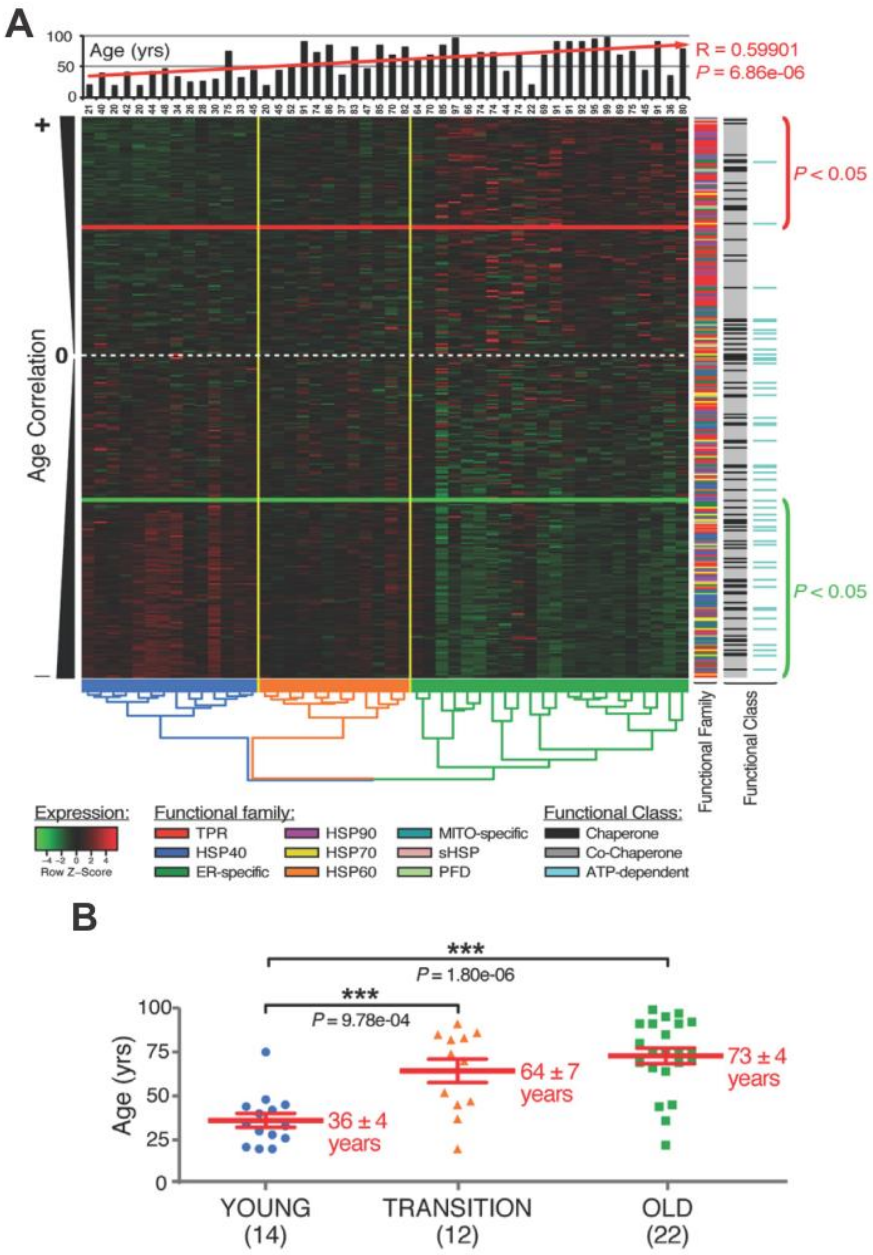


Figure 1.10. 'Chaperome' decline with ageing. (a) The heatmap illustrates 318 chaperones (y-axis) expressed in the human brain, ordered by declining correlation with the chronological age of the specimen. Transcripts above the red horizontal line get up-regulated upon ageing ($p < 0.05$); transcripts below the green horizontal line get down-regulated upon ageing ($p < 0.05$). Dendrogram based on hierarchical clustering of brain specimens. (b) Three groups deduced from hierarchical clustering in (a) are visualized in respective color code. The figure has been adapted from Brehme et al., 2014.

proteome level. In **Chapter 2** we will provide one of the first attempts of such a survey using a proteomics data on 94 male and 98 female mice of different genetic backgrounds [Chick et al., 2016], and describe protein modules that exhibit sex-specific abundances as well as complex stoichiometries. The underlying idea would be to recognize protein co-

variation patterns as potential bio-markers that could be used in context of a personalized treatment plan in the long run as well. The stratification of patients into male and female is just the first step towards that goal.

Another essential step involves the ability to stratify patient's proteotypes by age: The proteome is known to undergo several crucial re-arrangements as the organism grows older, particularly with regard to proteostasis. From the unfolded precursor state proteins are folded with the help of chaperones into their most efficient and functional structure; if a misfolding event occurs (e.g. due to mutation), there are degradation mechanisms in place to prevent misfolded protein species from aggregating [Houck et al., 2012]. Given that protein aggregates have profound consequences on cellular and organismal health, a major challenge for each cell is to maintain that subtle balance of protein folding in response to numerous signaling inputs, including heat shock responses, oxidative stress, and calcium signaling [Park et al., 2013; Yu et al., 2014]. The aging of an organism, however, is associated with a decline of that proteostasis capacity [Ben-Zvi et al., 2009], which a recent publication by Brehme et al. (2014) has connected to a decline of chaperones (**Figure 1.10**). Specifically the authors pinpoint at a 'core chaperome' of 21 genes decreasing in their overall expression in the human brain. The finding was further complemented with observations of an increased propensity of aggregate structures as well. While the mechanistic link still needs to be further disentangled, it is clear that a disbalance in the proteome leading to such an accumulation of misfolded species, represents the basis of cellular dysfunction [Douglas and Dillin, 2010]. Incidentally, the authors also made the distinction between chronological and 'proteostatic' age based on the level of chaperome-decline they observe, since the system is affected by an individual's genetic and environmental background [Brehme et al., 2014].

Another proteotypic hallmark of ageing concerns advanced glycation end products (AGEs) which are proteins that become glycated due to increased sugar levels in the blood. Once a critical threshold of glycated proteins in the blood is reached, AGEs can induce crosslinking to other proteins (e.g. collagen), thereby promoting vascular stiffening [Semba et al., 2009]. The presence of AGEs also leads to oxidized low–density lipoprotein (LDL) particles getting trapped in the artery walls, which could result in artherosclerosis [Prasad et al., 2012; Di Marco et al., 2013]. While the emergence of such products has been excessively linked to lifestyle [Uribarri et al., 2010; Vlassara, 2005], it could also be due to an inherent oncogenic re-wiring of the cellular metabolism. In tumor cells, for example, glucose metabolism get heavily up-regulated to sustain accelerated cell growth, and primarily gather energy from lactic acid fermentation in the presence of abundant oxygen rather than from mitochondrial oxidative phosphorylation. This process has been described as the Warburg effect [Vander Heiden et al., 2009] (**Figure 1.11**).

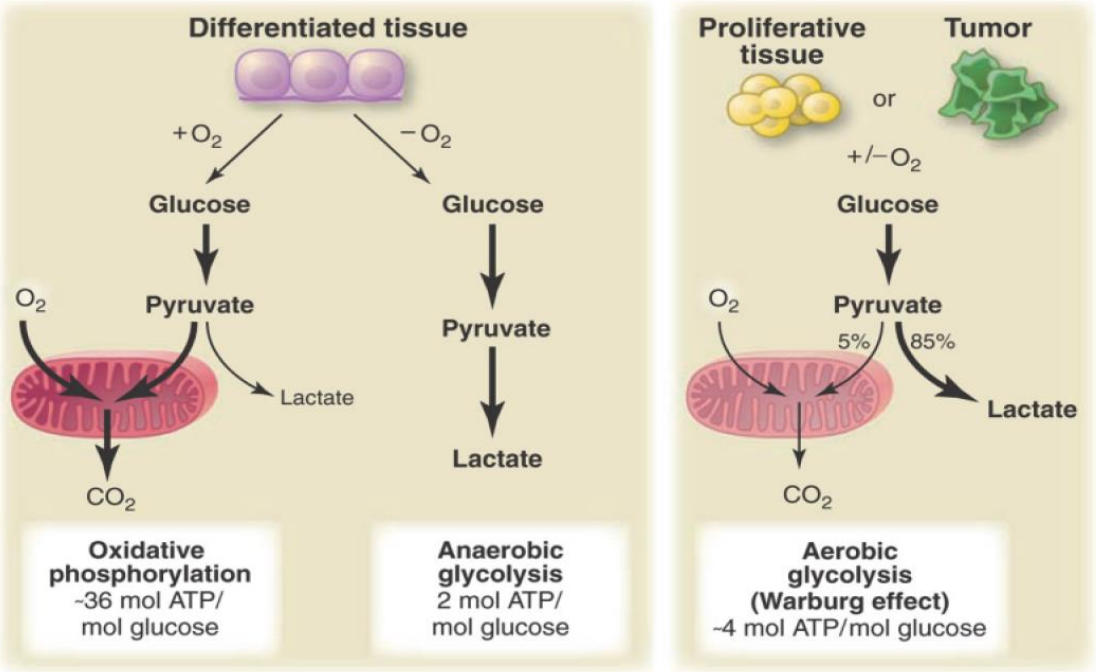


Figure 1.11. Scheme illustrating difference between oxidative phosphorylation, anaerobic glycolysis, and aerobic glycolysis (Warburg effect). (*Left*) Non-proliferating tissues metabolize glucose to pyruvate and completely oxidize pyruvate in the mitochondria to CO_2 (oxidative phosphorylation). In case of limited oxygen, pyruvate is converted into lactate (anaerobic glycolysis), yielding less ATP than in oxidative phosphorylation. (*Right*) Cancer cells convert most glucose to lactate regardless of oxygen presence (aerobic glycolysis), which is known as the Warburg effect. This process is less efficient than oxidative phosphorylation, but allows to redirect glucose into biosynthetic pathways. Graph adapted from Vander Heiden et al., 2010.

Such an aerobic glycolysis does not yield many more ATPs than its oxidative counterpart, but generates many additional metabolites, which in excess leads to the phenomenon of AGEs [Liberti et al., 2016]. Despite numerous studies on the Warburg effect, however, the precise causes for that phenomenon remain elusive. It could be a consequence of mitochondrial damage (during oncogenic progression) [Unwin et al., 2003], an adjustment to low-oxygen environments, or simply a consequence of cell proliferation that requires enhanced generation of metabolic building blocks derived from glycolysis [Lopez-Lazaro, 2008]. Some reports point towards over- expression of specific glycolytic enzymes to maintain high anaerobic glycolytic rates, such as for example the hexokinase enzyme [Bustamante et al., 1977], and the M2 splice isoform of pyruvate kinase [Unwin et al., 2003; Christofk et al., 2008]. Thus, it could be of ultimate use to monitor the abundance levels of a very specific sub-proteome of a human patient, namely proteins involved in glycolysis, to understand whether a potential Warburg effect could indeed arise from the given stoichiometric constellation, and target those specific enzymes to reduce their activity or

kinetic rates. **Chapter 4** is entirely dedicated to exploring the proteotype of young and old people in context of such metabolic rewiring.
One key question with regard to the usefulness of the proteotype as a model to establish its association with an eventual phenotype concerns the inference of metabolic flux [Madhukar et al., 2015]. It is assumed that absolute protein concentration- if accurately measured- gives an understanding of enzyme kinetics and the rate of metabolic processing as latter is (usually) correlated with enzyme abundance [Northrop et al., 1998]. It has been shown, on the other hand, that an increase in metabolic flux does not necessarily entail a shift in enzyme abundance, but that it is rather by means of allosteric control mechanisms that metabolic throughput is regulated, such as in carbon metabolism [Schwender et al., 2015]. Most studies, however, remain restricted to assessing the correlation of transcriptional variation of metabolic genes with the given flux, primarily in context of oncogenic transformations [Hu et al., 2013; Metallo et al., 2013; Deberardinis et al., 2008; Schulze et al., 2012]. In **Chapter 4** of this work, we will derive flux changes from robust measurements of protein abundances during the ageing process across different cell types of the human bone marrow. This particular aspect, however, remains to be further substantiated using actual metabolite measurements.

1.2.5. To what extent is the Genotype determining the Proteotype?

In the previous sections of this work we have established the necessity to understand how much of a given phenotypic variance is due to heritability (or explained by the genotype of an individual), and to what extent it is determined by environmental variability (section 1.2.3). We have also pinpointed at its relevance in context of personalized medicine to design more effective therapies for diseases (section 1.2.4). The underlying statistical approaches, however, remain to be discussed to understand how the proteotype model could be incorporated into established association-based methods. Today's computational approaches involve (A) linear models of the phenotype that use step-wise regression, (B) test-statistics for discovering so-called quantitative trait loci (QTLs), hence markers that are associated with a given quantitative trait, and (3) measurement of narrow sense (h^2) and broad sense (H^2) heritability, as well as environmental variance. While broad sense heritability describes phenotypic prediction by optimal arbitrary models, thereby revealing the actual complex molecular mechanism, narrow sense heritability describes the phenotypic prediction by a linear model and is thus efficient for genetic mapping studies [Vissher et al., 2008]. Hence, the phenotype is described as following:

$$p_i = f(g_i) + e_i$$

$i = individual\ [1 \dots N]$
$g_i = genotype\ of\ an\ individual$
$p_i = quantitative\ phenotype\ of an\ invidual\ i\ (single\ trait)$
$e_i = environmental\ contribution\ to\ p_i$

The simplest model for the function f is a linear function, and would effectively describe the narrow sense heritability:

$$f_a(g_i) = \sum_{j \in QTL} \beta_j g_{ij} + \beta_0$$

$i = individual\ [1 \dots N]$
$g_{ij} = marker\ j\ for\ individual\ i\ with\ values\ \{0,1\}$
$\beta_j = effect\ size\ for\ marker\ j$
$f_a(g_i) = additive\ model\ of\ genotypic\ components\ in\ g_i$

By restraining analysis to calculating the narrow sense heritability only in a standard QTL-approach, it has been found that the heritability h^2 for the human height is around 0.8 [Yang et al., 2010], whereas fitness traits, such as a wild animal life history, is around 0.3 [Gagliardi et al., 2010]. Thus, there are certain traits that have a higher genetic component than other phenotypic traits. It needs to be highlighted at this point that narrow sense heritability- while being useful- only explains a fraction of phenotypic variance by the additive model of markers $f_a(g_i)$. The difference between broad sense and narrow sense heritability then defines the fraction of so-called 'missing heritability'. Latter could be due to several factors, including incorrect heritability estimates, non-chromosomal elements, rare variants, structural variants, many common variants of low effect, and epistasis [Bloom et al., 2013]. In case of epistatic effects, hence actual gene-gene interactions, the effect on the phenotype would not necessarily pop up due to the linearity assumption. In the study conducted by Bloom et al. (2013), the authors examined interactions of significant QTLs with all other genes to reduce the burden on statistical power. They could observe that for most traits, missing heritability was not explained by pairwise interactions; yet for some traits, such as growth capability on maltose medium, missing heritability was explained up to 71% by pairwise interaction only. Given such findings, it would be interesting to see whether protein modules that are generated downstream of translation do have that information encoded in a similar way, with their respective genes interacting at the level of such trait-association analyses.

One of the most common QTL analyses is the identification of expression QTLs (eQTLs) with alleles explaining the variance in gene expression. Using microarray or RNA sequencing (RNA-seq), the mRNA levels of an individual are obtained and further compared to the

respective genotype of the individual (SNP genotyping or whole genome sequencing). By means of statistical methods, i.e. linear regressions, loci significantly contributing to RNA expression levels (eQTL) are identified, and quantified (effect size). Such mappings have been conducted in a large-scale manner across 53 human tissues (>10.000 samples), yielding up to 30.000 significant eQTLs [Carithers et al., 2015; Ward and Gilad, 2017]. Undoubtedly, the GTEx database provides a rich source of information required to bridge the genotype-phenotype gap and to interpret the findings from genome-wide association studies (GWAS) on various diseases [Lappalainen et al., 2013].

A fundamental caveat of all QTL-studies, however, is their inability to pinpoint at the precise biological mechanism that leads to a given association [Pearson et al., 2008]. A study by Battle et al. (2015), for example, suggested that only 65% of eQTLs get buffered, as their effects do not further propagate to downstream protein abundance levels. That finding prompts the idea of defining other types of QTLs that could explain variation at different regulatory levels (**Figure 1.12**), such as at the level of histone modifications (hQTLs) [Grubert et al., 2015], methylation (mQTLs) [Banovich et al., 2014], protein abundances (pQTL) [Melzer et al., 2008; Suhre et al., 2017; Battle et al., 2015; Chick et al., 2016] and even metabolite levels (meQTL) [Wu et al., 2014; Williams et al., 2016]. Integrating these different layers may offer clues for deriving mechanisms underpinning a certain phenotypic trait, and assessing the predictive power of the loci identified. By means of the so-called mediation concept such a QTL analysis could be further extended to entire protein modules.

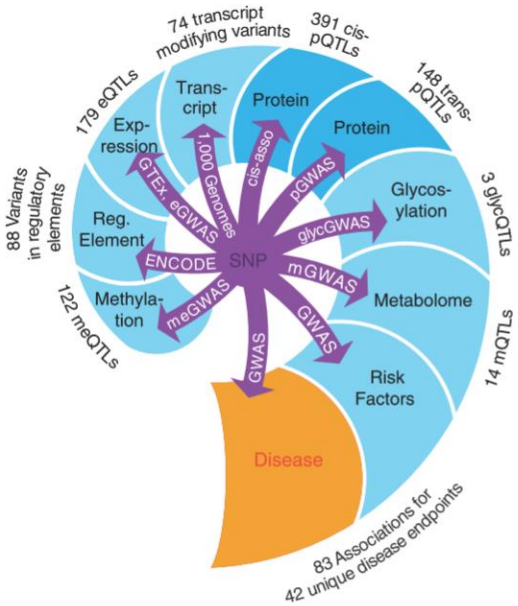


Figure 1.12. Overview on QTL interconnectivity. Data graph generated from http://proteomics.gwas.eu, illustrating number and type of inter-connectivity present in the field. Graph adapted from Suhre et al., 2017.

1.2.6. How do genetic effects propagate to protein modules?

Recent reports attempt to identify potential mechanisms of genetic control on protein abundances, and exploit the modular architecture of the proteotype [Chick et al., 2016; Goncalves et al., 2017; Roumeliotis et al., 2017; Ryan et al., 2017]. Protein complexes represent a particularly interesting set of modules as they exhibit a highly coordinated mode of regulation of their protein members. The stoichiometry of a complex needs to be rigorously maintained for it to function in the first place; however, the biological mechanism by which such a control is achieved remains elusive. Also it is not clear whether such a detection would apply to all protein complexes, and whether some complex members are more relevant than others during complex assembly. **Chapter 2** provides more insight into whether such a dissection is indeed possible by interrogating a large number of complexes for presence of stable and more variable subunits. Prior to that, however, it needs to be understood whether the stoichiometric maintenance is encoded at the genetic level, or an independent mechanism downstream of translation.

The work by Li et al. (2014), for example, indicates that *E.coli* has the rate of protein synthesis required for one particular complex encoded in one single operon. Using ribosome profiling and RNA-seq analysis the authors demonstrate that, while the mRNA levels do not show any stoichiometric balance whatsoever, the translation rate is reflective of the eventual stoichiometry of the complex (**Figure 1.13A**). The optimization of translation rates to achieve proportional synthesis in complexes is also demonstrated for yeast cells (**Figure 1.13B**); however, by what means the translation rates are adjusted is not clear. The strength of the Shine-Dalgarno site, as well as the local RNA structure does not account fully for the observed rates [Salis et al., 2009; Li et al., 2014]. Another constraint to the model established by the authors is the fact that it cannot be readily extrapolated to multi-cellular organisms. Genes coding for members of the same complex are not found to be particularly close to each other in their chromosomal location, and a similar ribosome profiling has not revealed any particular tuning of the translation rates [Ignolia et al., 2017; Stefely et al., 2016]. Instead, there have been proposals on (a) so-called 'mediation' events based on dependent genetic and transcriptional regulation of subunits of the same complex [Chick et al., 2016], and (b) attenuation of complex subunits by degradation of excessively produced proteins [McShane et al., 2016]. Although these two hypotheses are not mutually exclusive, they are characterized by very distinct biological mechanisms.

The mediation hypothesis (a) is an elegant means to link identified quantitative trait loci (QTL) to entire complexes or modules. Given a pQTL, it is assumed that it affects its local transcript or protein intermediate, which in turn influences the abundance of (a) distant protein(s); thus the pQTL would act in *trans*. Chick et al. (2016) illustrate the concept with the chaperonin-containing TCP1 (CCT) complex (**Figure 1.13C**), where all complex

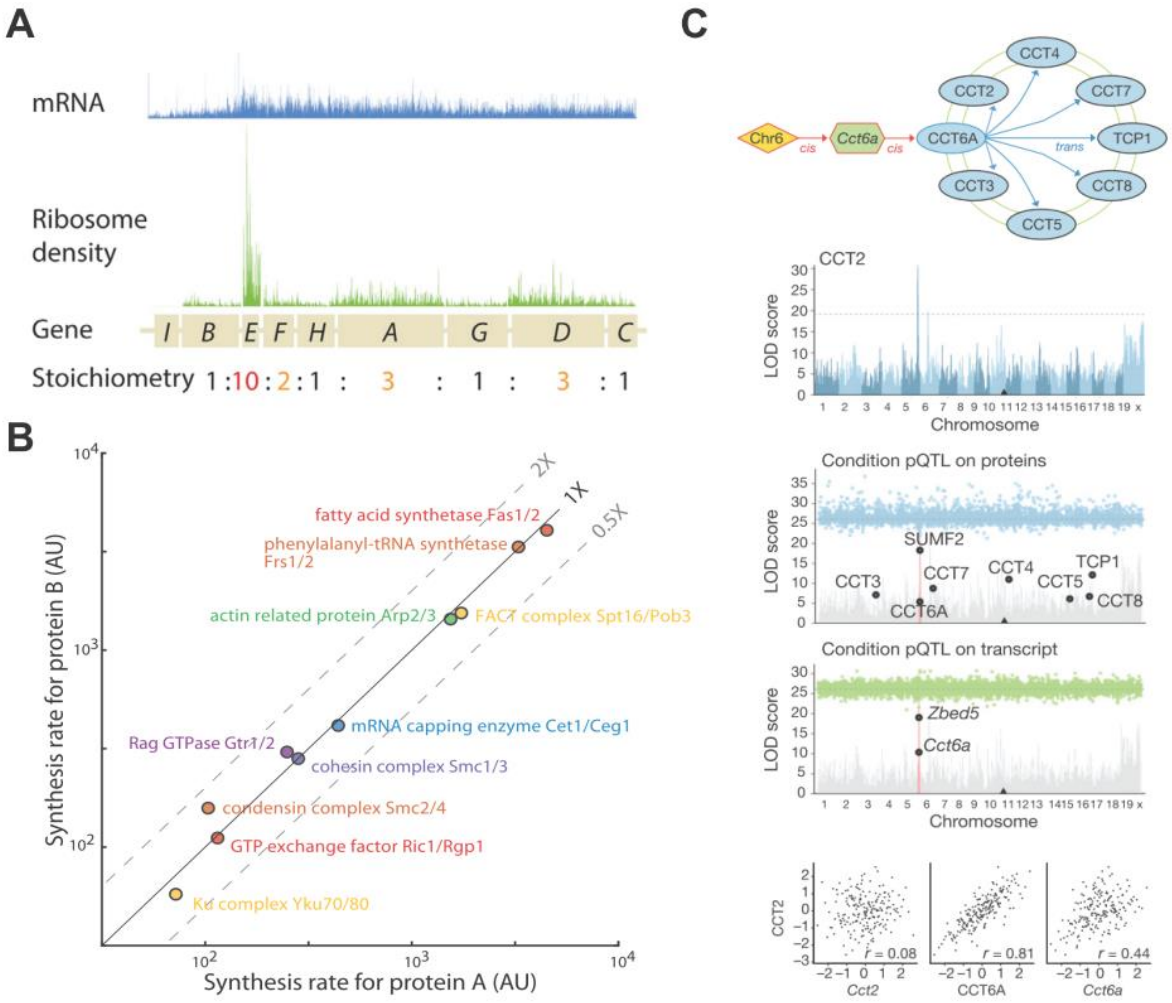


Figure 1.13. Models of complex stoichiometry maintenance by adjustment of translation rates and protein mediation. (a) Polycistronic mRNA (blue) expressing all subunits of F0/F1 ATP synthase, with translation rates (green) adjusted to stoichiometry of complex (data from ribosome profiling). (b) Heterodimeric complexes in *S.cerevisiae* are synthesized proportionally. Both graphs (a) and (b) were adapted from Li et al., 2014. (c) (From top to bottom) (1) Scheme illustrating mediation on chaperonin complex. (2) Variant on chromosome 5 shows distant effect on CCT2 abundance. (3) All members of chaperonin containing Tcp1 (CCT) complex are affected by latter variant on chromosome 5. (4) Variants acts proximally to Cct6a and acts on it at the transcript level. (5) Protein abundance of CCT2 is correlated with CCT6A protein and Cct6a transcript. Graph has been adapted from Chick et al., 2016.

members share a distant pQTL. Latter impacts the transcript and protein abundance of its local gene *Cct6a*, which is then further propagated to the other complex members through post-transcriptional mechanisms. It can thus be assumed that (physical) properties of CCT6A renders it a suitable mediator protein for the complex (i.e. scaffold protein), and that in case of perturbation in its abundance, it becomes more difficult to achieve the required stoichiometric balance. Chick et al. (2016) was one of the first studies to present a

comprehensive analysis on such mediatory events, followed by efforts from Roumeliotis et al. (2017) in colorectal cancer cell lines, where they demonstrated potential mediators in the BAF complex. Such studies remain scattered, however, and the power of mediation analysis still needs to be leveraged for all complexes and protein modules. If integrated with other resources on complex assembly [Levy et al., 2007], the bio-physical properties of discovered mediators could hint at potential progressions during assembly and disassembly of protein modules in general. Additionally, it needs to be further incorporated into our current understanding of the degradation mechanisms (b).

Studies of aneuploidy in yeast and human cells have revealed that the majority of autosomal gene duplications is reflected at the protein level as well; only subunits of multi-protein complex showed significant attenuation in their protein abundance levels [Stingele et al., 2012; Dephoure et al., 2014]. That finding has been also recovered in CPTAC and TCGA cancer panels by Goncalves et al. (2017) in context of copy number variations (CNVs). Generally, the observation further reinforces the hypothesis of excessive production of protein complex subunits and subsequent degradation of subunits that have not managed to get incorporated into their respective complex [Goldberg and Dice, 1974; Abovich et al., 1985]. One substantial piece of evidence for latter mechanism was provided by McShane et al. (2016), who investigated the degradation rates of young and old proteins by means of a proteome-wide pulse-chase SILAC set-up [Selbach et al., 2008]. Two distinct models were characterized, one being the exponential decay (ED) and the other being the non-exponential decay (NED), which could either coincide with age-dependent stabilization or de-stabilization of the protein. Using azidohomoalanine (AHA) in the media for further dissection into newly synthesized proteins ('young') and 'old' proteins, the authors could classify proteins according to their degradation rates. Around 50% were clearly following an exponential decay, and 10% were classified as non-exponentially degraded proteins (NED proteins) with most of them exhibiting an age-dependent stabilization. Those unbiasedly identified NED proteins were significantly enriched in complex-associated proteins (**Figure 1.14**); moreover, the authors reported an increase in the fraction of NEDs once protein complex formation is inhibited. Further testing in different human cell lines confirmed that the production of complex members in super-stoichiometric amounts is indeed a common and evolutionary conserved process with vital implications in protein complex formation, gene expression control and aneuploidy [McShane et al., 2016].

The model does however raise some issues regarding the necessity for the cell to do so, as such a mechanism is clearly not optimized and draining vital cellular energy resources. Several explanations are offered: (1) Proteins with exponential decay (ED) showed a preference of containing disordered structures, which could be harmful to the cell due to the aggregation propensity of such proteins [Vavouri et al., 2009]. Producing excess NED proteins could prevent such deleterious effects. (2) ED proteins could actually represent the

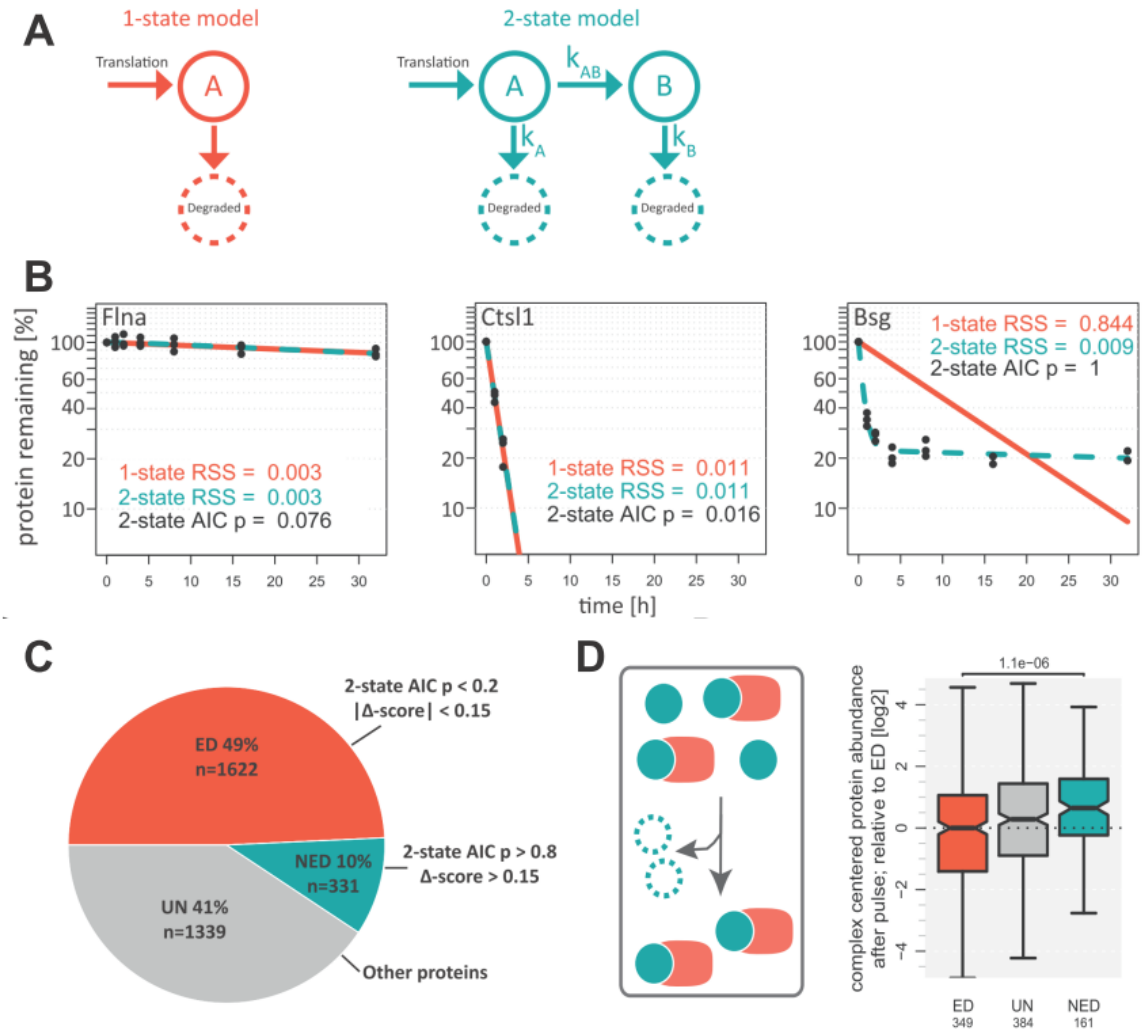


Figure 1.14. Dissection of proteins according to exponential and non-exponential degradation (graphs adapted from McShane et al., 2016). (a) Schematic representation of 1-state model (i.e. exponential decay), and 2-state model (non-exponential decay) using Marcov-chain models. (b) Examples of proteins fitting into models from (a). (c) Fraction of proteome degraded by either of the described modes (ED, NED) and undefined ones. (d) (Left) Simple model explaining non-exponential degradation in context of heteromeric complexes (NED=turquoise, ED=red). (Right) Production of NED proteins in super-stoichiometric amounts. Complex-normalized protein abundances are shown on the y-axis.

limiting factors (or 'mediators' as discussed above) that ultimately determine total protein complex abundance. It would be thus interesting to see whether known *trans*-pQTLs could be located at *cis*- sites of such ED proteins. (3) The over-production of NEDs might optimize for efficient complex assembly [Marsh et al., 2013]. It is possible that any of the presented theories applies to proteins to a different extent; it is equally possible that rather than interaction engagement, the particular localization of a protein or the emergence of a post-translational modification leads to its stabilization, and thus different degradation rate.

Incidentally, the mediation (a) theory and the super-stoichiometric production of complex-associated proteins (b) are not contradicting one another, but need to be further investigated to enlarge our understanding of how potential genetic mutations might be propagated to entire protein modules.

1.2.7. Functional consequences of modularity perturbations

More than 100.000 disease-causing mutations have been described in the OMIM database [Amberger et al., 2015]. Most of those mutations are missense-mutations and thus known to disrupt the overall structure of the protein, rendering it non-functional. On the other hand, there is also a proportion of mutations that induce an amino acid change in a disordered region (loops, or generally unfolded regions) and thus would not directly contribute to any obvious structural change. It could, however, affect a PTM-site, a protein's overall stability, or a protein's capability to engage in interactions. Meyer et al. (2017) explore latter hypothesis by assuming that a short linear motif located in the disordered mutation might be disrupted or gained due to a mutation, ultimately altering the interaction behavior. In an elaborate pull-down screening using wildtype and mutant synthetic peptides, the capability of all proteins to interact with those is interrogated using a SILAC set-up [Meyer et al., 2017]. One of the primary findings of the authors involves the identification of 619 specific interactions, with 180 being differential due to disease-causing mutations, either caused by loss or gain of interaction. One particular aspect concerned an enhanced interaction with clathrin proteins due to the formation of a dileucine motif (proline converted to leucine) that would not be present under wildtype conditions. Such a particular mutation in the disordered cytosolic tail of the glucose transporter GLUT1 effectively leads to clathrin-dependent endocytosis and mis-trafficking of GLUT1, ultimately resulting in the so-called GLUT1 deficiency syndrome [Lee et al., 2015]. The authors postulate that such dileucine motif gains in disordered cytosolic tails are significantly and specially linked to diseases (dileucineopathies). The findings are especially important in light of the mutated protein actually being functional; the phenotype is indeed only brought about by it being at the wrong location. We will further elaborate on the trafficking machinery and how it might be impacted by differential proteotypes of individuals in **Chapter 2**.

The above paragraph demonstrates all too well how mechanistic validation of mutations complements findings derived from GWAS and QTL-based studies. In the end, there are a number of ways a causal variant could affect protein function. It could modify gene expression by altering transcription factor binding sites [Mathelier et al., 2015; Melton et al., 2015; Kamanu et al., 2012; Martin et al., 1989] or by means of epigenetic mechanisms [Grubert et al., 2015]; it might as well affect physical properties of the gene product, such

as its structure and hence function [Azzollini et al., 2014], its stability [Studer et al., 2014; Prusiner et al., 1991; Collinge et al., 2001], or capability to get post-translationally modified [Wagih et al., 2015; Reimand et al., 2013]. From above we can deduce that disrupting short linear motifs (SLiM) could disrupt interactions between proteins [Muslin et al., 1996; Pandit et al., 2007], and effectively lead to mis-locations as well. Finally, mutations that affect protein quality control might affect an even larger fraction of proteins [Anczukow et al., 2008; Giannandrea et al., 2013; Amrani et al., 2006; Schloesser et al., 1991; Sun et al., 2017]. The proteotype model should therefore capture as many physical and physiological features of proteins as possible to understand the mechanism behind disease-causing mutations.

1.3 Other Features of the Proteotype- beyond Abundances…

The proteotype remains largely described in terms of abundance levels of individual proteins due to the straightforward protocols and interpretation of the MS-data. Measuring post-translational modifications, for example, renders issues usually more difficult, as there is a PTM-enrichment step involved (such as immobilized metal affinity chromatography (IMAC), and metal oxide affinity chromatography (MOAC)) which inevitably reduces the yield of the sample. Subsequent PTM searches are mostly based on expected fragment patterns, despite recent efforts to perform less biased MS-analyses using a mass-tolerant database to recover fragment with unknown modifications [Chick et al., 2015]. Other protein features are measured by other combinatorial MS-protocols: Protein interactions, for example, can be surveyed in an MS-setup using cross-linking reagents [Rinner et al., 2008]; kinetic rates (synthesis and degradation) can be monitored by a pulse-chase SILAC setup coupled to MS-analysis (as described in the previous section). Another pivotal characteristic of the protein is its stability…

1.3.1. Protein Stability

The function of a protein is primarily determined by its three-dimensional configuration. The sequence of amino acids is subject to electrostatic, van der Waals and hydrophobic forces between the atoms of the respective residues. By these forces the protein traverses through several fold states that can be characterized by the energy required to maintain and stabilize the entire structure. The eventual protein 3D-structure represents a stable and minimal energy structure (**Figure 1.15**), which coincides with the protein being fully functional. In order to effectively measure the thermodynamic stability of a protein, it is therefore necessary to monitor the unfolded versus the folded state of the protein. In terms

of biophysics, the stability of a protein can be calculated as the respective difference in the Gibbs free energy ΔG between the folded and unfolded state.

$$\Delta G = G_{folded} - G_{unfolded}$$
$$G = H - TS$$

G represents the Gibbs free energy, which is defined by the enthalpy (H), temperature (T) and the entropy (S) of the system (i.e. the protein). If ΔG is therefore negative, we can assume high stability of the corresponding protein.
Common experimental approaches to determine the stability of a protein usually entail denaturing the proteins from its native state using denaturing agents such as urea, guanidium chloride, or hot temperatures. Then one can measure ultraviolet light absorbency, fluorescence [Royer et al., 2006] or the catalytic activity [Daniel & Danson, 2013] of the denatured protein and compare it with the native one. Fluorescence might arise from the fact that hydrophobic aromatic residues (tyrosine, phenylalanine) get exposed when the protein gets denatured. Using such read-outs with different temperatures or other denaturing conditions allows to monitor how the protein unfolds and to calculate a protein's folding and unfolding rate. Their ratio then defines the equilibrium constant K_{eq} which can then in turn be used to calculate the aforementioned Gibbs energy ΔG.

$$K_{eq} = {k_{unfolding}}/{k_{folding}}$$
$$\Delta G = -RTlnK_{eq}$$

R represents the gas constant, T the temperature in kelvins.

The above described read-outs- while efficient- are limited to measuring the folding/unfolding behavior of one protein at a time.

1.3.2. Measuring protein thermal stabilities in a large-scale manner

A recently developed technology combines both the concept of monitoring the denaturation state of proteins upon heating with quantitative mass spectrometry and thereby provides a method to measure the thermal profiles of the entire proteome in a reasonable time frame [Savitski et al., 2014]. The technology can be applied to both cell extracts as well as living cells, which allows monitoring the effect on protein stabilities in the sample processing protocol, and assessing to what extent results derived from cell extracts are reflective of the

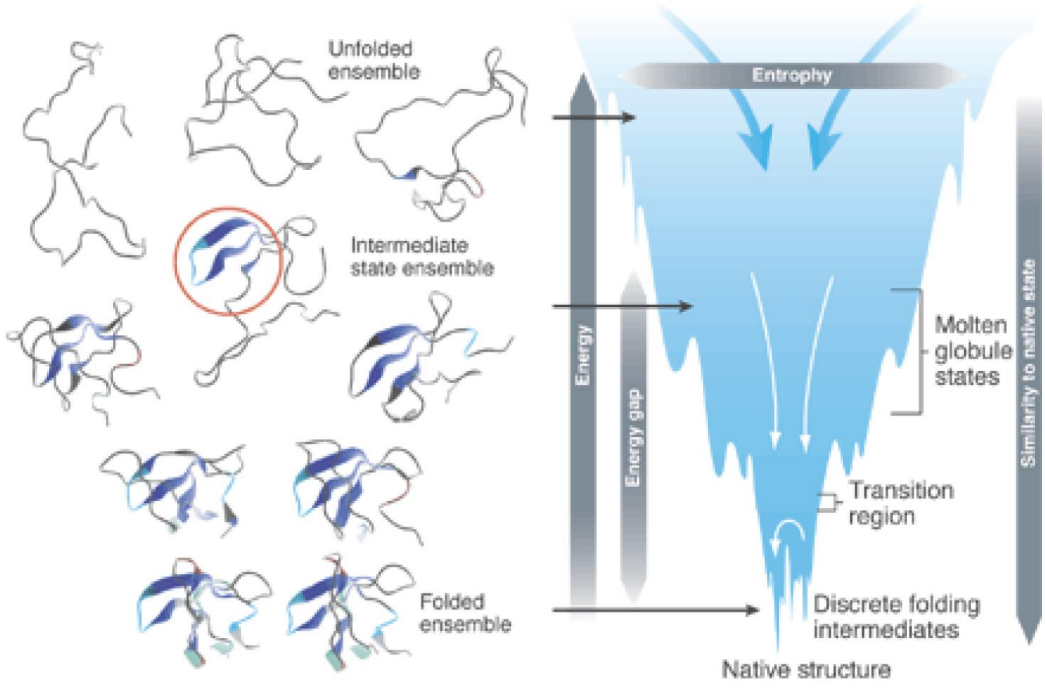


Figure 1.15. Energy landscape for protein folding (graph adapted from Brooks et al., 2001). While folding a protein organizes itself into intermediate structures (shown on the *left*) that are characterized by their enthalpy, and conformational entropy. Free energy is gained with appearing native interactions (*right*). The speed with which a protein transcends through the energy landscape is determined by the local topology of the landscape, relative to thermal energy.

protein stability pattern in living cells. The methodology entails heating cultured cells to 10 different temperatures thereby inducing denaturation of the entire proteome. Aggregates get removed, and the fraction of soluble proteins are extracted with a buffer and quantified by means of high-resolution mass spectrometry using TMT-10plex labeling for each respective temperature. Thus, one MS-run yields accurate thermal profiles or denaturation curves of the entire proteome in given cell culture (**Figure 1.16**). Clearly, this technology opens many avenues to understand how proteins traverse the energy landscape of possible stabilities under different conditions. One of the prime examples for an application described by Savitski et al. (2014) involves assessing drug effects on protein stabilities. For a palette of drugs it could thus be easily assessed which compound concentrations induce shifts in protein thermal profiles, and what proteins are directly or indirectly affected by compound addition.

At this point, the landscape of protein stabilities remains largely elusive even for very well-described processes such as the cell cycle. It is known that the cell cycle is accompanied by dramatic changes in the physico-chemical environment of proteins due to organelle

disassembly, etc. [Hernandez-Verdun et al., 2011]. While protein abundances might not necessarily be affected, protein stabilities and their solubility might very well be prone to dramatic change. In **Chapter 3** of the presented work, we will explore this matter more carefully, and present the power of the methodology exemplified on the cell cycle.
Undoubtedly, however, the range of applications is spanning a number of issues – from getting a better grasp on interaction-induced stabilities of protein modules and entire networks, to understanding the origin of proteopathic diseases characterized by protein aggregates, such as Alzheimer's disease [Irvine et al., 2008]. Such a cell-wide assessment of protein stabilities could indeed be integrated into the current understanding of the proteotype and also monitored in its association to genetic variation.

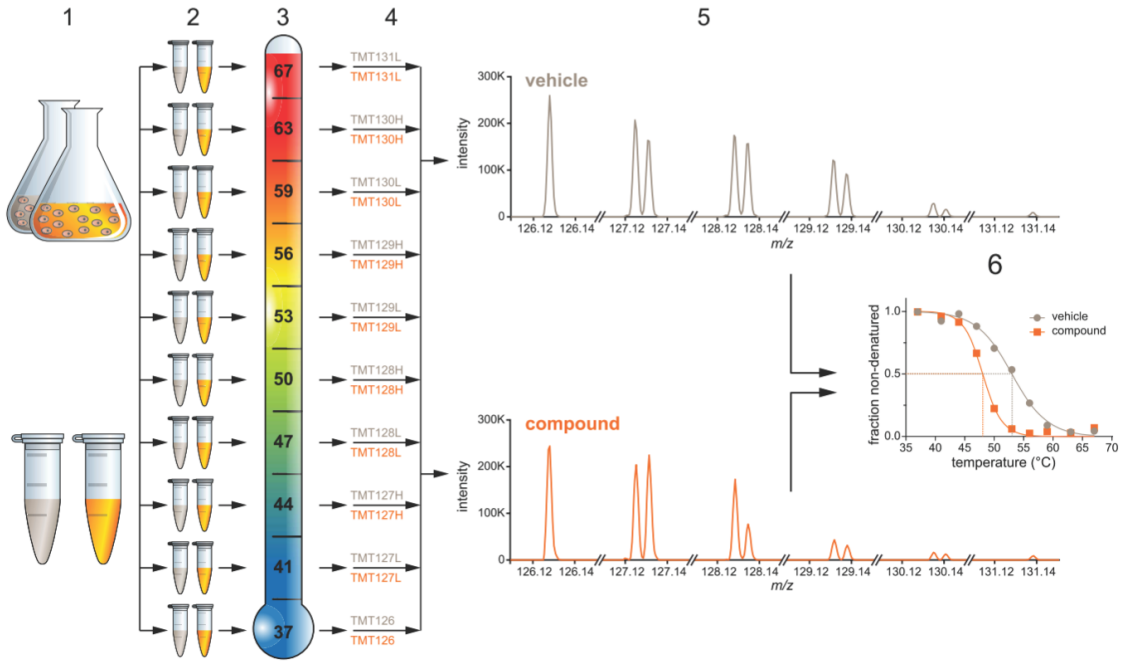


Figure 1.16. Quantitative profiling of protein thermal stability under different conditions in a proteome-wide manner. (1) Cells get cultured under different conditions, i.e. drug treatment. Alternatively, cells get extracted from respective cultures and get treated with drugs. (2) Samples get divided into 10 aliquots for each condition. (3) Aliquots get heated to indicated temperatures, and proteins are extracted using PBS, and further digested. (4) Peptides get labeled with different TMT-10 isotope tags. All samples from one condition get mixed and subjected to MS-analysis. (5) Reporter ion intensities are reported for each identified peptide/protein in the mass-spectrometry analysis. (6) Reporter ion intensities are used for fitting a melting curve and calculating the melting temperature T_m for each protein for each respective condition tested. The graph has been adapted from Savitski et al., 2014.

1.3.3. Could genetic variants influence protein stabilities?

A mutation leading to residue substitution might have a dramatic effect if it occurs at a critical position in the protein folding process, and therefore affect its function. The effect

of such a mutation could be measured as the difference in ΔG between the wildtype and the mutated condition, with values *>1.8* indicating high destabilization, and values *<-1.8* assuming further stability of the protein structure [Studer et al., 2014]. So far, many mutagenesis experiments have been conducted to assess the effect on stabilities, and summarized in databases like ProTherm [Bava et al., 2004], human genome mutation database (HGMD) [Stenson et al., 2014] and protein mutant database (PMD) [Nishikawa et al., 1994] which are manually curated. It would be of high relevance to overlay the available information from these resources with the unbiased analysis of thermal profiles. Latter could prove useful to see whether a mutation of one gene propagates to the stabilities of all other proteins interacting with its product, and whether this could represent a quick sensory mechanism for the cell to circumvent possible failures immediately. It is not surprising that disruption of protein stabilities could then have a wide range of clinical phenotypes associated with them, such as prion [Prusiner et al., 1991; Collinge et al., 2001], muscular [Boopathy et al., 2015] and neurodegenerative diseases [Lin et al., 2008].

1.3.4. Post-translational modifications as a feature of the Proteotype

Another representative characteristic to be taken into account when assessing the proteotype of a cell includes reversible chemical modifications on the proteins. These so-called post-translational modifications (PTMs) encompass a wide range of possible chemical groups that can get attached to certain amino acid residues of proteins, such as phosphorylation, ubiquitination, acetylation, methylation, sumoylation, glycosylation, etc. There are more than 200 types of PTMs that have been described so far [Minguez et al., 2012], with protein target sites being described in databases such as dbPTM [Lee et al., 2006] and PhosphoSitePlus [Hornbeck et al., 2015]. The addition of a chemical moiety to a protein site usually involves an active enzyme that recognizes a target site and attaches the modification (writer enzyme). On the other hand, another class of enzymes specialize in removing the modification if necessary (eraser enzyme). These mechanisms are pivotal for several processes that involve cellular signaling [Nusse, 2008], protein degradation [Wickner et al., 1999] and metabolism [Zhao et al., 2010], and affect protein structure and ultimately its function. It therefore comes as no surprise that mutations at residues prone to such modifications could be giving ground to disease phenotypes [Grasbon-Frodl et al., 2004]. A substitution T183A in prion protein, for example, prevents *N*-glycosylation which in turn is pivotal for the emergence of spongiform encephalitis [Grasbon-Frodl et al., 2004]. Mutations at PTM sites have generally been shown to be enriched in disease association [Reimand et al., 2015], and should be studied in context of genetic variation as well.

1.4 Aims of the Thesis

The above introduction has outlined particular aspects of the current proteomics research field, in context of establishing a systematic working model of the so-called proteotype to connect genotypic with phenotypic variability. The concept of the proteotype as a system involving networks of proteins is largely built on the principle of molecular cooperativity underlying cellular functionality in the first place. The presented work will interrogate protein connectivity in various contexts, in an effort to tackle the following questions:

Question 1: To what extent are cellular proteins organized in macro-molecular structures?

That question is one of the key focuses of **Chapter 2**, which basically demonstrates covariation of proteins that are part of the same module across human and mouse individuals, and conceptually dissects variation we observe for the abundances of entire modules as opposed to variation observed in the stoichiometric composition of modules. Such a dissection then allows to define core parts of protein modules, as well as auxiliary proteins that are more prone to variation across individuals relative to the core protein set of the corresponding module.

Question 2: What factors impact the variation in protein and module abundance and to what extent?

Chapter 2 further discusses for the first time the association of genetic and environmental factors to module abundance and stoichiometry. The sex of the animals in the studied cohort is taken as a proxy for genetic effects, whereas their diet conditions represent the measurable environmental impact. In this part of the work we established that the variability of the proteotype is indeed largely driven by variability of modular entities, such as protein complexes and pathways.

Question 3: Do changes in the modular organization of the proteotype have far-ranging consequences?

Finally, **Chapter 2** will additionally focus on the work conducted in collaboration with the Pepperkok group, elucidating the impact of stoichiometric adjustment of the COPI/COPII-transport machinery on receptor transport. The effort represents an interrogation of proteins on their biological and physico-chemical feature that can in theory be extended to any functional relationship between proteins in the cellular context. Therefore that particular chapter should be viewed as a potential resource and guideline for such investigations.

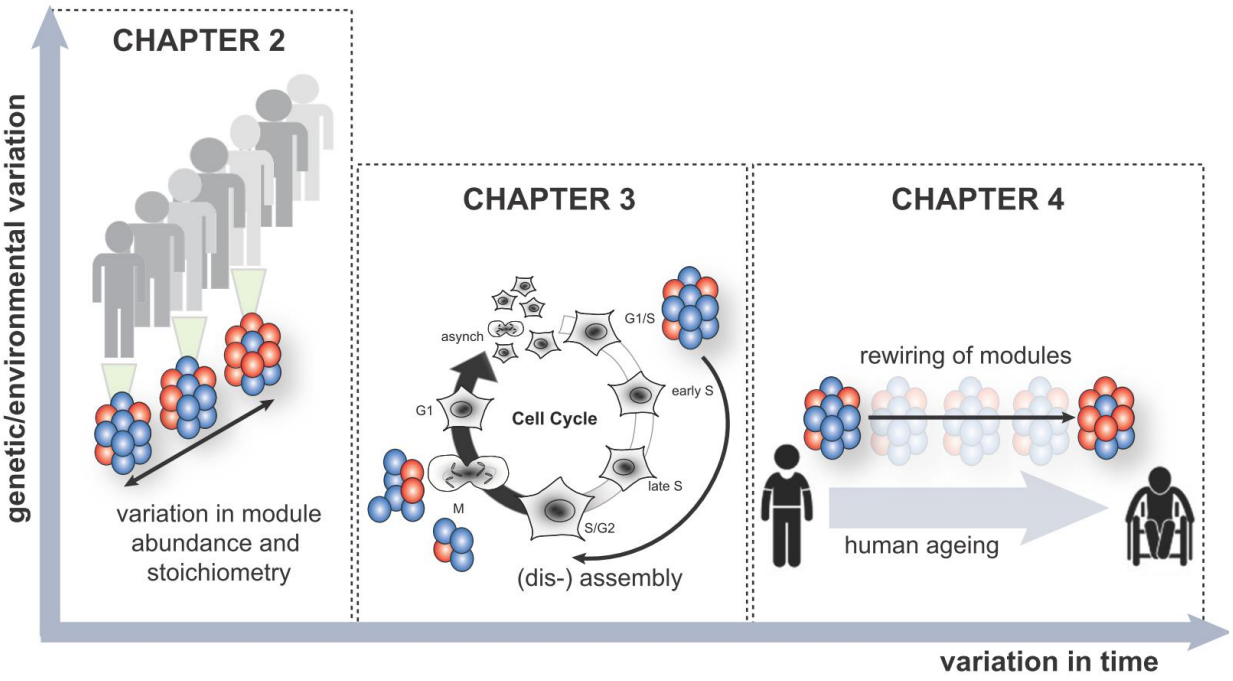


Figure 1.17. Schematic illustration of overall thesis structure and main themes. Functional modules are analyzed in context of their variation across individuals in Chapter 2, and subsequently in their temporal variation during cell cycle progression in Chapter 3 and human ageing in Chapter 4.

The second part of this work builds upon the concepts presented in **Chapter 2**, and examines the temporal variation of functional protein modules (**Figure 1.17**). We describe and assess to what extent modules are being re-organized on a short time scale, i.e. during cell cycle progression (**Chapter 3**), as well as during a long-term period of 40 years in humans (**Chapter 4**). Again there are specific questions to be tackled:

Question 4: What systematic changes in protein stability occur during the cell cycle? To what extent do they reflect molecular cooperativity?

Chapter 3 illustrates the modulation of protein networks and their dependencies on the cell cycle periodicity, surveying not only mere abundances of proteins, but also capturing the thermostability of proteins as well as their solubility. The data, which has been generated by in the Savitski lab, represents a valuable resource in that it combines such different metrics and protein properties. My contribution, on this part, has been the observation of correlated thermostability of proteins associated in complexes, indicative of general dis- and assembly pathways of entire complexes during the cell cycle. Furthermore, I aided in deconvolving the hypothesis on disordered proteins being pivotal in their stability transition during the cell cycle, and correlated it with mitotic phosphorylation sites that might be mediating this particular effect.

Question 5: How does ageing affect the proteomic landscape?

In line of what has been briefly discussed in section 1.2.4, exploring the proteome in context of ageing processes and the associated functional decline, is pivotal for future applications in personalized medicine as well. **Chapter 4** presents a collaborative effort in disentangling the effects ageing has on protein cellular networks in the healthy and diseases bone marrow. It also establishes the connection to other molecular layers, such as transcriptional data at the single-cell level, as well as metabolism. For the first time, this particular project could pinpoint at enzyme abundance changes in the upper glycolytic pathway as a potential source for the age-associated Warburg effect commonly encountered in cancerous cells.

Figure 1.18 below gives a visual summary for the given thesis and the major elements about to be discussed.

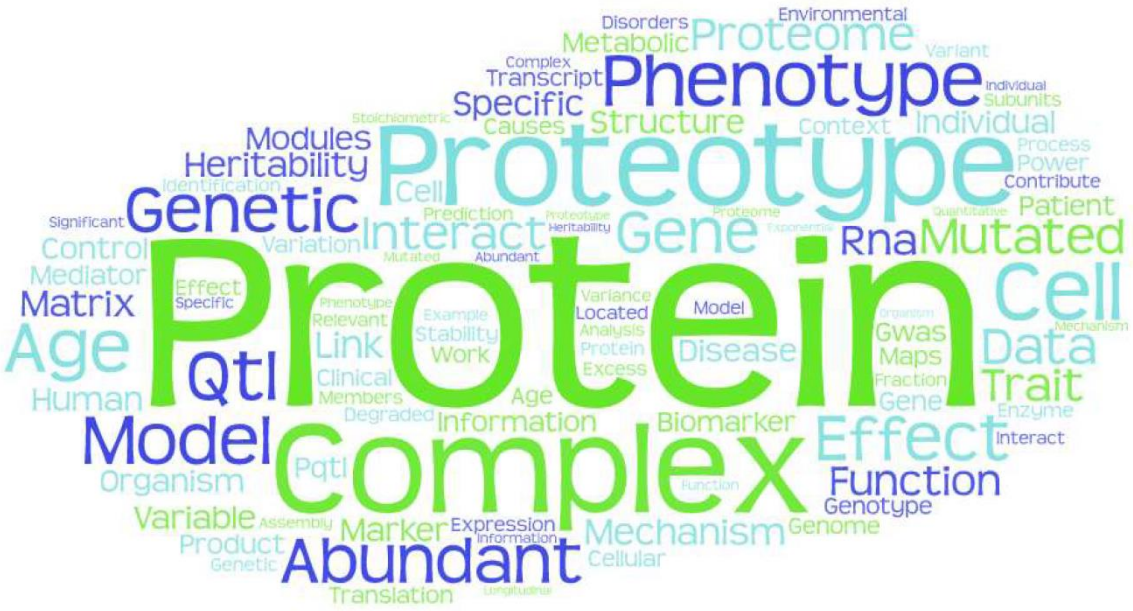


Figure 1.18. Word cloud on the subject of this work. (created by WordArt, https://wordart.com)

Chapter 2

Systematic meta-analysis of individual proteotypes reveals sex- and diet-specific functional modules

In this chapter, I describe a comprehensive analysis of protein organizational modules as maintained across mouse and human individuals and their variation due to sex-specific characteristics. The methodology behind the work has been conceived by me, and I carried out all the computational analysis. The data underlying this analysis was obtained from published articles and TCGA Cancer Panel (CPTAC), as specified further. The work has been conceptualized in the following manuscript:

Natalie Romanov, Alessandro Ori, Michael Kuhn, Martin Beck and Peer Bork (2018). Systematic meta-analysis of individual proteotypes reveals sex-specific functional modules. *In preparation.*

2.1 Introduction

Recent advances in the experimental throughput of mass spectrometry (MS)- based proteomics have enabled the first large scale studies of proteotypes measured for cell lines or tissues, defined as the proteome complement of a genotype [Picotti et al., 2013]. Although genotype and proteotype are poorly correlated [Willhelm et al., 2014; Liu et al., 2016; Payne et al., 2014; De Sousa et al., 2009], genetic variation has been shown to have a considerable impact on the abundance of proteins across yeast strains [Pictotti et al., 2013], mouse strains [Wu et al., 2014; Williams et al., 2016; Chick et al., 2016], and human individuals [Battle

et al., 2015; Wu et al., 2013; Liu et al., 2015]. While some rare diseases are 100% genetically determined, for most common ones, the genetic components are usually minor. In case of obesity for example, only about 6% of the phenotypic variance can be explained by the associated genetic [Speliotes et al., 2010]. The identification of functional traits in proteotypes therefore holds great promise to provide disease-associated fingerprints in individuals. Such traits should be a molecular reflection of both genetic and environmental factors, such as i.e. life style, and provide a basis for personalized treatments.

Establishing such connections from genetic or environmental factors to the individual proteotype remains challenging. This is due to technical limitations, in particular the variable experimental noise across studies but also biological buffering mechanisms [Stefely et al., 2016]. However, the modular architecture of the proteome, i.e. its organization into complexes, pathways and subcellular structures provides powerful means to overcome these issues by interpreting observed variations in the context of well-established biological functions [Stefely et al., 2016; Wang et al., 2017; Ori et al., 2016].

Protein complexes, in particular, represent key organizational units that – while robustly connecting protein subunits in a physical and functional manner- have also been shown to be adapting both in overall abundance and stoichiometry to cellular context, such as during temporal processes [de Lichtenberg et al., 2005] or between different cell types and tissues [Ori et al., 2013; Ori et al., 2016; Hansson et al., 2012; Frese et al., 2017; Arendt et al., 2016]. Other reports have detected similar variations during organism aging [Ori et al., 2015; Cellerino et al., 2017], upon physical exercise [Greggio et al., 2017], or as a consequence of cancer mutations [Goncalves et al., 2017; Roumeliotis et al., 2017]. Although the exact regulatory mechanism of such variation remains elusive, the functional consequences might be fundamental, i.e. the perturbation of protein complex stoichiometries can directly influence organism lifespan [Houtkooper et al., 2013; Miwa et al., 2014; Vilchez et al., 2012] or prevent cell differentiation [Hansson et al., 2012; D'Angelo et al., 2012; Lessard et al., 2007]. Due to structural requirements, that is the physical protein interfaces that are established within a complex and crucial for its function, the stoichiometry of protein complexes is tightly regulated. Both mRNA abundance and post-transcriptional mechanisms acting, e.g., at the level of protein synthesis and degradation, regulate protein levels and ultimately determine the stoichiometry of protein complexes [Ori et al., 2016; Goncalves et al., 2017; Roumeliotis et al., 2017; Liu et al., 2016; Ishikawa et al., 2017].

Several seminal studies have shown the variability of protein abundances across individuals in human and mice [Wu et al., 2014; Chick et al., 2016; Battle et al., 2015; Wu et al., 2013; Liu et al., 2015]. Although each study highlighted individual proteins and functional modules that were found to be variable, a systematic and unbiased analysis of functional modules across multiple studies is lacking. It remains unknown which type of modules (complexes, pathways, organelles) or cellular functions are affected. The extent to which

the proteome of individuals is variable and how this variability is linked to environmental and genetic factors remains difficult to estimate.
Here, we analyze public datasets to investigate proteotypes of healthy and diseased individuals from human and mice. We test the technical power of each dataset and systematically examine functional modules and their contribution to proteotype variation and association with genetic or environmental factors. Our unbiased analysis reveals that protein complex abundance and stoichiometry are indeed major determinants of an individual's proteome, while other functional units, such as e.g. molecular pathway stoichiometry are less variable. The consistently co-varying dynamic components in each protein complex reveal not only refined mechanistic functional details, but can also be associated with both genetic and environmental factors that influence the composition and abundance of a wide range of complexes within individuals. We demonstrate that sex explains the largest fraction of the observed variability in protein complexes and that different types of complexes are more abundant in males versus females. Our variation analysis further dissects core sub-structures and consistently dynamic components in each complex. We show that both genetic and environmental factors influence the composition and abundance of a wide range of complexes within individuals. Finally, stoichiometric alterations of modules are also put into the perspective of their potential role during the aging process of an organism. Our study thus adds another dimension to disentangle the dynamics of the proteotype. With the considerable effect sizes observed for only two examples of genetic and environmental factors, namely sex and diet, our study implies that protein modules serve as molecular sensors for the impact for a wide range of environmental factors on top of intrinsic genetic variation in an individual manner and calls for balancing factors, e.g. individualized diet [Zeevi et al., 2015].

2.2 Results

2.2.1. Interacting proteins are co-abundant across healthy Individuals

To elucidate functional traits in individuals, we first tested to what extent known functional modules or protein associations can be recovered in different proteomics datasets assuming that proteins do not function alone. Implicitly, we thereby test the power of each dataset and ensure consistent results. We examined datasets resulting from profiling proteins across cancer patients (TCGA panels, such as Ovarian Cancer [Cancer Genome Atlas Research Network, 2011; Zhang et al., 2016], Breast Cancer [The Cancer Genome Atlas Network, 2012; Mertins et al., 2016] and Colorectal Cancer [The Cancer Genome Atlas Network, 2012; Zhang et al., 2014]), healthy human individuals, healthy mouse strains that were exposed to different diets, and other proteomic datasets derived from cell types (entire list with details on each dataset given in **Table S2.1**) [Wu et al., 2014; Chick et al.,

2016; Battle et al., 2015; Wu et al., 2013; Khan et al., 2013] (**Figure 2.1A**). The respective studies differ with regard to the MS-technique employed for protein measurement, as well as the source organism, organ or cell type (see **Table S2.1**). While Wu *et al.* [Wu et al., 2014; Williams et al., 2016] (BXD80 mouse strains) and Chick *et al.* (2016) (DO mouse strains, and Founder mouse strains) recovered proteins from mouse livers from different mouse populations, Battle *et al.* (2015) extracted proteins from lymphoblastoid cell lines (LCLs) of the HapMap Yoruba individuals (Human Individuals). If available, we included transcriptional data as well as data derived from ribosome profiling on the same cell types or individuals.

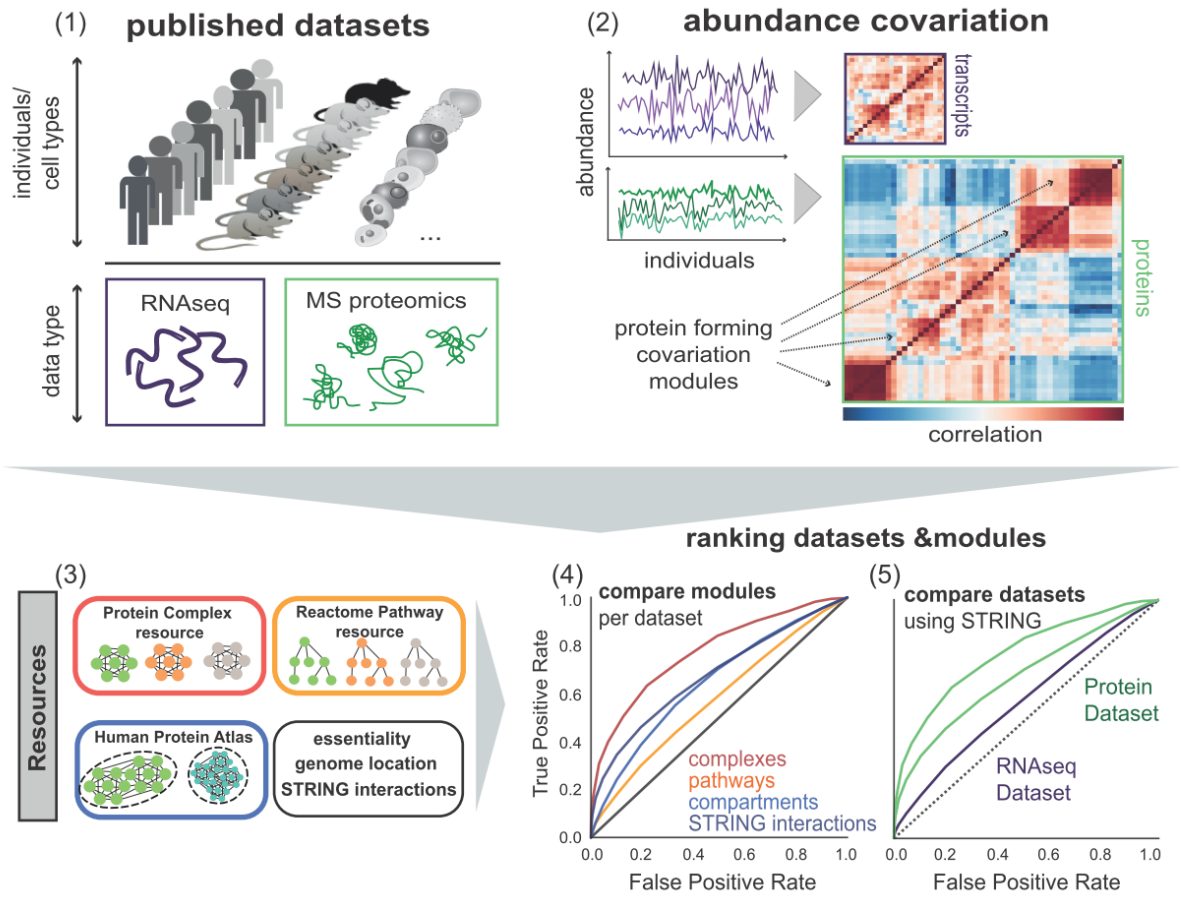


Figure 2.1. Schematic illustration of workflow. (1) Published proteomics datasets on human individuals, mouse strains and cell types are considered for the study. If available, RNA-seq datasets for the respective specimens are taken into account as well. (2) Co-variation of protein (transcript) abundances is calculated for each dataset. (3) Integration of resources on protein modules (STRING, complexes, Reactome, Human Protein Atlas, etc.) to understand features of co-varying proteins (transcripts). (4) Modules can be compared by ROC-statistics in each dataset. (5) Different datasets can be compared by the degree of recoverable known co-variation (STRING-interactions).

We assessed the power of each datasets for discovering functional modules by calculating the level of observed co-expression or co-abundance for known protein-protein interactions utilizing the STRING resource [Sklarczyk et al., 2017], and comparing the results to random associations(**Figure S2.1**). As expected, throughout all datasets, we recovered high-scoring interactions (STRING combined score >700) to be more co-abundant across conditions or individuals than proteins with no (or unknown) interactions. To further dissect the functional relevance of co-abundant protein sets, we added contextual information about chromosomal location, housekeeping roles [Eisenberg and Levanon, 2013], cellular compartment (Human Protein Atlas), essentiality [Wang et al., 2015], pathways (Reactome [Fabregat et al., 2016]) and protein complexes (**Figure 2.1, Table S2.2**). The latter were derived from a manually curated list of 279 largely non-overlapping protein complexes as defined by Ori *et al.* (2016). For each category of contextual information we assessed using ROC curves whether the respective dataset reliably recovered known functional entities, based solely on the co-abundance or -expression metric. This approach implicitly allows a dual assessment: i) the overall quality of each dataset based on the amount of co-abundance/expression, and ii) an unbiased assessment of the type of functional module yielding the highest level of co-abundance across datasets. With regard to (i) we observed datasets dervied from tissue samples to be noisier when compared to cell lines. Incidentally, the datasets generated by SWATH (BXD80 mouse strains, Kidney cancer cells) were both prone to recover less of known interactions than the other datasets, but were both derived from tissue samples.

As it has been reported for cell types and differentiation stages [Ori et al., 2013; Geiger et al., 2012; Siwak et al., 2015], we observed a consistently high level of co-abundance of members of the same protein complex in proteomics datasets (average AUC>0.6, **Figure 2.2A**). Proteins in other modules, such as pathways, organelles, the housekeeping proteome, etc. showed less coherence (**Figure 2.2A**). The shifts towards higher co-abundance were especially apparent in the TCGA proteomics panels, the healthy human Yoruban individuals [Battle et al., 2015] and the DO/Founder mouse strains [Chick et al., 2016] and less so in RNA-seq and ribosomal profiling datasets (**Figure 2.2A**). Recent reports [Goncalves et al., 2017; Roumeliotis et al., 2017] have demonstrated protein complex attenuation due to copy number variations common to cancer. Strikingly, the abundance shift for healthy individuals was in some cases even more pronounced than for cancer-derived datasets, confirming that co-regulation of protein complex members beyond the post-transcriptional level is indeed an inherent cellular mechanism, also in cells with natural gene copy number (**Figure 2.2B**). Several studies have indeed pointed towards tight regulation of protein complex stoichiometry by translational or post-translational mechanisms [Ori et al., 2016; Goncalves et al., 2017; Roumeliotis et al., 2017; Stefely et al., 2015; Ishikawa et al., 2017]. For example, for members of the F0/F1-ATP-synthase, gene

transcripts only have a mean correlation of 0.23, while effective protein abundances are strongly correlated (mean correlation of 0.79, **Figure 2.2C**).

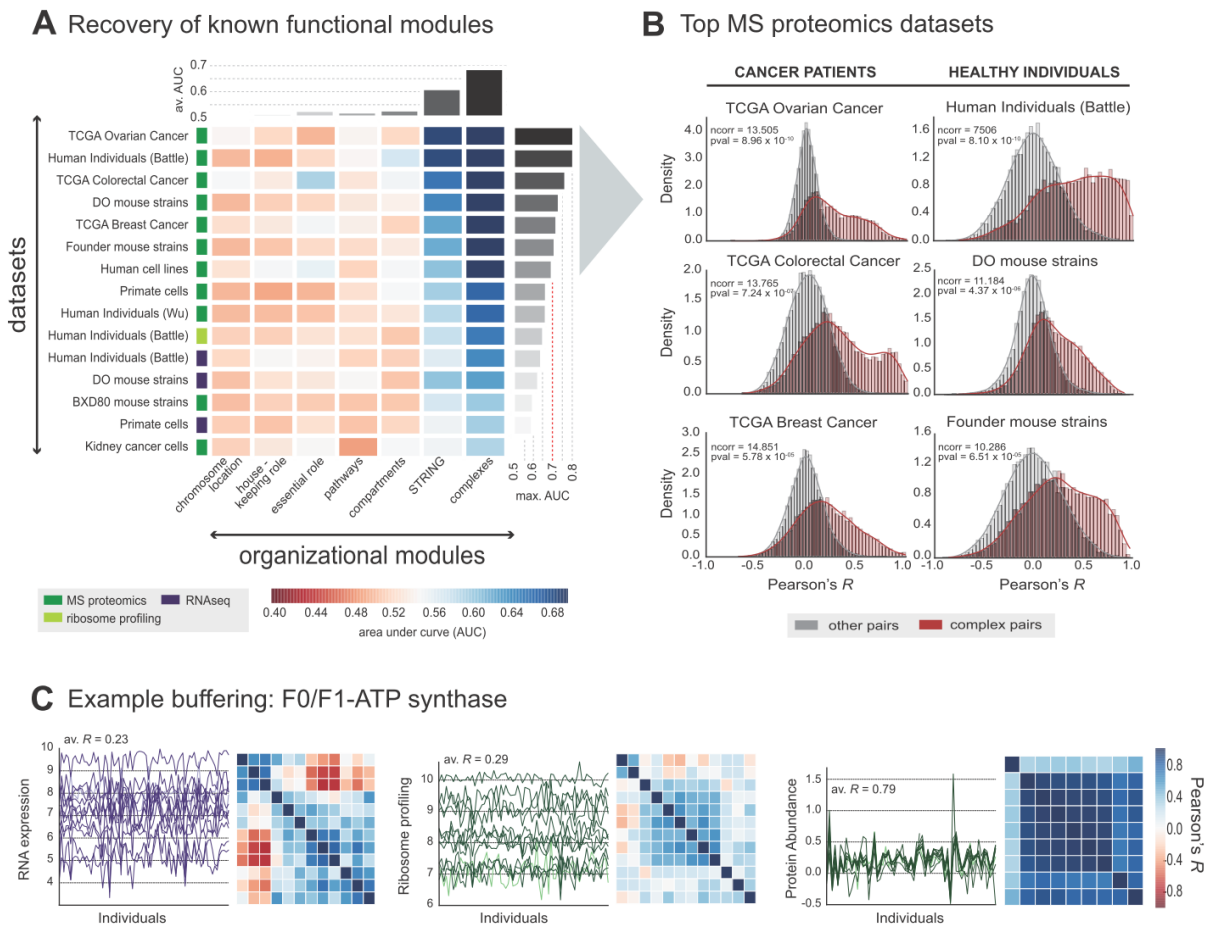


Figure 2.2. Strongest co-variation across individuals stemming from protein complexes. (a) Recovery of known functional modules by means of ROC-analysis (receiver operating characteristic). Each cell of the matrix displays the AUC (area under curve) value for the given module (*x*-axis) in the given dataset (*y*-axis). Modules are ordered according to the average AUC per module across datasets; datasets are sorted according to the maximum AUC per dataset across modules. The type of data is indicated next to the dataset. (b) The shift in Pearson correlation values for complex-associated proteins (dark-red) relative to the background correlation values (grey) is illustrated for the top 6 datasets derived from (a) as density graphs. Statistics are based on the Wilcoxon-rank test. (c) Buffering of transcriptional and translational variation of complex-associated proteins tends to be significantly higher at the level of protein abundances than in the upstream layers.

2.2.2. Protein Complexes vary in their Stoichiometry across Individuals

Given the strong signal of variation in complex abundance across individuals, we focused on a detailed analysis of protein complexes and their stoichiometry, in order to identify genetic and environmental factors associated with it. For this purpose, we examined only the proteomics datasets that were yielding the highest recovery of known modular entities

due to co-abundance (**Figure 2.2B**), namely all TCGA cancer datasets, and the datasets on human individuals [Battle et al., 2015] as well as Founder and DO mouse strains [Chick et al., 2016]. Using median co-abundance per complex as a proxy for stoichiometric variability across individuals and controlling for a number of technical biases and possible artefacts (see Methods, **Figure S2.2A**, **Figure S2.2B**), we recovered a protein complex variability landscape (**Figure 2.3**) that is highly consistent across the different proteomics datasets (average $R^2 = 0.645$). We ranked protein complexes according to their level of co-abundance across individuals and identified a subset that is rigorously maintained in stoichiometry (**Figure 2.3**). This implies that protein complex stoichiometry regulation is a general principle that is exploitable to reveal functional traits in different proteotypes.

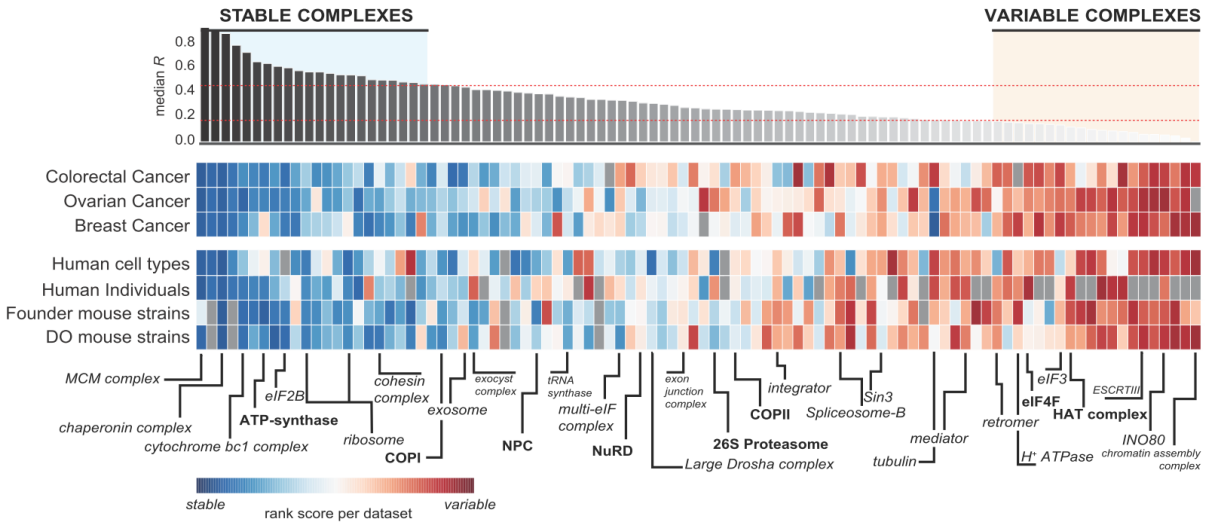


Figure 2.3. Co-variation landscape of protein complexes in different proteomics datasets. Manually curated complexes are shown on the x-axis and are sorted according to the median co-abundance observed of complex members in different datasets considered (*y*-axis). From original datasets analysed in Figure 2.2a, top 6 datasets and 'Human cell types' (used as a reference) are considered. In each dataset median co-abundance of each complex is calculated and ranked. The heatmap illustrates those ranked values, ranging from variable complexes (red) to stable complexes (blue).

Of 96 well-defined protein complexes with at least 5 subunits, 21 complexes exhibit a tight co-regulation of all subunits across diseased, as well as healthy individuals (median R^2/complex>0.46 (75th percentile)). They include the mini-chromosome maintenance (MCM) complex to complexes associated with the translational apparatus (ribosome, chaperonin complex, elongation factor eIF2F) and mitochondrial complexes within the electron chain transfer, such as the F0/F1 ATP-synthase, cytochrome *bc1* complex and the cytochrome c-oxidase. Variable complexes (median R^2<0.17 (25th percentile)), on the other hand, were enriched in chromatin-associated processes (see **Figure S2.3**, 4.68×10^{-34}, Fisher Exact Test) such as the RNA polymerase, the mediator complex, the BAF complex, etc.

The range of complexes in between represent instances where both co-regulated parts of complex, as well as more variable members are present, such as in the COPI/COPII, the nuclear pore complex, and the 26S proteasome. For these complexes where arguably both core sub-complexes and accessory subunits are present, we sought to extract consistent sub-structures from the different proteomic datasets in question, as well as defining subunits that primarily contribute to the overall variability of the complex. Briefly, complex subunits were normalized by the trimmed mean complex abundance across samples and their concurring variance relative to the other subunits was monitored [Ori et al., 2016] (see Methods). We observed a high consistency between datasets after the dissection of modules into stable and variable sub-parts (average R^2 of 0.23). Variable components - if identified at p-value<0.1- were making up 2-20% of the overall structure of the complexes (**Table S2.3**).

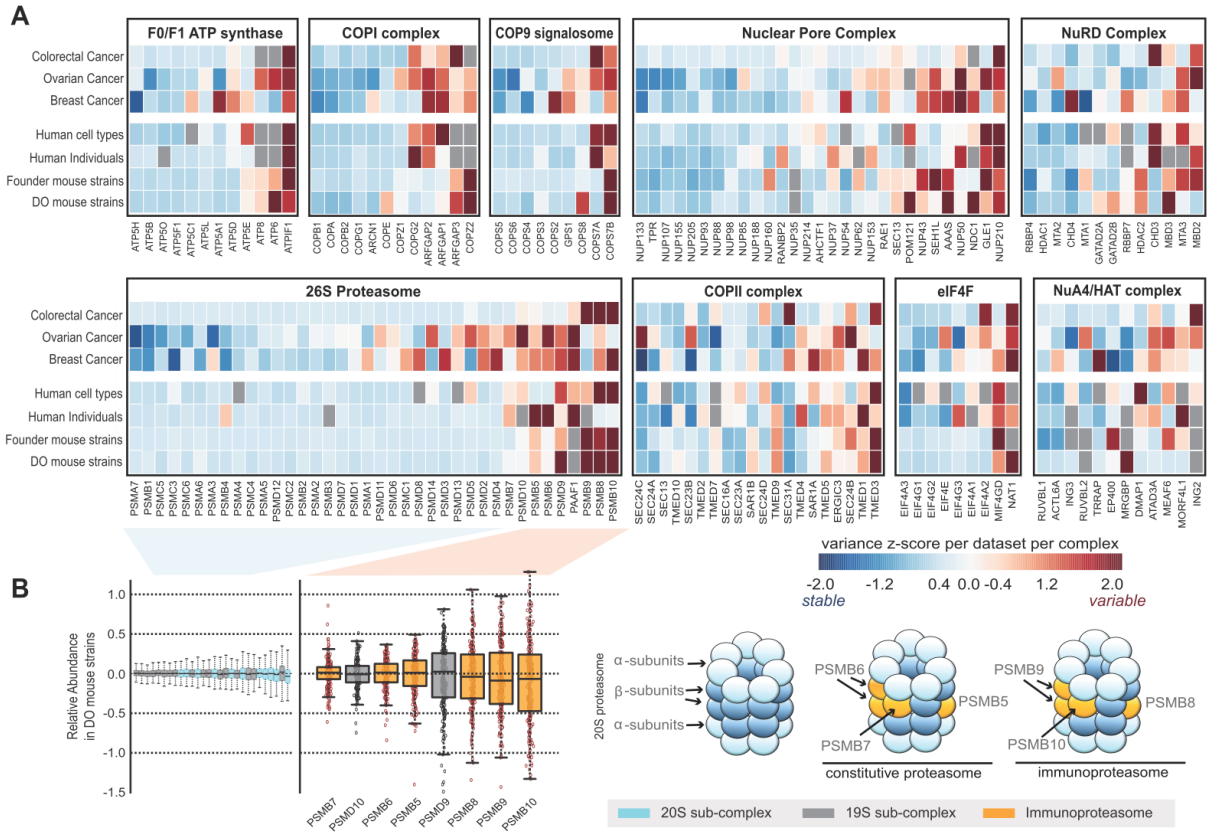


Figure 2.4. Consistent dissection of protein complexes in stable and variable components. (a) Illustration of stable and variable components in a number of exemplary complexes (x-axis: complex protein members, y-axis: datasets). Each cell of the heatmaps display a z-score which was calculated based on protein variances after complex-normalization (variable component (red): z-score > 1.5; stable component (blue): z-score < -1.5). If a given protein was not detected in a dataset in the field is left out grey in the heatmap. (b) Boxplot showing the relative abundance of proteins associated with the proteasome in the DO mouse strains. The colouring of the boxplots relates to the sub-structures of the proteasome as indicated in the right-hand cartoon (blue: 20S, grey: 19S, orange: immuno-proteasome).

Stable subunits tend to be strongly correlated with other stable subunits of the same complex, whereas this is not the case for variable components (**Figure S2.4A**, average R^2(stable) at 0.38 , average R^2(variable) at 0.17, *p*-value = 9.24×10^{-12}, Mann-Whitney U-test). Additionally, stable subunits tend to have more STRING interactions on average than the variable complex components (**Figure S2.4B**, average *ppi* (stable) at 48, average *ppi*(variable) at 32, *p*-value = 1.47×10^{-2}, Mann-Whitney *U*-test).

We found multiple instances of variable subunits consistent with known biology (**Table S2.3**). For example, the F1/F0 ATPase inhibitor ATPIF1 is consistently recovered as variable relative to the rest of the ATP synthase complex (**Figure 2.4A**, *p*-value<0.025). ATPIF1 is known to be the master regulator of the F0/F1 ATP-synthase [Garcia-Bermudez et al., 2016]. Its binding to the complex impedes the hydrolase activity of the ATP-synthase, effectively shutting down its activity to prevent excess wasting of ATP. The observed high variability of ATPIF1 across individuals can thus be explained by the variable energy requirements of the cell, for example [Sanchez-Cenizo et al., 2010; Sanchez-Arago et al., 2013]. Variable subunits of the nuclear pore complex (NPC) are peripherally associated to the core scaffold such as e.g. all three transmembrane nucleoporins (NUP210, NDC1, and POM121). Further variable members of the NPC were found to be ALADIN (AAAS), which potentially binds to transmembrane nucleoporins [Stavru et al., 2006] and has been linked to genetic disease, as well as Nup50- a subunit involved in active nuclear import [Beck and Hurt, 2017]. We also found that paralogous subunits are often variable, such as the ARFGAP-subunits of the COPI complex, the MBD2/3-paralogs involved in the NuRD complex, as well as COPS7A/COPS7B in the COP9 signalosome. This observation is in line with the report from Ori *et al.* (2016) where paralog switching between different cell types has been described as a major driver for complex re-arrangements.

Specific subunits of the 26S proteasome vary across healthy individuals, mouse strains and cancer patients (**Figure 2.4B**). Those are part of both the 20S and the 19S subunits, however, the 20S components PSMB8/9/10 of the immunoproteasome, a specific sub-module of the proteasome that is involved in the immune-regulatory response [McCarthy et al., 2015], stood out in terms of variability across individual (**Figure 2.4A/B**, average *p*-value of 6.57×10^{-2}). This occured alongside with the constitutive proteins subunits PSMB5/6/7 that they replace depending on the cellular context (**Figure 2.4B/C**). Furthermore, the PSMD9 proteasome component was also more variable relative to the other subunits of the complex, which could be explained by its known additional role as a transcriptional co-regulator besides its regulatory role during proteasome assembly [Sangith et al., 2014]. We conclude that proteotype data appear to be predictive of multi-functionality of sub-complexes or complex components.

To further dissect features of stable and variable complex components, protein modifications were mapped onto stable and variable complex components, such as acetylation,

methylation, phosphorylation, sumoylation and ubiquitination. For 6 out of 7 datasets we could identify an enrichment of proteins with ubiquitination target sites within the stable set of proteins (**Figure S2.5A**, *p*-value of 1.65×10^{-2}, Fisher Exact Test). We therefore sought to substantiate the hypothesis on more rigorous degradation control on the stable complex components by comparing protein turnover rates from stable versus variable components in all proteomic datasets (data not shown). Indeed, stable complex subunits seem to consistently show a higher turn-over rate on average than variable subunits (*p*-value of 0.14) [Ryan et al., 2017].
For variable complex subunits, on the other hand, we observed higher transcript-protein correlations relative to core protein complex components (*p*-value (ANOVA) = 1.2×10^{-7} for DO mouse strains, *p*-value (ANOVA) = 3.77×10^{-12} for Human Individuals, see **Figure S2.5B**), suggesting potentially stronger transcriptional control exerted on these proteins. This particularly manifests in the high protein-transcript correlations of the immuno-proteasome subunits in the mouse dataset, such as R^2=0.72 for PSMB8, R^2=0.66 for PSMB9 and R^2=0.55 for PSMB10 (**Figure S2.5C**).

2.2.3. Co-abundance of entire Protein Modules is explained by Sex Differences

To identify potential genetic and environmental determinants for above explained proteotypic features, we primarily leveraged the well-defined meta-data available on the DO mice strains, namely their sex and their diet, with half of the animals fed with rodent chow, and the other with high-fat diet [Chick et al., 2016]. We captured two different readouts, namely the variability in (a) complex abundances, and (b) complex stoichiometries (inset in **Figure 2.5A**, upper right). For (a) complex abundances, differences between male and female mice were evaluated using a standard *t*-test and Cohen distances [Sullivan et al., 2012] to yield effect size estimations for each complex (**Figure 2.5A**, upper part). From all 96 considered complexes, 21 complexes showed an overall higher abundance in male mice, while 36 were more abundant in females (*q*-value<0.01). Those complexes were enriched in complementary functional processes: Whereas complexes that were more abundant in males were part of the translational process (ribosome, eukaryotic translational factor 2B complex) and specifically protein transport processes involving COPI and COPII, complexes that were more abundant in females, were enriched in mRNA transport (nuclear pore complex), and RNA splicing processes such as the small nuclear ribonucleoprotein complex (**Figure S2.6**, FDR<1%). This functional complementary is indicative of a genetic and possibly environmental influence on the abundance of entire complex entities, as opposed to individual protein abundances. It has been shown for example, that female mice, unlike male ones, have a high constitutive activity of both mTORC1 and mTORC2, which are

master regulators of the energy metabolism, mitochondrial function as well as protein and lipid synthesis [Guergen et al., 2013].

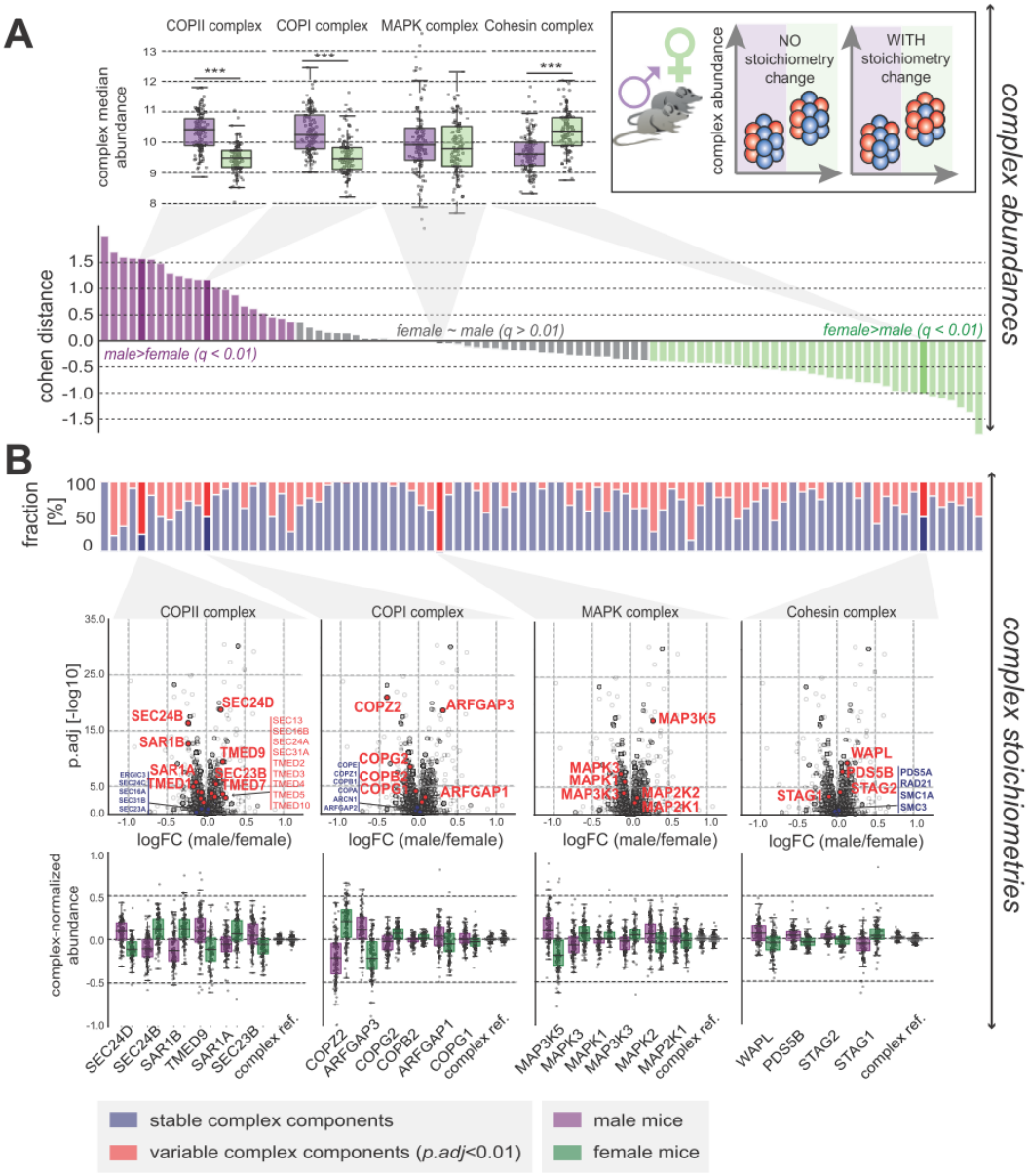


Figure 2.5. Sex-specific regulation of complex abundances and stoichiometry. (a) Delineation of differential abundance of complex structures between male (purple) and female (green) DO mice. The effect sizes (Cohen Distance) is shown across all 96 complexes with colors corresponding to significant effects (q-value < 0.01, t-test). Complex median abundances for selected examples are highlighted in the boxplot graphs above the Cohen-Distance barplots. The inset panel (upper right) illustrates the concepts of abundance variation and stoichiometry of protein modules. (b) For each complex, the fraction of stable components (blue, not changing in stoichiometry between male and female mice) and differential stoichiometric hits (red) are shown. The volcano-plots beneath illustrate the underlying data with log2 fold-changes (male/female, *x*-axis) and the adjusted *p*-value on the *y*-axis. Complex-normalized abundances are shown below, highlighting male and female stoichiometry within the complex.

2.2.4. Sex- and diet-specific Protein Complex Stoichiometries

We tested if diet and sex influence complex stoichiometry. To this end, a LIMMA-analysis [Ritchie et al., 2015] was performed on complex-normalized abundances for each complex separately (see Methods, **Table S2.4**). Generally, changes in complex abundance did not significantly correlate with the variability in complex stoichiometry (specified as the fraction of subunits affected, FDR<1%), R^2 at *-0.03*. For example, dense networks composed of MAP-kinases (MAPK1, MAPK3, MAP2K1, MAP2K2, MAP3K3, MAP3K5), did not yield any signal with regard to complex abundance (**Figure 2.5A**), but display a different complex stoichiometry between male and female mice (**Figure 2.5B**). In more general terms, a diverse range of functions was variable in complex stoichiometry, including ubiquitin protein ligase activity, mRNA splicing, catabolic processes and protein transport functions (**Figure 2.6A**). With protein transport, in particular, the COPI and COPII complex were largely affected in their relative stoichiometry between male and female mice (**Figure 2.5B**). Paralogous components, for example SEC24A/B/C/D – while unaffected between the different diet conditions (**Figure 2.6B**; **FigureS2.7A**) - contributed to very distinct sex-specific stoichiometry, with SEC24D being consistently more abundant than SEC24B in males, and vice versa in females (q-value<0.01, **Figure 2.5B**). SEC24A/C, on the other hand, had similar complex-relative abundance between male and female mice. Such sex-specific stoichiometric aberrations between individuals could indeed have severe functional implications, such as the efficiency and/or specificity of receptor transport, which has been shown to be affected by the absence and concentration of the specified paralogs [Scharaw et al., 2016; Adams et al., 2014].

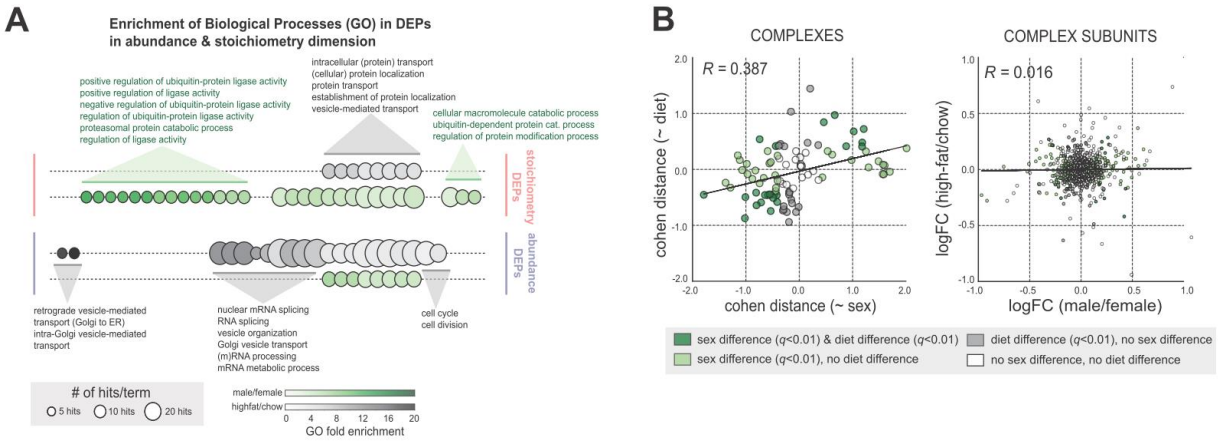


Figure 2.6. Module abundance and stoichiometry changes affect distinct functional processes. (a) GO enrichment of differentially expressed proteins (DEPs, FDR<1%, Benjamini-Hochberg) derived from LIMMA-analysis using original abundances (blue), and complex-normalized abundances (red) for both sex- and diet-specific differences (green, grey). The opacity of the coloring corresponds to the fold-enrichment and the size of the circles to the number of DEPs/terms. (b) (*left*) Scatter plot displaying the Cohen distances for sex- (x-axis) and diet-differences (y-axis) in complex median abundance, respectively. (*right*) LIMMA-derived log2 fold-changes for male/female differences (x-axis) and high-fat/chow differences (y-axis) are compared for all complex subunits.

2.2.5. Differential receptor transport mediated by COPI/COPII complexes

In order to understand what functional consequences differential stoichiometry of the COPI and COPII complex could have, we mapped known receptors and ligand molecules [Ramilowski *et al.*, 2015] onto the proteomic dataset published by Chick *et al.* (2016) on Diversity Outbred (DO) mice, and interrogated their abundance correlation with all complex components of COPI and COPII. To get a more comprehensive picture, we also included other elements of the transport machinery, such as ERGIC-sub-transport system, the lysosome, etc. Finally receptor molecules were also correlated against proteins resident in the endoplasmic reticulum and the Golgi apparatus to understand any bias arising from the molecule's original or final compartment (**Figure 2.7**).

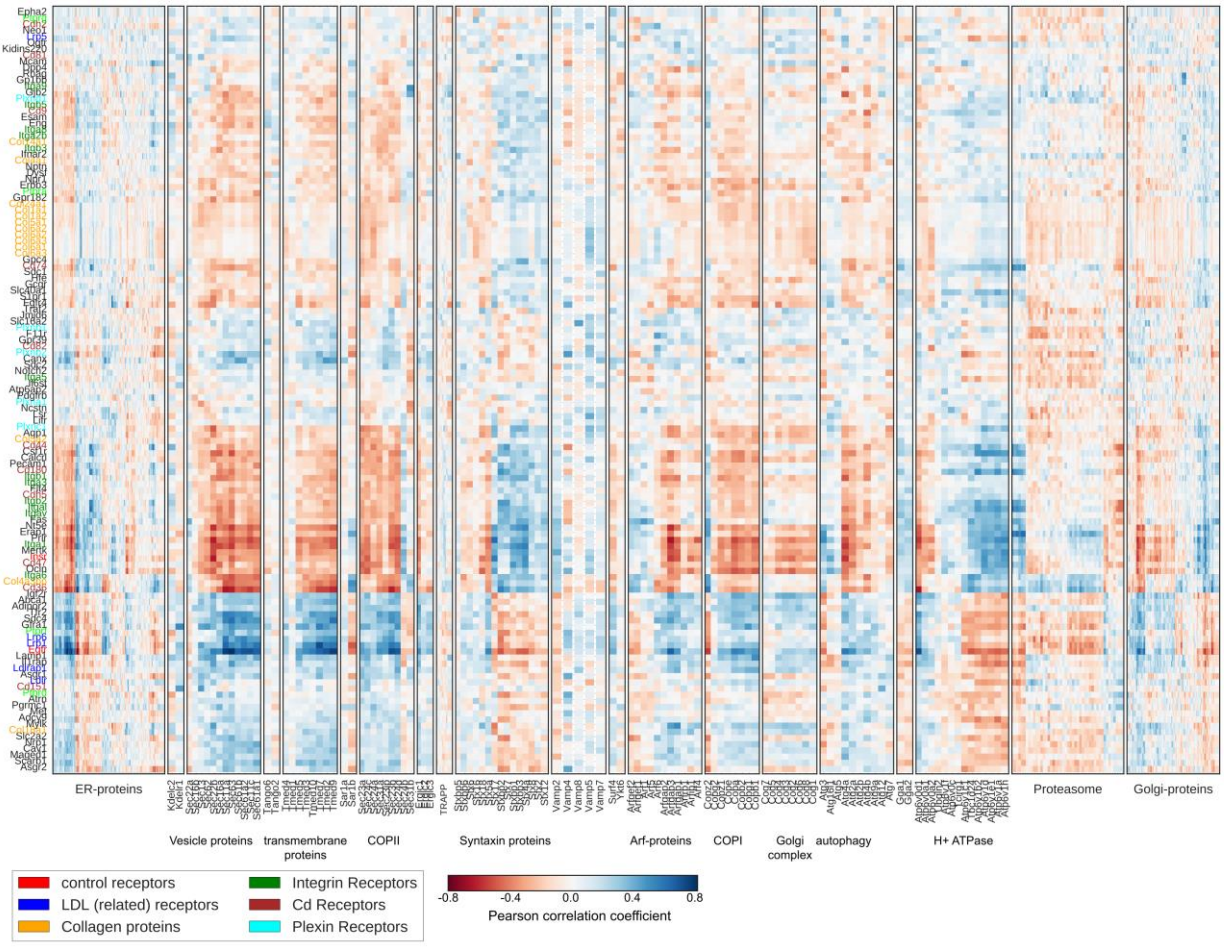


Figure 2.7. Clustering of receptors based on transport components between ER and Golgi. Heat map displaying correlation values calculated from protein abundance profiles transport components (*x*-axis) and each receptor (*y*-axis) that is sufficiently covered in the dataset (>50% coverage, hence ~90 mice). Only receptors with at least one correlation with an FDR<5% are displayed. The clustering is based on Euclidean metrics and specifically on the correlation values observed between receptors and the COPII components.

Clustering the receptor correlations with the individual transport components revealed two distinct clusters primarily driven by vesicle proteins, transmembrane proteins, COPII and COPI. The first cluster is characterized by receptors that are highly correlated with transport protein abundance, such as the EGF receptor (EGFR). The study by Scharaw et al. (2016) found that the transport of newly synthesized EGF-receptors (EGFR) from the ER to the plasma membrane coincides with the up-regulation of the isoforms SEC24B and SEC24D. We could independently recover a significantly high correlation of EGFR with those isoforms (**Figure 2.7, Figure 2.8**). The LDL-receptor-related proteins LRP1 and LRP6 were in the same cluster, suggesting a similar mechanism as with the EGF-receptor with regard to its transport.

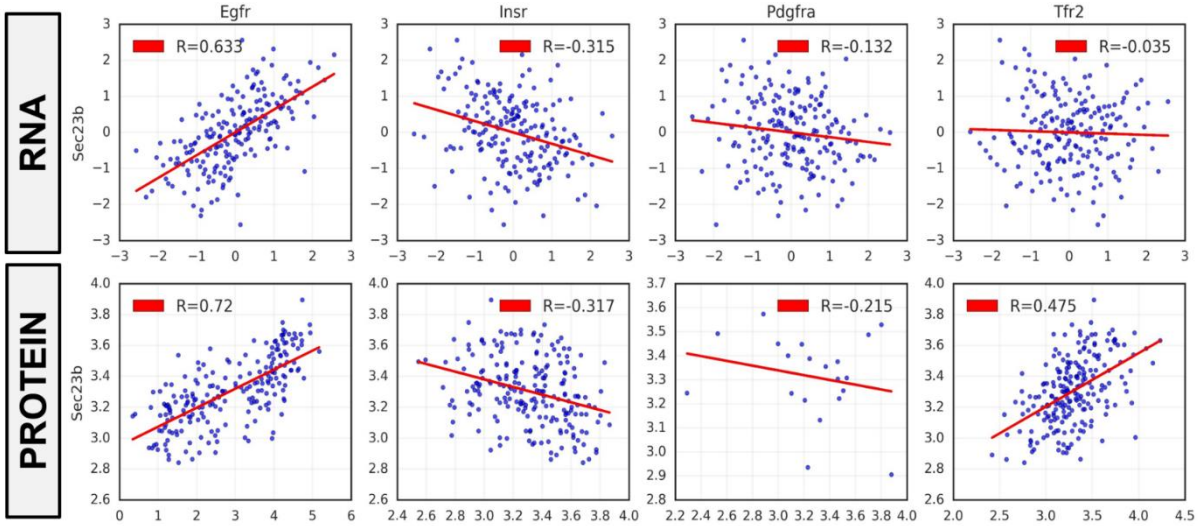


Figure 2.8. Comparison of expression or abundance levels of selected receptors and the COPII-component SEC23B in DO mice. Resolved picture on SEC23B RNA (top)- or protein abundance (buttom) correlation with RNA- or protein abundance profiles of receptors EGFR, INSR, PDGFRA or TRF2. Values are displayed as log2-values.

In contrast to the EGFR-cluster there are many other receptors who had the opposite behavior being anti-correlated with COPI/COPII components, suggesting that these receptors might work by mechanisms of saturation levels (**Figure 2.7**). Hence, if there is too much of a certain receptor, the transport system is signaled that it should not transport this particular receptor anymore. This cluster notably is enriched in integrin-receptors necessary for cell-cell interactions and adhesion, as well as cluster-differentiation CD-proteins relevant for immune signaling. Apart from the two main clusters, we can also note that certain receptors, such as the transferrin receptor TFRC do no exhibit any significant correlation with neither COPI- or COPII-subunits, which might have its recycling nature as a mechanistic underpinning. Indeed, some receptors, such as TFRC and LDL-receptors are simply retained in the cytoplasm again and re-used if necessary, and thus would not

elicit any up-or downregulation of the transport machinery. Indeed, the LDL-receptor LDLRC showed the same kind of behavior as the TFRC-receptor.
In summary, the observed clustering could therefore suggest that there are two signaling mechanisms that interpret abundance of respective receptors as a trigger for transport up- or downregulation (**Figure 2.7**). Further investigation would involve identifying distinct protein features of those clusters that might explain receptor transport variability. **Table 2.1** gives an overview on features that are currently explored.

Table 2.1: Overview

Tested Features	Description
Post-translational modifications	Mappings available for all proteins from PhosphoSitePlus [Hornbeck et al., 2015]
Degree of disordered protein structure	For each protein in this analysis a rank was calculated that gives a relative coverage of disordered regions in the entire sequence. The rank was calculated exactly as demonstrated in **Chapter 3**, and is based on data from d2p2 [Oates et al., 2013].
Other protein structural elements	For all proteins in the analysis structures (alpha-helices, beta-sheets, turns, coiled coils, etc.) have been mapped from PDB [Berman et al., 2000; www.rcsb.org].
Known protein domains	Domains are mapped from InterPro [Finn et al., 2017]
Known protein short linear motifs	Mappings are conducted for each protein from the ELM database [Dinkel et al., 2016]. The KKDEL (aromatic amino acid) motif is known to be important for transport, and can be screened for as well.
Unknown protein short linear motifs	Unbiased mapping is conducted with Dilimot from the Russel lab [Neduva et al., 2005]
Transmembrane (TM) screening	For estimating the length of the TM region the IUPRED database on hydrophobicity [Dosztanyi et al., 2005] is interrogated. This should serve as a proxy to also define Type1 and Type2 receptors (depending on whether their N-or C-terminus reaches into the cytosol or not).
Protein-protein interactions	Interrogation on common interaction partners for proteins in one cluster. Are there any pathways or other modules enriched?

The functional consequences of the observed changes in complex stoichiometry and their propagation to other cellular processes within an individual remain to be further explored in the future.

2.2.6. Sex- and diet- specific Variation in Module Abundance and Stoichiometry influence the Proteotype

We compared the number of stoichiometric hits that we could recover using complex-normalized abundances with the number of differentially expressed proteins when the original abundances were used (**Figure S2.8A**). While 1.207 complex-associated proteins were found to be differentially expressed between male and female mice using the original abundances, nearly half of it was not defined as differentially expressed using the complex-

normalized abundances. This effect is less apparent with pathway modules, where the pathway-normalized approach would still recover ~80% of the original hits (**Figure S2.8B**). The fact that pathway normalization does retrieve most of the difference might suggest that there is more variation of individual proteins within pathways while in complexes coordinated change of multiple members is more common. In addition using complex-normalized abundances actually recovers 219 novel hits that otherwise would have remained unnoticed. Thus, the method could be further leveraged to disentangle the slight fine-tuning for proper module functioning.

To have an understanding on the cumulative effect of sex and diet on the overall proteotype, we estimated the effect sizes of those two factors on the observed variation (**Figure 2.9A, Table S2.5**). The variation of individual proteins- regardless of their modular context- was on average less than 5% explained by sex differences, and even less so for diet differences (around 2%). Some proteins, however, were strongly influenced by the sex of the animal, i.e. SULT2A1 (63.65%) and PAPSS2 (64.82%) which are crucial for sulfation of the androgen precursor [Oostdjik et al., 2015]. We additionally examined whether the animal's sex is proxy for heritability by comparing the effect sizes directly with the results from Liu et al. (2015), where the authors measured to what extent the variation of 342 human plasma proteins can be explained by heritability, environment and the longitudinal dimension. 37% of the environmental effect on the human plasma proteome could be recovered in the DO mice as diet-dependent (ρ(Spearman) = 0.37). Heritability effect on the human plasma proteome and the impact of the animal's sex on protein variation correlated positively as well (ρ(Spearman) = 0.27).

We further explored to what extent the variation in the two above described metrics, module abundance (a) and module stoichiometry (b) are influenced by either factor, and cumulatively. A 'module' is hereby defined as either a complex entitity, or a pathway that is characterized by high (non-random) co-variation, such as the complement pathway (see Methods). Specifically, the sex of the mice explains up to 35% of the variance of entire module entities, as deduced from co-variate analysis controlling for diet (**Figure 2.9A**). The effect sizes of diet, on the other hand, on module median abundances was only reaching up to 15% and was generally affecting another set of modules, such as the Dsl1p complex, the HOPS complex and mitochondrial complexes, namely the cytochrome *bc1* complex (**Figure 2.9B**). On average, sex and diet would cumulatively explain around 11% of the abundance variation in complexes. Pathways, such as the Androgen and the Glucocorticoid Biosynthesis, had expectedly a large fraction of its abundance variation explained by sex differences (38.36% and 32%) (**Table S2.5**).

We also estimated the effect sizes of both sex and diet on module stoichiometry and observed less of an impact on stoichiometry than on module abundances (sex around 8%, diet around 4%)(**Figure 2.9A**). Most complexes, however, had less than 5% explained by sex differences

(**Figure 2.9A/B**). The same is true for diet effect sizes on complex stoichiometry, with most complexes being impacted less than 2% by diet. Interestingly, there was a subset of complexes that were considerably more affected in their stoichiometry by diet than by sex (8.6%), such as the mitochondrial pyruvate dehydrogenase (**Figure S2.7C**). In general, there was no significant correlation between stoichiometry changes between male/female mice and stoichiometry changes between high-fat/chow mice (R^2 = 0.016), again suggesting functional complementarity (**Figure 2.5C**).

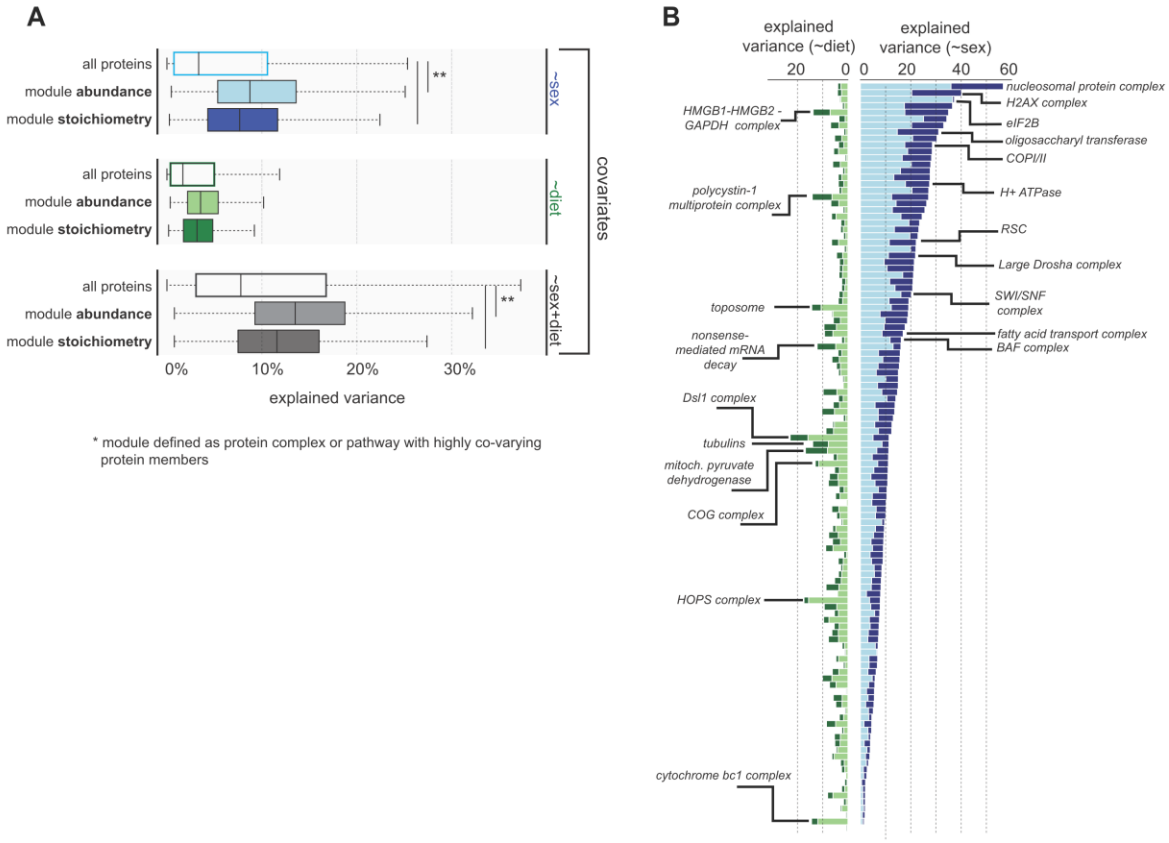


Figure 2.9. Effects of sex and diet on protein variation, as well as variation in module abundance and stoichiometry. (a) Distribution of the overall effect of sex, diet and the cumulative effect (~sex + diet) on protein abundance variation (all proteins), as well as abundance and stoichiometry variation of modules, including protein complexes and pathways with highly co-varying protein members. The lighter colors correspond to effects on abundances, whereas darker colors correspond to effects on module stoichiometry. (b) Distribution of sex- and diet-dependent effect sizes on all complexes (with ≥5 protein members), with lighter colors illustrating effects on abundances and darker colors effects on stoichiometry (see legend from A). Complexes with their variability being mostly explained by either sex or diet are highlighted.

2.2.7. The impact of ageing on sex- and diet-defined complex stoichiometries

Finally, we examined whether the sex- and diet-affected modules were somehow impacted in the longitudinal dimension, i.e. during ageing. For this purpose, we used data from a

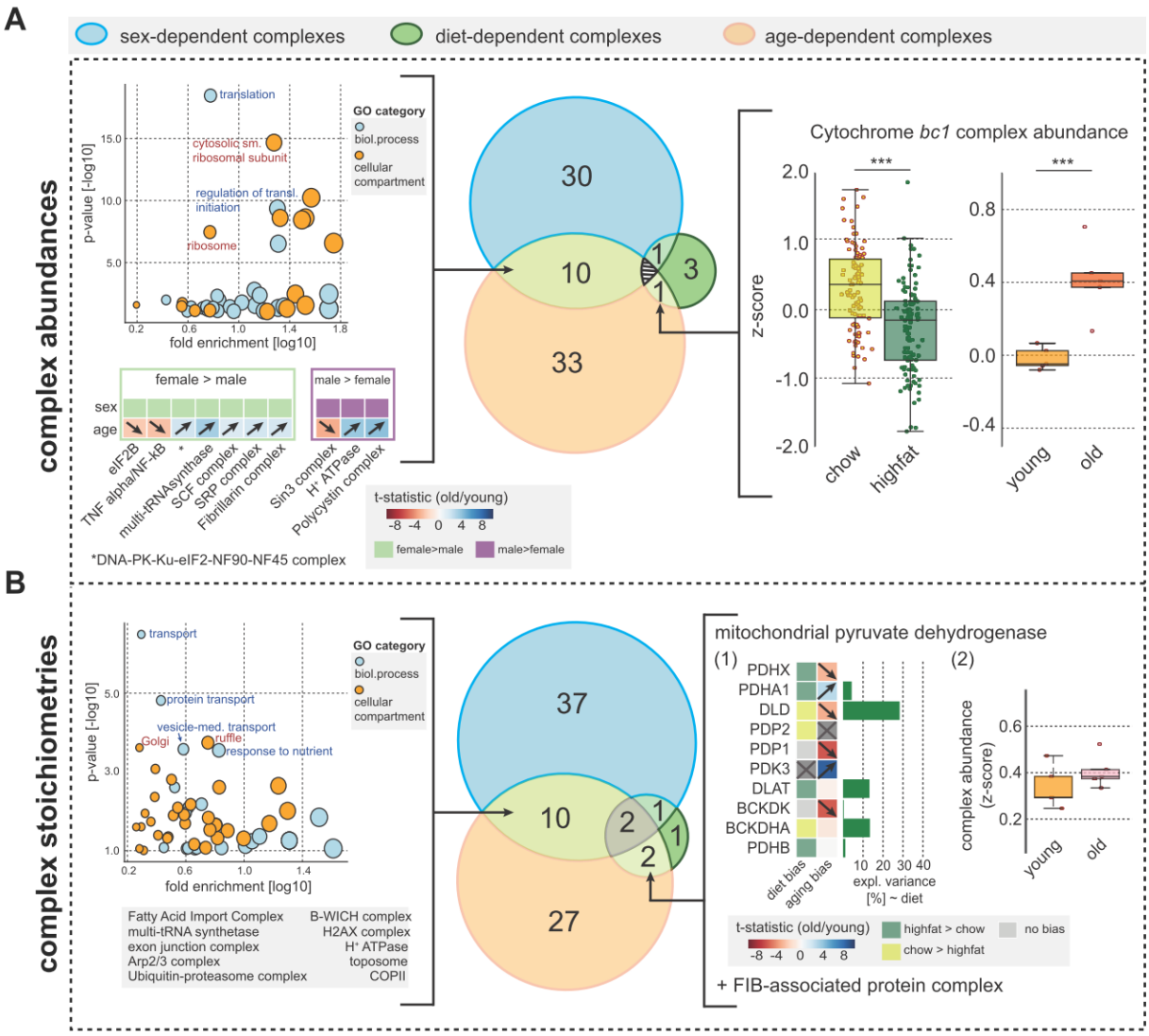


Figure 2.10. Age-dependencies of alterations in module abundance and stoichiometries. (a) Venn diagram showing the overlap of significantly sex-dependent (blue), with diet-dependent (green) and age-dependent (orange) complex abundances (q-value<0.1). The cytochrome *bc1* complex is affected in overall abundance in diet and age, as demonstrated on the right-hand side: First box-plot shows the z-score distribution of the median complex abundance in chow- and highat-DO mice. The second box-plot illustrastes the z-score distribution of the median complex abundance in young and old naked mole rats. For the overlap of complexes between sex- and age-dependencies, GO enrichment is shown as a scatter plot on the left-hand side, with blue circles showing biological processes, and orange circles showing cellular comparmtents. The heatmap below demonstrates the aging directionality of the affected complexes (old/young t-statistic, q-value<0.1). For each complex it is also indicated whether the complex is significantly more or less abundant in male or female mice on average (q-value < 0.1; male: purple, female: lightgreen). (b) Venn diagram showing the overlap of significantly sex-dependent (blue), with diet-dependent (green) and age-dependent (orange) complex stoichiometries (q-value<0.1). On the left-hand side of the Venn diagram the complexes affected by both sex and age are listed, and the GO enrichment is shown (same as in (a)). On the right-hand side, the ageing effect on the mitochondrial pyruvate dehydrogenase is shown: For each subunit of the complex, the bias towards chow (yellow) - and highfat (darkgreen) diet is shown in the first column, and the ageing directionality (old/young t-statistic, q-value < 0.1) in the second column. The bar plot on the side shows the amount of variance that can be explained by diet. On the right-hand side the box-plot shows the z-score distribution of the median abundance of the mitochondrial pyruvate dehydrogenase in the young and old rats.

recent study on ageing in the naked mole rat (NMR) livers by Heinze et al. (in preparation, 2018). We found 43 protein complexes to be affected in their overall abundance, and 37 protein complexes to show stoichiometric alterations between young and old NMRs (**Figure 2.10A/B, Table S2.6**). Modules involved in translational processes, such as eIF2B and the multi-tRNA synthetase, were both affected by the sex of an animal, and in ageing (q-value<0.1) (**Figure 2.10A/left**). Diet, on the other hand, had a limited effect on ageing modules; only the cytochrome *bc1* complex showed a significant abundance change due to differential diet conditions and ageing in NMRs (**Figure 2.10A/right**). The mitochondrial pyruvate dehydrogenase was another mitochondrial complex that had its diet-dependent stoichiometry influenced in the longitudinal dimension. The dihydrolipoyl dehydrogenase (DLD) in particular showed both a signficant down-regulation upon high-fat diet exposure and upon ageing (q-value<0.1) (**Figure 2.10B/right**). This moonlighting protein has been shown to be a regulator of several multi-enzyme complexes involved in the energy metabolism, and might therefore be a mediating factor for nutrition-dependent stoichiometric adjustment in the metabolic household.

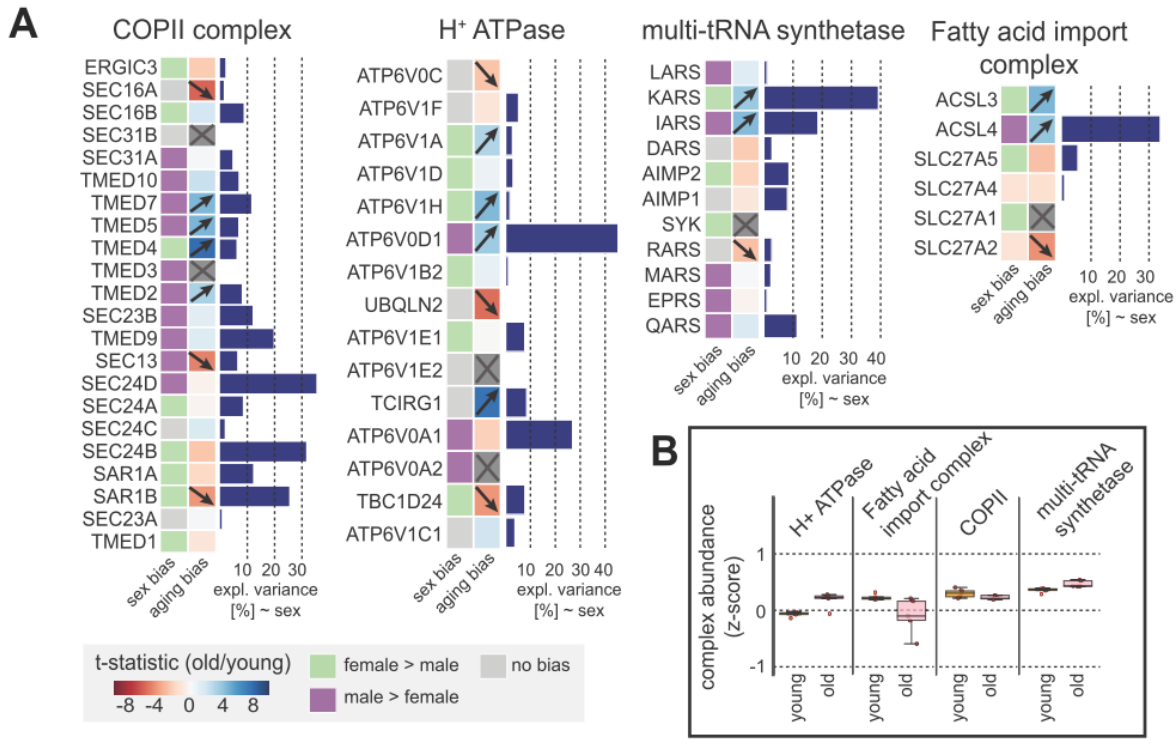


Figure 2.11. Sex- and age-dependent module stoichiometries. (a) Four sex- and ageing-dependent complexes are shown. For each complex, all subunits are shown, along with their sex bias (stoichiometrically more abundant in female (lightgreen), or in male (purple)) and how their relative stoichiometry changes between young and old NMRs (old/young t-statistic, q-value<0.1). The bar plot on the right-hand side illustrates the amount of variance that can be explained by the sex of the animal. (b) The box-plot shows the z-score distribution of the median abundance of the indicated complexes in young and old rats.

In contrast to translational processes being heavily affected by overall abundance changes during ageing, modules involved in molecule transport were impacted in their stoichiometry (**Figure 2.10B/left**), such as COPII, Fatty acid import complex, or the vacuolar H^+ ATPase. A more detailed view on age- and sex-dependent stoichiometric patterns for some of these complexes is shown in **Figure 2.11A**.

2.3 Discussion

The acquisition of large-scale proteotypes has become commonplace, and recent studies encompass larger cohorts of monitored individuals and more rigorous quantifications to assure reproducibility [Chick et al., 2016; Battle et al., 2015; Wu et al., 2013]. However, integrating these different proteomics datasets and thereby providing a comprehensive assessment on dataset quality as well as consistent biological features is scarcely attempted. Incidentally, understanding the effects that genetic and environmental factors have on the protein landscape and deducing robust conclusions remain hampered. Effects of an individual's sex on the proteotype, for example, remain largely centered around well-known mechanisms such as dosage compensation and differential expression of proteins due to their X/Y-chromosome location [Wu et al., 2013; Chick et al., 2016]; a large-scale view to what extent the sex of an individual actually influences the respective proteotype remains undescribed.

In this study, we comprehensively evaluated several aspects of the latter challenges by interrogating 11 unrelated MS-shotgun proteomic datasets on their modular architecture by co-variation analysis. Such an approach allowed us to both bench-mark the datasets relative to each other and to recover networks of proteins that were effectively co-varying across samples, specifically individuals. Protein complexes, for that matter, showed a particular consistency in the presence of stable sub-structures and variable components, such as the immuno-regulatory part of the proteasome complex. However, the underlying mechanisms leading to such differential stoichiometric robustness of modules remain to be further explored. So far both proteasomal degradation of unbound subunits [McShane et al., 2016; Ryan et al., 2017], as well as regulatory mechanisms at the RNA level [Wu et al., 2013] have been suggested as major drivers of stoichiometric robustness in complexes. Such mechanisms are probably not restricted to protein complexes only; for certain pathways, such as the complement pathway, we could observe very high and consistent co-variation of all its protein members involved. Also proteins in very coordinated processes such as the citric acid cycle showed a stable set-up in their relative stoichiometry across individuals. Notably, such pathways should be further interrogated to understand what variable members actually contribute to the differential proteotypes, and whether they consistently do so.

Our analysis further decomposes contributions to the proteotype variation from module overall abundance as well as module stoichiometry, and establishes to what extent both of these variations are determined by an individual's sex and diet. The effect sizes on variation in complex abundances were usually larger, with an average of more than 5%, than in complex stoichiometries (average <5%), which was expected due to the rigor in complex set-ups. Nonetheless, some clear-cut differential stoichiometry between male and female mouse strains was identified in transport-related processes, prominently involving the COPI- and COPII-complexes. At this point it remains unclear though whether the given stoichiometry has a truly genetic cause which can be traced back to X- and Y-associated gene expression, or whether it emerges from a different environmental effect, such as from hormone and receptor transport.

The latter question can incidentally be asked in reverse by suggesting that the given predefined stoichiometry of a functional module can ultimately change a cellular phenotype. Indeed, it remains to be seen what systemic effects such subtle stoichiometry changes can bring about, and how we can ultimately make use of it in a diagnostic and clinical context as well.

Chapter 3

Pervasive protein thermal stability variation during the cell cycle

This chapter presents part of the results of the collaborative work with Dr. Mikhail Savitski and Dr. Martin Beck on protein thermal profiles during the cell cycle. Not only does this work shed light on the changing biophysical parameters during the cell cycle such as stability and solubility, it also leverages the power of thermal profiles to understand protein complex assembly during the cell cycle. I was not involved in the generation of experimental data; experiments were primarily carried out by Dr. Isabelle Becher and Dr. Amparo Andres-Pons at the EMBL. The computational analysis, on the other hand, was handled by Dr. Frank Stein and me. My main contribution primarily encompassed characterizing the 'co-melting' behaviour of protein complexes throughout cell cycle progression, as well as examining disordered proteins in context of mitotic phase transition and a possible relation to mitotic phosphorylations.

The work in this chapter includes published material from the following article:

Pervasive protein thermal stability variation during the cell cycle. Becher I., Andres-Pons A., **Romanov N.**, Stein F., Maike Schramm, Florence Baudin, Dominic Helm, Nils Kurzawa, André Mateus, Marie-Therese Mackmull, Athanasios Typas, Christoph W. Müller, Peer Bork*, Martin Beck* and Mikhail M. Savitski*.*Cell*, 2018

3.1 Introduction

MS multiplexing technologies [Gillet *et al.*, 2012, Werner *et al.*, 2012, Werner *et al.*, 2014] have substantially broadened the scope of possible proteomic quantifications that can be made across a large number of biological conditions. Measuring protein melting or aggregation curves on a proteome- wide scale (Savitski *et al.*, 2014, coined as cellular thermal shift assay (CETSA)) is a particularly exciting way of exploiting this potential as it allows for the assessment of protein-drug, protein-protein interactions and post-translational modifications [Savitski *et al.*, 2014, Becher *et al.*, 2016, Reinhard *et al.*, 2015]. Thus, it adds another powerful dimension to the description of the cellular proteotype capturing protein stabilities in context of the biological processes they are involved in.
The typical experimental scenario starts with cells being cultured under differential conditions, and each culture being divided into 10 aliquots that get heated to different temperatures, followed by extraction with phosphate-buffered saline (PBS). Proteins- when reaching their melting temperature- then start to denature and gradually disappear from the PBS-extracted samples by aggregation at higher temperatures [Martinez Molina *et al.*, 2013; Asial *et al.*, 2013]. After protein extraction and trypsination, each sample gets labeled with a different isotope-coded isobaric mass tag (TMT10); subsequently all 10 samples are mixed and analyzed in a single liquid chromatography (LC) - MS run. In the resulting MS/MS fragment spectra, a curve is fitted then the reporter ion intensities and a protein-specific melting temperature (T_m) can be calculated and compared between conditions. The melting curves, on the other hand, correspond to the level of protein aggregation occurring at each temperature and thus directly reflect the protein-inherent unfolding behavior (Figure 1.16). This unfolding behavior could be influenced by ligand-binding or interactions with other protein, which should become evident in a respective shift of the melting point. Given that the method robustly recovers results from thermofluor unfolding assays and heat-induced precipitations in solution [Vedadi *et al.*, 2006; Asial *et al.*, 2013] it can be used to reliably map thermal stability features to each protein, and quantify differences across cellular conditions.
One particular cellular process that we explored with this methodology is the eukaryotic cell cycle- a temporal process encompassing many fundamental changes in cellular architecture. There are mainly four processes that need to be coordinated in order to ensure cellular survival: i) cell growth (G1-phase), ii) DNA replication (S-phase), along with chromosome condensation, iii) alignment and distribution of chromosomes to daughter cells (M-phase) and finally iv) cell division (cytokinesis) along with nuclear division which occurs already in the mitotic stage. While cell growth is a continuous process, the subsequent events (ii-iv) are sharply defined and bound to happen in a certain time-frame with cell

cycle checkpoints ensuring that the environmental conditions are appropriate for cell division [Morgan, 2007].
Using cell cycle synchronization protocols that enable to capture cells in one of those sharply defined stages has led to a wide range of discoveries about processes and mechanisms involved. For one thing, it is well-known that post-translational modifications (PTMs) play a crucial control in signaling via phosphorylation of the anaphase-promoting -complex, for example [Dephoure 2008, Olsen *et al.* 2010]. Ubiquitination also has been described as a means throughout the cell cycle [Nakayama and Nakayama, 2006, Pines 2006]. Changes in abundance levels due to degradation or stop of synthesis has also been well-characterized for a number of proteins. Such dramatic shifts in the protein landscape need to be tightly regulated to avoid "wasting" energy resources. De Lichtenberg *et al.* (2005) reported on a yeast mechanism to activate and de-activate protein complexes at specific cell-cycle phases by regulation of components of the complex critical for its functionality rather than the whole ensemble. That just-in-time tuning of protein complexes optimizes transcription of complex members while minimizing protein synthesis cost. It is in this chapter that we will further enlarge the idea to entire pathways, to see whether this could be a general cellular mechanism.
Arguably all these changes that have been described to occur during the cell cycle might not only affect *protein abundances* but also *protein stabilities*. There are several lines of evidence to suggest that: (i) mitotic processes lead to many proteins being exposed to varying biophysical environments [Jongsma *et al.* 2015]; (ii) extensive post-translational modifications might not only serve a signaling purpose but ultimately change the physical features of the protein and its stability [Bachant *et al.* 2002, Olsen *et al.* 2010, Pelisch *et al.* 2014]; (iii) protein complex rearrangements could induce stability changes to several members of the same complex [de Lichtenberg *et al.*, 2005]; (iv) the activity of several cell cycle regulators is controlled by selective degradation [Pines *et al.* 2006]. However, proteome organization during the cell cycle has never been explored with regard to protein stabilities, and it remains unknown to which extent they are affected and potentially regulated throughout the cell cycle.
In this work we have systematically measured protein melting curves during cell cycle progression on a proteome-wide scale *in situ*, revealing that protein stability regulation indeed affects various biological processes, such as transcription, spindle formation and key metabolic pathways. Protein stabilities are particularly affected during the mitotic stage, which coincides with the respective rearrangements occurring at this stage. Other phases, for that matter, are much less affected with the exception of early G1 that still recovers some of the protein stability changes that occurred during mitosis. The fact that mitosis and G1 are quite distinguished in their morphology emphasizes that the concept of protein stability is not confounded to morphological changes only, and is in fact a regulatory process

that needs to be maintained during cell cycle progression. We also observed a tendency of intrinsically disordered proteins to be stabilized during mitosis, coinciding with extensive sumoylation and mitotic phosphorylation. Such a behavior could suggest that stability regulation in general could serve the purpose of preventing protein aggregation during mitotic spindle formation and chromosome separation. In summary, this work has provided the research community with a comprehensive dataset that covers various aspects of transcription, structural biology, intrinsically disordered proteins, metabolism and cell cycle regulation.
This chapter will outline the main results of this work, and mainly the aspects that the author of this thesis has been involved in. In line with the global theme of this thesis, a particular focus will be the rearrangements of complexes during cell cycle progression.

3.2 Results

3.2.1. Profiling the thermal stability, abundance and solubility of proteins during the cell cycle

To examine potential protein stability changes across different cell cycle stages in situ, the thermal proteome profiling technology (TPP) was employed allowing to both measure thermal stability and abundance [Savitski *et al.*, 2014]. To this end synchronized HeLa cells were collected in six different cell cycle stages, namely G1/S transition, early S phase, late S phase, S/G2 transition, mitosis and early G1 phase (see Materials & Methods). For control purposes asynchronous cells were collected as well (**Figure 3.1**). As described in the introduction to this chapter, intact cells would then be heated to temperatures ranging from 37°C-66.3°C. Subsequently they were lysed with a mild detergent (NP-40) and soluble proteins were quantified in the LC-MS setup using the tandem mass tag (TMT) labeling [Werner *et al.* 2014].
To make sure that we can compare stabilities and abundances across cell cycle stages, a two-dimensional TPP (*2D-TPP*) setup was established, multiplexing different cell cycle stages at specific temperatures (Becher *et al.*, 2016, see Materials & Methods) with G1/S phase as the reference condition. Thereby out of 10.064 identified proteins, 4.970 proteins passing quality control requirements would effectively have a two-dimensional data matrix across cell cycle stages and temperatures (**Table S3.1**). Such a set-up made it then possible to calculate abundance and stability scores for each protein and to assess significance based on three biological replicates (see Materials & Methods) (**Table S3.2**). To further disentangle whether protein abundance shifts might be due to altered protein solubility (i.e. transition from organelle-bound to a free state due to cell-cycle dependent organelle disintegration), the total protein amount (for 5.835 proteins) was quantified after

solubilizing with a strong detergent (SDS), and solubility scores were calculated accordingly (see Materials & Methods) (**Table S3.3**). Finally, in order to capture actual protein melting curves, the TPP temperature range (*TPP-TR*) approach was applied to mitotic and G1/S-cells, multiplexing ten different temperatures for each stage[Huber *et al.* 2015, Reinhard *et al.* 2015, Savitski *et al.* 2014] and generating melting curves for 5.263 proteins (**Table S3.4**). Thus, for at least half the proteome usually expressed in human cell lines [Beck *et al.* 2011, Nagaraj *et al.*, 2011] this work provides comprehensive information on abundances, stabilities and solubilities in context of cell cycle progression.

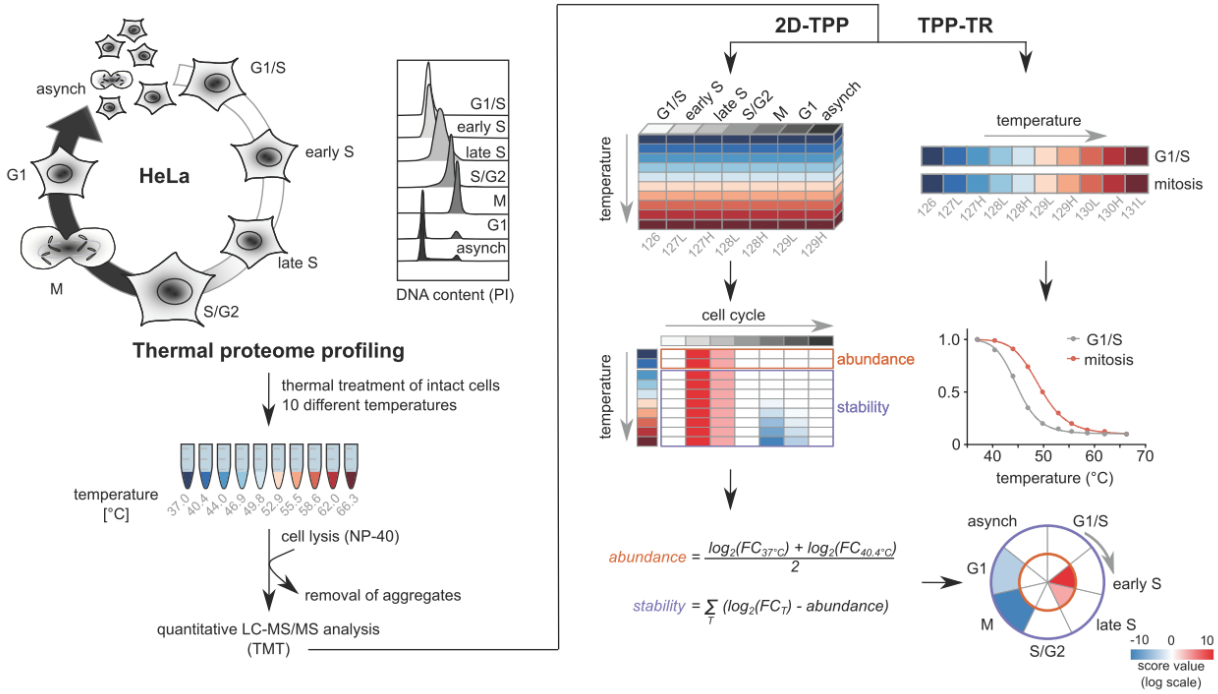


Figure 3.1. Overview on experimental design to assess protein abundance and stability changes across the cell cycle (graph primarily designed by Isabelle Becher). After arrest of HeLa cells in respective cell cycle stage (G1/S, early S, late S, S/G2, M and G1), samples were heated to 10 different temperatures followed by cell lysis with NP-40. Protein aggregates get removed and soluble protein fractions are labelled with TMT10-plex for further protein quantification. In TPP-TR mode (right panel), samples are combined to allow assesing melting curves for G1/S or mitosis. In 2D-TPP mode (left panel), samples are combined according to a 2D-matrix with cell cycle stages pooled for one thermal treatment (TMT7-plex de facto for each temperature). G1/S is always used as a reference for fold-change calculation. The resulting matrix for each protein gets condensed to a respective abundance and stability score. Latter are used to deduce significant changes, visualized in circle plots for each protein (inner circle (orange): abundance changes; outer circle (purple): stability changes). An additional aliquot of cells is lysed with the strong detergent SDS (not shown in scheme) in order to quantify protien abundance changes independent of protein solubility.

To make sure that the dataset indeed reflects characteristic biological features of the cell cycle, we first checked on cyclin proteins as well as cyclin-dependent kinases (CDKs). For example, we could recover the expected abundance profile of cyclin-A2 (CCNA2) –

increasing in late G1 and declining in prometaphase [Morgan, 2007]; yet no stability variation occurred with CCNA2 (**Figure 3.2A, Figure S3.1A**). CDK1, on the other hand, did not show any significant changes in relative abundance across the cell cycle stages at lower temperatures, but did so at higher temperatures (**Figure 3.2A, Figure S3.1A**). Thus, our data reveals that the thermal stability of CDK1 dramatically declines after G2. Apart from these two examples, many other proteins that are known to play functional roles during the cell cycle, were found to be either affected in their abundance or their stability.

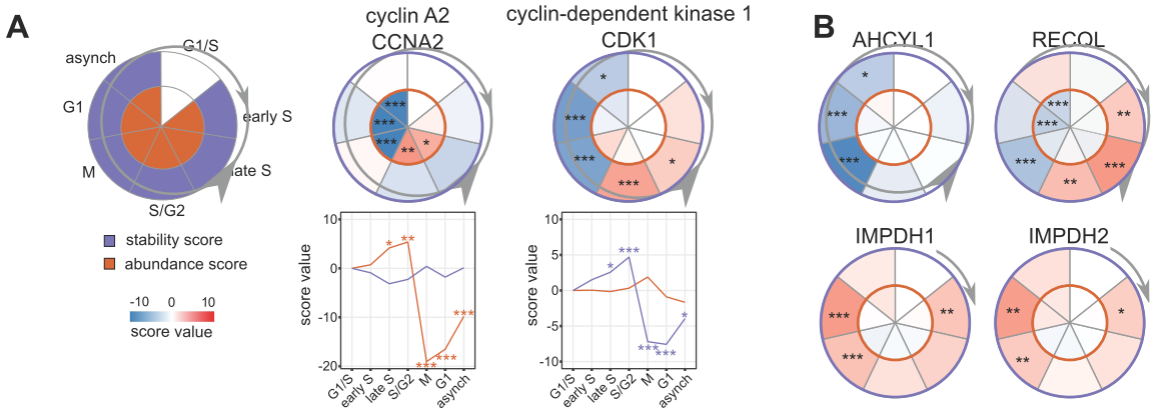


Figure 3.2. Abundance and stability changes of established cell cycle markers (graph primarily designed by Frank Stein). (a) CCNA2 and CDK1 are shown in circular plots (outer circle (purple): stability scores; inner circle (orange): abundance scores). Color (blue to red) indicates decrease and increase of score, respectively. Line plots are depicted as well for additional illustration of the signal. All data presented is based on 3 independent biological replicates, with * p <0.05, ** p <0.01, *** p <0.001, unless otherwise indicated. (b) Further exemplary circle plots of cell-cycle related proteins with observed stability changes.

3.2.2. Protein abundance and stability vary independently from each other during the cell cycle.

Significant cell-cycle dependent changes in protein thermal stability and abundance has been detected for 1.592 and 1.044 unique proteins, respectively, with most abundance changes occurring in early G1 and most stability changes occurring in mitosis (**Table S3.2** and **Figure 3.3A**). Changes that occurred in both latter phases were notably much more pronounced compared to any other cell cycle stage.

To understand which cellular processes and sub-structures were most prominently affected by cell-cycle dependent changes in stability, we performed a gene ontology (GO) analysis using DAVID (see Materials & Methods). Various biological processes, such as the nuclear envelope breakdown, the purine salvage pathway and the pentose phosphate pathway [Fridman *et al.*, 2013] were identified to be affected (**Table S3.5**). During mitosis in particular, a stabilization of signaling-, DNA- and chromatin-associated proteins occurred,

whereas destabilization was primarily affecting proteins involved in metabolic processes. We found no obvious differences in stability distributions of different subcellular compartments, with the exception of the endoplasmic reticulum (**Figure S3.2A**). Strikingly, the endoplasmic reticulum contained relatively more proteins with a variable stability, especially stabilized proteins, than other organelles (p-value=3.04x10^{-5}, Fisher Exact test, **Figure S3.2A/B**). That might reflect morphological changes of the endoplasmic reticulum [Schwarz and Blower, 2016] and/or the down-regulation of the secretory pathway during mitosis [Yeong, 2013].

In order to explore a potential relationship between stability and abundance, abundance and stability scores were combined and clustered using Euclidean hierarchical clustering using the ward-method **(Figure 3.4**, see Materials & Methods, **Table S3.2**). The resulting 21 clusters strongly delineate proteins that are either affected in their stability or abundance, indicative of those features being controlled independently (**Figure 3.4**). To further substantiate that stability-affected proteins are confounded by their abundance metrics, we compared our data with previously determined protein half-lives [Boisvert al., 2012]. Proteins significantly changing in abundance across cell cycle stages tend to turn over faster, as opposed to proteins that are affected in their stability (**Figure 3.5A**). Furthermore, protein melting points are not correlated with protein half-lives (**Figure 3.5B**), suggesting that cell-cycle dependent stability variation cannot be recapitulated from standard protein abundance measurements.

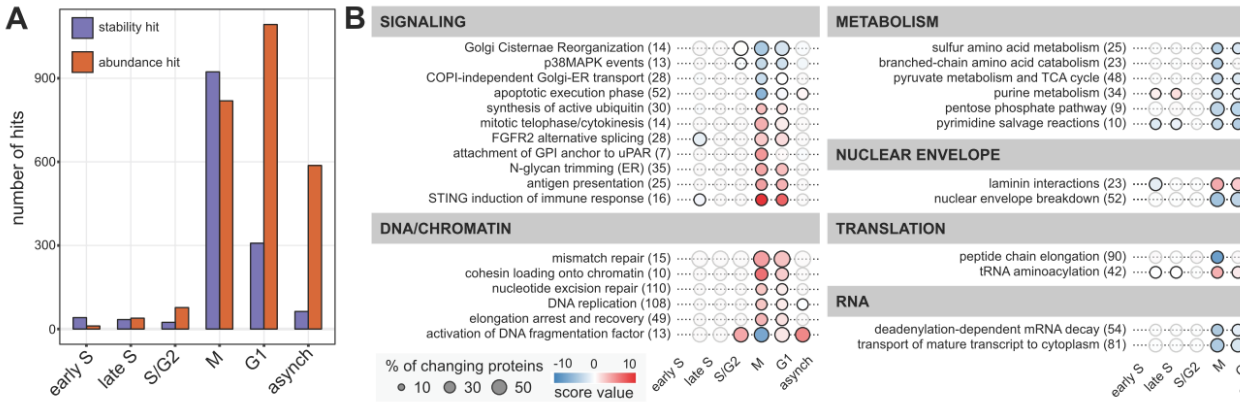


Figure 3.3. Overview on abundance and stability hits, and affected pathways. (a) Number of significantly affected proteins in either stability (purple) or abundance (orange). (b) Selected pathways from Reactome database (see Materials & Methods) with prominent protein stability changes (number of proteins associated with pathway given in brackets). Size of circles relates to the fraction of the pathway affected in at least one cell cycle stage; circle colour indicates the mean stability of the pathway components that significantly change their stability (full opacity: p<0.05; transparency: no significant stability hits in given cell cycle stage despite protein quantification). Pathways are categorized into broader functional groups.

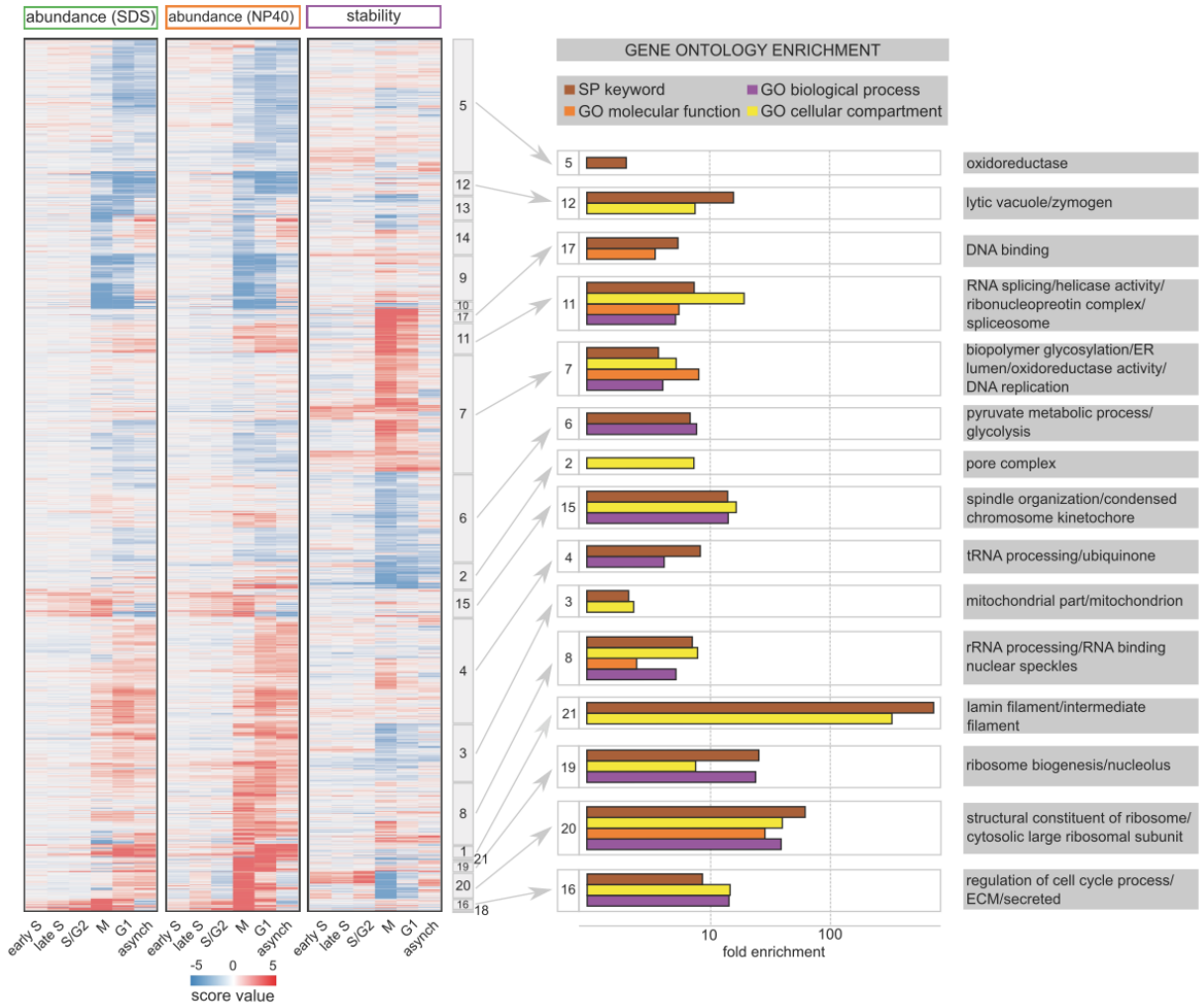


Figure 3.4. Clustering and Gene Ontology of Cell Cycle Hits. Hierarchical clustering of stability, abundance and expression matrix of significantly changing proteins, resulting in 21 individual clusters. Each cluster was analysed with DAVID (https://david-d.ncifcrf.gov) using all quantified proteins used as a background. For each cluster the top GO result for different GO-categories are illustrated (FDR<0.05) on the right-hand side.

Recovering stability as an inherent feature of the protein proves useful to tackle challenges in identifying metabolic enzymes that are more active due to increased substrate availability rather than increased enzyme abundance. It has been established previously that substrate binding indeed increases the thermal stability of enzymes [Feng *et al.*, 2014; Savitski *et al.*, 2014], hence stability features in general can serve as a proxy for enzyme activity. In our data we indeed observed cell-cycle dependent variation in enzyme stability in several interconnected metabolic pathways (**Figure 3.6, Table S3.5**). The fatty acid biosynthesis pathway, for example, which is required for mitotic exit, has its key enzymes significantly stabilized in mitosis and early G1 [Scaglia *et al.*, 2014]. The non-oxidative branch of the pentose phosphate pathway (PPP) also has its respective enzymes significantly stabilized,

which is in line with previous reports on increased metabolic flux at G1/S [Fridman *et al.*, 2013].

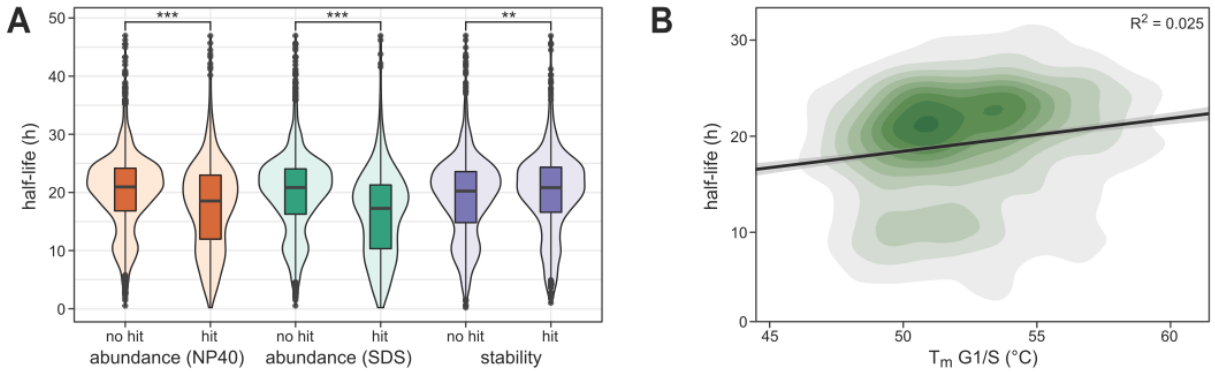


Figure 3.5. Analysis on protein half-lives in context of cell cycle stability changes (graph primarily designed by Mikhail Savitski and André Mateus). (a) Violin plots comparing protein half-lives [Boisvert *et al.*, 2012] of proteins changing significantly in abundance (orange (NP-40), green (SDS), and stability (purple))in comparison to non-changing proteins. Significance levels obtained from Wilcoxon-rank test: * p<0.05, ** p<0.01, *** p<0.001. (b) 2D-density plot of protein half-lives in comparison with protein melting points in G1/S. Linear model is fitted to it.

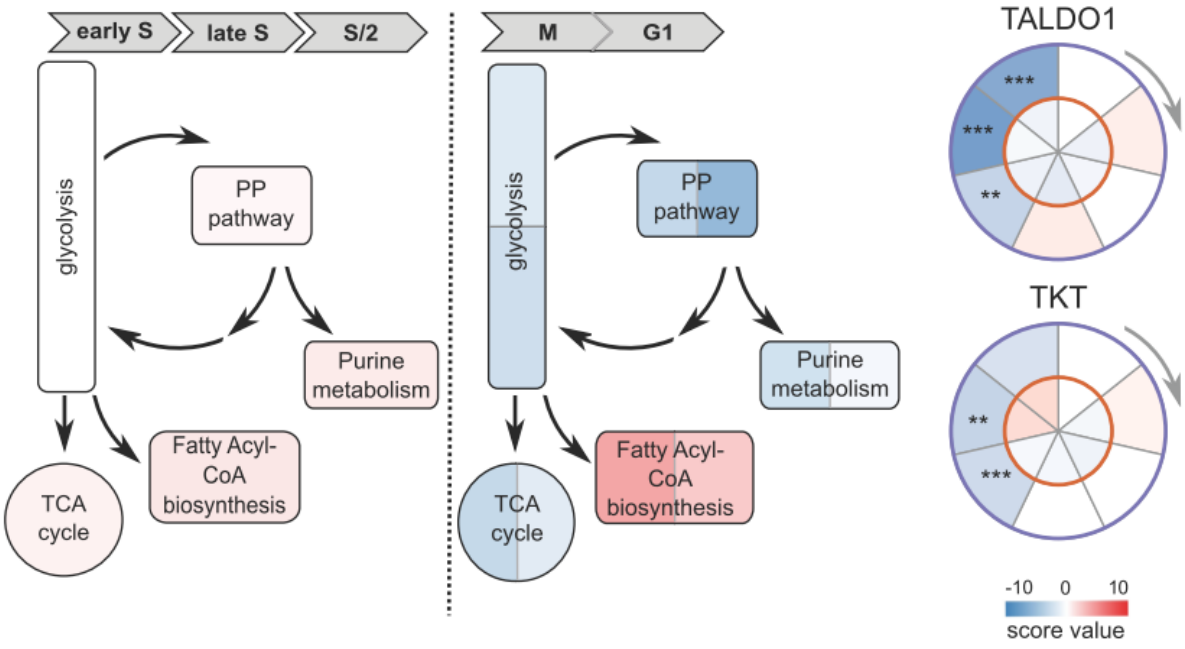


Figure 3.6. Cell-cycle related changes in abundance and stability of metabolic pathways. On the left-hand side, the pentose-phosphate (PP) pathway is depicted along with connected pathways, coloured according to average stability across the stages from early S to S/G2. On the right-hand side, the same constellation is shown, coloured in stability values derived from mitosis (*left*) and G1-phase (*right*) for each protein. TKT and TALDO1 (circle plots shown) are two enzymes of the PP pathway that get significantly destabilized upon the onset of mitosis.

3.2.3. Complex-dependent variation in stability across the cell cycle

In the following analysis we examined how members of protein complexes are coordinated during cell cycle progression both in terms of their abundance and stability features. We calculated the correlation of the abundance and stability scores of proteins that are members of the same annotated complex (as defined by Ori *et al.*, 2016) and compared the resulting distribution to correlation values stemming from all other proteins that are not part of any of the annotated complexes (**Figure 3.7A**). For both abundance and stability profiles of protein complex subunits we observed a shift towards higher correlation values as opposed to random protein-protein interactions (**Figure 3.7B-C**). The temporal adjustment of complex subunit abundances has already been demonstrated before in Ori *et al.* 2016; however, our observations on correlated protein stabilities actually suggest that protein complexes mostly melt as a whole unit *in situ* once a critical temperature is reached. Indeed, protein complex subunits have a significant tendency towards coherent melting behavior (**Figure S3.3A**). Some complexes displayed strongly correlating subunits for stability while others correlated better for abundance, which also held true for annotated cell cycle complexes with temporally regulated assembly [Jensen *et al.* 2006] (**Figure S3.3B**). Combining the stability and abundance values for each protein yielded the best discrimination between proteins that are part of complexes from those that are not (**Figure 3.7D**).

Going into more detail we discovered that complex member correlation patterns were actually reflecting stable sub-complexes, and vice versa loosely associated or potentially moonlighting subunits (Mendeley graphs, http://dx.doi.org/10.17632/xrbmvv5srs.2). The exosome complex provides an illustrative case in point (**Figure S3.4A**): While its core components were collectively destabilized in mitosis and generally behaved in sync throughout all stages, its two catalytic subunits, DIS3 and EXOSC10, were stabilized in mitosis and fell out of pattern. Notably, latter stabilization is in agreement with previous reports for fission yeast [Murakami *et al.* 2007] and Drosophila [Graham *et al.*, 2009] where those two subunits have been identified as uniquely required for mitotic progression. Such sub-structural patterns could also be recovered for other complexes, i.e. the 26S proteasome, and the condensin complex (**Figure S3.4B, Figure S3.4C**), substantiating the notion that the presented method is indeed able to detect sub-complex specific changes.

That becomes especially apparent with the nuclear pore complex (NPC), which needs to be dis-and re-assembled during cell cycle progression and consequently is subject to fundamental stoichiometric changes. While membrane-bound in interphase, it resolves into soluble sub-parts during mitosis after a phosphorylation trigger [Laurell *et al.*, 2011]. It is still not clear, however, what part of the nuclear pore complex is affected first, and how

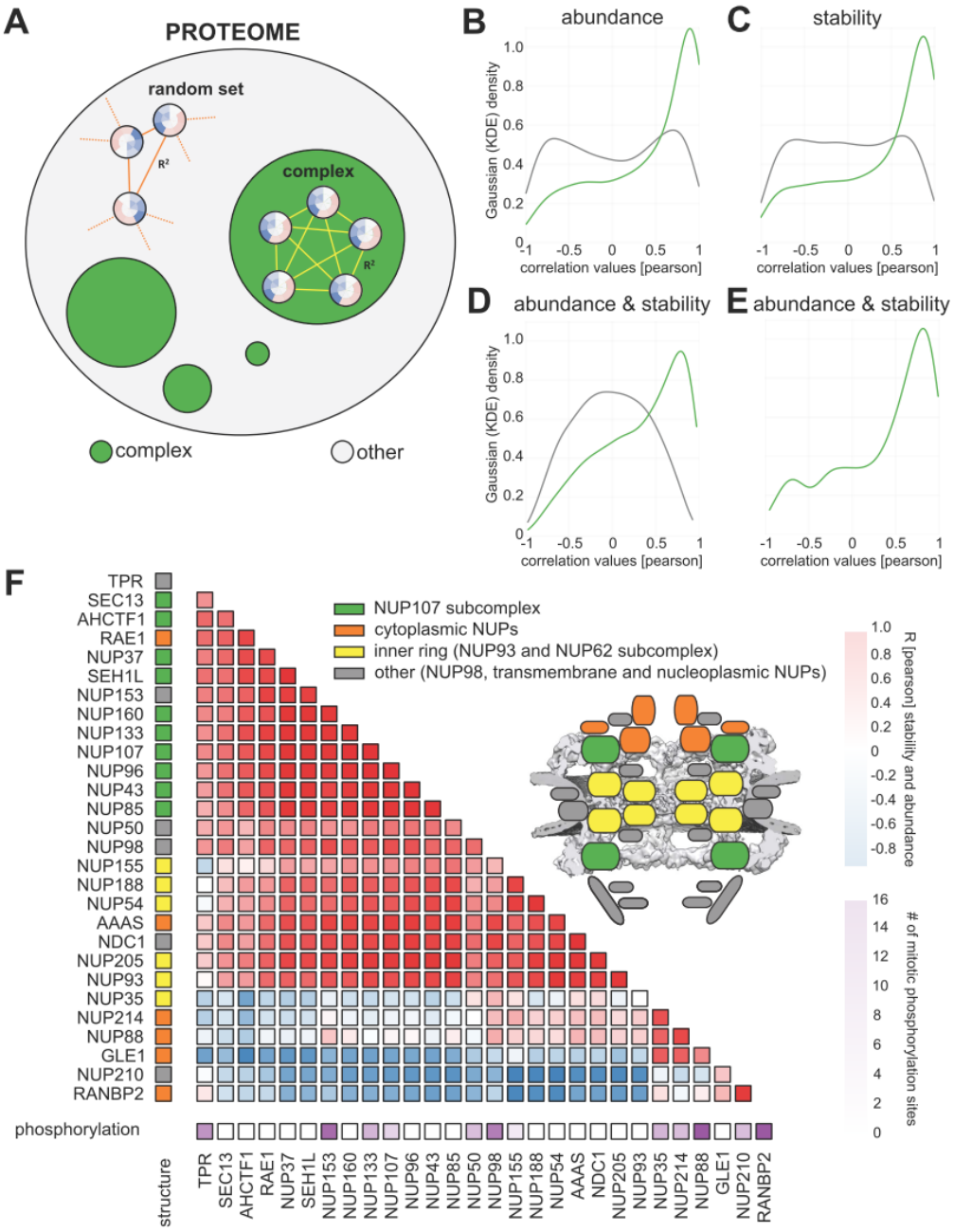


Figure 3.7. Co-stability of known protein complexes and sub-modules of the NPC. (a) Schematic illustration of correlation analysis (see Materials & Methods for further details). (b/c/d) Density graph of correlation coefficient values (Pearson) calculated from abundance (b), stability (c), and concatanated abundance-stability (d) profiles between proteins known to be members of the same complex (green). They grey density shows correlation values from all combinations of proteins that are not associated to any complex. (e) Density graph of correlation valeus (Pearson) calculated from concatanated abundance-stability profiles of all subunits of the Nuclear Pore Complex (NPC). (f) Correlation matrix of NPC-subunits based on their concatanated abundance-stability profiles. The colours on the left-hand side indicate their association with a specific sub-structure of the NPC, as coloured in the respective cartoon. The scale on the bottom of the matrix indicates how many mitotic phosphorylation sites the corresponding protein subunit contains as described by Olsen *et al.*, 2010.

such a dis-assembly would progress. Shedding light on the different stability and abundance profiles of potential sub-complexes of the NPC might provide a means to tackle this challenge (**Figure 3.7E, Figure 3.7F**). Indeed, combined stability- and abundance profiles of nucleoporins (Nup) neatly cluster the proteins into known sub-complexes, such as the NUP107, the inner ring sub-complex and accessory proteins. The major scaffolding complexes show a minor destabilization in mitosis, coherent with it being detached from the NPC scaffold; notably, the effect sizes for the so-called Y- and inner ring complexes are slightly different, which is probably due to architectural disparities and subsequent influence on extractability (**Figure 3.7F, Figure S3.5**). The most intrinsically disordered Nups, on the other hand, remained stable in mitosis or were even strongly stabilized, such as Nup358 (RANBP2). Given that many of the Nups with high IDP content are prone to aggregation [Lemke *et al.*, 2016], this is indeed an unexpected observation.

3.2.4. Disordered proteins are stabilized during mitosis

There are two important characteristics of NUP358 that render its strong stabilization during mitosis quite intriguing: For one thing, it is large, multifunctional proteins with several folded domains linked by long, intrinsically disordered regions relevant for nucleocytoplasmic exchange in interphase. On the other hand, NUP358 dissociates from the Y-complex and localizes to the spindle region during mitosis [Joseph *et al.*, 2004; Joseph *et al.*, 2002]; this translocation is crucial as a depletion of NUP358 notably leads to spindle defects [Hashizume *et al.*, 2013]. We therefore sought to investigate whether (i) spindle-binding, and (ii) disordered proteins had a distinct (de-)stabilization behavior during the cell cycle.

Similarly to NUP358, other confirmed spindle-binding proteins showed a strong mitotic stabilization, such as the chromodomain-helicase-DNA-binding protein 4 (CHD4, Yokoyama *et al.*, 2013), chromosome-associated kinesin (KIF4A, Kurasawa *et al.*, 2004), as well as SMARCA4 (Yoyokama *et al.* 2009), as is illustrated **Figure 3.8A** and **Figure 3.8B** (**Figure S3.6A and B**). Examining all annotated mitotic spindle-binding proteins (Sauer *et al.*, 2005) revealed that these proteins in general have a significantly lower thermal stability than other proteins (**Figure S3.6C, Table S3.6**). Given that the aggregation of several of these proteins is already substantial at 47°C (around 30% in case of NUP358, **Figure 3.8A**), we assumed a severe dis-balance on spindle integrity due to these features, and experimentally validated that (**Figure 3.8C**).

Generally, proteins that were severely stabilized in mitosis while their inherent stability was rather low, were significantly enriched ($FDR<5\%$) for GO-terms related to DNA-binding, chromosome- and chromatin organization and transcription-related processes (**Figure S3.7,**

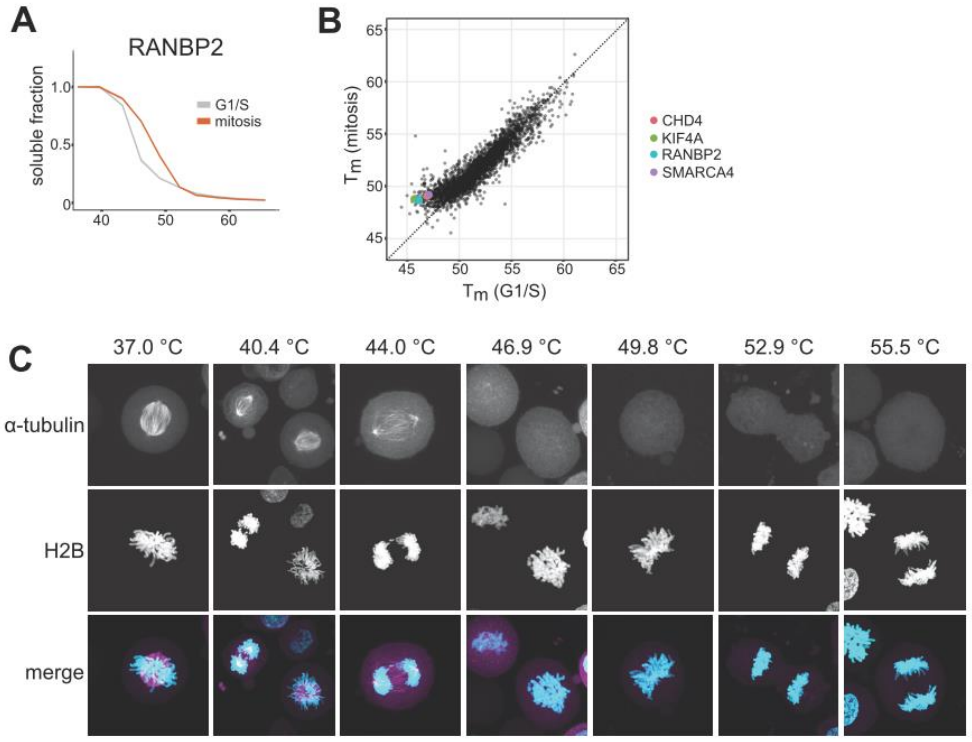


Figure 3.8. Stabilization of spindle-associated proteins and the mitotic spindle (graph primarily designed by Amparo Andres-Pons and Frank Stein). (a) Melting curves of RANBP2 in G1/S and mitosis (data again based on 3 biological replicates). (b) Scatter plot comparing the melting temperatures (Tm) for proteins in G1/S (x-axis) and M (y-axis). Shift towards higher melting points in mitosis for spindle-related proteins CHD4, KIF4A, RANBP2 and SMARCA4 are indicated with colouring. (c) Microscopy images of the mitotic spindle at different temperatures. Mitotic HeLa Kyoto EGFP-alpha-tubulin/H2B-mCherry cells were imaged after heat treatment. Z-stacks maximum intensity projections of samples treated at the indicated temperatures are shown. On the merged images, α-tubulin and H2B are shown in magenta and cyan, respectively. After heat treatment at 47°C the mitotic spindle was barely visible, and at higher temperatures was completely disintegrated.

Table S3.5). More specifically, however, we found that this particular set of proteins was also enriched in proteins with large proportions of disordered regions (**Figure 3.9A**, see *Materials & Methods*). We gathered that such a synchronized stability change of disordered proteins during mitosis might be due to a regulatory switch, i.e. mitotic phosphorylations which are known to occur in disordered regions [Dephoure *et al.*, 2008; Olsen *et al.*, 2010; Typanova *et al.*, 2013], thereby instantiating folding of disordered protein parts [Bah *et al.*, 2015; Desjardins *et al.*, 2014].

We indeed found proteins with annotated mitotic phosphorylation sites (Olsen *et al.*, 2010) to be stabilized in mitosis and detected a considerable overlap of disordered and mitotically phosphorylated proteins that exhibit the strongest stabilization (**Figure 3.9B, Figure 3.9C, Table S3.6**). Notably, such a pattern could not be recovered for non-mitotic phosphorylation, ubiquitination, acetylation or methylation sites (**Figure S3.8A**, Hornbeck *et al.*, 2015). Solely proteins with sumoylation sites showed a similar trend, emphasizing its

previously described connection to cell cycle progression and mitosis [Bachant *et al.*, 2002; Pelisch *et al.*, 2014], crosstalk with phosphorylation [Yao *et al.*, 2011], as well as phase separation [Lallemand-Breitenbach and de The, 2010].
Overall, our findings indicate that chromatin- and spindle-associated proteins with low stability and disordered regions are significantly stabilized in mitosis. This effect might also be driven by mitotic phosphorylation sites in the first place.

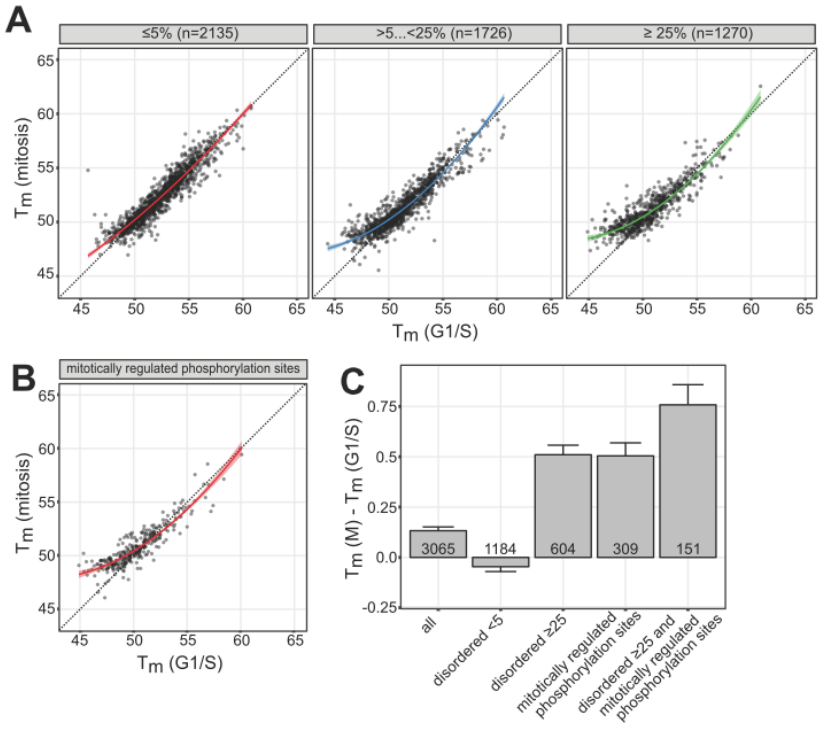


Figure 3.9. Stabilization during mitosis is prominent for disordered proteins and proteins containing mitotically regulated phosphorylation sites. (a) Scatter plots depict the melting point shift from G1/S (*x*-axis) to M (*y*-axis) at different cut-offs of the disordered state of a protein. (b) Shift of melting point for proteins containing mitocally regulated phosphorylation sites as described in Olsen *et al.*, 2010. (c) Melting point difference for proteins with different levels of disordered regions in their sequence and containing mitotic phosphorylation sites.

3.2.5. Persisting stability change at the transition between mitosis and G1

We found 984 proteins that show a significant stability change in mitosis that persists to G1 as well. These are primarily components of the RNA-processing machinery including RNA splicing, as well as the ribosomal complex. For the ribosomal complex, we already described the likely biological underpinning for its "leaking" destabilization (part of the subunits are recovered immediately in their stability; some subunits are not (primarily from the small ribosomal subunit)). Besides RNA-binding proteins, there were two other enriched processes worth pointing out: First, we found a persistent stabilization of proteins involved

in histone ubiquitination, such as 4 out of 5 components of the polymerase associated factor (PAF1) complex (PAF1, CDC73, LEO1, CTR9), which associates with the RNA polymerase II subunit POLR2A and with a histone methyltransferase complex. The complex plays a role in histone modifications, primarily ubiquitylation and methylation. Along with that we see a persistent stabilization of two E3 ubiquitin-protein ligases (RNF20, RNF40), one E2 ubiquitin-conjugating enzyme (UBE2E1), and CBX8 (chromobox 8), which has been related to ubiquitin protein transferase activity. Lastly, SUZ12 – a component of the very histone methyltransferase complex the PAF1 complex attaches to, is among these proteins that display increased stability in G1 as well. Both ubiquitination of histones, as well as the specific methylation marks set by latter complex, are known to lead to transcriptional repression. Since stability might serve as a proxy for enzyme activity, as we demonstrated for POLR2A (see manuscript), we assume that this coherent stabilization pattern that leaks from mitosis into G1, might be indicative of the onset of transcriptional repression as well. The other process to be enriched in proteins with persisting stability patterns, is the pyruvate metabolic process. Most prominently, the entire pyruvate dehydrogenase complex, is being kept destabilized in mitosis and G1 (PDHA1, PDHB, DLAT, DLD), as well as metabolic enzymes that lead to the production of pyruvate, such as the lactate dehydrogenases LDHA, LDHB and the malic enzyme ME3, or use pyruvate for further reaction, i.e. the pyruvate carboxylase PC. Notably, basigin (BSG) is also being kept destabilized, which is normally important for targeting the monocarboxylate transporters SLC16A1, SLC16A3 and SLC16A8 to the plasma membrane. These transporters are pivotal in getting lactic acid and pyruvate into the cell from outside; we quantify only SLC16A1 and SLC16A3 and both of them show the aforementioned "leaked" de-stabilization. It would seem as if the cell is trying to put a number of safeguards into place to ensure that no pyruvate is being processed and the energy-yielding machinery is completely shut down.

3.2.6. Solubility changes capture cell-cycle-dependent phase transitions

As specified in 3.2.1, protein abundances were measured under conditions using mild and strong detergent, respectively, in order to disentangle protein expression from solubility. In this manner the entire proteome could be de-convoluted into well-defined sub-proteomes due to strong solubility transitions at specific cell cycle stages. For example, ribosomal proteins, nuclear lamins, vimentin, plectin and their interactors [Lechertier *et al.*, 2009] displayed a strong transition from a non-soluble to soluble state in mitosis (Clusters 19, 20, and 21; **Figure 3.4, Figure 3.10A**). These observations fall in line with properties of cellular sub-structures observed during at this cell cycle stage: Nuclear lamins and vimentin both get phosphorylated which leads to de-polymerization of filaments, and plectin transfers into a soluble state by the same trigger [Chou *et al.*, 1990; Foisner *et al.* 1996]. Remarkably,

nuclear lamins remain insoluble in early G1 as well with high abundance and reduced stability (**Figure 3.10A/B**).
Such strong solubility transitions primarily captured components of phase-separated, membrane-less organelles that get dissolved in mitosis, such as nucleolar proteins [Hernandez-Verdun, 2011] (**Figure 3.10A**) or components of nuclear speckles (**Figure 3.4** cluster 8, Spector and Lamond, 2011; **Table S3.2**). To assess the fraction of proteins in soluble and insoluble state in more detail, separate TPP-TR experiments were conducted in mitosis and the G1/S-stage under mild and strong detergent conditions (as outlined in 4.2.1, **Figure S3.9A, Table S3.7**).

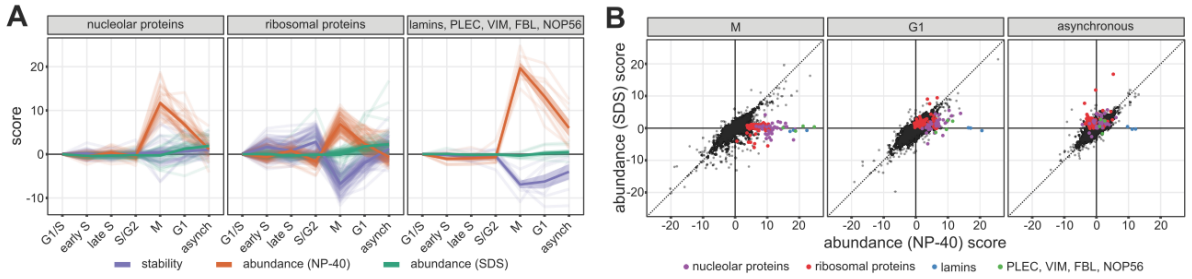


Figure 3.10. Solubility transition of nucleolar, ribosomal and lamin proteins. (a) Abundance (orange: NP-40, green: SDS), and stability (purple) profiles of three types of proteins (nucleolar, ribosomal, lamins and others from cluster 19,20,21 (Figure 5.4)). (b) Scatter plot comparing NP-40-based abundance score (x-axis) with SDS-based abundance score (y-axis) for each protein. The functional groups from (a) are indicated. Lamins notably remain soluble during the G1 phase.

The experiments revealed that a small fraction of the proteome remains consistently insoluble throughout G1/S and mitosis, including core members of the chromosomal passenger complex (CDCA8, INCENP, and AURKB, **Figure S3.9A**), as well as members of the FACT complex (SUPT16H, SSRP1). Both these complexes are known to play key roles during mitosis, with the chromosomal passenger complex being a key regulator [Carmena *et al.*, 2012] and the FACT complex being essential for microtubule growth and bundling [Zeng *et al.*, 2010] (**Table S3.7**). Curiously, we did not observe a single protein that would *transition* significantly from a soluble to an insoluble state upon mitotic entry; most of the proteins were in fact becoming more soluble in mitosis. Latter were enriched in RNA-binding functions (**Figure S3.10A**), mitotic phosphorylations (**Figure 3.11A/B**) and sumoylation (**Figure S3.10B**) (**Table S3.6**). As expected, many of these solubilized proteins were part of the nuclear envelope [Thul *et al.*, 2017; Wilkie *et al.*, 2011], as well as membrane-less organelles [Thul *et al.*, 2017] (**Figure 3.11C/D**). The solubilized sub-proteome of the membrane-less organelles were strongly enriched in disordered regions when compared to other proteins of the same organelles (**Figure 3.11A**), such as nucleostemin (GNL3, Romanova *et al.*, 2009), fibrillarin (FBL, Tiku *et al.*, 2016), nucleolar protein 7

(NOL7, Kinor and Shav-Tal, 2011), nucleolar protein 11 (NOL11, Freed *et al.*, 2012) and SON, which has recently been proposed to be a key scaffolding factor for nuclear speckles [Sharma *et al.*, 2010] (**Figure 3.11C**). That very distinct regulated set of proteins could in principle suggest that membrane-less organelles have a less extensive core proteome than previously thought [Thul *et al.*, 2017]. This could be of relevance given that the nucleolus is the site of ribosome assembly [Kressler *et al.*, 2010], which is also impacted by such strong solubility transitions upon mitotic entry. Part of the observed ribosomal proteins solubilizes in mitosis, which is probably represent non-assembled species (**Figure 3.11D**). Interestingly, the insoluble interphase sub-populations of the ribosomal small (40S) and large subunit (60S) components have very different effect sizes (**Figure 3.11D/E**), which is in agreement with 40S proteins residing a shorter time period in the nucleolus than the 60S proteins [Lam *et al.*, 2007]. The ribosomal associated complex (defined by Havugimana *et al.*, 2012) that ensures the maturation of the 60S proteins showed the largest effect size suggesting that the complex is primarily localized in the nucleolus in G1/S (**Figure 3.11D/E**).
This sub-part of the project has outlined a particular aspect of the proteome that so far remains under-explored. The solubility landscape of proteins during different cell-cycle stages, as well as the de-convolution of sub-proteomes that differ in their solubility, is useful in light of current discussion on phase separation that occurs upon mitosis. It also remains a challenge to understand exactly why disordered proteins are primarily affected, and whether it stems from their inherent physique or a higher-order regulatory switch.

3.3 Discussion

The work presented encompasses an in-depth and large-scale analysis of protein thermal stability and solubility in the eukaryotic cell cycle. Thus, the study adds substantial layers to our understanding of the proteome, and its variability and re-organization during the cell cycle. While we found most of the abundance changes to occur in G1, the mitotic stage was particularly characterized by a large proportion of protein stability changes, mirroring the respective morphological changes that occur at this stage. For example, the extensive re-organization of the ER during mitosis coincides with strong stability changes of ER-residing proteins [Schwarz and Blower, 2016], which has never been demonstrated before.
Another crude morphological change concerns the nuclear dis- and subsequent re-assembly at the end of M-phase. We observed a pronounced delay of nuclear lamina proteins to get insolubilized (**Figure 3.10A/B**), which could entail that lamin needs to be fully polymerized before final re-assembly [Dechat *et al.*, 2010; Moir *et al.*, 2000]. Many subunits of the NPC remain de-stabilized in early G1; thus, delayed lamin assembly could benefit NPC assembly that occurs from inside out through the nuclear envelope [Otsuka *et al.*, 2016].

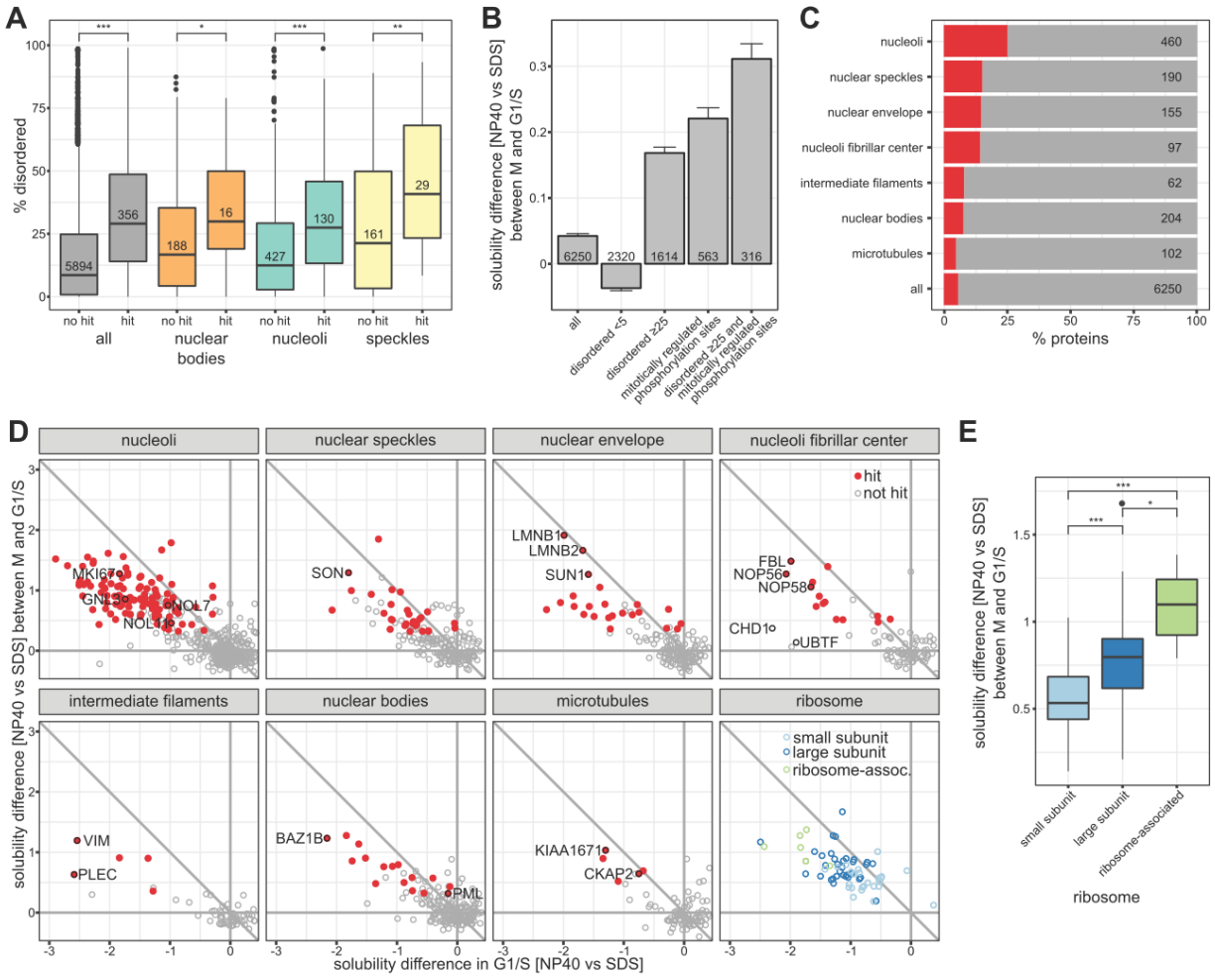


Figure 3.11. Sub-proteome transitioning in solubility in a cell-cycle dependent manner. (a) Proteins were stratified by compartments (all, nuclear bodies, nucleoli, speckles, based on the Human Protein Atlas (HPA)). For each category proteins are further dissected into hits (changing in solubility in mitosis vs. G1/S) and compared to other proteins in the category in terms of their fraction of disordered regions. (b) Solubility differences in mitosis vs. G1/S for proteins with different levels of disordered regions in their sequence and containing mitotic phosphorylation sites. (c) Fraction of organelle-specific sub-proteome significantly changing in solubility in mitosis vs. G1/S. Localization was defined by HPA, except for nuclear envelope, which combines nuclear membrane annotated proteins as defined by HPA and proteins annotated as inner nuclear membrane proteins by Wilkie *et al.*, 2011. (d) Scatter plots comparing the solubility of proteins in G1/S and the relative change in solubility of proteins in mitosis *vs.* G1/S, for different organelles, for the cytoplasmic ribosomal subunits, andthe ribosome associated complex determined by Havugimana *et al.*, 2012 (BOP1, RRS1, GNL3, EBNA1BP2, FTSJ3, MKI67). Proteins with negative x-axis values and close to zero y-axis values are insoluble in G1/S and remain insoluble in mitosis. Proteins with negative x-axis values and positive y-axis values are insoluble in G1/S and become more soluble in mitosis. Proteins that are significantly more soluble in mitosis compared to G1/S are marked in red. (e) Solubility differences in mitosis *vs.* G1/S compared between proteins from the small and large cytoplasmic ribosomal subunits and the ribosome associated complex (BOP1, RRS1, GNL3, EBNA1BP2,FTSJ3, MKI67) determined by Havugimana *et al.*, 2012.

Specifically looking at the dis- and assembly process of the NPC, we observed that intrinsically disordered Nups which usually tend to aggregate fairly easily, were stabilized in mitosis, such as NUP358, which facilitates scaffold Nup oligomerization in interphase [von Appen *et al.*, 2015]. Generally the stabilization phenomenon was predominantly occurring with highly disordered proteins that tend to have low stabilities and are targets of mitotic phosphorylations [Olsen *et al.*, 2010]. These observations are on par with the extensive morphological changes that occur in these stages, exposing proteins to radically different biophysical environments. One could speculate that IDPs get integrated into specific protein complexes or other modularities in interphase such that local concentration is restrained (i.e. NPC scaffold, Lemke *et al.*); in mitosis, however, such restraints would be lifted allowing for IDP aggregation. Mechanisms such as phosphorylation and separation from the membrane might help in preventing such aggregation scenarios during mitosis [Jiang *et al.*, 2015; Nott *et al.*, 2015].

The broad picture on differential biophysical features of the proteome becomes even more de-convoluted when taking into account actual solubility transitions and protein solubility. Solubility transitions occurred for 356 proteins, which were primarily disordered and mitotically phosphorylated, and primarily associated with the nucleolus. Notably, such radical solubility changes confirm the scale of morphological alterations, such as the dissolution of the nucleolus as shown by light microscopy. However, identifying the actual molecular factors responsible proves difficult since phase-separating, membrane-less organelles cannot be readily purified even from interphase cells.

Clearly, the presented landscape of biophysical properties of proteins varies extensively across the cell cycle, and arguably affects metabolic activity, complex composition and transcriptional activity (see manuscript). We demonstrated that stability is indeed a characteristic that cannot be recovered from protein abundance variation, and used as a proxy for protein activity and protein – protein interactions.

Indeed, we carefully studied protein-protein interactions in context of the cell cycle and revealed that the combined stability and abundance variation of complex subunits can be used for delineating major subcomplexes. In general, the subunits of well-annotated complexes (Ori *et al.*, 2016) exhibit concerted changes in their stability behavior. We identify clear examples of complex subunits that have known moonlighting functions and exhibit different cell cycle-dependent stability changes than other core subunits. This highlights the utility of the present dataset for structural biologists, as it can be used for hypothesis generation of cell cycle-dependent complex composition and subunit function (de Lichtenberg *et al.*, 2007).

Chapter 4

Cell-specific proteome analyses of human bone marrow reveal molecular mechanisms of age-dependent functional decline

In this chapter, I describe a comprehensive effort to describe the changes that occur in the proteomic landscape during the human ageing process and at the onset of myelodysplastic syndrome (MDS). It has been a major effort of Dr. Anne-Claude Gavin and Dr. Anthony Ho to gather the necessary data in the course of 3-4 years within the SyStemAge-grant framework. The samples were obtained by Dr. Patrick Horn, and the major proteomics analysis has been performed by Dr. Marco Hennrich. The computational handling of the data was largely my responsibility, under the supervision of Dr. Anne-Claude Gavin and Dr. Peer Bork. The chapter includes information from the following manuscript:

Cell-specific proteome analyses of human bone marrow reveal molecular mechanisms of age-dependent functional decline. Marco L. Hennrich, **Natalie Romanov**, Patrick Horn, Samira Jaeger, Volker Eckstein, Violetta Steeples, Fei Ye, Ximing Ding, Laura Poisa-Beiro, Mang Ching Lai, Benjamin Lang, Jaqueline Boultwood, Thomas Luft, Judith Zaugg, Andrea Pellagatti, Peer Bork, Patrick Aloy, Anne-Claude Gavin*, Anthony D. Ho*

4.1 Introduction

The hype is big when it comes to discussions about ageing in connection to human stem cells. Ageing of stem cells, in particular, has been considered to be the underlying cause for the functional decline of tissues and organs that accompany this process and eventually bring about the altered phenotype [Iscove and Nawa, 1997; Schlessinger and Van Zant, 2011]. The biological system maintained by stem cells encompasses haematopoiesis which is characterized by a high turnover rate. If failing it might lead to several hallmarks of ageing features, including anaemia, decreased competence of the adaptive immunes system, an expansion of myeloid cells at the expense of lymphopoiesis and a higher frequency of hematologic malignancies [Liang et al., 2005; Offner et al., 1999; Rossi et al., 2005].

These age-associated phenotypes have their origin at the very top of the haematopoietic hierarchy, namely in the so-called haematopoietic stem and progenitor cells (HPCs) [Geiger et al., 2013; Iscove and Nawa, 1997]. Previous research has shown that not only does the number of stem cells in the HPC population decline, cells gradually lose their ability to repopulate the bone marrow [Liang et al., 2005; Morrison et al., 1996; Offner et al., 1999; Sudo et al., 2000]. Transcriptomic data have provided a blueprint of potential molecular causes, including the HPC-specific up-regulation of genes associated with cell cycle progression, myeloid lineage specification, as well as myeloid malignancies when becoming older [Doulatov et al., 2010; Pang et al. 2011; Rossi et al. 2005]. However, these findings have been made mostly, if not exclusively, in murine ageing models, and remain to be validated in human subjects.

The effect that ageing might have on the microenvironment surrounding the HPC- coined the bone marrow 'niche'- also remains to be investigated. While specific adhesion molecules that are essential for homing and maintenance of HPCs have been thoroughly described, it is not clear to what extent they are affected by ageing processes as well [Beerman et al., 2017; Calvi et al., 2003; Ellis and Nilsson, 2012; Lo Celso et al., 2009; Schofield, 1978; Wagner et al., 2008a]. Similar to the intrinsic changes in HPC, the studies in that particular area have also been highly restricted to transcriptomic and epigenetic alterations, especially in the human mesenchymal stem and stromal cells (MSCs), which make up for the largest part in the bone marrow niche [Bork et al., 2010; Wagner et al., 2009]. Other cellular elements in the bone marrow such as monocytes and macrophages have also been shown to be implicated in the restructuring of the niche upon ageing [Chow et al., 2011; Ehninger and Trumpp, 2011; Winkler et al., 2010]. To date, however, mechanisms of ageing remain restricted to individual cell populations of the bone marrow and are not analyzed in context of the entire niche.

In the following work we provide a systematic study on the molecular mechanisms upon ageing throughout a number of distinct cell populations constituting the human bone

marrow, thereby dis-entangling intrinsic mechanisms (in the HPCs) from extrinsic ones as well as identifying inter-dependencies. Beyond transcriptional analysis, we performed a comprehensive, quantitative proteomics survey of the HPCs and their niche in a large cohort of healthy human subjects from different age groups. We further consolidate our findings using single-cell analysis to subdivide the HPCs into sub-populations, and dissect what enzymes are affected by ageing or rather by hematopoietic lineage skewing, which is characterized by the expansion of the myeloid sub-population of HPCs relative to the lymphoid sub-population.
The underlying datasets should not only represent a valuable resource for mechanistic analyses and for validation of knowledge gained from animal models, but they also provide a first atlas of the proteomic signature of human ageing process within the cellular network of the bone marrow. The systemic approach should build a foundation for a better understanding of age-related diseases such as myelodysplastic syndromes (MDS) in the future.

4.2 Results

4.2.1. Multi-scale, quantitative proteomics profiling of the human bone marrow

Bone marrow samples from 59 healthy human subjects (45 male and 14 female) with ages ranging from 20 to 60 years (median age ~ 33.2 years), were of sufficient and high quality and therefore submitted to proteomics analysis (**Figures 4.1A/B**, **Table S4.1**). More specifically, 6 cell sub-populations that constitute 94.2% (+/- 2.8%) of all mononuclear cells were isolated from each bone marrow sample, namely lymphocytes and respective precursors (LYM), monocytes/macrophages and respective progenitors (MON), granulocytic (GRA) and erythroid (ERP) precursors, mesenchymal stem/stromal cells (MSC) and finally HPCs by enriching for CD34 as a positive marker. These different cell populations were analyzed separately, and tryptic digests were labeled with tandem mass tag (TMT-6plex, Thompson et al. 2003) with 5 different human subjects in one batch, plus the population-specific internal standard for accurate quantifications across all donor (**Figure 4.1C**, see Method for details). In total, 12,158 proteins were identified throughout all human subjects and cell-populations, amounting to around 77% of the currently detectable human proteome [Willhelm et al., 2014] (**Figure 4.1D/E, Figure S4.1B**, **Table S4.2**). The number of proteins differed however between the cell populations, from 6,340 in ERPs to 9,454 in MSCs.

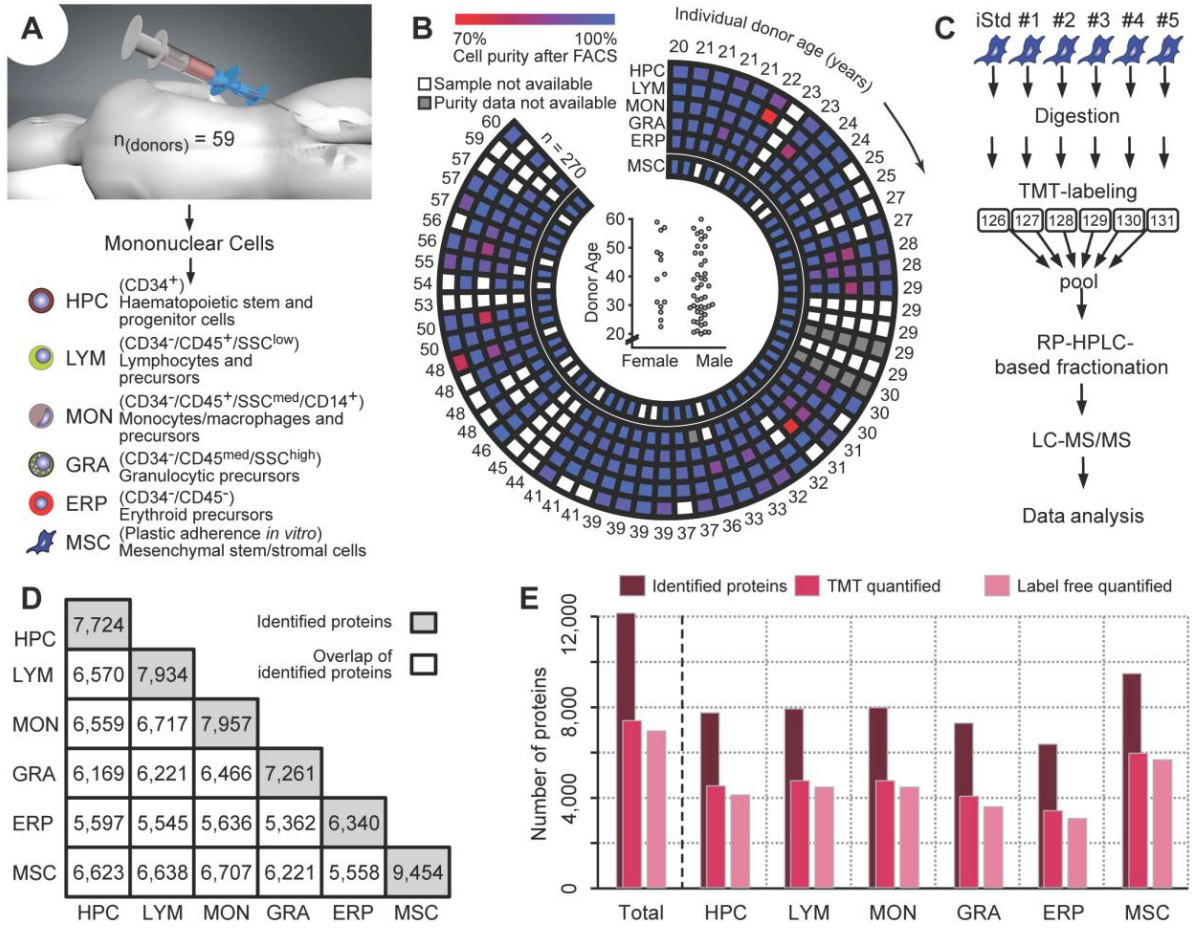


Figure 4.1. Overview on experimental design of the study (graphic designed by Marco Hennrich and Patrick Horn). (A) bone marrow samples were aspirated from 59 healthy human subjects: mono-nuclear cells were separated by FICOLL gradient centrifugation. Different cell populatioins were isolated by FACS; MSCs were expanded in cell culture. (B) Purity of 270 samples visualized from 70 (red)-100% (blue), according to cell population (circles). The age of the respective donor is indicated at the periphery. The jitter graph inside the circle shows the age distribution of the 59 subjects stratified by gender. (C) The 6 cell populations were processed separately prior to proteomics analysis: Samples from given human subjects, plus a cell-type specific internal standard, were handled in one batch. Cells were lysed, digested, TMT-labelled, and pooled. The pool was then subjected to reversed-phase HPLC fractionation, and LC-MS/MS analysis. (D) Matrix showing the pairwise overlap of identified proteins between cell populations. The diagonalindicates the total number of proteins identified in the respective cell population. (E) Total number of proteins identified, TMT-quantified, and LF-quantified (label-free) in each cell population.

For quantification purposes, we exploited TMT-labeling to capture molecular changes in the temporal dimension (ageing). The technical reproducibility for the TMT-based quantification was above 0.9 (average 0.94; median coefficient of variation = 4.2%). In order to simultaneously assess the proteomic differences between the different cell populations, we specifically developed another MS-based quantification methods largely derived from the label-free technique introduced by Schwanhausser et al. (2011) (see Methods). After

removing technical outliers (**Figure S4.1C**), a total of 270 samples were retained, with 7,375 protein abundance profiles across the donor cohort (TMT quantification) and 6,952 across the cell populations (label-free (LF) quantification), **Figure S4.1C**.
A small fraction of inter-individual variability quantified in the TMT-dimension (3.1-21.8% depending on the cell type) could be attributed to ageing (biological variability). At the same time, however, proteins within specific pathways or protein complexes showed coherent changes across donors (**Figure S4.2A/B**) [Ori et al., 2016; Ori et al., 2015]. To additionally examine the quality of the label-free analysis, we looked at whether hierarchical clustering of protein abundances would actually recover known lineage relationships, and could confirm that (**Figure S4.2C**). Furthermore, the abundances from our label-free analysis should recover profiles of known cell type specific markers of the corresponding sub-populations, which is indeed the case as shown in **Figure S4.2D**. Given the quality of the quantifications, the proteomics datasets were deemed reliable and appropriate to address the questions on age-dependent differences across cell populations.

4.2.2. The proteomic landscape of HPCs, and five other cellular elements of the human marrow niche

Examining the six major cell populations of the human bone marrow revealed only a small fraction (8.3%; 578 proteins) to be expressed in a strictly cell type- specific manner, which might be associated with specific functions and therefore serve as novel markers for isolation. For HPC for example, 17 proteins- including relevant lymphoid and myeloid markers such as DACH1, DNTT, BCL11A and BSPRY – had their expression restricted to the HPC population only, which could be suggestive of their role in earlier stages of haematopoiesis. MSCs, on the other hand, had the biggest set of proteins with unique expression (7.7% of the quantified proteome, 452 proteins), with most being involved in the organization of the extracellular matrix (ECM) and MSC-mediated HPC homing [Li and Wu, 2011]. Given that these proteins were mostly expressed at the cell surface, they represent potential candidates for the characterization of MSCs, which so far lacks clear-cut markers for isolation. Remarkably we also found specific and highly abundant expression of nestin (NES) in the human MSCs, which has been reported to play a pivotal role in HPC maintenance and homing in mice [Mendez-Ferrer et al., 2010]. To our knowledge, our results reflect for the first time that nestin is specifically expressed in MSCs of the human bone marrow as well.
The vast majority (73.3%) of the quantified proteins were present in more than one cell type, with 950 proteins (18.6%) consistently quantified in all six cell populations (**Table S4.3**). Commonly expressed proteins made up 70% of the total abundance of the quantified proteome (**Figures S4.3A/B**), and were primarily enriched in essential, housekeeping functions. The abundance patterns of these housekeeping proteins, i.e. their relative

stoichiometry, could be used to define the six different populations to the extent of recovering the actual lineage (**Figure 4.2B**). This observation might reflect the specific metabolic requirements for lineage commitment and adaptation to cell-specific processes and functions [Ori et al., 2016]. To understand which functional process required stoichiometric adjustments in different cell populations, we quantified the fraction of mapped pathways that differed between cell types (see Materials & Methods). Pathways involved in metabolic processes were in general more prone to such stoichiometric arrangements, when compared to pathways involved in translational processes (**Figure 4.2C**). One of the largest assembly of enzymes affected was found to be the glycolytic pathway, converting glucose to pyruvate, allowing for subsequent ATP- and carbon substrate production. The relative abundances of the relevant enzymes different between cell populations, yet the actual stoichiometric ratios were rigorously maintained across the different human donors (**Figure S4.3C, Table S4.4**). Interestingly, ERPs had the most divergent enzyme stoichiometry with all enzymes downstream of GAPDH being less abundant than in other cell types (**Figure 4.2D**). Such a deviation from the other cell types could be explained by the Luebering-Rapoport glycolytic shunt in erythrocytes converting 1,3-biphosphoglycerate (BPG) to 2,3- BPG, which ultimately regulates oxygen release from haemoglobin and its delivery to tissues [Benesch and Benesch, 1967; van Wijk and van Solinge, 2005]. The observation of decreased abundance of downstream enzymes in the glycolytic pathway in the production of 2,3-BPG could therefore be indeed an indicator for an early-on specialization of the ERPs. Moreover, we found a cell type-specific expression of isozymes of the hexokinase (HK1/2/3) representing the initial rate-limiting enzyme in the pathway controlling the fate of glucose-6-phosphate (G6P) [Wilson, 2003] (**Figure 4.2E**). While HPCs, ERPs and MSCs primarily expressed HK1 channeling G6P to glycolysis, MONs and GRAs mainly expressed HK3, which direct G6P to anabolic pathways such as the pentose phosphate pathway. Given the supposed heightened flux into the pentose phosphate pathway, we expected the corresponding enzymes to be more abundant in MONs and GRAs as well: Indeed, enzymes in the oxidative and non-oxidative branches of the pentose phosphate pathway alike were highly abundant in these particular cell populations (**Figure 4.2D**). Since the pentose phosphate pathway is primarily responsible for generating NADPH which serves as a precursor for nucleotide synthesis, our observations are consistent with reports of nucleotide synthesis being pivotal for neutrophils, granulocytes and macrophages [Azevedo et al., 2015]. While a thorough flux analysis and metabolic validation is key to further widen our understanding of the metabolic adjustments during cell differentiation in the bone marrow, our dataset provides a reasonable proxy by depicting proteome adaptations to cell type- specific functions.

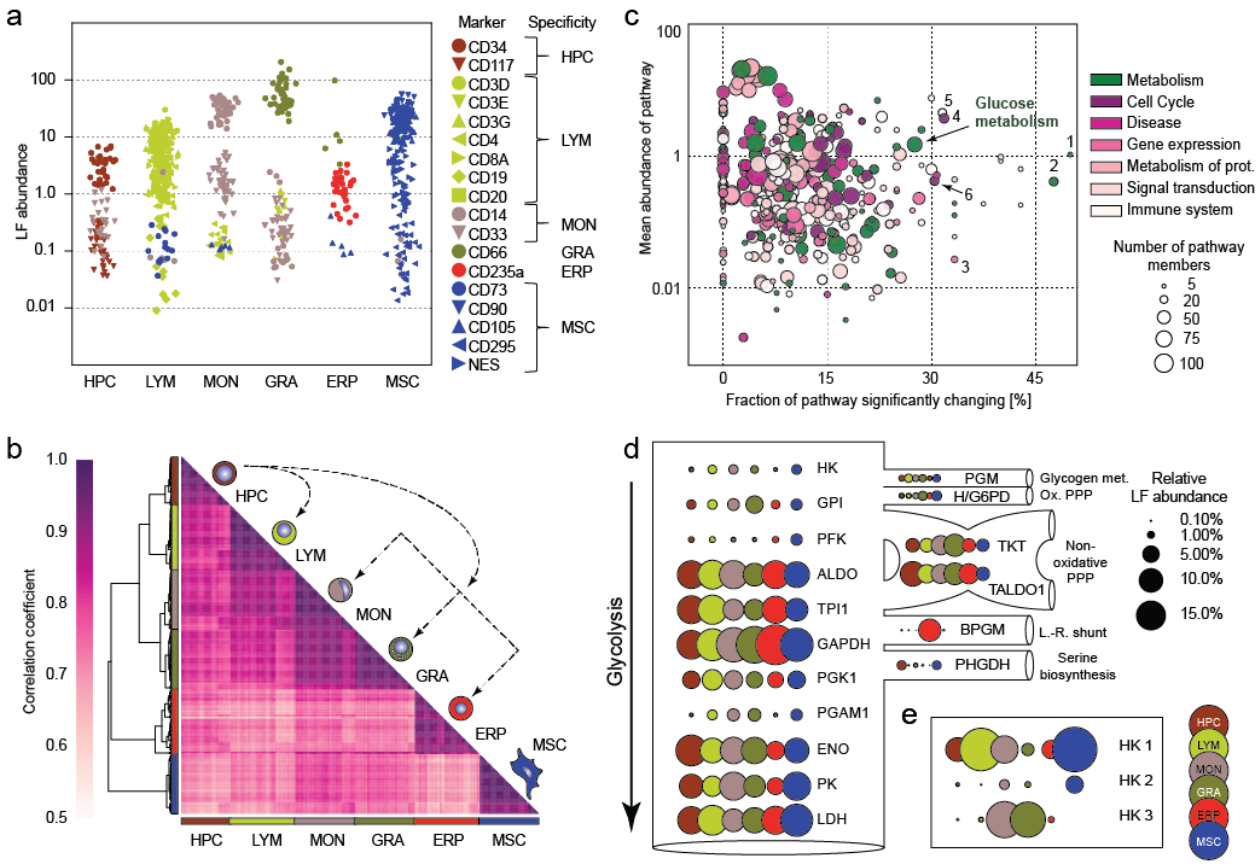


Figure 4.2. Quality of the dataset and differences between the cell populations. (a) Jitter plot displaying the relative label-free (LF) abundance of established cell type specific markers. Each symbol represents an individual sample. (b) Correlation matrix of the Spearman correlation coefficients between all 270 samples and depiction of the haematopoietic subpopulations examined. The correlation coefficient was calculated based on the LF abundances of proteins covered in ≥ 85% of the samples of each cell population (950 proteins). Characteristic proteome clusters were identified in samples derived from the same cell population. (c) Bird-eye view on the pathways that are present at different stoichiometry in the different cell populations. The fraction of a pathway that changed significantly between at least two cell types (x-axis) is plotted against the mean label-free (LF) abundance of an entire pathway across the different cell populations. The circles in the scatter plot signify Reactome pathways and their size corresponds to the number of quantified proteins per pathway. The colour refers to the highest hierarchy of the particular pathway according to Reactome. The highest hierarchies were filtered to have at least 30 pathways quantified in the dataset with each containing 5-100 protein members. The numbers indicate the following pathways; 1 ethanol/oxidation; 2 His/Lys/Phe/Tyr/Pro/Trp catabolism; 3 RNA polymerase II transcription termination; 4 DNA replication pre-initiation; 5 ER-phagosome pathway; 6 activation of ATR in response to replication stress. (d) The glycolytic pathway and branching points to connected pathways e.g. the oxidative pentose phosphate pathway (Ox. PPP) or the Luebering-Rapoport shunt (L.-R.-shunt) are depicted. Each bubble represents an enzyme with the area of the bubble encoding for the cell type specific, relative LF abundance with 100% representing the sum of all proteins per cell population depicted in (d). The colour codes for the cell population as depicted in the lower right corner. (e) Relative abundances of the hexokinase isoenzymes in the respective subpopulations are shown analogous to Fig. 2d, but 5 fold enlarged to better illustrate the difference.

4.2.3. Impact of ageing on proteome landscapes

Apart from differences between cell-populations, we quantified proteomic alterations associated with human ageing within each cell population, thereby capturing alternative ageing procedures of cell types as well. We performed a Spearman correlation analysis between the abundance of proteins (measured in >15% of all subjects, see Methods) and the corresponding donor ages. An association with a p-value < 0.05 would be considered significant (**Table S4.5**): For ERPs 175 proteins (5.2%) were affected, in HPCs 411 proteins (9.1%), in MSCs 737 (12.4%) (**Table S4.6**). Interestingly, proteins affected by age in different cell types would only partially overlap, which might be both suggestive of distinct ageing phenotypes of those cell types or their varying half-lives. HPCs and MSCs, for that matter, are relatively long-lived, persisting progenitor cells, while all the other cell populations primarily represent lineage-committed precursors with high turnover rates and hence considerably shorter half-lives.

To understand which biological processes and pathways were affected by ageing in each of the different sub-populations, we examined protein alterations in context of associated pathways as defined in the Reactome database (http://www.reactome.org, **Figure 4.3**, **Figure S4.4**, and **Table S4.7**). Thereby we detected a significant increase in glycogen breakdown, synthesis of prostaglandins and thromboxanes (arachidonic acid metabolism) and metabolism of nitric oxide in older HPCs. In the other cell populations, named processes remained largely unaffected. Strikingly, MSCs displayed very prominent alterations in protein abundance patterns in pathways associated with cellular response to stress, replicative senescence, white adipocyte differentiation and extracellular matrix organization. Specifically our data also captures some of the few established ageing markers identified by transcriptomic analyses, such as the interferon regulatory factor 8 (IRF8) in HPCs. The abundance of the protein becomes considerably reduced in older HPCs, which is reported to be associated with dysregulated proliferative activity and biased myeloid differentiation [Stirewalt et al., 2009]. We also detected a significant reduction in abundance of DNA methyltransferase 1 (DNMT1), which is responsible for maintaining DNA methylation [Hermann et al., 2004]. It has been shown earlier that if down-regulated, a dysregulation of methylation might lead to several age-related diseases [Benetatos and Vartholomatos, 2016], such as acute myeloid leukaemia (AML) and myeloidysplastic syndromes (MDS). In mouse models DNMT1 has been reported to be critical for HPC maintenance and their self-renewal after transplantation [Trowbridge et al., 2009].

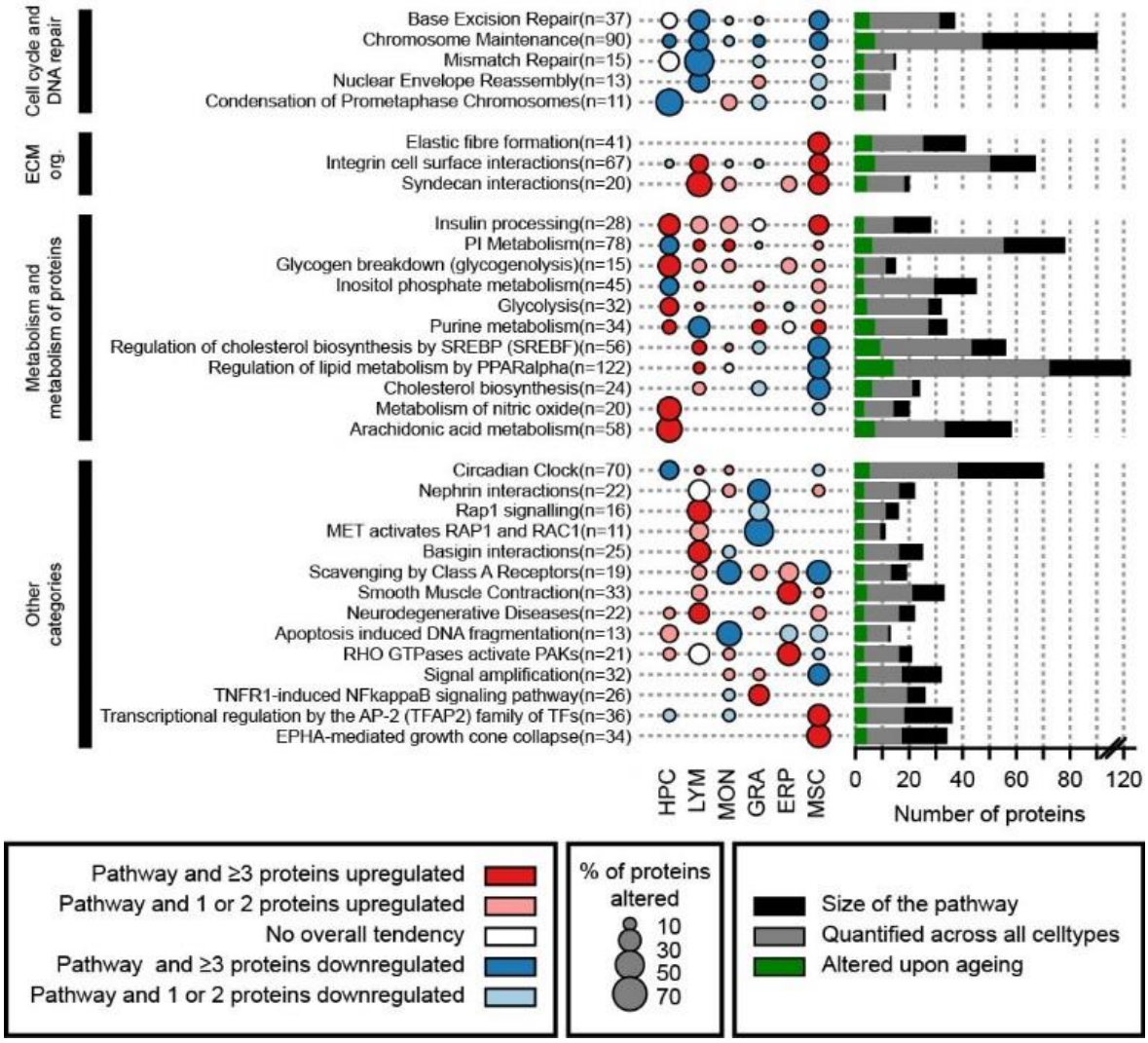


Figure 4.3. Age-affected pathways in the individual cell populations. A selection of pathways from the Reactome database that show the most prominent changes upon ageing are depicted. Pathways were required to have between 5 and 150 members, to be sufficiently covered in at least one cell population (>30% of the proteins are quantified), and to have at least 20% of its quantified components being significantly (p-value < 0.05) altered upon ageing. The figure illustrates the five most up- and down-regulated pathways per cell population (see Methods). The area of the bubbles represents the percentage of proteins quantified by TMT that are significantly (p-value < 0.05) altered. If no bubble is shown, no protein of the pathway has been observed to be significantly altered in the respective cell population or no protein of the pathway has been quantified. The colour of the bubbles codes for the direction of the alteration, with red indicating an overall increase of the pathway members with at least three proteins being upregulated and pink an overall increase with one or two proteins being upregulated. Blue codes for pathways with an overall trend towards downregulation, with strong blue coding for pathways with at least three proteins being downregulated and light blue containing one or two proteins being downregulated. The colour white indicates that no overall tendency for the proteins associated with the corresponding pathway could be observed. The bars on the right-hand side of each pathway illustrate the number of proteins being significantly altered upon ageing regardless of the cell population (green), being quantified by TMT (grey), and the total number of members of the pathway (black) as also mentioned in the pathway annotation (n). The grouping of the pathways on the left side is based on the highest hierarchy levels defined in Reactome, e.g. extracellular matrix organisation (ECM org.).

4.2.4. Ageing affects central carbon metabolism in HPCs

In line with metabolic adjustments between different cell populations, we sought to investigate whether the metabolic household could be somehow affected by ageing. For HPCs one of the most remarkable age-dependent changes indeed encompassed enzymes involved in glycolysis, glycogen catabolism and fatty acid beta-oxidation (FAO). Briefly, the changes were reflective of an enhanced metabolic and specifically anabolic activity of old HPCs versus young HPCs, with protein abundances again serving as a reliable proxy for metabolic flux and activity (**Figure 4.4**). Notably, the alterations were highly specific to the HPCs, as corresponding enzymes in all the other cell populations remained unaffected (data not shown, see manuscript).

The upstream and rate-limiting part of the glycolytic pathway showed a significant age-associated increase in enzyme abundance, i.e. hexokinase 1 (HK1), phosphofructokinase M (PFKM), as well as aldolase C (ALDOC) and triosephosphate isomerase 1 (TPI1) (**Figure 4.4**). At the same time enzymes involved in glycogen catabolism, glycogen phosphosphorylases PYGB and PGYL, as well as the glycogen debranching enzyme (AGL) were significantly more abundant in older HPCs as well. Our assumption that older cells seemed to be more prone to fuel glycolysis with catalytic products of glycogen rather than other glucose sources, was further supported by the increased abundance of phosphoglucomutase 1 (PGM1). On top of actual resource shift, we also identified an increase in abundance of transaldolase 1 (TALDO1) which plays a crucial role in the non-oxidative branch of the pentose phosphate pathway. Given a similar increase of glycerol-3-phosphate dehydrogenase (GPD2), an dihydroxyacetone kinase (DAK), it became clear that the upstream part of the central carbon metabolism experienced a full blown increase in activity, primarily consuming ATP and converting glucose to dihydroxyacetone phosphate (DHAP) and D-glyceraldehyde 3-phosphate (GA3P). Notably, we did not see any age-dependent changes in the subsequent part of the carbon metabolism, characterized by the production of ATP, NADH and pyruvate in the so-called Krebs cycle (**Figure 4.4A**). Such an abundance pattern and deduced metabolic flux is reminiscent of the Warburg effect where excess of glycolytic carbons becomes redirected to pathways that branch out of the glycolysis/Krebs cycle axis, thus producing co-factors and intermediates for anabolism (nucleotides, lipids, amino acids, …) [Heiden et al., 2009] and epigenetic processes [Smith et al., 2016]. To provide orthogonal evidence to our observations at the proteomics level, we additionally measured relative amounts of metabolites that play a role in glycolysis (**Figure 4.4B**). Two metabolites in particular - namely ribose-5-phosphate as well as ribulose-5-phosphate show a tendency to get accumulated in old HPCs, suggesting that glucose is increasingly shuttled through the pentose phosphate pathway for anabolic purposes.

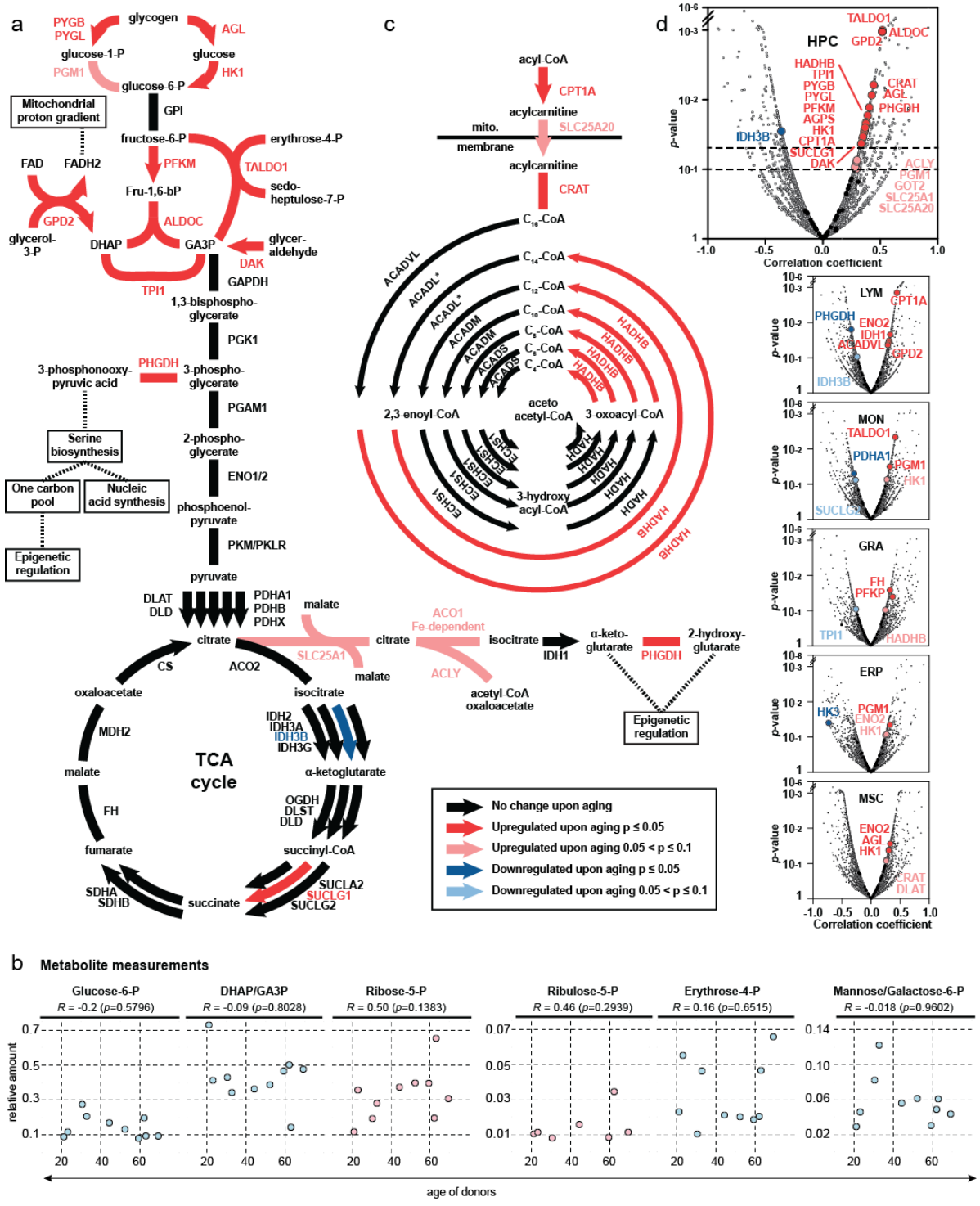


Figure 4.4. Prominent changes upon ageing in the central carbon metabolism (graphic designed primarily by Marco Hennrich). (a) The glucose metabolism and the tricarboxylic acid (TCA) cycle of HPCs are depicted with arrows representing unidirectional reactions and strokes representing bidirectional reactions. The gene names of the respective enzymes are written in capital letters and the colour codes for changes upon ageing, as described in the legend. A star indicates that the particular protein was not covered for quantification. Apart from the main glycolytic pathway, glycogen break-down, as well as a sub-part of the pentose phosphate pathway is shown. (b) Measurements

of relative amounts of phosphorylated metabolites relevant to the preparatory phase of the glycolytic pathway and the pentose phosphate pathway. The age of the respective donor (x-axis) is plotted against the relative amount of metabolites (y-axis) after normalizing for the cumulative amount of the given metabolites detected in each donor. The *p*-value (p) is based on the given Spearman correlation coefficient. (c) Similar to (a) the scheme illustrates the effects of ageing on a specific set of enzymes involved in the mitochondrial beta oxidation of fatty acids. (d) Volcano plots of all proteins quantified in the respective cell populations. Proteins represented in (a) and (b) are colour coded according to the legend. All other proteins are coloured grey. The dashed lines indicate *p*-values of 0.1 and 0.05.

Phosphoglycerate dehydrogenase (PHGDH), which we found to be increased in abundance as well upon ageing, is such a branching enzyme converting 3-phospho-D-glycerate into 3-phosphoonooxypyruvate for the serine biosynthesis. As a matter of fact, this enzyme is often over-expressed in tumors [Locasale et al., 2011] as serine is required for nucleotide synthesis, NADPH production and biosynthesis of S-adenosyl methionine (SAM), which eventually is critical for epigenetic-mediated control of gene expression via DNA methylation [Heiden et al., 2009; Maddocks et al., 2016].

Consistent with an increased metabolic branching, other enzyme abundances were increased to adjust for the ageing flux pattern, e.g. the mitochondrial citric acid transporter SLC25A1. Specifically, this enzyme shuttles citrate to the extra-mitochondrial periphery, where it can be metabolized to acetyl-CoA and oxaloacetate by the soluble aconitase (ACO1) and ATP citrate lyase (ACLY). Notably, both latter enzymes were found to be up-regulated as well; ACLY in particular has been shown before to be a key regulator between aerobic glycolysis and amino acid as well as *de novo* lipid synthesis involve in proliferation of tumor cells [Zhao et al., 2016]. Thus, having these enzymes affected during ageing of the cells further substantiates our hypothesis of an increased Warburg effect.

Another substantial pathway which is highly connected to glycolysis, and subject to age-dependent alterations, is fatty acid beta-oxidation (FAO), encompassing enzymes, such as the tri-functional enzyme subunit beta (HADHB), peroxisomal bifunctional enzyme (EHHADH), propionyl-CoA carboxylase beta chain (PCCB), carnitine O-acetyltransferase (CRAT), as well as carnitine palmitoyl transferase 1 (CAPT1A), and the carinitine-acyl-carnitine transporter SLC25A20 (**Figure 4.4B**). Thus, the up-regulation involved the actual fatty acid import machinery, the conversion of acyl-CoA to acyl-carnitine (CPT1A) in the cytosol, its transport into mitochondria via SLC25A20, and the reconversion to acyl-CoA by CRAT [Houten and Wanders, 2010], as well as the major enzymes of the FAO (HADHB in mitochondria, EHHADH in peroxisomes).

Together with glycolytic catabolism, FAO has been reported to be an important hallmark of HPC maintenance and quiescence [Ito et al., 2012; Shyh-Chang et al., 2013]. We can thus conclude from the observed abundance alterations that ageing in HPCs is highly associated with a rewiring of central metabolic pathways and the rerouting of metabolic intermediates for the synthesis of co-factors pivotal for anabolic and epigenetic processes.

4.2.5. Granulocytic, megakaryocytic differentiation at the expense of lymphoid differentiation with ageing

Apart from metabolic rewiring, we also interrogated HPCs on their ability to self-renew and differentiate into all functional blood cells when getting older. We found some of the specifically expressed 17 proteins (see 4.2.2), to be decreased significantly upon ageing, such as DNTT and BCL11A, which have been linked to lymphoid development and function (**Figure 4.5A**). In contrast, the heterodimeric soluble guanylate cyclase (GUCY1A3 and GUCY1B3), i.e. downstream signaling effectors of nitric oxide (NO), increased significantly in abundance in older HPCs. NO/cGMP signaling has indeed been shown to modulate haematopoiesis and might well be correlated with an increased differentiation bias towards the myeloid lineage, coined lineage skewing [Ikuta et al., 2016]. To further substantiate this hypothesis, we looked at potential lymphoid and myeloid markers that we manually assembled from various previous publications, e.g. Doulatov et al., 2010. Lymphoid markers, such as MME (CD10, official lymphoid markers for human haematopoetic progenitors), IKZF1 (regulator of lymphocyte differentiation), and EBF1 decreased in abundance with age. Several other proteins described by Doulatov et al., 2010 to be characteristic for human lymphoid development, also tend to be down-regulated with age (**Figure 4.5B**). In sharp contrast, proteins associated with myeloid lineages, such as PTGS1 (prostaglandin-endoperoxide synthase 1), PSTPIP2, TBXAS1, as well as PML show elevated abundance levels in older HPCs (**Figure 4.5C**). We conclude therefore that using the proteomic abundance patterns of lymphoid and myeloid markers, the increased tendency of lineage skewing with age can successfully be delineated.

Given this observation, however, it is critical to understand whether the increase of glycolytic enzymes reported in section 4.2.4 is indeed due to ageing of cells or rather the expansion of the myeloid sub-population in the $CD34^+$ HPC cells. To address the issue, the transcriptome of 519 single-sorted HPCs originating from young and old human subjects (each n=2) was analyzed. Given the expression levels of the mRNA markers associated with lymphoid or myeloid differentation (**Figure 4.5C**) (consistently measured in single cells), we categorized cells into myeloid- or lymphoid-primed subsets. We generally observed that the mRNA levels of age-increased glycolytic enzymes were higher in myeloid-primed than in lymphod-primed HPCs. Transcripts of age-unaffected enzymes, on the other hand, tended to be similar between sub-populations (**Figure 4.5D**). To delineate the effect of cell ageing on enzyme transcript levels, we compared gene expression between young and old donors in myeloid- and lymphoid-primed cells, respectively. We observed that enzymes of the 'preparatory' phase of the glycolytic pathway were mostly affected by ageing in the myeloid-primed sub-population (**Figure 4.5E**). Thus, while the lineage skewing of the $CD34^+$ cells towards myeloid differentiation upon ageing can explain the increase of enzyme

abundances to some extent, we also observed a lineage-independent ageing effect (**Figure 4.5E**).

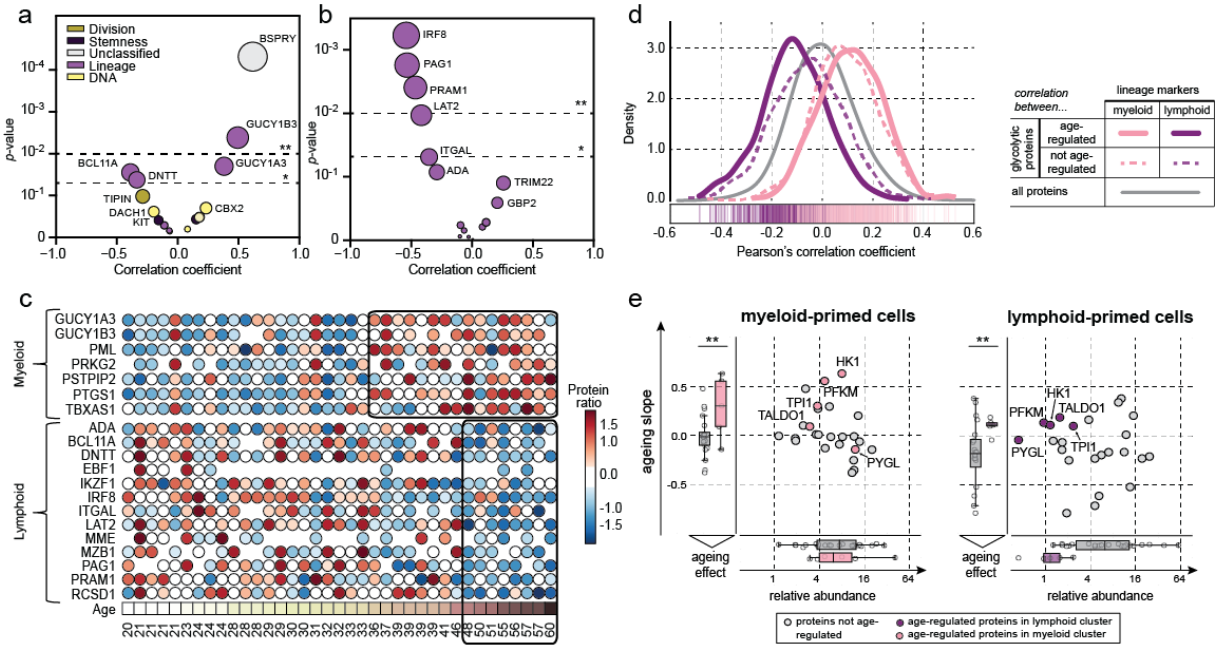


Figure 4.5. Age-related alterations of lineage specific proteins in HPCs. (a) Scatter plot of the relationship between the correlation coefficient (x-axis) and the corresponding *p*-value of proteins specifically expressed in HPCs (y-axis). The colour assigned to each bubble is indicative of its function, as explained in the upper left legend. The area of the bubbles represents the significance level (*p*-value [-log10]). The stars at the dashed lines indicate a *p*-value of 0.05 (*) and 0.01 (**). (b) Scatter plot similar to (a), with proteins associated with common lymphoid progenitors (CLP) compared to megakaryocytic and erythroid progenitors (MEP), common myeloid progenitors (CMP) and haematopoietic stem cells/multipotent progenitors (HSC/MPP) in accordance with the GO annotation of immune system process (GO:0002376) [Doulatov *et al.*, 2010]. (c) Significant alterations of the proteome landscape of HPCs upon ageing. Each circle represents a successful quantification in an individual human subject, with the respective age noted at the bottom. The colour codes for the z-score of relative intensity versus the internal standard (TMT protein ratio) with red indicating an increase and blue a decrease in abundance upon ageing. The boxes indicate the age boundaries where the majority of the respective age-dependent increase or decrease in protein abundance takes place. (d) The density plot shows distributions of correlation coefficients between lineage markers and glycolytic enzymes. The calculation is based on the results from the single cell RNA-seq data. Density distributions based on correlations between myeloid (pink) or lymphoid (purple) markers with glycolytic enzymes that are up-regulated upon ageing, are shown with thick lines. The corresponding individual correlation values are displayed as lines in the box below the plot in the respective colours. Density distributions based on correlations of the respective markers with glycolytic and TCA-related proteins that did not change upon ageing, are shown as dashed coloured lines. The grey distribution represents the correlation values of all proteins against myeloid and lymphoid markers (e) Scatter plot illustrating the effect of ageing on glycolytic enzymes in lymphoid- and myeloid-primed cells, as deduced from single-cell analysis. The dots correspond to proteins in glycolysis that were not affected by age (grey), and glycolytic proteins that were altered upon ageing according to the proteomics data. The x-axis illustrates the relative abundance of those enzymes in myeloid-primed cells (left), and lymphoid-primed cells (right). The y-axis corresponds to the ageing slope derived as a measure of age-effect in lymphoid- and myeloid-primed cells across donors (Mann-Whitney U-test, * p-value < 0.05). The data from the scatter plots are collapsed into box plots on both axes.

4.2.6. Alterations in the bone marrow niche and their relationships to changes in HPCs as they become older.

As discussed in the Introduction to this chapter, the bone marrow is not only defined by the cell types it is populated with, but also by the microenvironment surrounding the cells, controlling for HPC maintenance and regulating haematopoeitic homeostasis. To understand to what extent the microenvironment might be affected by ageing and metabolic shifts that go along with it, we sought to specifically investigate essential factors and adhesion molecules produced by the cellular niche and responsible for homing and egress of HPCs [Dykstra et al., 2011; Mendelson and Frenette, 2014; Nakamura-Ishizu and Suda, 2014], such as SDF1, VCAM1, and fibronectin (FN1) [Ley et al., 2016]. These factors notably decreased in abundance in older MSCs (**Figure 4.6**), whereby other proteins involved in glycosaminoglycan and collagen metabolism were significantly decreasing as well. These observations suggested that the extracellular matrix might indeed get substantially reorganized during ageing, which in turn would affect the entire bone marrow niche.

Since ageing probably encompasses coordinated, concerted alterations in this network of cell communities within the bone marrow, we reckoned that a direct correlation of extracellular protein-ligand pairs between different cell types might be indicative of a direct or indirect functional relationship mediated through the bone marrow niche (**Figure S4.5**). Our correlative analysis based on the STRING database [Szklarczyk et al., 2017] demonstrated that 28% of receptor-ligand pairs directly interacted with one another (p-value=0.0498), while 62% showed an indirect functional relationship (p-value=0.047). A decrease in VCAM1 and FN1 in MSCs would then for example be complementary to their corresponding ligand, α4/β1-integrin (ITGA4/ITGB1), which decreased in HPCs upon ageing. Interestingly, αL/β2 integrin (ITGAL/ITGB2) followed a very similar pattern (**Figure 4.6** and **Figure S4.5**), supporting the notion that β2-containing integrins on HPCs show synergy with α4/β1-integrins, as reported previously [Papayannopoulou et al., 2001].

Soluble factors that get secreted were also affected as the bone marrow niche became older, which might be linked to the functional attenuation of HPCs. TGF-beta1 – a secretory factor that has been proposed to contribute to lineage skewing by stimulating myeloid-biased HPCs [Challen et al., 2010; Mendelson and Frenette, 2014]- was elevated in LYMs and ERPs from older subjects. Current literature also points towards another vital factor released into the microenvironment to play a pivotal role in triggering lineage skewing and hence serving as a major messenger molecule- nitric oxide (NO). While we did not interrogate the metabolite directly, we again used our proteomics data as a proxy for metabolic fluxes. We found a significant decrease in abundance of the nitric oxide (NO) synthase inhibitor (NOSIP) in MSCs, while DDAH1 and DDAH2 (dimethylargininase 1 and

2) - enzymes that degrade ADMA, an inhibitor of NO synthase- were significantly increased in abundance in HPCs (**Figure 4.6**). These complementary changes suggested that the secondary messenger NO-molecule gets increasingly abundant in the ageing niche, and its downstream signaling effectors, i.e. the heterodimeric soluble guanylate cyclase (GUCY1A3, GUCY1B3), and the cGMP-dependent kinase 2 (PRKG2) were significantly elevated in older HPCs (**Figure 4.6**). These observations might indeed indicate that elevated levels of both TGF-beta1 and NO in the bone marrow might represent initial triggers for lineage skewing in general. Although our observations need to be further substantiated with respective metabolic experiments, our analysis to comprehensively describe complementary changes of extrinsic and intrinsic factors between different cell types could represent fertile ground for further hypothesis on the ageing of the bone marrow niche and its connection to the ageing of HPCs.

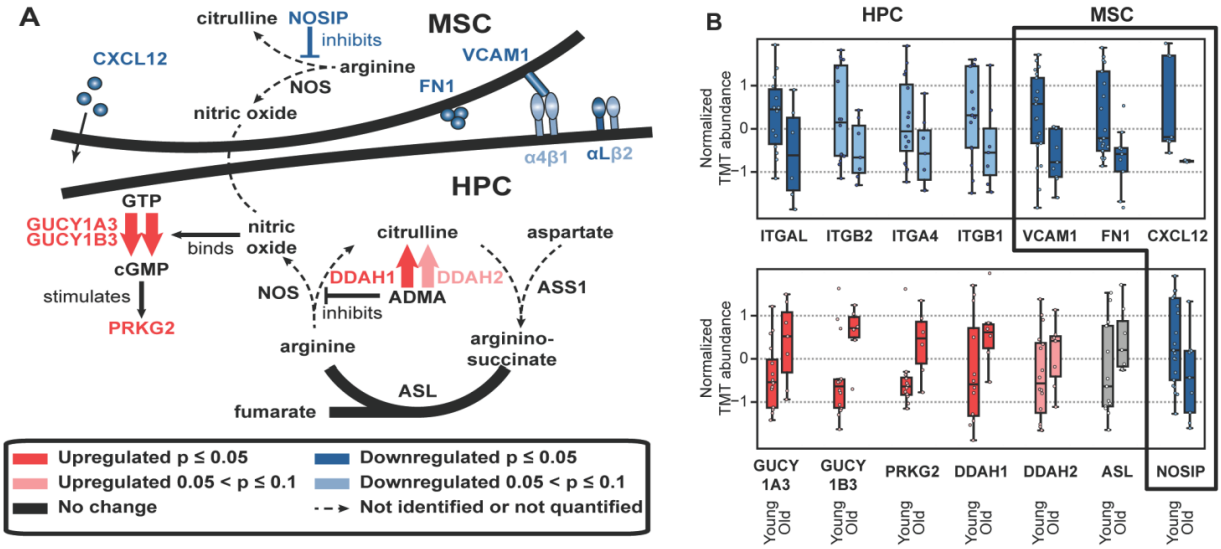


Figure 4.6. Alterations of protein abundance in the haematopoietic stem cell niche with age (graph primarily designed by Marco Hennrich). (a) CXCL12, VCAM1 and FN1 in MSCs and the integrins alpha4, alphaL, beta1 and beta2 (α4, αL, β1, and β2) in HPCs decrease in abundance upon ageing. Age-related changes in connection with nitric oxide (NO) synthesis, the urea cycle and potential NO crosstalk between MSCs and HPCs is depicted by arrows for unidirectional reactions and strokes for bidirectional reactions. Dotted lines or dotted arrows illustrate reactions for which no enzyme was detected or quantified in this study. The gene names of the respective enzymes are written in capital letters. The colour encodes changes upon ageing, as described in the legend. (b) Box-plot representation of the proteins depicted in (a). The dots represent the individual results from younger (age <30 years) and older (age > 50 years) human subjects. Proteins known to have a direct effect on homing or egress of HPCs from or into the bone marrow are shown in the upper and proteins associated with NO signalling and the urea cycle are plotted in the lower graph.

4.3 Discussion

In this collaborative effort an atlas of age-associated alterations in the proteomic landscapes of human HPCs as well as five other sub-populations in the bone marrow niche has been presented. While most of the work has been primarily aimed at understanding how proteins are affected by ageing (40 years of time span), the application of orthogonal quantitative methods allowed us to also directly compare abundance levels between the different cell populations and appreciate stoichiometric adjustments. These were most apparent in metabolic pathways, reflective of the respective adjustments that need to take place in the changing microenvironment.

Among significant abundance changes of proteins upon ageing, the most prominent ones suggested an enhanced metabolic and anabolic activity of older HPCs, with enzymes of the upstream glycolytic pathway being particularly affected. Glucose metabolism has indeed been shown to play a pivotal role in governing stem cell fate in terms of proliferation, differentiation or dormancy [Shyh-Chang et al., 2013]; however, no evidence has been provided so far on its role during the ageing process. Our observations point towards a Warburg effect [Heiden et al., 2009; Smith et al., 2016] and is therefore compatible with reports of increased proliferation rates of aged HPCs in the bone marrow [Geiger et al., 2013]. Such a high proliferation of HPCs generally goes along with a diminished competence of the adaptive immune system, an expansion of myeloid cells [Geiger et al., 2013] and an increased platelet priming and functional platelet bias [Grover et al., 2016]. Our proteomic analyses of HPCs supported previous findings, and in addition allowed for a quantitative assessment of proteins pivotal to lineage skewing.

Moreover, investigating five other sub-populations in the bone marrow in context of their interactions with HPCs substantiated a hypothetical triggering mechanisms for lineage skewing, with key factors responsible for homing, egress and differentiation of HPCs declining in abundance, whereas soluble factors instantiating myeloid expansion significantly increased with age [Michurina et al., 2004; Mujoo et al., 2011]. Additional changes in the proteomic make-up of MSCs suggested changes in the overall architecture of the extracellular matrix in the bone marrow niche. Such a contextual analysis and profiling of the niche could also prove useful to further provide a new foundation for better understanding age-related diseases such as myelodysplastic syndromes (MDS) in the future. Overall, the presented analysis captured for the first time the proteomics signatures of the ageing process in a human tissue broadening the scope of our understanding regarding the role of metabolic pathways, lineage skewing and niche elements.

Chapter 5

Conclusions & Outlook

The advent of mass spectrometry technology provides us with the ability to comprehensively describe the proteome, and interrogate a broad palette of its features. Beyond measuring protein abundance levels, various techniques as well as combinations of those have allowed the assessment of protein stabilities, post-translational modifications, interactions and degradation rates in a large-scale manner. So far, however, the integration of published proteomic datasets, along with an evaluation of the consistency have been lacking [Collins et al., 2017]. In this work I have attempted to provide a comprehensive picture on such datasets, and pointed at a protein complex landscape that is highly consistent across different studies. Such efforts need to be extended to derive mechanistic insights that can prove useful in extrapolation and clinical appliances. Beyond the characterization of protein modules and their variation across individuals, I scrutinized potential modular rewiring on the temporal scale- both short- and long-term. In the following section, I will give an overview on the insights gained from each chapter, as well as potential caveats. Finally, these elements will be connected for the purpose of gaining a more refined understanding of the proteotype model.

5.1 Proteome modularity in context of temporal processes and variation between individuals

To give the reader a better understanding of the actual findings and advancements in this work, the novelty and potential answers to the questions posed in 1.4 are listed and discussed here. At this point, however, it has to be iterated that the presented solutions and answers to the problems tackled in this thesis, are more of an approximate nature, and do need further validation to test for broad consistency.

Question 1: To what extent are cellular proteins organized in macro-molecular structures?

The work presented in **Chapter 2** dissected proteomic data from various published resources on both mice and human individuals, as well as cell types, according to the observed co-variation in protein abundance. We found proteins to be consistently co-abundant in protein complexes, when compared to other functional modules, including pathways. Stoichiometric variation of those complexes proved to be highly consistent across datasets as well, with chromatin-associated complexes prone to strong variation, and translation-associated complexes more rigorous in their stoichiometric make-up. Moreover, we were able to define complex components that were contributing most to the complex variation, and identified the immunoproteasome as a highly variable entity of the proteasome complex.

Question 2: What factors impact the variation in protein and module abundance and to what extent?

Chapter 2 further discusses for the first time the association of genetic and environmental factors to module abundance and stoichiometry. The sex of the animals in the studied cohort is taken as a proxy for genetic effects, whereas their diet conditions represent the measurable environmental impact. We observed that the variability of the proteotype was driven to a large extent by the variability of protein module abundance and stoichiometry, with sex differences primarily impacting translation- and chromatin-associated complexes. Interestingly, differential diet conditions affected a complementary set of complexes associated with mitochondrial functions, such as cytochrome *bc1* complex, etc.

Question 3: Do changes in the modular organization of the proteotype have far-ranging consequences?

Given our observation on dramatic stoichiometric re-arrangements in transport-associated protein complexes COPI and COPII, we investigated how such alterations could potentially impact receptor transport in general. We indeed observed a clear-cut clustering of receptor molecules clearly correlated with one particular COPI/COPII constellation of paralogs, while not being correlated with another constellation. We observe for example that the abundance of the EGF receptor positively correlates with nearly all components of the transport machinery, whereas integrin-associated receptors tend to be negatively correlated with latter. Though correlative in nature, the analysis offers exciting new hypotheses regarding distinct receptor transport mechanisms depending on the prevalent stoichiometric composition of the COPI/COPII complexes. Further investigation with regard to particular bio-physical features of the receptor clusters are currently underway as well.

Question 4: What systematic changes in protein stability occur during the cell cycle? To what extent do they reflect molecular cooperativity?

In **Chapter 3** we discussed using a thermal profiling approach how protein thermal stability gets extensively modulated during the progression of the cell cycle. Given that no other method can capture protein thermal stabilities in a large-scale manner, we could recover some important biological insights. One of the key observations concerned protein complex assemblies, such as in the nuclear pore complex. The NPC assembly status apparently affects the stability of its sub-complexes in a differential manner. The NPC scaffold represents a grafting surface for FG-nucleoporins (NUPs) in order to maintain a defined local concentration of those NUPs to prevent aggregation [von Appen et al., 2015]. Releasing those proteins during mitosis should thus render them aggregation prone. Interestingly, the exact opposite was observed, along with a concurring enhanced stability of those NUPs during mitosis.

That stabilization phenomenon was generally observed with highly disordered proteins that tend to have low stabilities, which is in line with the radical changes in the bio-physical milieu surrounding the protein. To further enhance our understanding of how the proteome is dynamically adjusted to such abrupt changes upon entering the mitotic stage, we also monitored its solubility transitioning. Morphologically that phenomenon becomes visible with the dissolution of the nucleolus by light microscopy [Stevens, 1965]; pointing at the proteins driving those events is challenging, however, due to difficulties in purifying phase-separating, membrane-less organelles. The presented data for the first time captures the specific sub-proteome of around 300 proteins that is part of the membrane-less organelles, especially the nucleolus. More specifically, these proteins- while being insoluble during interphase- get solubilized in mitosis, and thus less aggregation-prone. Strikingly, we also observed that the subset of proteins was predominantly characterized by highly disordered regions (IDPs, or, intrinsically disordered proteins). Thus, work conducted in that collaborative project effectively revealed a potential mechanism by which the mitotic cell prevents aggregation in the wake of phase separation, which has arguably never been demonstrated before on this scale.

With regard to our interpretation of molecular cooperativity as outlined in the Introduction, the observed stability- as well solubility transitioning do allow for exciting speculations about how the cell could have possibly evolved a mechanism to forestall aggregation. For example, could the IDPs be deliberately installed into specific protein complexes, such as the NPC, to manage their local concentration? [Lemke et al., 2016] Could biophysical changes of sub-complexes (i.e. stability) or post-translational modifications be responsible for triggering the 'solubilization' of IDPs? [Jiang et al., 2015; Nott et al., 2015] Such a

cellular feat at the level of molecular cooperativity could indeed shed more light on the role of IDPs in signaling pathways.

Question 5: How does ageing affect the proteomic landscape?

Chapter 4 delineates a proteomic map of different hematopoietic cell populations and bone marrow stromal cells, and characterizes its age-associated changes. While recovering to a large extent know alterations previously reported in murine studies, such as lineage skewing, and the decrease of homing receptor-adhesive interactions with ageing, the work has also captured some exciting novelties. We found that in HPCs the abundance levels of several enzymes making up for the up-stream glycolytic pathway ate increased with age, hence for the first time capturing a potential causative mechanism for metabolic rewiring in human stem cells. To exclude that these ostensible age-dependent changes could be caused by lineage skewing, hence the expansion of the myeloid sub-population, we layered our analysis with an orthogonal methodology, namely single cell RNA-seq. The method allowed us to truly dissect protein abundance changes according to lineage skewing effects and the age of the human subjects, and effectively confirmed that the protein abundance changes observed were coherently caused by the longitudinal factor.
Yet again that chapter also allows us to add another feature to the proteotype model, pinning certain age-dependent abundance effects to very particular glycolytic factors. In light of what has been discussed in **Chapter 2** it would be exciting for example to explore the effect the genetic composition and the environment has on that very particular set of proteins that could be the causative agent for the Warburg effect, resulting in an aged-cell phenotype.

5.2 Relevance of Exploring the Proteotype for Personalized Medicine

One of the prime goals for the 21st century is set to be the individualized health care model for each patient, with a customized treatment plan and management [Chin et al., 2011]. The idea goes beyond clinical data from patients, expanding to their molecular profiles, such as their genomic composition, as well as phenotypical measurements, i.e. metabolite levels. Not only would that kind of information prove useful to speed up treatments in general, it is also meant to decrease the frequency of so-called adverse drug effects (ADRs). The cause of ADRs could be manifold, ranging from actual inherited factors [Philips et al., 2001; Goldstein, 2003] and hepatic insufficiencies, to interactions with other proteins (i.e. lipoprotein) and metabolites [DeVane, 2002]. Assessing what might be the causative agent for the emergence of ADR-related symptoms remains strikingly difficult for each patient, and is subject to the so-called Naranjo algorithm, which is nothing but a survey of 10

yes/no-questions [Narano et al., 1981]. Paired with the WHO causality term assessment criteria, the Naranjo score gives a likelihood of an ADR-reaction [Davies et al., 2011]. Given the amount of inconsistencies in the expert judgements of that score, however, the robustness of that causality assessment is highly questionable [Davies et al., 2011]. Even more so that it does not allow to predict a patient's potential reaction to a drug in advance, to avoid ADR symptoms in the first place. Such symptoms could- depending on their severity- entail death, hospitalization, disabilities and permanent damages, in general.
For all those reasons, the advent of technology to assess a patient's personal metabolism and biological data is critical. One of the most renowned efforts to boost our knowledge on that matter is the Cancer Genome Atlas project (TCGA) which involves collecting and analyzing tumor samples from around 10,000 cancer patients [Weinstein et al., 2013]. For those specimens a multitude of omics-data was collected, such as copy number variations, exome sequencing, somatic mutations, DNA methylation, gene expression, miRNA expression, as well as clinical data. More efforts in recent years have also led to the accumulation of data on the proteotype of those samples; yet due to technical limitations and caveats in the quantification strategies using mass spectrometry (1.1.2), proteomic data remains relatively scarce in that context.
We can assume, however, that the proteotype of an individual will be paramount to interpreting the genetic information of a patient; not least because of the sophisticated buffering mechanisms that the human cell has installed [Stefely et al., 2016]. This was also apparent from the data analyzed in this work, and previous studies [Liu et al., 2016] (1.1.3), demonstrating that transcript levels serve as a sufficient yet not accurate proxy for protein abundance. Moreover, we can also assume that the cell needs to have rescuing mechanisms in place to prevent a genetic mutation to propagate through entire cellular network, i.e. in case of one pathway rendered inefficient, another possibly cross-talking pathway could partly recover the original phenotype [Louden et al., 2013; Jamieson et al., 2000].
Another aspect to be considered in that regard is a particular feature of the proteotype model as discussed in 1.2.1: The proteotype is already the result of the integrational process involving transcriptional and translational control as well as RNA interference and modulation. Knowing a somatic mutation and the gene expression level, however, does not guarantee that a given signal is truly propagated to the downstream signaling layers as well. Thus, for clinical purposes to quickly and (accurately) assess the proteotypic fingerprint could eventually prove powerful indeed.
In **Chapter 2** we enhance that notion with our observation on proteotype-dependent module abundances and stoichiometries. Although we could effectively only point at 10-15% of their variation to be explained by sex and diet, it leaves us with the intriguing question of what the remainder of that variation could be attributed to. It has previously been demonstrated by Ori et al. (2016) that the differential stoichiometry of protein complexes in particular

could successfully distinguish cancer from benign cells. Thus, modular stoichiometries could indeed prove to be a molecular footprint to exploit for diagnostic purposes in cancer research as well. That hypothesis gets even more substantiated in **Chapter 3**, where we managed to successfully distinguish modular sub-complexes due to their differential stability patterns. We could thus envisage the application of thermal protein profiling (TPP) in personalized diagnostic procedures to characterize the functionality of such modulated cellular processes. With regard to **Chapter 3** there are two other exciting issues to be put into perspective at this point: (a) The first one relates to the idea of associating inherent protein stabilities with genetic properties of an individual in a similar way as in GWAS studies. Given the large-scale monitoring and reliable quantification of thermal stabilities of thousands of proteins in a single MS-run, the idea is straightforward: For a given set of e.g. Yoruban individuals (to overlay with existing protein abundance data as described in **Chapter 2**), protein stabilities are measured in a TPP-setup. To our knowledge there has been no study pinpointing at potential 'stability-QTLs'. Could they possibly explain a large number of eQTLs that seemingly do not have their effect propagated to the protein abundance level? Could a QTL propagate its effect by changing the stability of a trans-target by binding to its locally expressed gene product? It is intriguing to speculate that such an analysis could indeed boost our understanding of what possible signaling layers might be affected by genetic variation. (b) The second exciting issue relates to our findings of the drastic solubility changes of intrinsically disordered proteins (IDPs), which needs to be discussed within the scope of drug design in general.

5.3 Data Integration as a Major Challenge in the Future

Finally, data integration represents one of the key elements of this work; from the integration of data from several published resources (**Chapter 2**) and overlapping transcript and proteomic data (**Chapter 4**), to integrating information about a protein's disordered structure with its stability and solubility state (**Chapter 3**). It has become clear that while each method or assay offers a peek into a complex biological system, the signaling layers and events are inter-dependent or interactive [Huang et al., 2017]. To determine coherent biological signatures, it is therefore crucial to combine different *omics* sources.
Methods capable of such integration are getting more and more prominent, yet generally focus on well-characterized pathways and are not necessarily performed in an unbiased manner. It is noteworthy that the methods presented in this thesis are also not offering a pipeline that can readily be applied to other problems. MOFA, or Multi-Omics Factor Analysis [Argelaguet et al., 2017], on the other hand, could provide an exciting means to explore further sources of protein abundance variation. The method is based on an

unsupervised dimensionality reduction that identifies the key sources of variation in multi-omics data. For a future hands-on application, MOFA could be applied to protein abundance data to infer the extent of heritability, ageing and diet effects, in a similar fashion as it has been illustrated in **Chapter 2**. One could also envision to scrutinize data acquired in **Chapter 4** more carefully in an integrative manner. For example, is a protein varying with age due to transcript variation, or are post-translational mechanisms primarily responsible for that phenomenon? In fact, on top of the analysis that has been conducted in regard to ageing, one could further explore whether age-dependent proteins are actually enriched in certain bio-physical features. Could the degree of the disordered state of the protein, for example, render it more prone to ageing? Clearly, there is more potential to be investigated on that part as well.
We can also assume that the studies presented in this thesis could prove useful as a resource that could be considered for applications on big data in biology. One particularly exciting outlook concerns surveying such big data using deep-learning algorithms- an approach that has been vividly discussed in 440 publications on the bioRxiv platform so far [Webb, 2018]. Though not without pitfalls such as prediction accuracy and sample quality and size, we can probably expect great breakthroughs from that mix of computational and biological sciences [Webb, 2018].

Taken together, this thesis presents a compendium of findings and hypotheses that could help advance our understanding of proteome dynamics and the proteotype as a model to incorporate into the current gap between the genotype and the phenotype.

Appendix A

Computational Materials & Methods

Code Dependencies and availability

Most of the presented computational analysis was performed in Python version 2.7.10 and R (version 3.3.1), and all of it is available under GNU General Public License V3 as GitHub projects (https://github.com/natalierom). Unless further specified, most of the figure design and plotting was performed using Matplotlib version 1.4.3 [Hunter, 2007a] and Seaborn version 0.8.0 [Waskom et al., 2014]. For numerical and statistical analysis the Python modules Scipy 0.19.1 [Pedregosa et al., 2011] and Numpy 1.13.1 [der Walt et al., 2011] were employed, as well as Pandas version 0.20.3 [McKinney and Others, 2010] for handling of data frames. For R packages such a Limma version 3.30.0 [Ritchie et al., 2015] and DESeq2 version 1.18.1 [Love et al.,2014] have been used for differential expression analysis and RNAseq analysis, specifically.

Related to Chapter 2:

Information Resource and Integration of Data. Protein-protein interactions were obtained from the STRING database (version 10.5, Szklarzyk et al., 2017); interactions were considered to exist if the combined score > 0, to be confident if the combined score > 500, and high-confident interactions if the combined score > 700. The database of complexes was manually compiled and curated from COMPLEAT and CORUM by Ori et al. (2016) and quantified proteins from all published datasets considered for the analysis were mapped accordingly. Pathways were obtained from the Reactome Pathway Database (downloaded in February, 2017; Fabregat et al., 2016, https://reactome.org/download-data/); cellular locations, on the other hand, were extracted from the Human Protein Atlas (downloaded February 2017, Uhlen et al., 2015) considering protein mappings only if this assignment has been either validated, supported or approved by antibody analysis. Chromosome locations

were mapped using the Python package *mygene* (https://pypi.python.org/pypi/mygene) using the *hg19* GenBank assembly for human and the *mm9* genome assembly for mice, respectively. Finally, essentiality of genes was defined based on the genetic screen performed in the human cell lines KBM7, K562, Jiyoye, and Raji by Wang et al. (2015) (Table S2, "Identification and characterization of essential genes in the human genome"); genes of a housekeeping role were obtained from the supplementary files of the report by Eisenberg & Levanon (2013).

Large-scale proteomic datasets. For the delineation of protein abundances across individuals, we considered primarily large-scale shotgun proteomics studies on human individuals, cancer patients and mouse strains. For control purposes we also included the proteomic profiles of 11 human cell lines generated by Geiger et al. (2012). Technical specificities of each dataset (i.e. sample number, MS-acquisition, …), as well as the number of quantified proteins, as well as all required module mappings are given in **Table S2.1**. For the 60 Yoruban HapMap individuals [Battle et al., 2015] not only proteomic, but also the respective data from RNASeq-analysis and ribosome profiling was available and therefore included as well for control purposes. Also data on DO (diverse outbred) mouse strains were available at the proteomic as well as transcript level [Chick et al., 2016]. The cancer proteomics datasets were downloaded from the TCGA CPTAC project [Cancer Genome Atlas Research Network, 2011; Zhang et al., 2016; The Cancer Genome Atlas Network, 2012; Mertins et al., 2016; The Cancer Genome Atlas Network, 2012; Zhang et al., 2014].

AUROC analysis. For the Receiver Operation Curve (ROC) analysis across different types of modules in different datasets, true positive hits were defined based on the databases as outlined above (STRING, complex interactions, pathway membership, etc.). For pathways, in particular, we removed interactions that are known to occur in a complex context (i.e. ribosome complex in the broader translation-related pathways). Notably, for the category 'chromosome location' we would consider true positive 'interactions' to exist between genes encoded on the same chromosome. For the categories 'essentiality' and 'housekeeping role' the true positive interactions would be defined as any interaction that might occur between essential genes and housekeeping genes, respectively. The set of false positives, on the other hand, was defined by protein-protein pairs that were not shown to exist in the particular category in question; since the number of false positives thereby is much higher than true positives, we randomly sampled from the potential false positives the same number of true positives for ROC curve calculation.

Complex co-abundance. The database of complexes was manually compiled and curated from COMPLEAT and CORUM by Ori et al. (2016) and quantified proteins from all

published datasets considered for the analysis were mapped accordingly. Notably, we also mapped a complex as 'well-defined' in case of increased literature content for the respective complex. For further analysis only protein complexes with at least 5 quantified members were considered. Proteins assigned to the same complex were correlated (*Pearson* method); as a control proteins that were not part of any complex assembly were randomly assigned into artificial complexes and cross-correlated as well (**Figure S2.2B**). In addition, the data was also permuted and proteins subunits were again tested for co-abundance using the Pearson correlation (**Figure S2.2B**).

Gene ontology analysis. For all gene ontology (GO) analyses in this study, respective genes were analyzed using DAVID (version 6.7, Huang da et al., 2009). The GO-terms 'Biological Process', 'Molecular Function' and 'Cellular Compartment' were considered; the background for the GO-analysis was represented by all quantified proteins in a given dataset. Results were filtered according to FDR (Benjamini-Hochberg) of less than 0.01; the fold-changes associated with those significantly enriched GO-terms were usually shown (**Figure S2.3; Figure 2.5C**).

Identification of stable and variable protein complexes. As a general principle we used median correlation of protein complex subunits as a proxy to differentiate between stable and variable complexes. To cross-compare the extent of complex stability and variability, the correlations were ranked in each dataset; finally the median correlation of each complex as recovered from each considered dataset was calculated, and complexes were sorted accordingly. The top quantile (25%) of these complexes were considered to be highly stable ($R^2 > 0.4$), whereas the lowest quantile were considered highly variable ($R^2 < 0.2$). Fisher's exact test was used to assess the significance of the overlap of both variable complex members and complexes between the datasets of reprogramming and 11 cell lines (**Figure 2.2B**).

Protein Complex Stoichiometry Analysis. To allow for an assessment of compositional rearrangements of protein complexes as opposed to their overall abundance changes, a module-wise normalization was performed, as previously described [Ori et al., 2013; Ori et al., 2016]. Basically, protein belonging to the same complex were normalized by the respective trimmed mean of the complex across all individuals/samples. In case of proteins involved in multiple complexes, the average value from all the corresponding complexes was taken into account. Given the complex-normalized abundances, the variance of each subunit in a given complex was calculated. To cross-compare these variances between vastly different proteomics datasets and approaches, those variances were converted to z-scores

per complex (**Figure 2.4**). Proteins were considered 'stable' or 'variable' in case of the associated p-value < 0.05.

Protein Modification Site Enrichment Analysis. Protein modifications sites were extracted from PhosphoSitePlus (downloaded February 2017) [Hornbeck et al., 2015], and mapped against all quantified proteins within the datasets in question. To perform a Fisher exact test, contingency tables for each modification separately; briefly, proteins would be grouped according to whether they tend to have less of a certain modification site ($<$ median of number of modification sites/protein), or more. For these two groups then, we checked whether more or less variable complex subunits were enriched. For visualization purposes (**Figure S2.5A**), we considered emphasizing enrichments with a p-value < 0.05, and to specify whether stable or variable subunits are enriched in having more of a certain modification site.

Sex- and diet-specific abundance and stoichiometry changes. To assess the differences in abundances of entire complex structures between male/female mice, and mice exposed to high-fat and chow diet, the median abundances of each complex was calculated in each individual/sample (protein subunits were required to be quantified in at least 50% of samples). For each complex it was then assessed via a t-test whether median complex abundances in male mice were significantly different from the ones in female mice; the effect size was monitored as the Cohen distance [Sullivan et al., 2012]. P-values were further adjusted using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] (significance $\alpha = 0.05$), and complex structures were considered significantly different in case of q-value < 0.01. For internal rearrangements of the complex (stoichiometry), we performed a separate analysis applying the R-package LIMMA (Linear Models for Microarray data analysis, [Ritchie et al., 2015]) using the complex-normalized protein abundances as input. Contrasts were set accordingly to identify differences between male/female mice and high-fat/chow mice, respectively. P-values were adjusted using the false discovery rate (FDR) as described by Benjamini Hochberg, and protein complex members were considered differentially expressed (termed DEP (differentially expressed protein)) in case of q-value < 0.01. The corresponding fold-changes are highlighted in volcano plots in **Figure 2.5B**. We also ran the LIMMA-analysis for each complex on the not complex-normalized proteins to assess the extent of truly stoichiometric rearrangements. The analysis was also performed for Reactome pathways, and can be readily applied to any specified protein set/module.

Explained variance of module abundance and stoichiometry. To understand to what extent both protein complex abundance and stoichiometry is affected by either sex or diet,

a co-variate analysis has been performed using the scikit Python package (http://scikit-learn.org).

Comparison of module stoichiometry across datasets. In order to cross-compare module stoichiometries between the proteomics datasets, a LIMMA analysis was performed with a different set-up. Specifically, for each complex the trimmed mean vector across all given samples was calculated and compared against the non-complex-normalized abundances of each complex member. For each complex (n=73) that were consistently quantified across all given datasets we could then recover the respective fold-changes of involved subunits. For each protein we then computed the variance in module-specific fold-changes discovered in human and mouse proteomics datasets, respectively, and also assessed the level of variance between human and mouse datasets (data not shown).

Software and Data Availability. Statistical tests and visualizations were performed using Python version 2.7.10 and R version 3.3.1. All scripts and underlying data are available on demand.

Related to Chapter 3:

MS Computational Analysis and Normalization (short summary on analysis prepared by Frank Stein). Raw mass spectrometry files were processed with IsobarQuant [Franken et al., 2015]; the Mascot 2.4 (Matrix Science) search engine was used for peptide- and protein identification (human UniProt database with reversed protein sequences). Search parameters are described in the manuscript. In total three independent datasets were analyzed: i) 2D-TPP data, ii) SDS-data and iii) TPP-TR data. Potential batch effects were removed using *limma* [Ritchie et al., 2015] and all datasets were normalized using variance stabilization (*vsn*, Huber et al., 2002).

Thermal proteome profiling (TPP) analysis (short summary on analysis prepared by Frank Stein). In order to identify melting points the normalized TPP-TR data was subjected to the TPP package [Franken et al., 2015]. For the 2D-TPP dataset, on the other hand, fold changes were calculated against 37°C and the G1/S data points with *limma*, which are the base for abundance and stability scores. Finally, for SDS data, fold changes were similarly calculated relative to the G1/S stage, and transformed into z-scores corresponding to expression scores. For each expression score the global FDR was yielded using *fdrtool* [Strimmer, 2008]; the local FDR by correcting *limma* p-values using the Benjamini Hochberg procedure [Klipper-Aurbach et al., 1995] with the *p.adjust* method.

Regarding the abundance and stability cores, an additional bootstrapping was performed with 500 iterations for accurately quantifying a distribution of abundance and stability values per protein and estimating the likelihood that the distribution is different from 0 (no change). The averages were transformed into the z-distribution, reflecting more accurate stability and abundance scores, and the global FDR was calculated using *fdrtool*. Proteins are considered to change significantly in abundance, expression or stability in a given cell cycle stage if the corresponding local and global FDR is < 0.01.

Clustering of proteins. Only proteins with at least one significantly changing score in abundance, expression or stability, were considered for clustering. Abundance-, expression- as well as stability vectors were concatenated for each proteins and used to calculate the Euclidean Distance using the Ward.D2 clustering method. 21 different clusters were yielded, describing the data most accurately.

Solubility. Similar to above calculations, the solubility scores and corresponding q-values were calculated by comparing solubility in G1/S (abundance difference between NP40 and SDS) and mitosis (abundance difference between NP40 and SDS). Additional requirements were an FDR cut-off of 0.05, an effect size of >0.322 (25% regulation) for the log2 fold-changes and a solubility in G1/S of <0.

Reactome pathway analysis. Using the *mygene*- package implemented in Python (https://pypi.python.org/pypi/mygene), quantified proteins from this study were mapped to Reactome pathways (http://reactome.org/pages/download-data/, February 2017; Fabregat et al., 2016). Corrected p-values (local FDR) of the quantified protein components of a pathway were combined for each pathway using the Empirical Brown method (EBM) to account for dependent p-values [Poole et al., 2016]. Further, these combined p-values were corrected using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). In order to display pathways of biological relevance in the context of thermal stability, we ensured that the pathway is sufficiently covered (>60% pathway coverage with 5-150 quantified protein members) and that at least 30% of its components change significantly in at least one cell cycle stage (local FDR (stability) < 0.01). That way, 106 pathways were considered and manually curated for proteins that would appear in several pathways; overall, 30 non-redundant pathways were gathered that were visualized accordingly in **Figure 3.3** (**Figure S3.2, Table S3.5**). Full opacity in this figure occurs in case the combined adjusted p-value for a pathway in a given stage is less than 0.05, otherwise bubbles would stay transparent. The color signifies the overall tendency of a pathway to be stabilized or de-stabilized as derived from the median stability score of its significantly changing protein components (local FDR (stability) < 0.01) in each cell cycle stage.

Gene ontology analysis. Proteins changing significantly during the cell cycle in either stability or abundance were subjected to gene ontology analysis using DAVID (version 6.7) [Huang da et al., 2009a, b]. The GO-terms 'Biological Process', 'Molecular Function' and 'Cellular Compartment', as well as the SP-PIR keywords were considered. The background for the GO-analysis was represented by all quantified proteins. Results were filtered according to FDR (Benjamini-Hochberg) of less than 0.05 and shown accordingly in **Figure S3.2** (**Table S3.5**). **Figure 3.4**, on the other hand, represents a snapshot of fold enrichments of the top-hits in each GO-category for each of the 21 clusters.

Complex correlation analysis. The database of complexes was manually compiled and curated from COMPLEAT and CORUM by Ori et al. (2016) and quantified proteins from the experiments were assigned accordingly. Stability- as well as abundance z-scores were concatenated for each protein, or remained separate vectors, respectively. Proteins assigned to the same complex were correlated (Pearson method); as a control proteins that are not known to be part of any complex assembly were correlated with other proteins of this type (**Figure 3.7**, **Figure S3.3-S3.5**).

Complex sub-clustering. In order to define sub-complexes in known complexes, we again considered concatenated stability- and abundance vectors for each protein. For each complex k-medoids were defined iterating from 2 to 5 potential clusters per complex, and the Euclidean distance between those medoids was monitored at each iteration step. The best possible clustering per complex is recovered from the maximum distance between the respective medoids.

Analysis of disordered proteins. We hypothesized that disordered proteins might be particularly affected in the observed shift in melting points in mitosis, and therefore interrogated proteins with this particular phenotype on their disordered state. The fraction of disordered parts of a protein is described in the d2p2-database, which combines predictions of several algorithms and makes calls on regions if they have been found to be disordered in at least 75% of those libraries (PONDR VL-XT PONDR VSL2b, PrDOS, PV2, IUPred (+sub-versions of it), Espritz (+sub-versions if it)) (http://d2p2.pro/, February 2017). Given the relative number of amino acid residues that are positioned in supposed disordered regions, proteins were ranked (notably, the disordered region was supposed to span at least 5 amino acids to be valid). Additionally, the relative disordered rank was also corrected taking into account known structural motifs that might overlap with the predicted disordered region using the PDB (see **Table S3.6** and **Table S3.7**).

Mapping of proteins to posttranslational modifications. Information on known acetylation, methylation, ubiquitination, sumoylation and phosphorylation sites was extracted from PhosphoSitePlus (http://www.phosphosite.org, March 2017; Hornbeck et al., 2015), and cross-correlated with stability and abundance patterns of quantified proteins across the cell cycle stages. Mitotically regulated phosphorylation sites, on the other hand, were taken from Olsen et al., 2010 (**Table S3.6** and **Table S3.7**).

Organelle/Compartment Analysis. Cellular compartments were extracted from the GO-terms (only experimental evidence is considered, as opposed to inference from computational methods) and the Human Protein Atlas (http://www.proteinatlas.org/, May 2017; Uhlen et al., 2012). For latter only data supported, validated or approved by antibody analysis is considered. Notably, for the evaluation of the solubility transition data, we took into account Human Protein Atlas data, as well as data from Wilkie et al. (2017) for nuclear envelope annotations.

K-means clustering of thermal profiles of complex-subunits (data not shown). Clustering of thermal profiles of protein subunits that have been previously assigned to protein complexes (279 in total) with the *k*-means algorithm allowing up to 10 clusters at each iteration step, respectively. For each iteration step we perform 250x resampling to acquire *cluster agreement probabilities*. Thereby it can be assessed how likely a particular complex could indeed be divided into X clusters. In addition we also calculate the probability that a particular subunit is indeed in one cluster or the other (*membership probability*).

Data and Software Availability. All analyzed data is made available in supplementary tables and figures. Clustering of individual complexes can be found at Mendeley: http://dx.doi.org/10.17632/xrbmvv5srs.2. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD008646. All code used for data analysis is available on demand.

Related to Chapter 4:

MS data analysis. For the analysis of the aging MS-data two different strategies were employed: i) TMT quantification to assess changes in protein abundance upon ageing within the same cell population, and ii) a label-free (LF) approach to determine differences in protein abundance across cell populations

Data analysis for TMT based quantification (short summary from analysis as done by Marco Hennrich). MS raw files were processed through a search pipeline set up in the Thermo Proteome Discoverer (1.4.1.14) with spectra getting de-isotoped, deconvoluted and the mass-range from 126-131.3 m/z being excluded prior to database search in Mascot (version 2.5.1, UniProt database). Search parameters considered trypsin cleavage specificity, 1 permitted missed cleavage site, precursor mass tolerance $<$ 20 ppm, fragment mass tolerance of 0.02Da, carbamidomethylation and TMT as fixed modifications and methionine oxidation as a variable modification. The false discovery rate (FDR) was calculated with Percolator (version 2.04). For further analysis unprocessed peptide spectrum matches (PSMSs) with FDR$<$1% were exported and processed.

PSM filtering and TMT ratio calculation. Prior to calculation of TMT ratios, PSMs were filtered according to the following parameters: i) search rank $\leq$1, ii) isolation interference $<$30%, iii) sum of intensities across TMT channels 127-131 $>$30,000, iv) no missed cleavages. In case quantification values would be missing in the channels 127-131, they were replace by the representative minimum value detected in the corresponding TMT experiment. For each PSM the intensities of the TMT channels were used to calculate the ratios of the donor samples against the internal standard (TMT channel 126). The ratios were normalized by multiplication with the median ratio as derived from in between the internal standard and the respective donor channel. To determine peptide ratios, the median ratio of the three PSMs with highest precursor intensity was calculated. After assigning peptides to proteins, latter were grouped according to gene names; for simplicity these groups are referred to as proteins throughout the report. The eventual TMT ratio for each protein is defined as the median ratio of peptides that are uniquely assigned to the protein group. We only consider proteins with at least two unique peptides for quantification.

Data analysis for label-free (LF) quantification. In order to perform accurate comparisons of protein abundances across different cell populations, all MS raw-files were additionally analyzed with MaxQuant 1.5.3.17 [Cox and Mann, 2008] specifically with label-free (LF) quantification (absolute protein abundance estimation). Search parameters were set as mentioned above for Thermo Proteome Discoverer. For each peptide, the respective precursor intensities were extracted in each LC-MS analysis and all associated PSMs collected. Further detailed information on the PSMs were taken from the corresponding evidence files, to the end that for each identified peptide in each LC-MS run one intensity value can be assigned. For further analysis, only PSMs common to both the TMT quantification approach as described above the MaxQuant analysis were taken into account and filtered as described above. The TMT-ratio for a peptide were derived from the median

of its top three PSMs with the highest sum of intensities, as it has been the case in the above described procedure.
Given the MS1 intensity of an individual peptide species, we fractioned the total precursor area according the TMT ratios as derived from the reporter ion intensities of the six TMT channels. Area intensities per channel were then corrected for potential sampling aberrations by multiplying them with the median ratio determined between the internal standard and the respective donor channel. These normalized area intensities per channels were then divided by the number of potentially observable unique tryptic peptides per protein (criteria: peptide length 8-25 amino acids, no missed cleavage allowed). The label-free score per protein then basically represent the sum of the label-free scores of the corresponding unique peptides. To allow for comparison between donors across the different cell populations, the LF-scores per protein were log-transformed, the median was calculated in each TMT-6plex experiment and was subtracted from each log-transformed label-free score in each donor. Notably, to summarize these normalized LF-scores across all donors in one cell type, the sum of the unlogged LF-score was divided by the number of donors available for the respective cell type (denoted as the LF sum-value in **Table S4.3**).

Quality assessment. *Reproducibility.* For reproducibility assessment one MSC lysate was split into 3 aliquots prior to digestion, and treated as individual samples subsequently. TMT-ratios were highly reproducible between these technical samples (mean R^2=0.941, median CV = 4.2%). *Labeling Efficiency.* All experiments were also searched for with TMT-label defined as a variable modification; to estimate the labeling efficiency the number of all completely labeled PSMs was divided by the total number of PSMs as identified in the search-run (98.5%). *Sample Exclusion.* To determine outlier samples, principle component analysis was performed using log2-transformed TMT-ratio and using Hyndman visualization, samples were discarded when there was >97% probability that they were indeed outliers. Thereby 14 samples of the original 284 samples had to be removed from the study.

Analysis of protein expression with age and statistical significance calculation (TMT). For identifying age-associated proteomic changes within each cell population, we performed a Spearman correlation analysis to detect proteins whose expression changes with age. For each protein with a donor coverage above 15% we calculated the Spearman correlation between the quantified TMT-ratios and the respective donor ages to assess its behavior with respect to age. Positive correlations indicate an increase of the abundance of a protein with advanced age, while negative (or anti-) correlations indicate a decrease of its abundance with age. Proteins with a p-value < 0.05 are considered to be significantly altered upon ageing. We also checked for a possible effect of the gender disparity in the samples by re-

analyzing the proteomic data after removing all female samples. The criteria for the male only analysis was identical with the analysis of all samples. The comparison of the results from the male only versus all samples is visualized in **Figure S4.6**.

Hierarchical clustering related to Figure 4.2B. Hierarchical clustering as depicted in **Figure 4.2B** was performed using the scipy-python package (python.org) with the linkage matrix based on the correlation metrics and using complete clustering.

Statistical methods to determine changes between cell populations related to Figure 4.2C (LF). Label-free abundances for proteins were leveraged to understand differences between the six cell populations, and not for age differences as these are more reliably analyzed by TMT ratios. Filtering criteria for protein inclusion were as strict as for TMT-quantification, requiring proteins to be quantified in at least 15% of available donors of a given cell population. In order to understand whether pathways differ in their abundance or stoichiometry across the different cell populations (**Figure 4.2C**), we applied the following pipeline: We considered proteins quantified in the LF-approach (7585 proteins) and mapped those against the *Reactome* database (http://www.reactome.org/download-data/, February 2017) after filtering the database for pathway sizes of 5-100 proteins. For each cell population, the abundance of a pathway was approximated by the median LF-abundance of proteins associated with it. The average of those medians results in the mean abundance of the pathway across the different cell populations (y-axis of **Figure 4.2C**). To estimate the fraction of proteins in a pathway that change in their stoichiometry, we proceeded as follows: For all 270 samples, protein abundances were normalized to the median pathway abundance to avoid any significance stemming from abundance change of the entire pathway. These normalized values if at least 15% of donors were quantified in an individual cell population, were cross-compared between the cell populations (*Wilcoxon* test). From the set of p-values obtained from those comparisons, the mean p-value is calculated. This procedure is iterated across all proteins associated with pathways, and all p-values are adjusted thereafter using the Benjamini-Hochberg procedure. To finally obtain the fraction of the pathway that shows a significant alteration between the cell populations, we considered the number of proteins per pathway with an adjusted p-value of less than 0.05. Note that this procedure does not take into account cell-type specificity of proteins, and does not restrict itself on proteins that are expressed throughout all cell populations. The Supplementary Table 4 contains details for further exploration.

Complex & pathway co-abundance analysis (TMT). Complex annotations were based on Ori et al. (2016) who manually curated a list of 279 non-redundant protein complexes derived from CORUM AND COMPLEAT protein complex sources. Pathway annotations

were sourced from Reactome as subsequently for the pathway analysis (http://www.reactome.org/download-data/, February 2017; Fabregat et al., 2016). Proteins assigned to the same complex were correlated against each other using their TMT-ratios across donors per cell type. The densities represent the distribution of all correlation coefficients as derived from Pearson correlation analysis from in between subunits of the same complex; the same procedure is applied to pathways, with the exception that protein pairs known to be in complexes were specifically excluded. The background distribution, on the other hand, was derive from correlations of randomly selected proteins who are not known to occur together in functional modules such as complexes or pathways.

Pathway enrichment related to Figure 4.3 (TMT). The Reactome Database (http://www.reactome.org/download-data/, February 2017) was the basis for the analysis of pathways. The displayed pathways of **Figure 4.3** were selected based on their size, with the requirement for at least 5 and a maximum of 150 proteins. More than 30% of these needed to be quantified by TMT in at least one cell population and out of the quantified more than 20%, but at least three proteins need to be significantly altered upon ageing (p-value < 0.05). Thereby, we obtained 28 pathway hits for HPC, 44 for LYM, 3 for GRA, 4 for MON, 2 for ERP and 221 for MSC. In order to avoid redundancies, we removed pathways whose significantly altered proteins were completely covered in another pathway. If a pathway, however, contained at least one unique significantly altered protein, the pathway was kept. In a scenario of pathways containing exactly the same altered proteins, the largest pathway was reported. In case of equal pathway size, the pathway with a higher hierarchy level was taken. Thereby, we obtained 109 pathways in total that were largely non-overlapping and exhibiting considerable changes upon ageing. In order to decide whether the proteins of a pathway have a general tendency of increasing or decreasing upon ageing, we first calculated the slopes of all proteins based on the linear regression between the donor to internal standard ratio (average normalized per TMT-6-plex experiment) and the age. For the slope calculation normalization by studentization of the protein ratios was avoided and average normalization was applied instead for normalization to avoid artificial high slopes for slightly altered proteins (see section 'Summary of the normalization steps of the TMT data'). The slopes of all significantly altered proteins within a pathway were averaged. Pathways with an average slope of the altered proteins between -0.001 and 0.001 per year of life were reported as having no tendency. A slope of 0.001 translates to an estimated average increase of 4% in protein abundance in a life span from 20 to 60 years (40 years). The results of the 109 pathways were displayed in **Figure S4.7**. **Figure 4.3** is a selection of these 109 pathways, which is based on taking the five most up- and down-regulated pathways per cell population.

Cross-cell population correlation related to Figure 4.6C. Extracellular protein-ligand pairs extracted downloaded from http://fantom.gsc.riken.jp/5/ [Ramilowski et al., 2015] and overlapped with the TMT quantifications. To be further considered for quantification, receptor and ligand proteins were required to change upon ageing in the respective cell population (p-value < 0.1). For each cellular interface (MSC to HPC, MSC to LYM, MSC to GRA, MSC to MON, MSC to ERP), protein profiles (based on normalized TMT ratios) of at least 85% overlapping donor individuals were correlated using the Spearman method. Correlation coefficients were visualized in **Figure 4.6C**, highlighting significant correlation results (p-value < 0.1).

Data analysis of the transcriptomics data of total HPC population. Reads were trimmed for Nextera, Smart-seq2 adapter sequences using skewer-v0.1.125. Trimmed read pairs were mapped to human genome hg38.ERCC using HISAT2 version 2.0.0-beta. Uniquely mapped read pairs were counted using featureCounts subread-1.5.0, using exons annotated in ENSEMBL annotations, release 75. The subsequent analysis is performed in the programming language R. For the analysis of the raw counts retrieved from RNA-seq experiments, we used the *DESeq2* package (version 1.18.1). We applied a minimal pre-filtering to remove rows that have only 0 or 1 read, as suggested in the *DESeq2* manual. To assess the quality of the data PCA was used after a 'regularized log' transformation (*rlog*) of the data, accounting for the library size of each sample. Using HDR-plots on the derived principal components similar as described in the section 'Quality assessment and sample inclusion of the proteomics analysis', outliers were detected for HPC, GRA, MON and MSC and removed. For comparison with the corresponding proteomics data, we defined young as ≤ 30 years, and old as ≥ 50 years (similar to **Figure 4.6B**). For each fold-change (old/young) a weighted p-value was calculated based on the IHW-package that takes into account number of reads as a covariate for the adjustment of p-values. From the output-table the *log2* fold-change was extracted for **Figure S4.6**, which gives an estimate for the effect size.

Single-cell data pre-processing. The single-cell data pre-processing is performed using the programming language R. Raw reads were processed using the recent version of the Salmon pipeline (v0.9.1), with the index derived from transcriptome data from the hg38 build for mapping purposes (http://ftp.ensembl.org/pub/release-87/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz).
The count matrix generated for individual transcripts across cells in each sample was then subjected to further processing using the Bioconductor package *tximport*. Thereby, the transcript-specific count-tables were converted into gene-specific count tables across cells. To filter for qualitative cells, we only retained cells where at least 1000 genes have been

found to be expressed at a minimum of 10 reads each, and where the total read count is at least 150.000. That filtering step has been adapted from Velten et al. 2017. Additionally, only genes with at least 10 reads in at least 5 cells were kept for further processing. The resulting count tables were analyzed using the Bioconductor package *simpleSingleCell* (version 1.2.0). The pipeline was applied with the following steps: (a) additional quality control on cells and filtering due to library size and possible batch effects, and (b) normalization of cell-specific biases using computed size factors. For details on individual samples, the table below is to be consulted. The normalized log-expression values were further adjusted to the mean expression in each cell.

List containing information on donors used for single-cell analysis. The columns list the age and gender of the donor, the number of cells analyzed per donor (# of cells retained), the median library size, the number of unique genes, the average number of genes per cell, as well as the fraction of lymphoid- and myeloid-primed cells per donor.

Sample ID	meta-data	# of cells retained	% of cells retained	Median library size	# of unique genes	average # of genes/cell	# of lymphoid cells (fraction of all cells)	# of myeloid cells (fraction of all cells)
1	male/59.2 years old	101	52.6%	686,990	11,976	5,819	9 (8.9%)	49 (48.5%)
2	female/30.6 years old	152	79.2%	1,662,115	13,374	5,711	35 (23.0%)	58 (38.2%)
3	female/62.2 years old	127	66.1%	1,656,706	13,093	5,991	32 (25.2%)	48 (37.8%)
4	female/21.2 years old	139	72.4%	1,212,586	12,693	5,274	28 (20.1%)	41 (29.5%)

For the last two columns consider the Methods below (Clustering of single cells into lymphoid and myeloid clusters.).

Clustering of single cells into lymphoid and myeloid clusters related to Figure 4.5E. For clustering of single cells, we used Python version 2.7. We first determined whether lymphoid and myeloid markers as delineated in **Figure 4.5** for the proteomics data were yielding signal in the single-cell RNA-seq dataset. To ensure signal consistency, we excluded markers that did not correlate with the other lymphoid or myeloid markers, respectively (p-value < 0.01). Thereby, ITGA6 had to be removed from lymphoid markers (remaining 12 genes), and IKZF1, ITGAL, PRAM1 and BCL11A from myeloid markers (remaining 8 genes). The lymphoid and myeloid markers also had a significant correlation with other known lineage markers, such as TFRC (CD71) and CD19 (data not shown). For further analysis CD71 and CD19 were also included. To cluster cells into lymphoid/myeloid lineage, cells were required to have at least half of the respective markers stably expressed (>0), and none of either CD71, or CD19 (**Figure S4.8**). That way we could characterize cells in a more conservative manner as being lymphoid-primed, or myeloid-primed cells; cells that did not fall into either of those categories, were labeled as 'undefined'. When compared to the entire set of cells per donor, we could see that cells defined as lymphoid- or myeloid-

primed were significantly different in their marker constellation (Fisher exact test: average p-value (lymphoid) = 9.77 x 10^{-10}, average p-value (myeloid) = 1.47 x 10^{-2}), whereas this was not the case for the undefined cells (Fisher exact test: average p-value= 7.46 x 10^{-1}).

Deriving lineage- and age-dependencies for gene expression changes in glycolysis. Genes derived as age-dependent from the proteomics dataset (**Figure 4.4**) were subsequently analyzed on whether they are affected by lineage, or the age of the donor in the single-cell RNA-seq dataset. We correlated genes involved in the glycolysis, TCA cycle, and fatty acid oxidation (FAO), with lymphoid and myeloid markers, respectively. We found that age-regulated genes (p-value < 0.1, **Figure 4.4**) had a stronger disparity between lymphoid and myeloid correlation distribution than genes not found to be age-dependent (**Figure 4.5D**) (age-dependent: effect size (Cohen)=1.59, p-value=7.29x10^{-8}, age-independent: effect size (Cohen)=0.9, p-value=3.06x10^{-3}). This lineage effect was further examined by calculating the ratio between expression levels in lymphoid versus myeloid cells for each protein. Age-regulated enzymes of the upper glycolytic pathway (p-value < 0.05) were found to be significantly affected by lineage (p-value < 0.01, Mann-Whitney U-test) across samples, as opposed to enzymes that had no age-dependency. Ageing effects on enzyme expression levels were tested in lymphoid and myeloid cells, respectively, to remove lineage effects. The slope calculated from the median expression levels across samples indicated that age-upregulated glycolytic enzymes are indeed more prone to becoming higher expressed with age, at the single-cell level as well (ANOVA-test: p-value = 0.079) (**Figure 4.5E**).

Data & Software availability. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD007048. Raw data for both the single-cell RNA-seq and bulk RNA-seq experiments have been uploaded to the Gene Expression Omnibus (GEO), and are accessible by the accession number GSE115353. Result tables from the bulk RNA-seq analysis are available as Supplementary Material to the manuscript. Code and pipelines are available on demand; the computational pipeline for this chapter is outlined in **Figure S4.9**.

Appendix B

Experimental Materials & Methods

Related to Chapter 3:

Experiments were performed by Isabelle Becher and Amparo Andes-Pons from the Savitski and Beck Lab. Please note that the following descriptions of experimental procedures can be followed up in detail in the corresponding manuscript and is meant to provide a short overview on the methodology underlying the data used in this report.

Cell culture and cell cycle arrest. HeLa Kyoto cells were cultured at 37°C (5% CO_2, DMEM with 1mg/ml glucose, 10% FBS, 1mM glutamine) and synchronized in G1-S using a double-thymidine block (2mM thymidine, Calbiochem) overnight (o/n). They were then released for 8 hours, and blocked a second time o/n. After the second release cells were collected at 0 hours (G1-S), 2 hours (early S), 4 hours (late S) and 6 hours (S-G2). For synchronization in mitosis, cells were treated o/n with 2mM thymidine, released for 4 hours and then treated with 100ng/ml nocodazole (Millipore) o/n. After a release of 0.5 hours mitotic cells were collected. For synchronization in G1, cells synchronized for mitosis as previously described, were collected by shake-off and released into fresh medium for 4.5 hours. Only attached cells were then collected (G1). A culture of asynchronous cells was used as a control.

Flow cytometry. The synchrony of cells was checked upon by FACS with two bivariate analyses: i) DNA/Proliferating Nuclear Antigen (PCNA, Sasaki *et al.* 1993) and ii) DNA/phosphorylated serine 10 of histone H3 (pH3). For i) $1x10^6$ cells were permeabilized for 10 min on ice with 0.1% Triton X-100/1% bovine serum albumin (BSA)/PBS, and fixed with methanol at -20 °C for 3min. After pelleting, cells were re-suspended and stored in PBS at 4 °C. For ii) $2x10^6$ cells were fixed with 1 % formaldehyde for 15 min at RT. Cells were then re-suspended in PBS and stored in 70 % ethanol at 4 °C. Both samples were spun

and the cell pellet was re-suspended for 15min in PBS at RT before blocking in 1% BSA/PBS for 10min prior to incubation with the primary antibodies (anti-PCNA PC10 conjugated with Alexa Fluor 488, Cell Signaling Technology or anti-Phospho-Histone H3 (Ser10) conjugated with Alexa Fluor 647, Cell Signaling Technology) for 1 hour (R). Cells were washed and DNA was stained by 20 µg/ml propidium iodide (PI) in 0.1 % Triton X-100 containing 0.2 mg/ml RNase. Samples were analyzed on a BD LSR Fortessa instrument with the laser line of 561nm (75mV) used for PI detection (BP-filter of 610/20nm). The actual cell cycle stage assessment was performed with FlowJo v10 software. Bivariate analysis of PCNA vs DNA content presented an inverted U shape; cells with higher signal were considered to be in S-phase. Cells with a G2/M DNA content, and positive for pH3 antibody (ii) were considered mitotic. The % of cells in G1 was calculated by fitting the PI signal to a cell cycle distribution using the Watson pragmatic model approach.

Thermal protein profiling (TPP) and sample preparation for mass spectrometry. TPP experiment was conducted as described in Becher *et al.*, 2016. Briefly, cells were harvested, washed with PBS and 10 aliquots were distributed in PCR tubes that were heated for 3min to different temperatures, respectively. Lysis buffer (final concentration 0.8 % NP-40, 1.5mM MgCl2, protease inhibitor, phosphatase inhibitor, 0.4 U/µl benzonase) was added and cells were incubated at 4 °C for one hour, whereupon protein aggregates were removed. From the supernatant the protein concentration was determined and for sample preparation 10 µg of proteins were reduced, alkylated, digested with trypsin/Lys-C with a modified SP3 protocol and labeled with TMT10plex (Thermo Fisher Scientific). Here, samples were combined in two ways: either all temperatures of one cell cycle arrest were combined allowing the determination of melting curves (TPP-TR), or all cell cycle arrests were combined for each temperature resulting in seven conditions analyzed together (2D-TPP). Pooled samples were fractionated on a reversed phase C18 system running under high pH conditions, resulting in twelve fractions. In addition, both samples from NP-40 lysis (37°C) and SDS lysis from G1/S or M arrested samples were combined, allowing the determination of protein solubility differences.

LC-MS/MS measurement. Peptides were separated on an UltiMate 3000 RSLC nano LC system (Thermo Fisher Scientific, pre-column of C18 PepMap 100, 5µm, 300 µm i.d. x 5 mm, 100 Å, and an analytical column (Acclaim PepMap 100, 75 µm x 50 cm C_{18}, 3 µm, 100 Å)). The LC-system was directly coupled to a Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific) via a Nanospray-Flex ion source (spray voltage 2.3kV, 320°C). Solvent A was 99.9% LC-MS grade water with 0.1 % formic acid and solvent B was 99.9% LC-MS grade acetonitrile with 0.1 % formic acid. The elution scheme is described in detail in the Materials & Methods in the publication. The MS was operated in positive ion mode with

full-scan MS spectra collected within a mass range of 375-1200 m/z (resolution of 70,000) and fragmenting top 10 peaks per MS-scan (normalized collision energy = 32, dynamic exclusion = 30s)

RNAi experiments. HeLa Kyoto cells were transiently transfected using siRNA spotted cell culture plates and CELLviewTM slides following a previously described [Erfle et al., 2007] reverse transfection protocol using pre-designed Silencer® Select siRNAs (Ambion). A non-transfected control (NTC) was included for each experiment, where cells were seeded on non-spotted wells. The knock-down of target genes was validated at the level of mRNA by qRT-PCR after 2 days of transfection, as well as at the level of protein by LC-MS/MS analysis after 2, 3 and 4 days of transfection. Cell cycle stages were checked upon by FACS using PI staining; in addition, apoptosis was measured using the Annexin V-FITC Detection kit (Promokine). For assessment of live and dead cells a commercial assay was used (Promega, MultiTox-Fluor Multiplex Cytotoxicity Assay).

Tubulin heatshock and microscopy. HeLa Kyoto EGFP-alpha-tubulin/H2B-mCherry were synchronized at the G1/S transition. After the second thymidine block, cells were released for 8 to 10 hours and mitotic cells were collected, washed, re-suspended in PBS, and further subjected to a heat treatment as described above. After fixation (4% PFA, 15min RT), an aliquot of each sample was set onto a well of a CELLview slide (Greiner Bio-One) previously coated with poly-L-Lys, and further imaged on an inverted Leica TCS SP8 STED3x microscope (Leica Microsystems, Mannheim, Germany), using a HC PL APO CS2 63x/1.40 OIL objective and a pulsed White Light Laser (WLL, emission from 470-670nm). Standard, diffraction limited confocal images were recorded for mRFP (594 nm, BP 600-690 nm) and EGFP (488 nm, BP 495-541 nm) and z-stacks were acquired (z-step size of 1µm). Images were processed using the ImageJ (Fiji).

Related to Chapter 4:

Experiments were performed by Marco Hennrich, Fei Ye and Ximing Ding and members of the Gavin lab. Please note that the following descriptions of experimental procedures can be followed up in detail in the corresponding manuscript referenced at the beginning of Chapter 4 and is meant to provide a short overview on the methodology underlying the data used in this report.

Specimen and donor cohort. Bone marrow (BM) aspirates were collected from 59 healthy human subjects (20-60 years old, male & female), as approved by the Ethics Committee for Human Subjects at the University of Heidelberg with written informed consent obtained from each donor. BM samples were harvested through puncture at the posterior iliac crest using a Yamshidi needle, with aspirations at 5 to 7 different levels of approximately 10 ml at each level. For validation experiments (metabolomics and single cell RNA-seq) ten additional subjects were collected for isolation and analysis of $CD34^+$ cells. Mononuclear cells derived from umbilical cord blood (UCB) were used as internal standards for the proteomics analyses. The acquisition and isolation have been described in complete detail in previous publications [Wagner et al., 2004; Wagner et al., 2009].

Cell isolation. BM aspirates were processed by FICOLL density fractionation to isolate mononuclear cells (MNCs). 5 different cell populations were isolated by staining CD34-APC, CD45-FITC, and CD14-PE and sorting on a FACSAria II flow cytometry cell sorted (BD Biosciences). Cells were checked for purity by mass spectrometry-based proteomics on tiny aliquots of the 5 purified cell populations, and stored at -80°C. For each cell population an individual internal standard was prepared from FACS sorted cells from umbilical cord blood and bone marrow, respectively. For single-cell RNA-seq of CD34 positive cells, single CD34 positive cells were sorted directly into 96-well plates containing 4.4 µl of lysis buffer per well. The lysis buffer contained 0.2% Triton X-100 (Sigma), RNase inhibitor (Takara), oligo-$dT_{30}VN$ primer (Sigma) according to Picelli et al. 2014 and 2.2 mM dNTP (Invitrogen). The lysed cells were frozen on dry ice cooled ethanol and kept at -80°C until further processing.

Sample generation of different cellular subsets. Frozen cells were suspended and lysed with lysis buffer containing protease inhibitors (Sigma P8340), RapiGest SF surfactant (Waters) and 200mM HEPES buffer, and buffered to pH=8 with NaOH. The sample was kept at 90°C (5min) and subjected to sonication for 20min afterwards. After centrifugation the supernatant was mixed with dithiothreitol (Biomol, 2mM) for reduction of di-sulfide bonds, followed by iodoacetamide (Merck, 5mM) for carbamidomethylation of cysteine side-chains. Proteins were then digested with Lys-C (Wako Chemicals) for 3 hours (37°C) and then trypsin (trypsin gold, Promega Corporation, 37°C) overnight.

Isolation, culture and sample preparation of human mesenchymal stem/stromal cells (MSCs). Contrary to the 5 described cell populations above, MSC from the respective human subjects were isolated from an in vitro culture. MNCs were seeded in a low FCS-MSC medium (medium detailed in publication, density ~ $1x10^6$ cells/cm^2) in tissue culture flasks coated with 10ng/ml fibronectin (Sigma). Culture medium was changed twice per

week. After initial colony formation (10-14 days, ensuring 80% confluence), cells were trypsinized, counted and re-seeded at 10^4 cells/cm^2 for further expansion. At passage 2, the cells were scratched off, washed, and cell material stored at -80°C as a pellet for proteomics analysis.

Quantification Labeling. Five samples of the same cell type and one internal standard were subjected to tandem mass tagging (TMT 6-plex, Thermo Fisher Scientific). After labeling, the reaction was quenched with hydroxylamine and the acid cleavable detergent RapiGest was cleaved with tri-fluoro-acetic acid; the lipophilic part of RapiGest reagent precipitates was pelleted by centrifugation. The supernatant was desalted on a C18-reversed phase material. By first running an LC-MS/MS analysis on a small aliquot of the six TMT-labeled samples, the mixing of the samples was adjusted to approximate an equimolar amount of each sample as much as possible. After mixing, any residual organic solvent was removed by a vacuum pump, and the pH of the total sample was adjusted to >10 with 25% ammonia (total volume = 50µl). The sample was subjected to separation on an Agilent 1260 infinity HPLC system (Waters XBridge C18; 3.5µm; 1 x 100 mm reversed phase column, 75µl/min) with buffers of 20mM ammonium formate at pH 10 and 100% acetonitrile. After the collection of 90 samples, organic solvent were again removed under vacuum and desalted and pooled into 18 fractions in total.

Liquid chromatography coupled to mass spectrometry (LC-MS). 50% of each of the 18 pooled fraction was analyzed on a Waters nanoAcquity UPLC system (C18, 5 µm, 180 µm x 20 mm trapping column (Waters), and nanoAcquity BEH C18, 1.7 µm, 75 µm x 200mm analytical column (Waters)) coupled to an Orbitrap Velos Pro (Thermo Fisher Scientific). Mobile phases A and B were composed of 0.1% formic acid and acetonitrile, respectively, changing from 3% to 85% mobile phase B within 120 minutes (flow rate=300nL/min). The spray voltage was set to 2.2kV. The MS operated in positive ion mode, with full-scan MS spectra operating in a range of 350-1500m/z (resolution = 30,000), and the 10 most intense precursor ions being subjected to HCD-fragmentation (normalized collusion energy = 40). Singly charged ions were excluded, and the MS/MS-resolution set to 7,500.

Single-cell RNA-seq of HPCs. Sequencing libraries from 192 single CD34$^+$ cells per donor were generated based on the smart-seq2 protocol of Picelli et al. 2014 and the tagmentation procedure of Hennig et al. 2018 with slight modifications. Briefly, single CD34 positive cells were FACS sorted directly into 96-well plates containing 4.4 µl of lysis buffer per well as described above. The lysates were incubated for 3 min at 72°C and kept on ice while adding reverse transcription mix. Our reverse transcription (RT) mix had a final volume of 6.35 µl to have the final reaction volume of 10 µl and in contrast to the smart-seq2 protocol a final

concentration of 10 mM $MgCl_2$. Twenty-two cycles were applied for the PCR. The subsequent washing procedure was optimized. 25 µl nuclease-free water and 30 µl of SPRIselect (1:0.6 ratio) were added and no ethanol wash was performed. After incubation, removal of supernatant, and drying 13 µl nuclease-free water was applied for elution and 11 µl were taken for a second purification step. 40 µl nuclease-free water and 25 µl of SPRIselect (1:0.5 ratio) were added and after incubation, removal of supernatant, and drying 13 µl nuclease-free water was applied for elution. 1.25 µl of the supernatant were used for tagmentation, which was performed as previously reported. Briefly, Tn5 was mixed with equal amounts of Tn5ME-A/Tn5MErev and Tn5ME-B/Tn5MErev and incubated at 23°C for 30 min. Loaded Tn5 and sample were incubated for 55°C for 3 min in 10 mM Tris-HCl pH 7.5, 10 mM MgCl2 and 25% dimethylformamide. The mixture was cooled to 10°C and the reaction was stopped with 0.2% SDS for 5 min. at room temperature. KAPA HiFi HotStart ReadyMix was used for PCR amplification. One µl of each of the 192 samples of one donor was combined and bead purification using 0.8 vol. of SPRIselect was performed including two ethanol washing steps. The elution volume was 50 µl. Sequencing was performed on an Illumina NextSeq 500 with 75bp single-end reads.

RNA sequencing of the total HPC population. RNA was extracted with trizol (Invitrogen) using a linear acrylamide carrier. RNA was then treated with DNase I and purified using Agencourt RNAClean XP beads. RNA quality and concentration were assessed by using an RNA 6000 bioanalyzer pico kit (Agilent). Samples with concentrations less than 30 pg/µl and/or an RNA integrity number (RIN) less than 6 were excluded from further analysis. A cDNA library was produced using the Smart-Seq2 protocol. Sequencing was performed on an Illumina HiSeq4000 with 75 bp paired-end reads with the aim to achieve coverage of 25 million reads per sample.

Appendix C

Supplementary Figures

Related to Chapter 2:

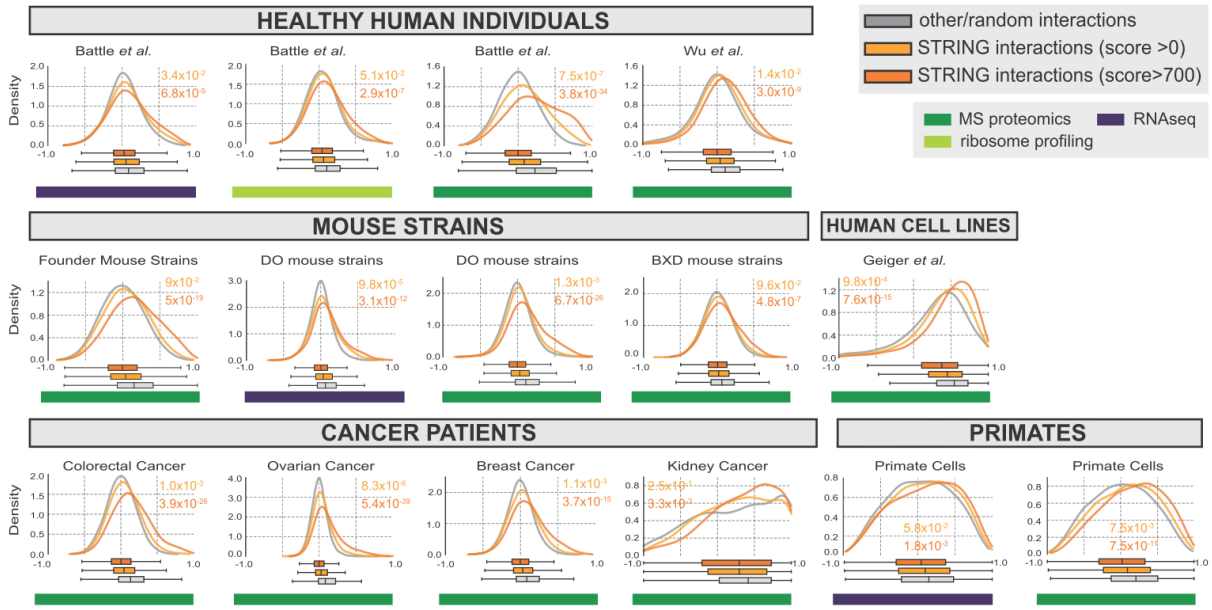


Figure S2.1. Recovery of known STRING interactions in different published datasets (related to Figure 2.1). For all datasets as considered for the ROC-analysis, the distribution of Pearson correlation coefficients for protein-protein pairs with STRING interaction score > 700 (combined score, orange), known interaction (combined score >0, yellow) and random protein-protein pairs (grey) is shown. The Mann-Whitney *U*-test was used to assess significance of the respective shifts (indicated in colored p-values).

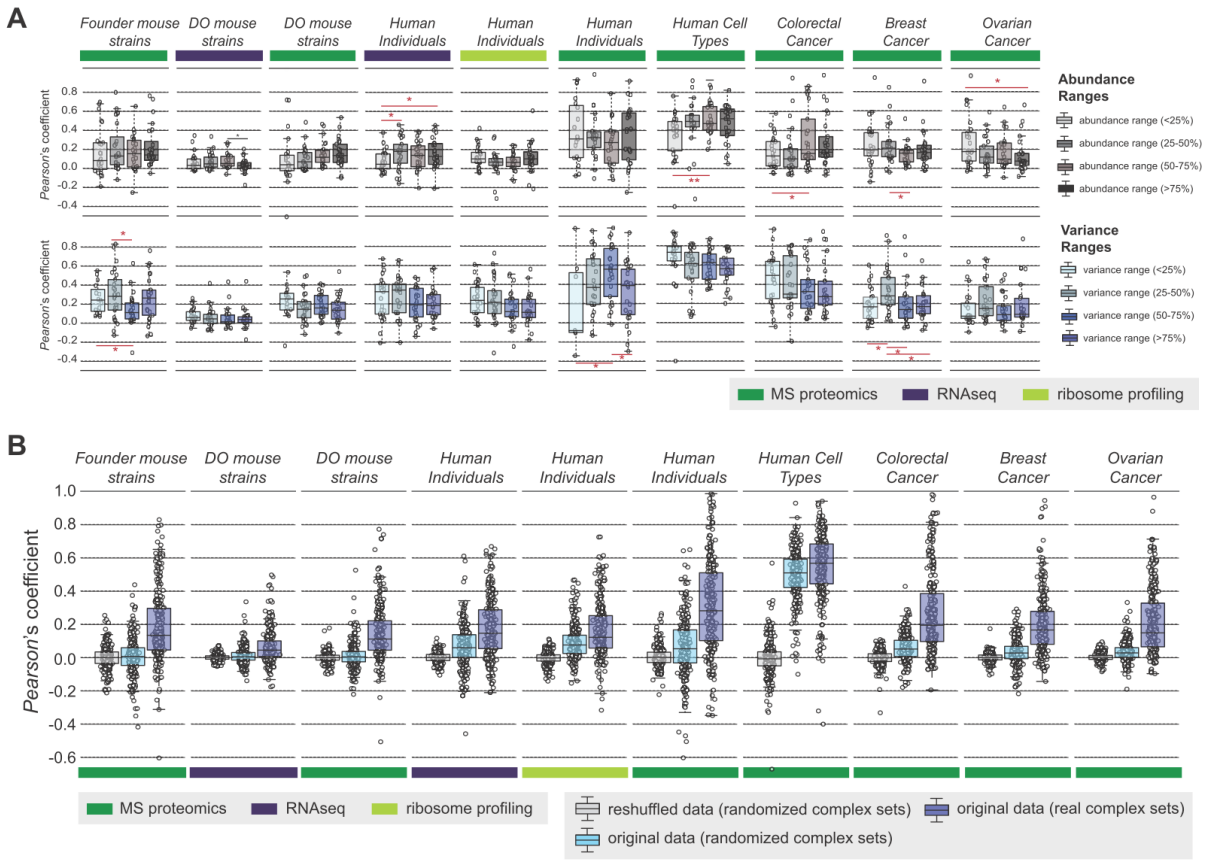


Figure S2.2. Technical bias in abundance assessment and complex correlations (related to Figure 2.2). (a) For the datasets as indicated by their respective labeling, it is asssessed whether the complex median correlation (Pearson) is biased by the abundance of the respective complex (first row, grey shading) or by the complex variance (second row, blue shading). For comparability, abundances and variances are rank-sorted and further split into 25%-bins, the median correlation is then monitored in each bin. While they were significant differences between some bins (Wilcoxon-test: * p<0.1, ** p<0.05, *** p<0.01), no general trend could be observed and also those significancies could not be recovered consistently across datasets. (b) For the same datasets as above, median correlation values (Pearson) were monitored for randomly assembled complexes (decoy complexes) from permuted data (reshuffled data) (grey, first boxplot), decoy complexes from original data (light-blue), and real complex sets from original data (purple).

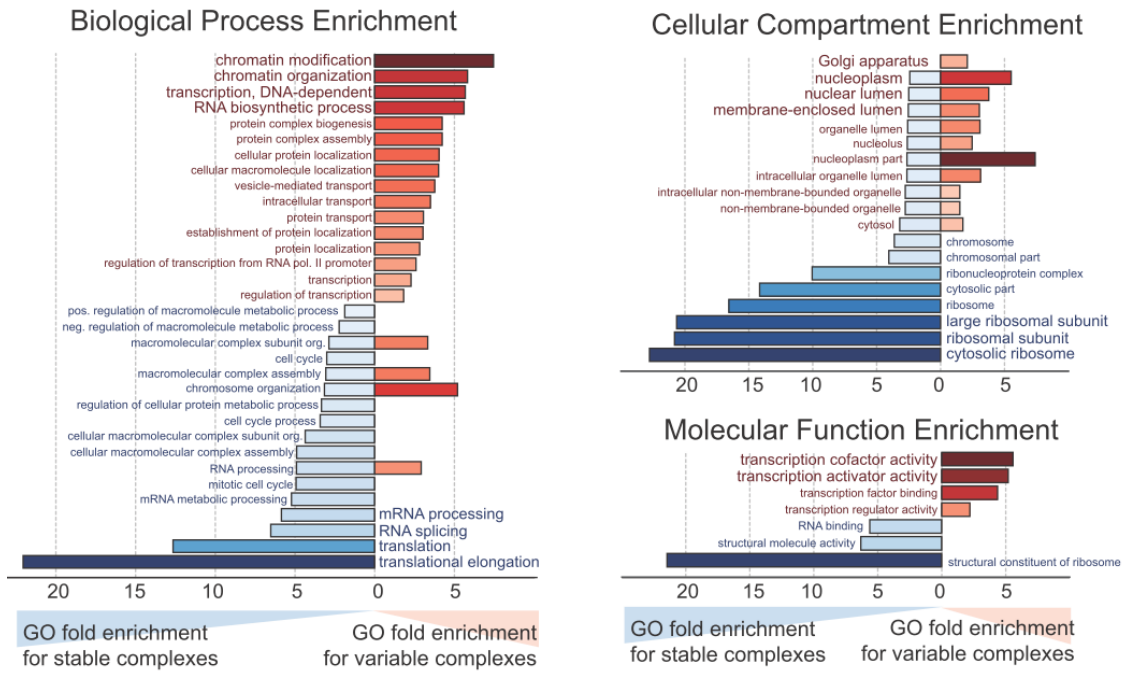


Figure S2.3. GO-enrichment for variable and stable modules (related to Figure 2.3). GO-enrichment analysis in 3 categories (Biological Processes, Cellular Compartment and Molecular Function) delineating the functional differences between stable and variable complexes as recovered from Figure 2.2a. The x-axis represent fold-enrichment for stable complexes to the left (blue), whereas to the right fold-enrichments are shown for variable complexes (red). Color opacity correlates with the fold-changes. Only GO-terms with an FDR<1% (Benjamini-Hochberg) are shown.

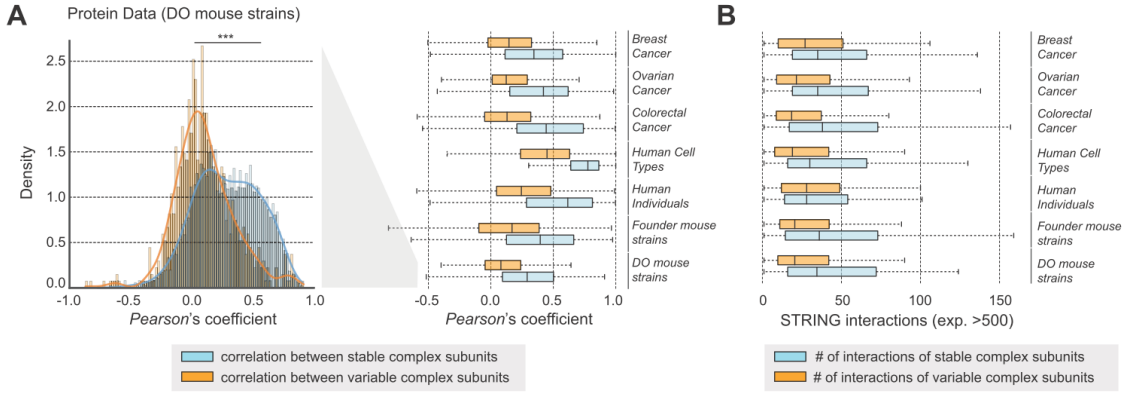


Figure S2.4. Comparing features of stable and variable module components (related to Figure 2.4). (a) Stable complex members (defined in Methods) tend to be consistently more correlated with each other than variable components between each other within the same complex (Wilcoxon-test: *** p<0.01). The trend was consistent for all shown datasets (boxplots on the right-hand side); the data for the proteomics data for the DO mouse strains [Chick et al., 2016] is enhanced on the left-hand side as a density plot. (b) The number of STRING interactions (very confident interactions, experimental score >500) tends to be consistently higher for stable complex members than observed for variable complex members. Again this is a consistent trend in all given datasets.

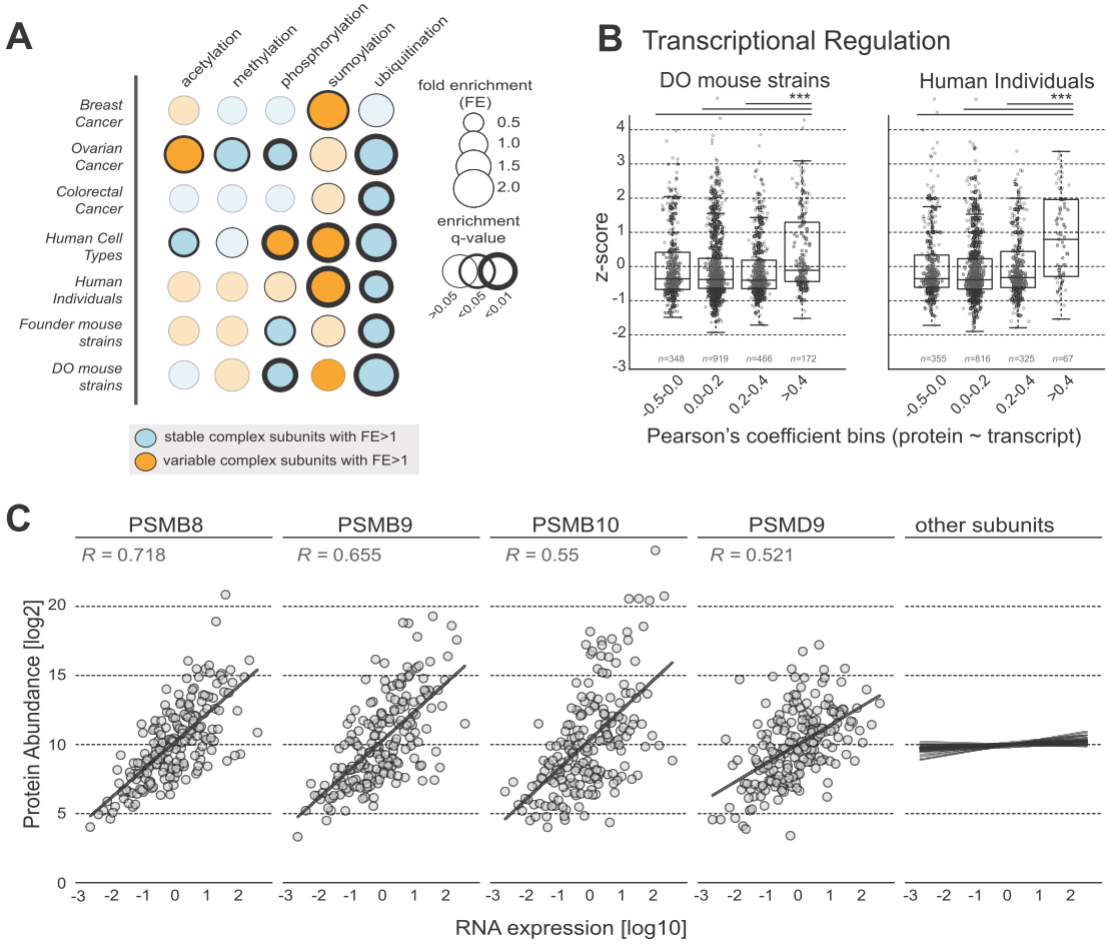


Figure S2.5. Stable complex components tend to get ubiquitinated, variable components to be transcriptionally regulated (related to Figure 2.4). (a) Matrix delineating enrichment of stable/variable complex members in protein sets with certain modification sites. The size of the circle indicates the fold-enrichment; the edge thickness and opacity relate to the significance of the given enrichment. For example, stable complex members are consistently found to be enriched in proteins with more ubiquitination target sites (mean p-value=1.65×10^{-2}, Fisher Exact Test). (b) Protein-transcript correlation pairs are binned on the x-axis; the y-axis shows the z-score distribution in each bin indicative of the variability of subunits per complex. Variable subunits (z-score > 1.5) were found to be enriched in thelast bin (protein-transcript $R^2>0.4$, *p*-value (ANOVA)=1.2×10^{-7} for DO mouse strains [Chick et al., 2016], *p*-value (ANOVA)=3.77×10^{-12} for Human Individuals [Battle et al., 2015]). (c) DO mouse strains: example of high protein-transcript correlations for subunits of the immuno-proteasome, PSMB8/9 and 10, as well as the variable subunit PSMD9. On the right, trend-lines for protein-transcript correlations of all other subunits of the proteasome complex are shown.

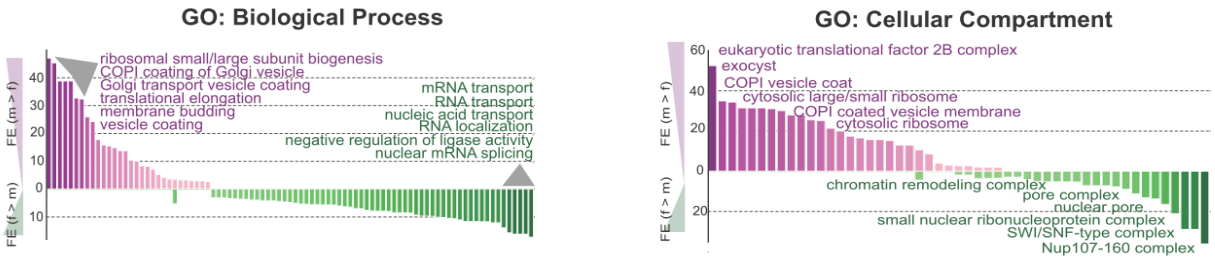


Figure S2.6. (related to Figure 2.5). GO-enrichment analysis (biological processes and cellular compartments, see Methods) for complexes that are either more abundant in male (purple) or female (green) mice as an entire structure.

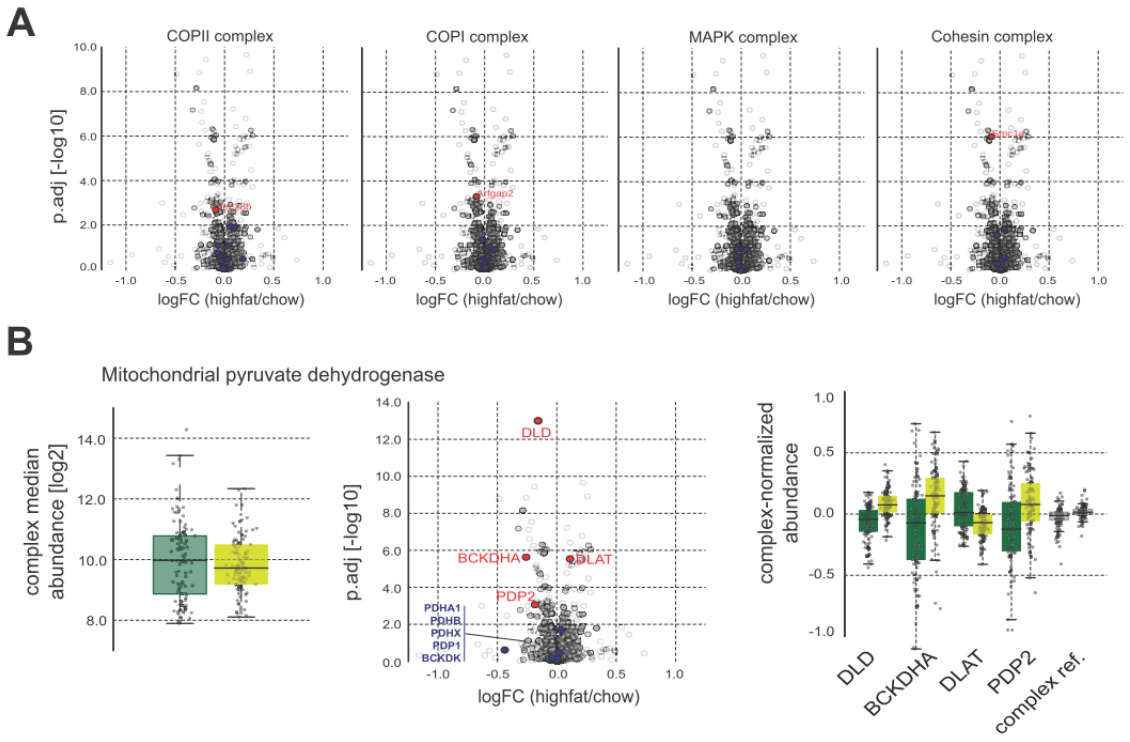


Figure S2.7. Stoichiometry effects are sex- and diet-specific (related to Figure 2.5). (a) Sex-specific stoichiometry of complexes is not influenced by diet differences. Volcano plots are equivalent to Figure 2.5b, but illustrate diet differences in stoichiometry. (b) Diet-specific stoichiometry of the mitochondrial pyruvate dehydrogenase: (left) the overall complex median abundance is not affected, (center) volcano plot highlighting the complex-specific fold-changes of particular subunits of the complex, (right) complex-normalized abundances with enhanced differentially expressed proteins (high-fat = dark green, chow = light green).

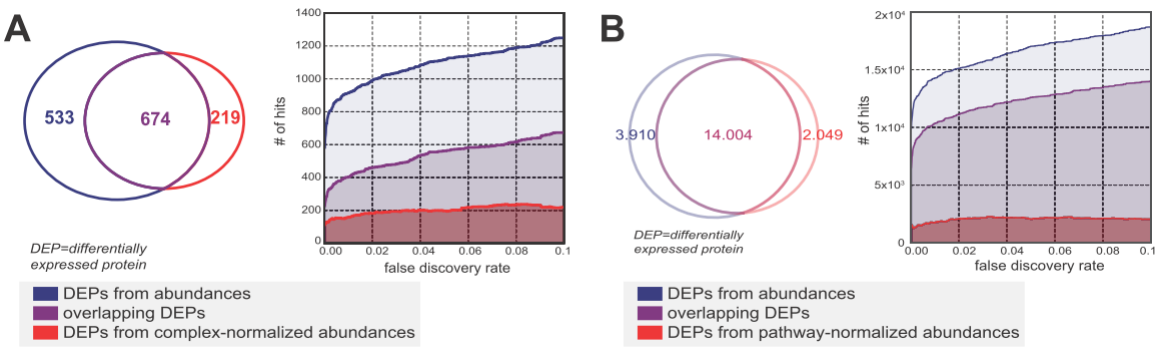


Figure S2.8. Leveraging module normalization for discovery of differentially expressed proteins in context of modules (related to Figure 2.5). (a) (*left*) Number of DEP (differentially expressed protein) hits from the LIMMA analysis (q-value<0.1) between male and female DO mouse strains, using original data as input (blue, abundances), and complex-normalized abundances (red). (*right*) The line plot shows the number of hits as a function of the false discovery rate. (b) (*left*) Number of DEP (differentially expressed protein) hits from the LIMMA analysis (q-value< 0.1) between male and female DO mouse strains, using original data as input (blue, abundances), and pathway-normalized abundances (red). (*right*) The line plot shows the number of hits as a function of the false discovery rate. Note that many proteins can be part of several pathways at the same time, hence the number of hits is much higher than in case of the complex analysis where latter are much more strictly defined.

Related to Chapter 3:

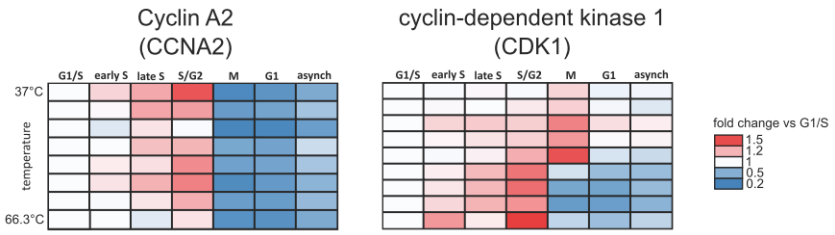


Figure S3.1. Checking data quality with known cell cycle markers (related to Figure 3.2). Heat map for CCNA2 and CDK1, from which abundance and stability scores were calculated. The vertical direction indicates the increase in temperature and the horizontal direction progress in cell cycle.

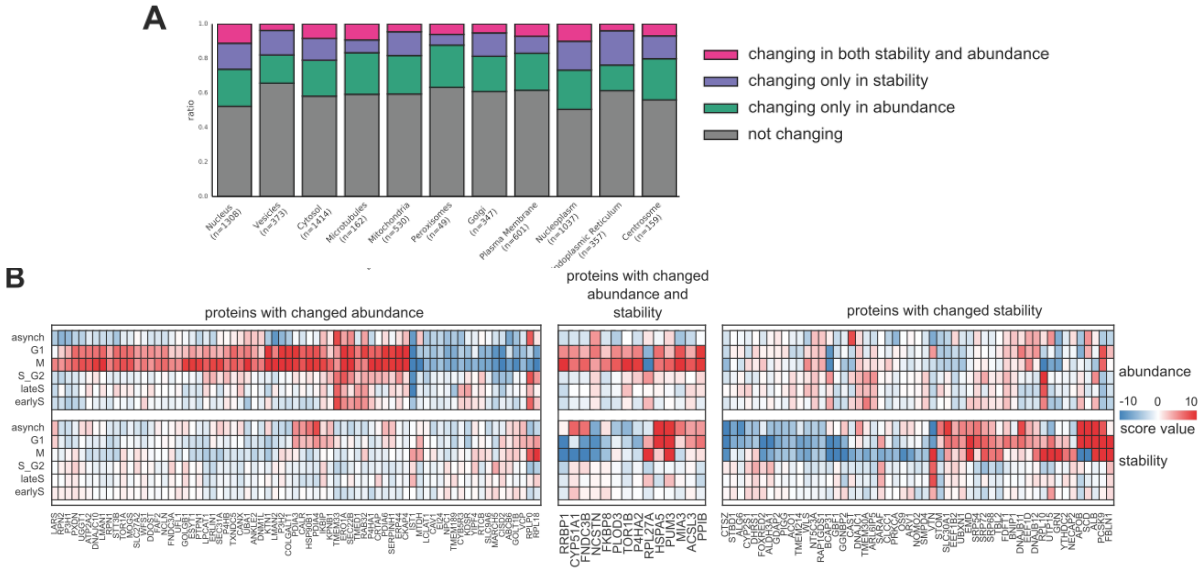


Figure S3.2. Analysis of cell cycle-dependent stability effects on organelle-specific proteins (related to Figure 3.3). (a) Bar plots displaying the fraction of organelle-specific proteins (GO-based) affected in either stability or abundance, or both. (b) Heatmap depicting ER-associated/localized proteins that are significantly affected in abundance (*left*), stability (*right*), or both (*middle*) in at least one cell cycle stage.

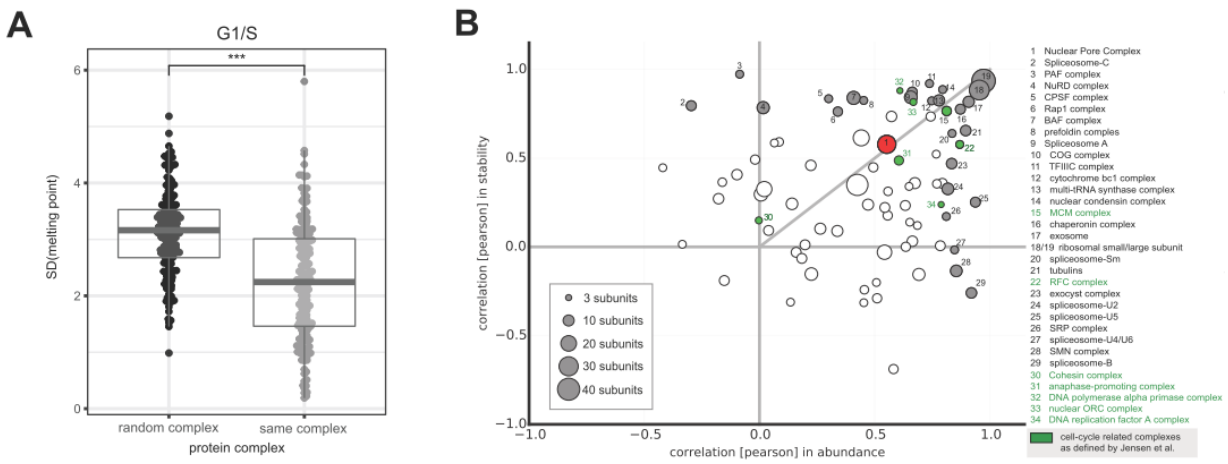


Figure S3.3. Overview on complex co-stability and co-abundance (related to Figure 3.7). (a) Melting temperatures of proteins associated to the same protein complex in G1/S, compared to random complex assignment. (b) Scatter plot showing complex median co-abundance (across members of given complex, *x*-axis), versus complex median co-stability (*y*-axis). Complexes related to the cell cycle are labeled in green [Jensen et al., 2006].

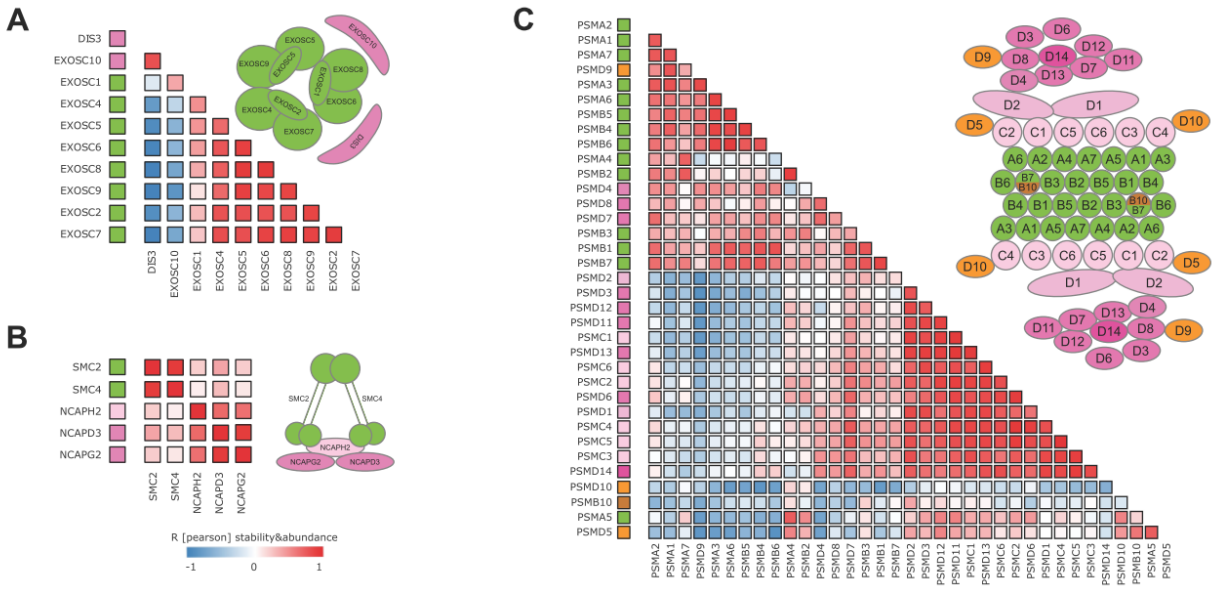


Figure S3.4. Differential stability pattern of moonlighting subunits of complexes (related to Figure 3.7). (a) Exosome structure with indicated catalytic subunits, DIS3 and EXOSC10 (light red). Correlation matrix calculated from concatanated abundance- stability profiles of all members of the exosome complex. (b) Correlation matrix of condensin II-subunits based on their concatanated abundance-stability profiles. The colours on the left indicate their association with a specific sub-structure of condensin II, as coloured in respective cartoon. (c) Correlation matrix of proteasome subunits based on their concatanated abundance-stability profiles. Colours on the left-hand side indicate sub-structure of the proteasome as indicated in the respective cartoon.

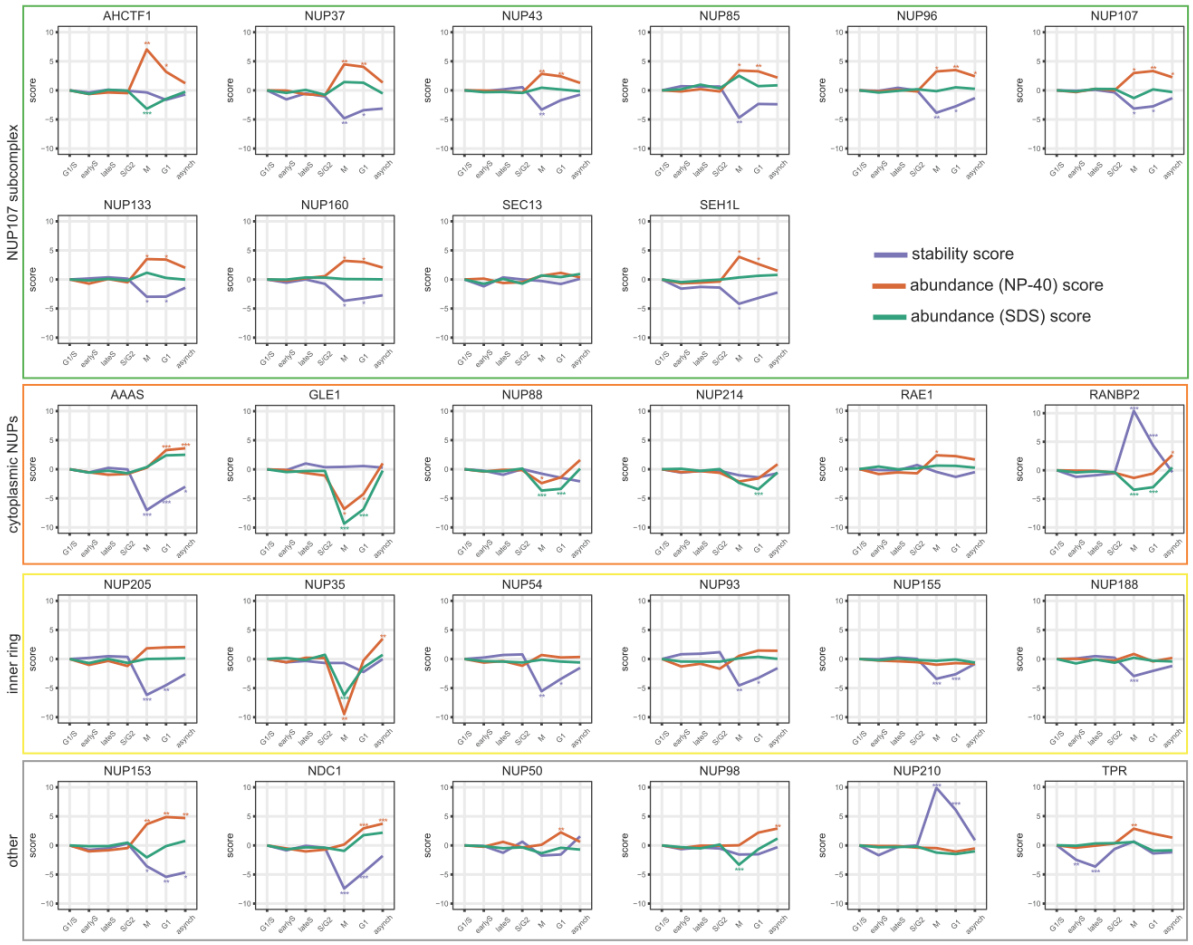


Figure S3.5. Detailed line plots for all measured subunits of the NPC-complex (related to Figure 3.7). Stability is shown as a purple line, abundance (NP-40, orange) and abundance (SDS, green).

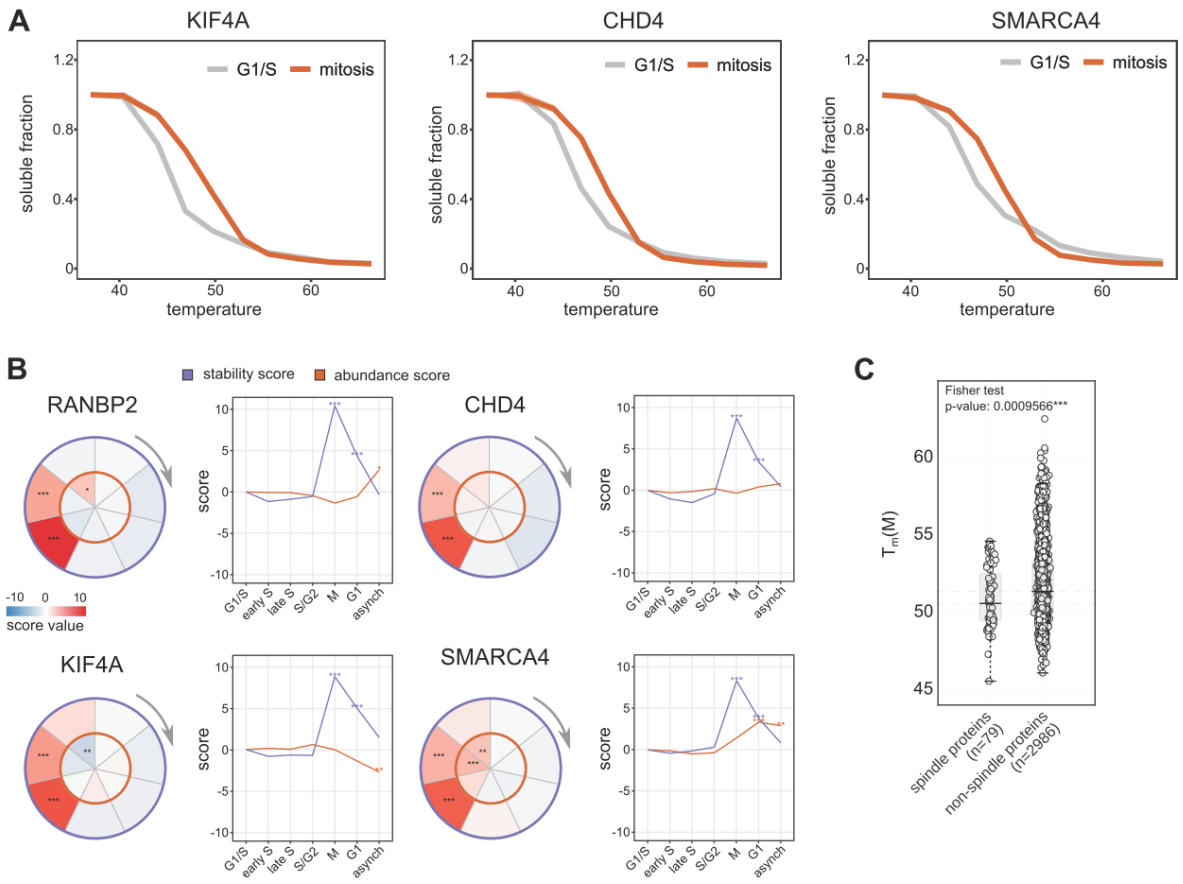


Figure S3.6. Detailed melting patterns of spindle-associated proteins (related to Figure 3.8). (a) Melting curves for CHD4, KIF4A, and SMARCA4 (data based on 3 replicates). (b) Circle- and line plots for RANBP2, CHD4, KIF4, SMARCA4 (purple: stability, orange: abundance). (c) Boxplot comparing melting temperatures in mitosis of spindle proteins annotated from Sauer et al. (*left*) against non-spindle proteins (*right*).

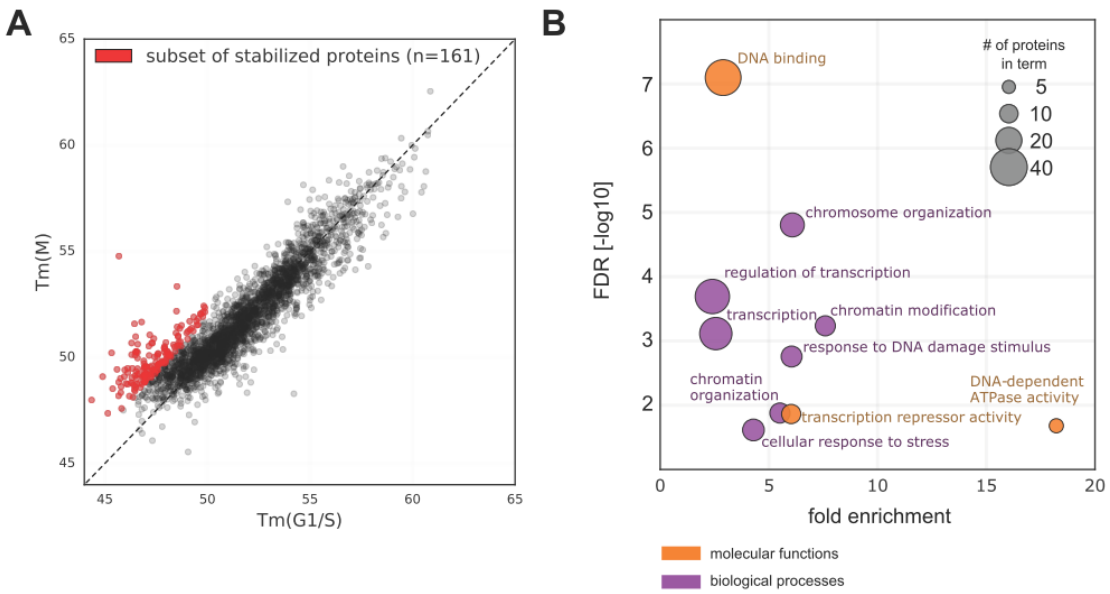


Figure S3.7. Selection of stabilized proteins, GO-analysis for stabilized set of proteins (related to Figure 3.9). (a) Melting temperatures for G1/S versus M. Proteins in red were selected for further GO-analysis shown in (b). (b) GO for stabilized set (a). GO analysis using DAVID (https://david-d.ncifcrf.gov) was conducted on the protein set that is stabilized in mitosis relative to the G1/S reference point. For filtering we considered proteins that were significantly stabilized ($p<0.1$), and with their melting point in mitosis below 55°C and in G1/S below 50°C. The resulting 161 proteins were primarily found to beinvolved in DNA-binding activity and chromatin-associated processes. The scatter illustrates the fold-enrichment against the respective FDR for terms derived from the broad categories 'molecular function' and 'biological processes'. The size of each bubble relates to the number of proteins identified for each term.

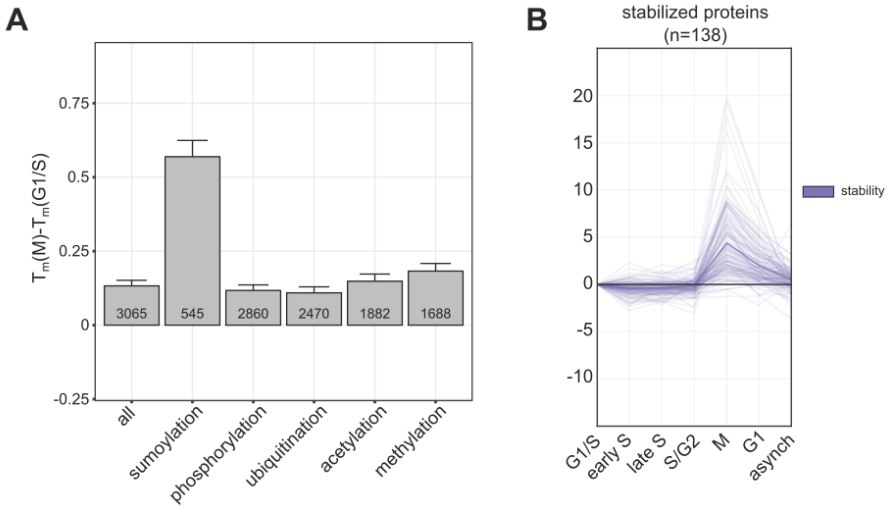


Figure S3.8. Stabilization of sumoylated proteins, and leak-over (related to Figure 3.9). (a) Difference in melting points (G1/S versus M) of all proteins known to be modified with specific PTMs (specified in Materials & Methods). (b) Leak-over of stabilized proteins. Mitotically stabilized protein sert used for Figure S3.7b (n=161) was overlapped with 2D-TPP dataset. The graphs show the line plots of the 138 overlapping proteins, and the mean stability pattern as a thicker line. The set of stabilized proteins tends to remain stabilized in G1 as well (leak-over).

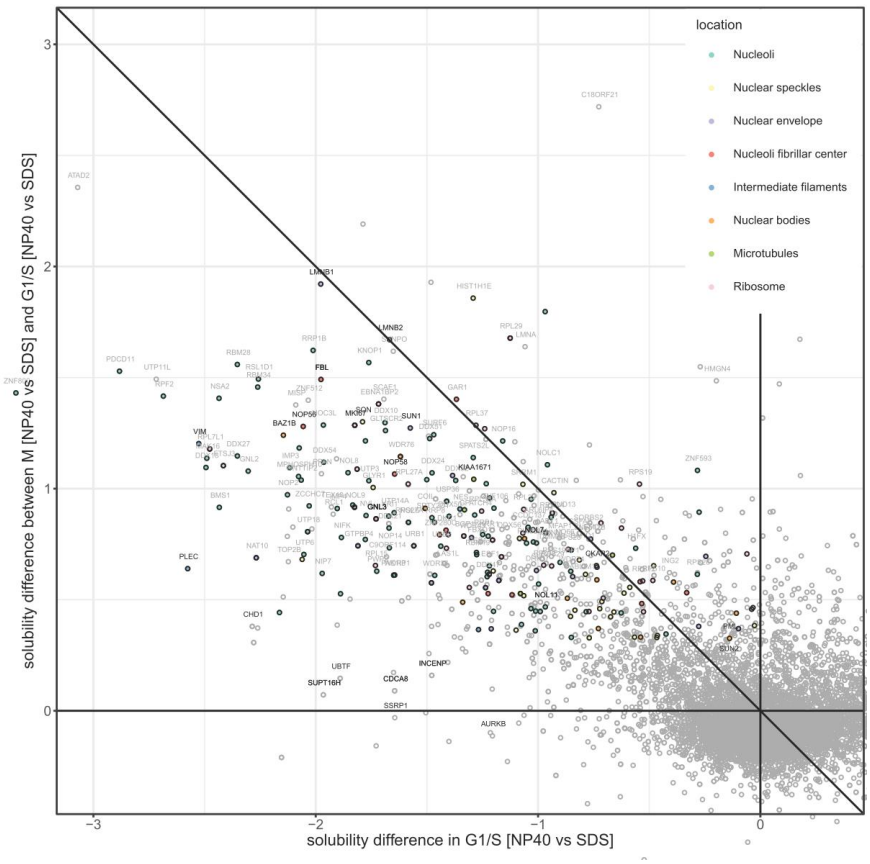


Figure S3.9. Detailed solubility track of proteins (related to Figure 3.10). Scatter plot comparing the solubility of proteins in G1/S (*x*-axis) and the relative change in solubility of proteins in mitosis versus G1/S (*y*-axis). Proteins with negative x-axis values and positive *y*-axis values are insoluble in G1/S and become more soluble in mitosis.

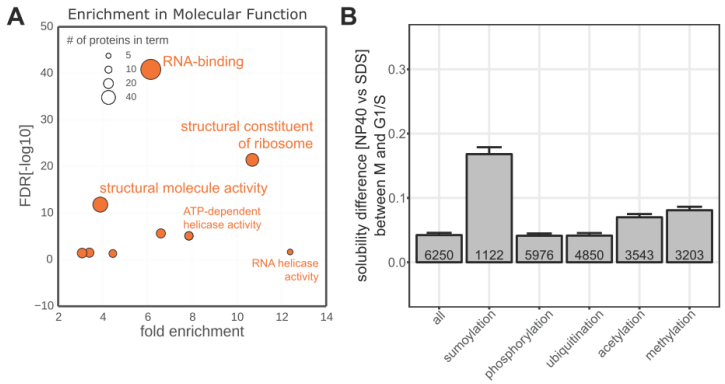


Figure S3.10. Functions and modifications of proteins with prominent solubility transition during the cell cycle (related to Figure 3.11). (a) GO-analysis using DAVID (https://david-d.ncicrf.gov) was performed on the protein set that is significantly more soluble in mitosis versus G1/S. x-axis: fold-enrichment, y-axis: FDR for terms derived from broad category 'molecular function'. Sizes of circles correspond to the number of proteins identified for each term. (b) Solubility differences in mitosis versus G1/S of all proteins that are known to have target sites for specific PTMs.

Related to Chapter 4:

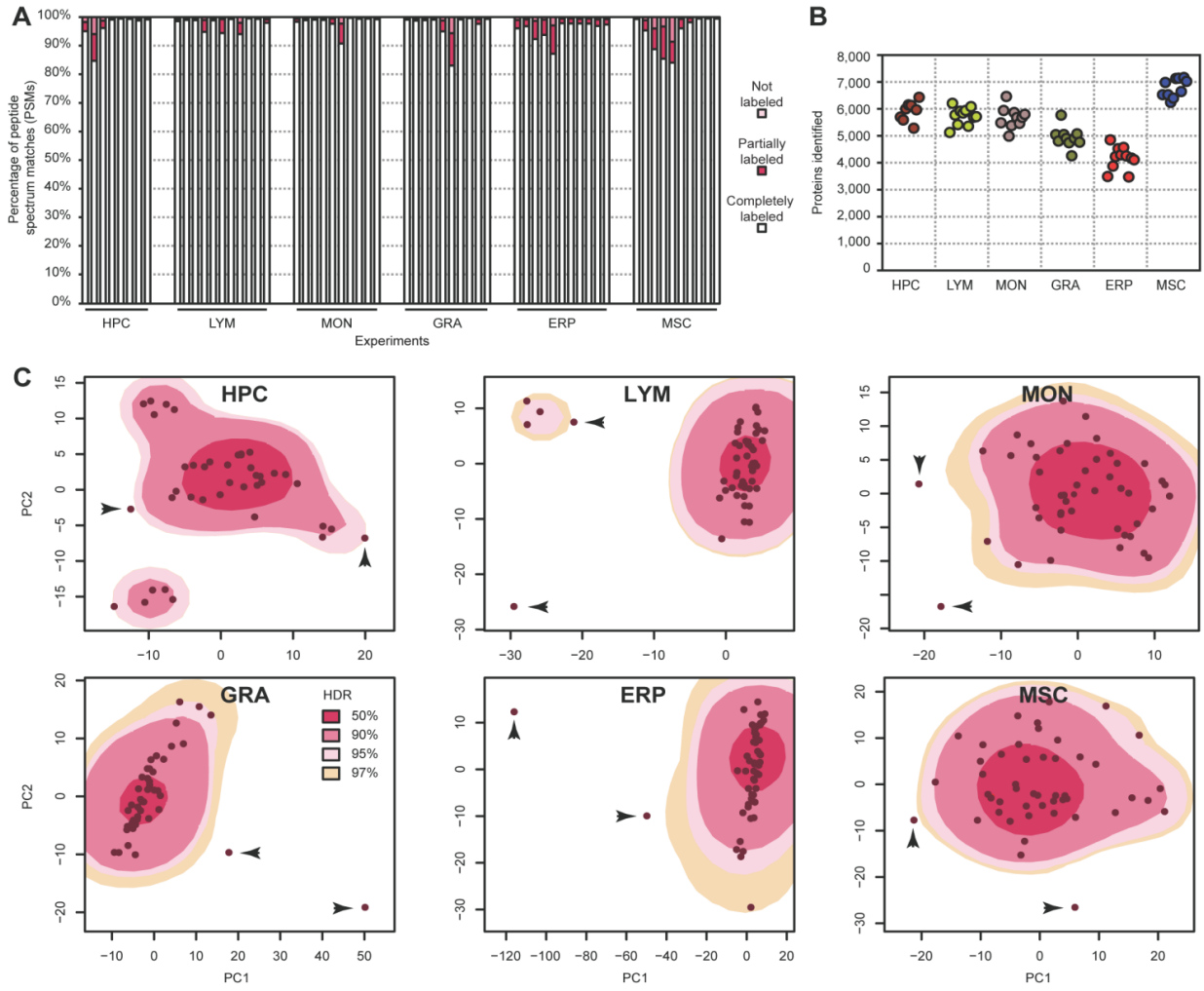


Figure S4.1. Labeling efficiency, protein number and outlier analysis as part of the quality control analysis (related to Figure 4.2). (a) Completely labeled, partially labeled and unlabeled peptide spectrum matches (PSMs) expressed as a percentage of all PSMs are stacked. (b) Protein identification number for all experiments. Each point indicates the number of proteins identified in a TMT-6-plex experiment. (c) A principal component analysis (PCA) was performed on the log2-transformed data and the first two principle components (PC1 and PC2) were plotted against each other. Highest density regions (HDR plot) of 50, 90, 95 and 97 percent probability were visualized based on Hyndman and samples with >97% probability were defined as outliers to be discarded. The discarded samples are marked by arrows.

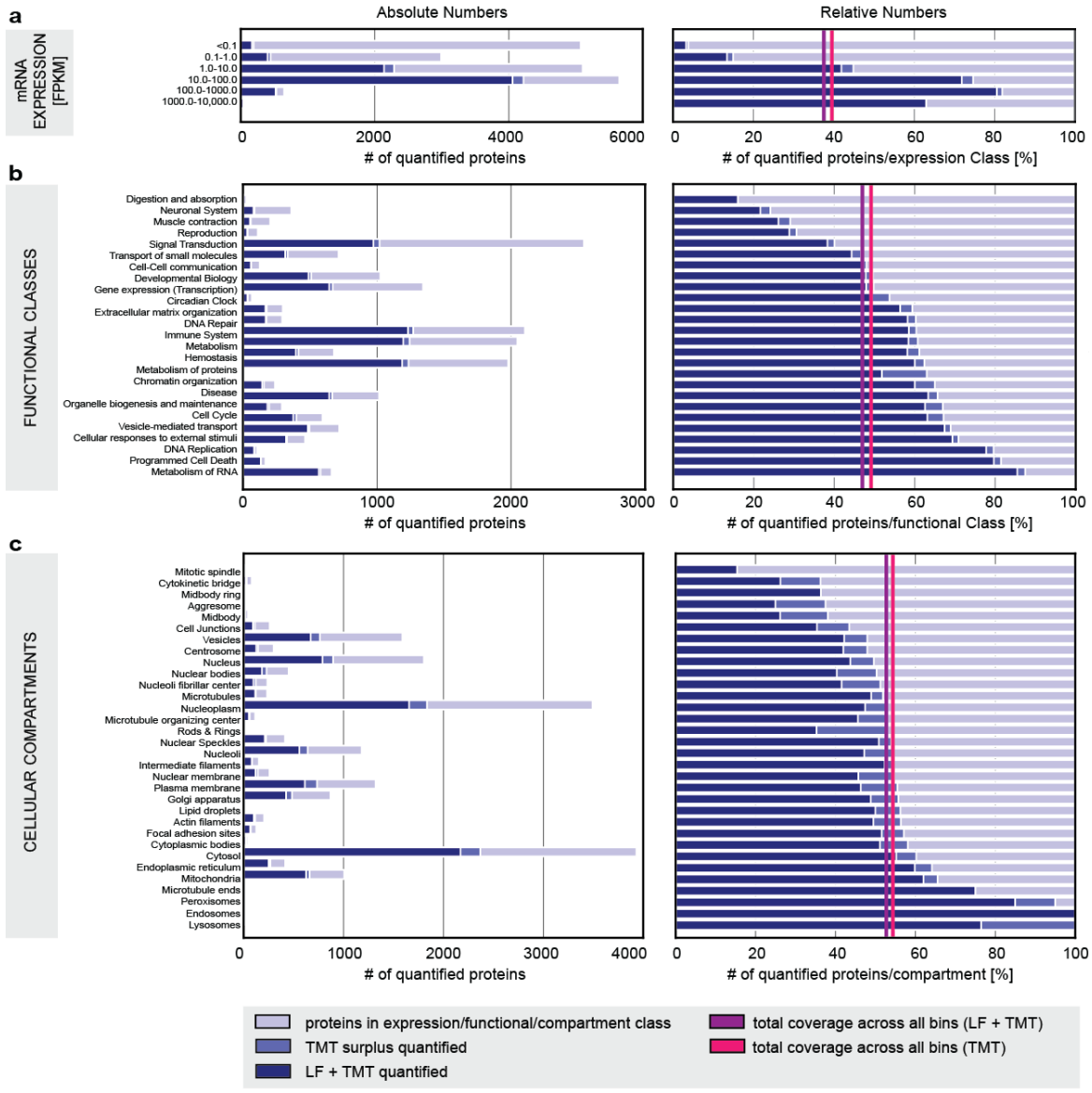


Figure S4.2. Overview of the mRNA expression levels, functional classes, as well as cellular compartments of all quantified proteins in comparison with the total human proteome (related to Figure 4.2). (a) Label-free (LF) and TMT quantified proteins are grouped by their mRNA expression levels and their absolute and relative number are visualized as fraction of the total human proteome. The expression classes are defined according to transcript levels measured in the bone marrow tissue by Uhlen et al. (2015). (b + c) The quantified proteins are grouped by function (b) and cellular compartment (c). The functional classes are defined by the highest hierarchical level of the Reactome pathway database (http://www.reactome.org) and the compartment classes are defined according to the Human Protein Atlas [Uhlen et al., 2015]. The red and violet vertical lines illustrate the total coverages of the quantified proteins across all bins in the presented class.

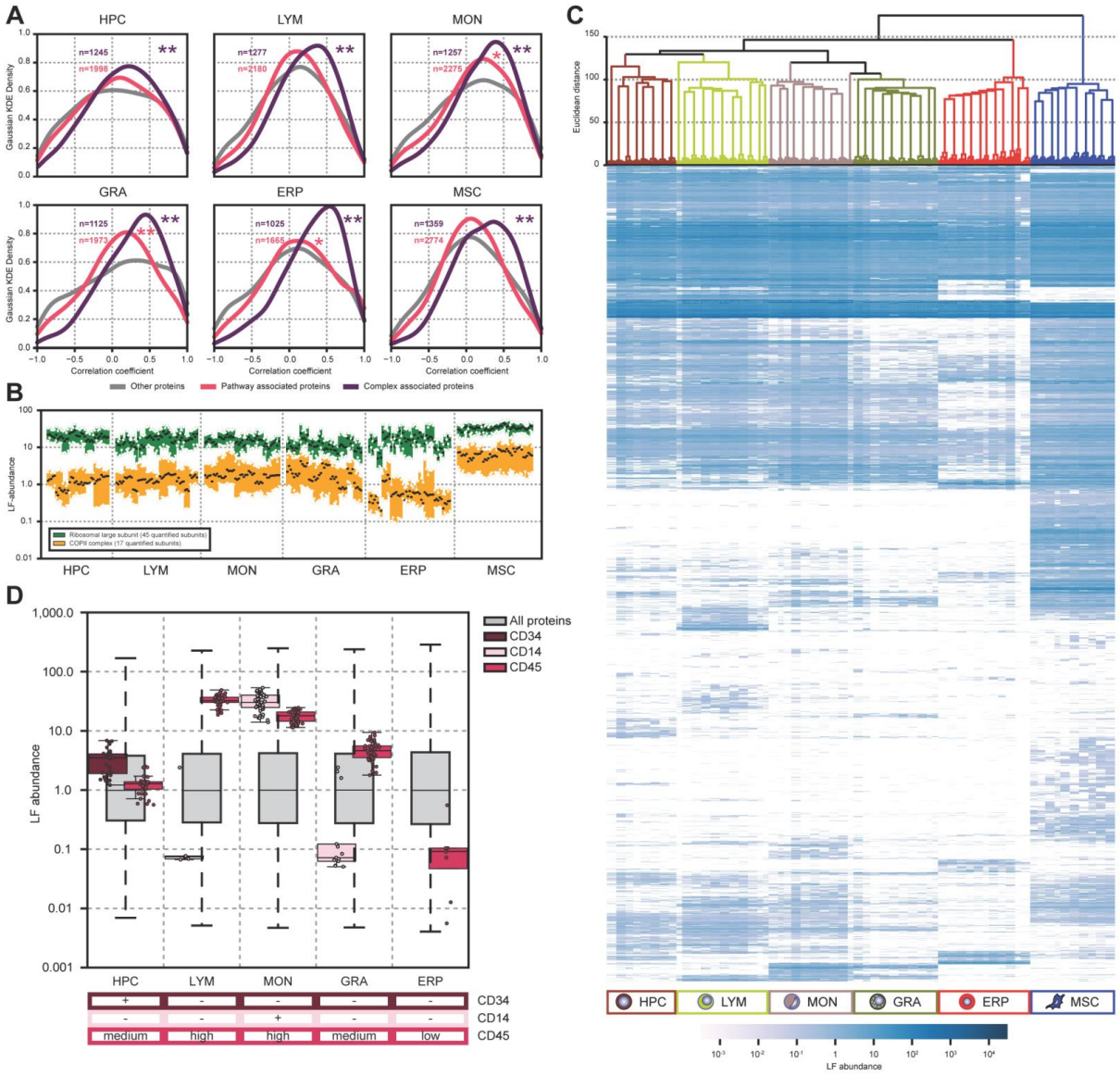


Figure S4.3. Comparison of the proteomes of the six investigated cell populations (related to Figure 4.2). (a) Density plots for each cell population visualizing the co-variation of proteins within a complex or pathway. Proteins within a complex (purple) have a clear shift towards higher correlation compared to proteins that are not annotated to a complex or pathway (other proteins). *P*-values were calculated using random sampling of the distributions and calculating the Kolmogorov-Smirnov test statistic 1,000 times to then estimate the median *p*-value. Stars indicate the significance value with (*) equaling a p-value < 0.05 and (**) a p-value < 0.01. The total number of proteins associated to complexes, and pathways in each cell population, is presented in each subplot in the respective colour (n). (b) The average label-free (LF) abundance of the members of the ribosomal large subunit (green) and the COPII complex are plotted for all samples. (c) Hierarchical clustering of all 270 samples quantified by label-free (LF) quantification. The LF abundance (blue colour scale) of all proteins from all samples was used for clustering using Euclidean distances. White colouring defines absence of a quantification event. The samples cluster according to their cell population. (d) Visualization of the relative label-free (LF) abundance of surface markers used in FACS for sorting the cell populations. The dots represent the results of the individual samples, if the protein was quantified. Protein abundance (top) correlates well with gates (bottom) used for FACS with CD34 being specific for HPC, LYM and MON being $CD45^{+}$ (PTPRC), GRA being $CD45^{med}$, ERP being $CD45^{-}$, and MON being gated for $CD14^{+}$. For comparison the results of all quantified proteins are represented as a grey box plot for each individual cell population.

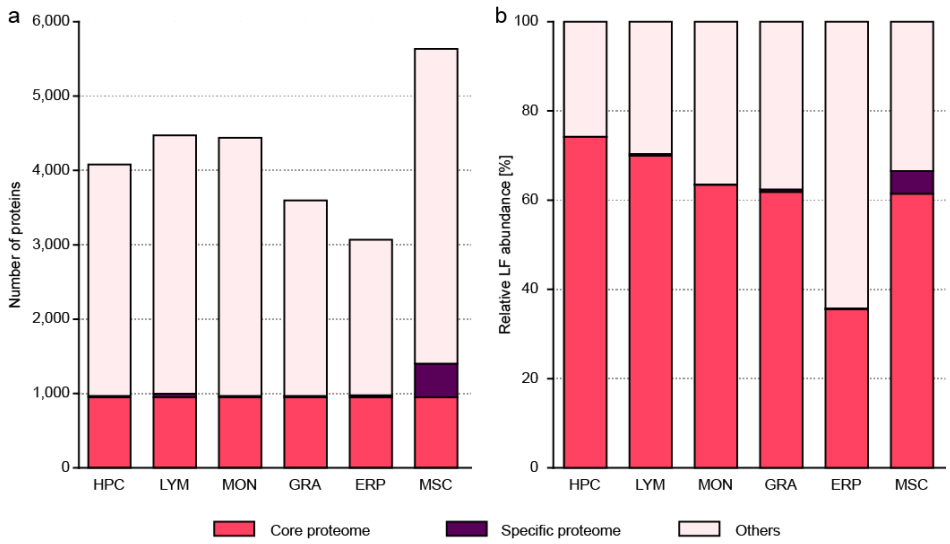


Figure S4.4. Characterization of the number and relative abundance of the core and specific proteome in the individual cell populations (related to Figure 4.2). (a) Overview on the number of proteins quantified by label-free quantification in each cell population. Colours indicate the fraction of proteins corresponding to the core proteome (≥ 85% sample coverage in all cell populations) or specific proteome (≥ 85% sample coverage in a single cell population and < 15% sample coverage in all other cell populations), and remainder (others). (b) Overview on the relative label-free (LF) abundances of the corresponding proteome categories in each cell population relative to the sum of the abundances of all proteins.

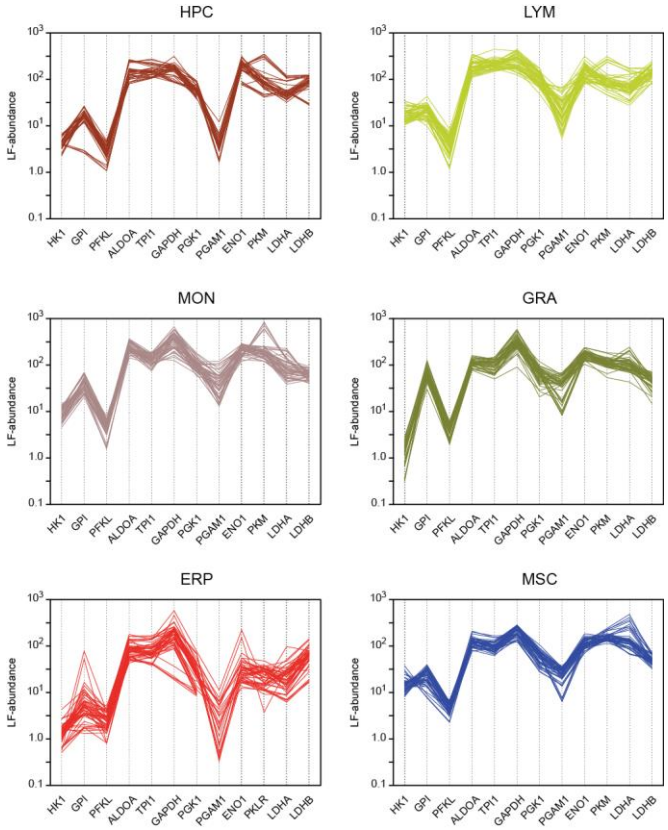


Figure S4.5. The stoichiometry of proteins of the glycolytic pathway are maintained across donors in the different cell populations. Each line connects the label-free (LF) abundances of the major proteins of glycolysis for one individual donor. The glycolytic enzymes are annotated at the x-axis and each subplot represent the results for one of the six different cell populations.

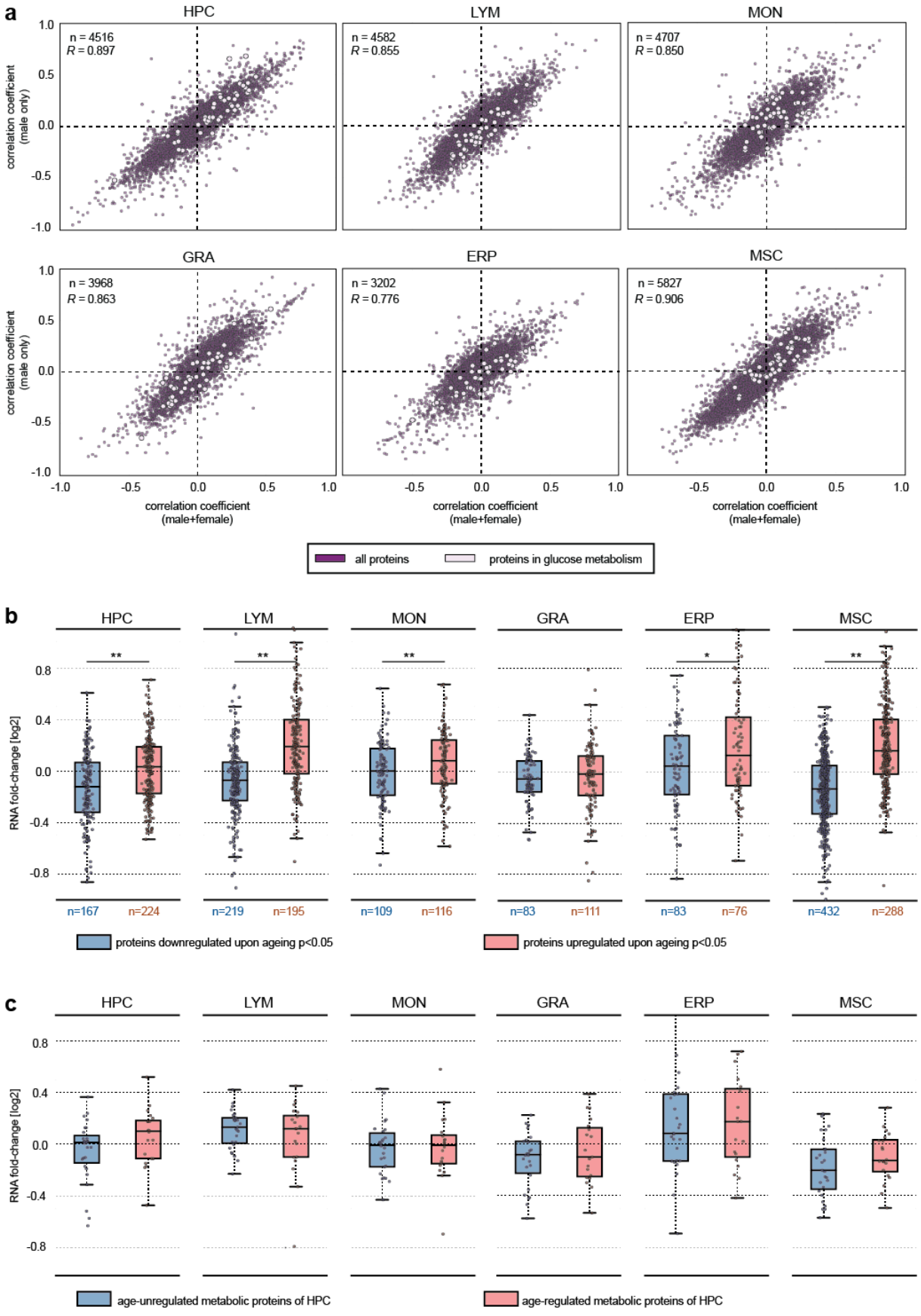
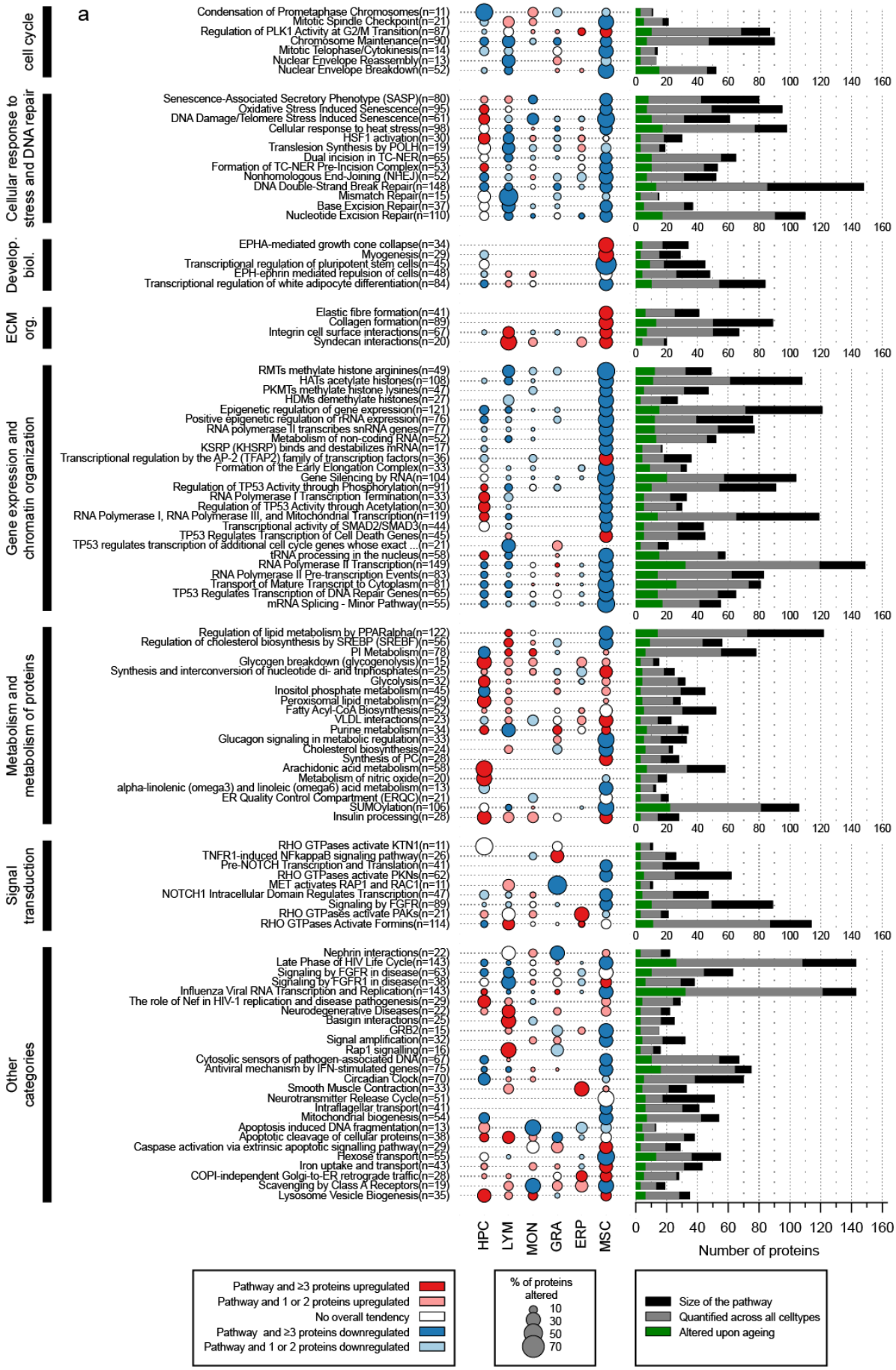

a
HPC
LYM
MON
GRA
ERP
MSC
n = 4516
R = 0.897
n = 4582
R = 0.855
n = 4707
R = 0.850
n = 3968
R = 0.863
n = 3202
R = 0.776
n = 5827
R = 0.906
correlation coefficient (male only)
correlation coefficient (male+female)
all proteins
proteins in glucose metabolism
b
RNA fold-change [log2]
n=167
n=224
n=219
n=195
n=109
n=116
n=83
n=111
n=83
n=76
n=432
n=288
proteins downregulated upon ageing p<0.05
proteins upregulated upon ageing p<0.05
c
RNA fold-change [log2]
age-unregulated metabolic proteins of HPC
age-regulated metabolic proteins of HPC

Figure S4.6. Overview on gender effects in ageing, and transcript expression changes upon ageing. (a) The Spearman correlation coefficients for protein changes upon ageing are calculated for all samples and for only male samples per each cell population. The correlation coefficients of protein changes of all samples (x-axis) are plotted against the male only correlation coefficients (y-axis). Proteins that are related to the glucose metabolism are highlighted as grey dots, and all other (purple) dots represent all other quantified proteins (n is given in the upper left corner, respectively). (b) The box-plots depict RNA fold-changes of proteins that are down-regulated (blue) or up-regulated (red) with age according to the proteomics data (p-value<0.05). Fold-changes were calculated between old (>50 years) and young (<30 years) donors (Mann-Whitney U-test: p-value < 0.01 (**), p-value < 0.05 (*)). (c) The box-plots depict RNA fold-changes of glycolytic and TCA-related proteins that are not altered upon ageing in HPC (blue) and that are altered upon ageing (red). Proteins from Figure 4a+c were taken into account. Significance was assessed using a Mann-Whitney U-test (p-value for HPC = 0.0522).


a
cell cycle
Condensation of Prometaphase Chromosomes(n=11)
Mitotic Spindle Checkpoint(n=21)
Regulation of PLK1 Activity at G2/M Transition(n=87)
Chromosome Maintenance(n=90)
Mitotic Telophase/Cytokinesis(n=14)
Nuclear Envelope Reassembly(n=13)
Nuclear Envelope Breakdown(n=52)
Cellular response to stress and DNA repair
Senescence-Associated Secretory Phenotype (SASP)(n=80)
Oxidative Stress Induced Senescence(n=95)
DNA Damage/Telomere Stress Induced Senescence(n=61)
Cellular response to heat stress(n=98)
HSF1 activation(n=30)
Translesion Synthesis by POLH(n=19)
Dual incision in TC-NER(n=65)
Formation of TC-NER Pre-Incision Complex(n=53)
Nonhomologous End-Joining (NHEJ)(n=52)
DNA Double-Strand Break Repair(n=148)
Mismatch Repair(n=15)
Base Excision Repair(n=37)
Nucleotide Excision Repair(n=110)
Develop. biol.
EPHA-mediated growth cone collapse(n=34)
Myogenesis(n=29)
Transcriptional regulation of pluripotent stem cells(n=45)
EPH-ephrin mediated repulsion of cells(n=48)
Transcriptional regulation of white adipocyte differentiation(n=84)
ECM org.
Elastic fibre formation(n=41)
Collagen formation(n=89)
Integrin cell surface interactions(n=67)
Syndecan interactions(n=20)
Gene expression and chromatin organization
RMTs methylate histone arginines(n=49)
HATs acetylate histones(n=108)
PKMTs methylate histone lysines(n=47)
HDMs demethylate histones(n=27)
Epigenetic regulation of gene expression(n=121)
Positive epigenetic regulation of rRNA expression(n=76)
RNA polymerase II transcribes snRNA genes(n=77)
Metabolism of non-coding RNA(n=52)
KSRP (KHSRP) binds and destabilizes mRNA(n=17)
Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors(n=36)
Formation of the Early Elongation Complex(n=33)
Gene Silencing by RNA(n=104)
Regulation of TP53 Activity through Phosphorylation(n=91)
RNA Polymerase I Transcription Termination(n=33)
Regulation of TP53 Activity through Acetylation(n=30)
RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription(n=119)
Transcriptional activity of SMAD2/SMAD3(n=44)
TP53 Regulates Transcription of Cell Death Genes(n=45)
TP53 regulates transcription of additional cell cycle genes whose exact ...(n=21)
tRNA processing in the nucleus(n=58)
RNA Polymerase II Transcription(n=149)
RNA Polymerase II Pre-transcription Events(n=83)
Transport of Mature Transcript to Cytoplasm(n=81)
TP53 Regulates Transcription of DNA Repair Genes(n=65)
mRNA Splicing - Minor Pathway(n=55)
Metabolism and metabolism of proteins
Regulation of lipid metabolism by PPARalpha(n=122)
Regulation of cholesterol biosynthesis by SREBP (SREBF)(n=56)
PI Metabolism(n=78)
Glycogen breakdown (glycogenolysis)(n=15)
Synthesis and interconversion of nucleotide di- and triphosphates(n=25)
Glycolysis(n=32)
Inositol phosphate metabolism(n=45)
Peroxisomal lipid metabolism(n=29)
Fatty Acyl-CoA Biosynthesis(n=52)
VLDL interactions(n=23)
Purine metabolism(n=34)
Glucagon signaling in metabolic regulation(n=33)
Cholesterol biosynthesis(n=24)
Synthesis of PC(n=28)
Arachidonic acid metabolism(n=58)
Metabolism of nitric oxide(n=20)
alpha-linolenic (omega3) and linoleic (omega6) acid metabolism(n=13)
ER Quality Control Compartment (ERQC)(n=21)
SUMOylation(n=106)
Insulin processing(n=28)
Signal transduction
RHO GTPases activate KTN1(n=11)
TNFR1-induced NFkappaB signaling pathway(n=26)
Pre-NOTCH Transcription and Translation(n=41)
RHO GTPases activate PKNs(n=62)
MET activates RAP1 and RAC1(n=11)
NOTCH1 Intracellular Domain Regulates Transcription(n=47)
Signaling by FGFR(n=89)
RHO GTPases activate PAKs(n=21)
RHO GTPases Activate Formins(n=114)
Other categories
Nephrin interactions(n=22)
Late Phase of HIV Life Cycle(n=143)
Signaling by FGFR in disease(n=63)
Signaling by FGFR1 in disease(n=38)
Influenza Viral RNA Transcription and Replication(n=143)
The role of Nef in HIV-1 replication and disease pathogenesis(n=29)
Neurodegenerative Diseases(n=22)
Basigin interactions(n=25)
GRB2(n=15)
Signal amplification(n=32)
Rap1 signalling(n=16)
Cytosolic sensors of pathogen-associated DNA(n=67)
Antiviral mechanism by IFN-stimulated genes(n=75)
Circadian Clock(n=70)
Smooth Muscle Contraction(n=33)
Neurotransmitter Release Cycle(n=51)
Intraflagellar transport(n=41)
Mitochondrial biogenesis(n=54)
Apoptosis induced DNA fragmentation(n=13)
Apoptotic cleavage of cellular proteins(n=38)
Caspase activation via extrinsic apoptotic signalling pathway(n=29)
Hexose transport(n=55)
Iron uptake and transport(n=43)
COPI-independent Golgi-to-ER retrograde traffic(n=28)
Scavenging by Class A Receptors(n=19)
Lysosome Vesicle Biogenesis(n=35)
0 20 40 60 80 100 120 140 160
HPC
LYM
MON
GRA
ERP
MSC
Number of proteins
Pathway and ≥3 proteins upregulated
Pathway and 1 or 2 proteins upregulated
No overall tendency
Pathway and ≥3 proteins downregulated
Pathway and 1 or 2 proteins downregulated
% of proteins altered
10
30
50
70
Size of the pathway
Quantified across all celltypes
Altered upon ageing

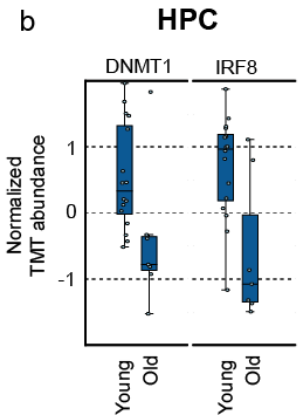
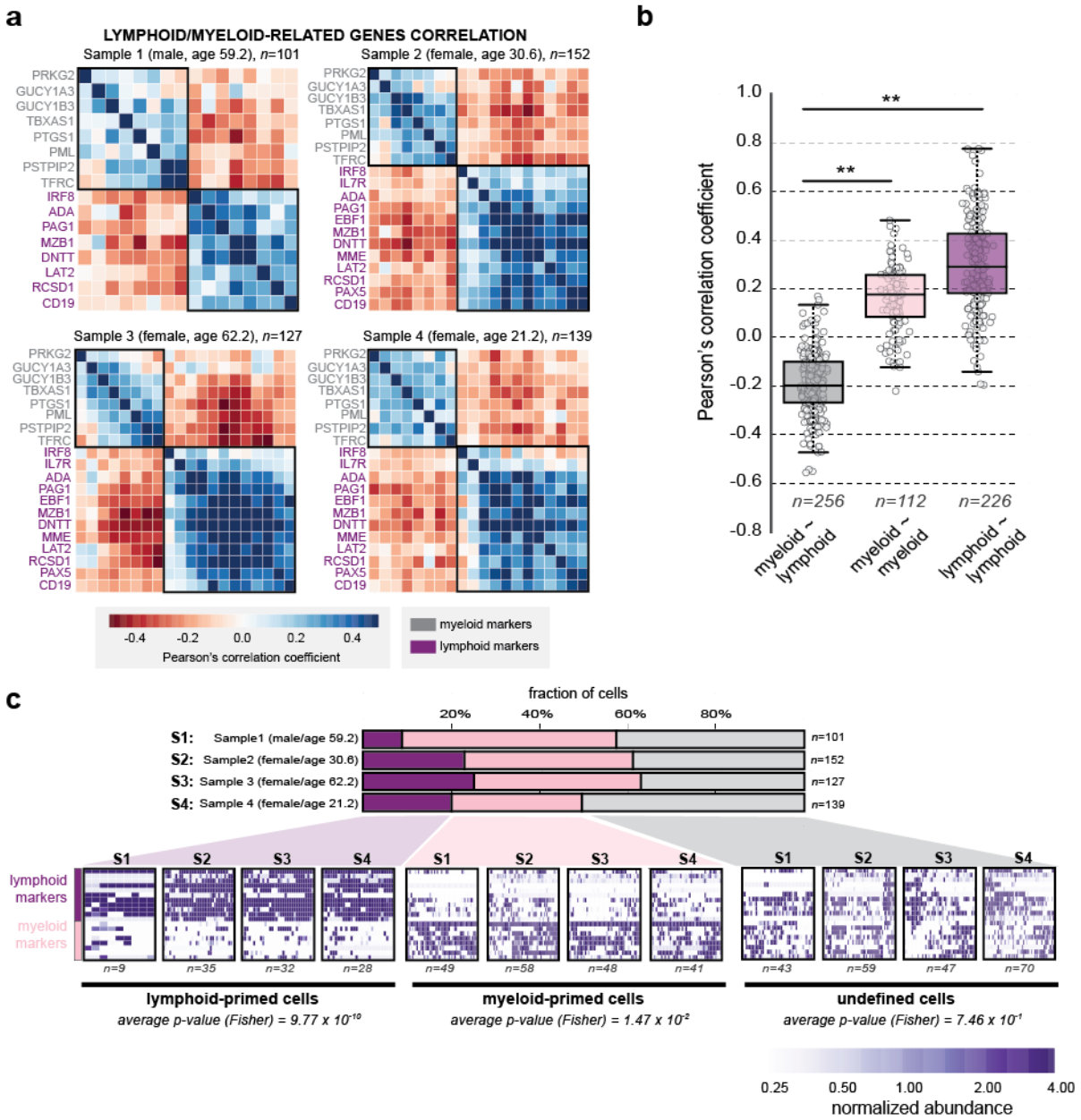


Figure S4.7. Age-affected pathways in the individual cell populations (related to Figure 4.3). (a) A selection of pathways from the Reactome database that show prominent changes upon ageing are depicted. Pathways were required to have between 5 and 150 members, to be sufficiently covered in at least one cell population (> 30% of the proteins are quantified), and to have at least 20% of its quantified components being significantly (p-value < 0.05) altered upon ageing. The area of the bubbles represents the percentage of proteins quantified by TMT that are significantly (p-value < 0.05) altered. If no bubble is shown, no protein of the pathway has been observed to be significantly altered in the respective cell population or no protein of the pathway has been quantified. The colour of the bubbles codes for the direction of the alteration, with red indicating an overall increase of the pathway members with at least three proteins being upregulated and pink an overall increase with one or two proteins being upregulated. Blue codes for pathways with an overall trend towards downregulation, with strong blue coding for pathways with at least three proteins being downregulated and light blue containing one or two proteins being downregulated. The colour white indicates that no overall tendency for the proteins associated with the corresponding pathway could be observed. The bars on the right-hand side of each pathway illustrate the number of proteins being significantly altered upon ageing regardless of the cell population (green), being quantified by TMT (grey), and the total number of members of the pathway (black) as also mentioned in the pathway annotation (n). The grouping of the pathways on the left side is based on the highest hierarchy levels defined in Reactome, e.g. extracellular matrix organisation (ECM org.) and developmental biology (develop. biol.). (b) Illustration of ageing effect on DNMT1 and IRF8 protein abundance in HPC. Box-plot representation of DNMT1 and IRF8. The dots represent the individual results from younger (age < 30 years) and older (age > 50 years) human subjects.


a
LYMPHOID/MYELOID-RELATED GENES CORRELATION
Sample 1 (male, age 59.2), n=101
Sample 2 (female, age 30.6), n=152
Sample 3 (female, age 62.2), n=127
Sample 4 (female, age 21.2), n=139
PRKG2
GUCY1A3
GUCY1B3
TBXAS1
PTGS1
PML
PSTPIP2
TFRC
IRF8
IL7R
ADA
PAG1
EBF1
MZB1
DNTT
MME
LAT2
RCSD1
PAX5
CD19
-0.4 -0.2 0.0 0.2 0.4
Pearson's correlation coefficient
myeloid markers
lymphoid markers
b
Pearson's correlation coefficient
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
**
**
n=256
n=112
n=226
myeloid ~ lymphoid
myeloid ~ myeloid
lymphoid ~ lymphoid
c
fraction of cells
20%
40%
60%
80%
S1: Sample1 (male/age 59.2) n=101
S2: Sample2 (female/age 30.6) n=152
S3: Sample 3 (female/age 62.2) n=127
S4: Sample 4 (female/age 21.2) n=139
S1 S2 S3 S4 S1 S2 S3 S4 S1 S2 S3 S4
lymphoid markers
myeloid markers
n=9 n=35 n=32 n=28 n=49 n=58 n=48 n=41 n=43 n=59 n=47 n=70
lymphoid-primed cells
average p-value (Fisher) = 9.77 x 10-10
myeloid-primed cells
average p-value (Fisher) = 1.47 x 10-2
undefined cells
average p-value (Fisher) = 7.46 x 10-1
0.25 0.50 1.00 2.00 4.00
normalized abundance

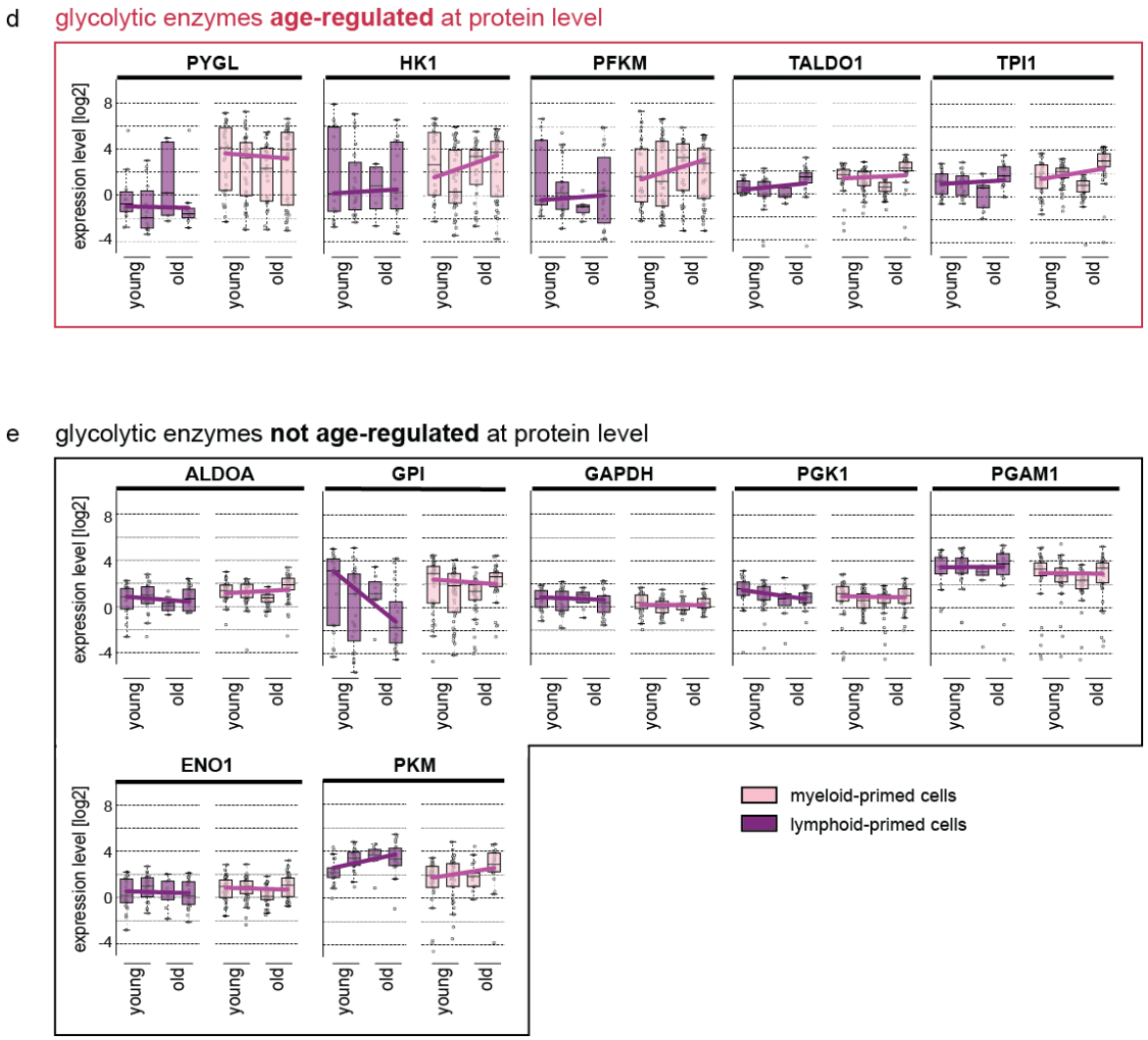


Figure S4.8. Single-cell analysis reveals lineage- and age-dependent increase of glycolytic enzymes (related to Figure 4.5). (a) Matrices demonstrate correlations of lymphoid- and myeloid-associated marker genes across single cells per sample. Marker genes (derived from Figure 5c) and single cells were *a priori* filtered for quality (see Methods). Sample 1 had less of the given marker genes quantified than the other samples. (b) The distribution of correlation values between myeloid and lymphoid markers (grey), between myeloid markers only (pink), and between lymphoid markers only (purple) is presented as box-plot. Correlation values are taken from all four samples. Significance was assessed using a Mann-Whitney U-test (p-value < 0.01 (**)). (c) Cells are classified into lymphoid- and myeloid-primed cells based on markers (see Methods for details). The bar-plot gives an overview on the fraction size of lymphoid-primed (purple), myeloid-primed (pink) and undefined (grey) cells in a given sample ('undefined' meaning that cells could not be classified as myeloid or lymphoid). For each fraction and each sample (S1-S4) the heatmaps below show the expression levels of lymphoid- and myeloid-associated marker genes, as scaled in the colour bar. Significance of marker distribution for each fraction was assessed by a Fisher Exact test (p-values depicted). (d + e) Box-plots representing expression values of a gene in lymphoid- primed (purple) and myeloid-primed (pink) cells in a given sample (2 young (S4+S2), and 2 old (S3+S1)). A trend line connects the respective box-plot medians; the respective slope is further used for **Figure 4.5e**. (d) Box-plots representing genes of the preparatory phase of glycolysis that were found to be age-regulated (p-value < 0.05), and quantified in the single-cell RNA-seq data. (e) Box-plots representing genes of glycolysis that were found to be not age-regulated (p-value > 0.05), and quantified in the single-cell RNA-seq data.

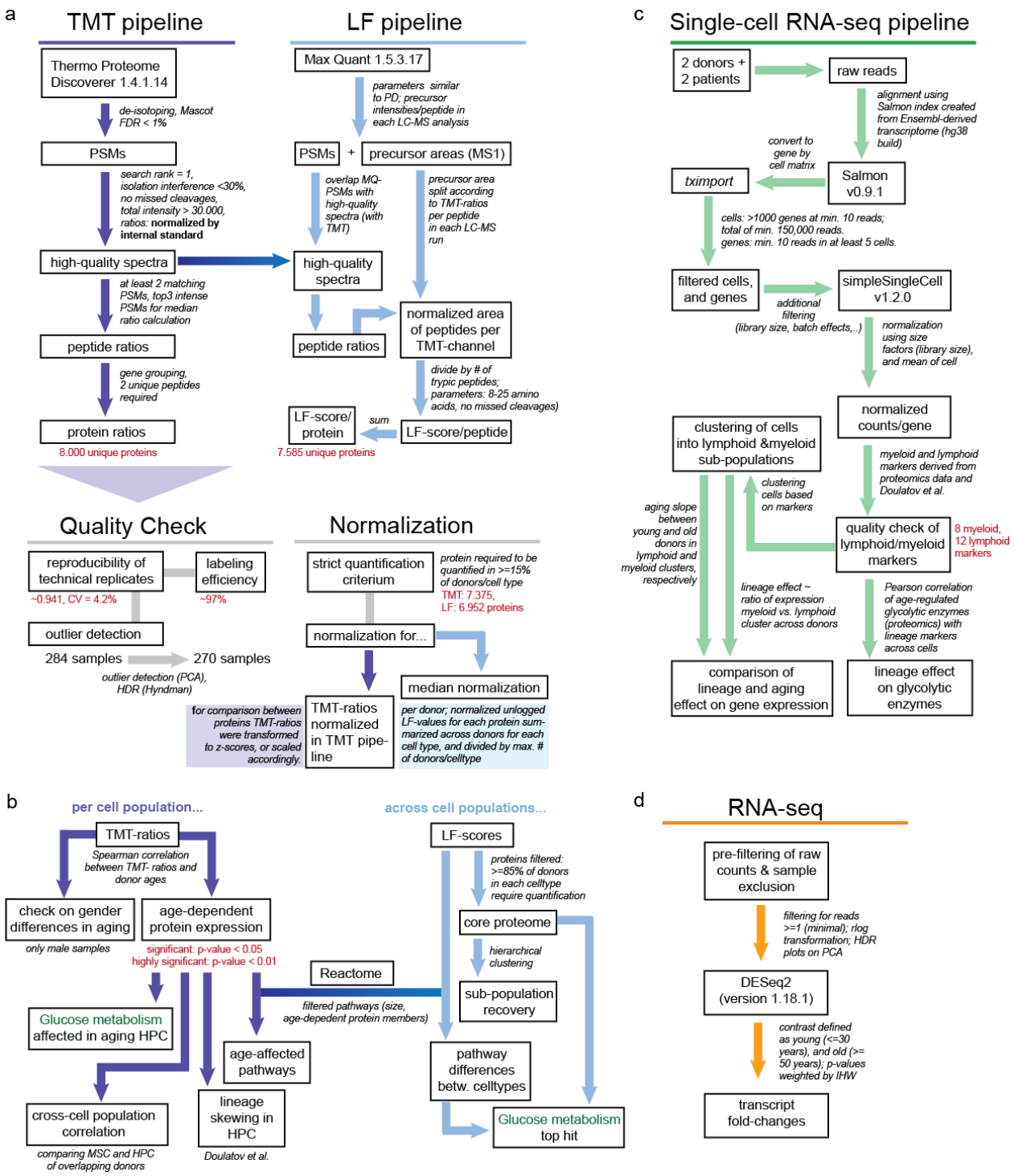


Figure S4.9. Overview on the computational data processing (related to Methods). (a) Delineation of the processing pipeline after acquisition of mass-spectrometry data. (b) Data-processing after obtaining TMT-ratios and LF-scores in (a). (c) Single-cell RNA-seq processing pipeline after acquisition of raw reads. (d) Bulk transcriptome analysis with DESeq2, as explained in Methods.

Bibliography

Abovich, N., Gritz, L., Tung, L., Roshbash, M. (1985). Effect of RP51 gene dosage alterations on ribosome synthesis in Saccharomyces cerevisiae. Molecular Cell Biology, 5(12):3429-3435

Adams, E.J., Chen, X.W, O'Shea, K.S., Ginsburg, D. (2014). Mammalian COPII Coat Component SEC24C Is Required for Embryonic Development in Mice. Journal of Biological Chemistry, 289, 20858-20870.

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. Nature, 422(6928):198–207.

Aebersold, R., and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. Nature, 537(7620):347-355

Alberts, B., Johnson, A., Lewis, J., and Raff, M. (2008). Molecular Biology of the Cell. Garland Science

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. Nucleic Acids Research, 43(D1):789-798

Amrani, N., Sachs, M.S., Jacobson, A. (2006). Early nonsense: mRNA decay solves a translational problem. Nature reviews. Molecular cell biology, 7(6):415-425

Anczukow, O., Ware, M.D., Buisson,M., Zetoune, A.B., Stoppa-Lyonnet, D., Sinilnikova, O.M., Mazoyer, S. (2008). Does the nonsense-mediated mRNA decay mechanism prevent the synthesis of truncated BRCA1,CHK2, and p53 proteins? Human mutation,29(1):65-73

Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M.D., Wagner, G.P. (2016). Nature Review Genetics, 17(12):744-757

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O. (2017). Multi-Omics factor analysis disentangles heterogeneity in blood cancer. bioRxiv, https://doi.org/10.1101/217554

Ashley, E.A. (2016). Towards precision medicine. Nature Reviews Genetics, 17:507-522

Asial, I., Cheng, Y.X., Engman, H., Dollhopf, B., Wu.P, Nordlung, P., Cornvik, T. (2013). Engineering protein thermostability using a generic activity-independent biophysical screen inside the cell. Nature Communications, 4:2901

Azevedo, E.P., Rochael, N.C., Guimaraes-Costa, A.B., de Souza-Vieira, T.S., Ganilho, J., Saraiva, E.M., Palhano, F.L., and Foguel, D. (2015). A Metabolic Shift toward Pentose Phosphate Pathway Is Necessary for Amyloid Fibril-and Phorbol 12-Myristate 13-Acetate-induced Neutrophil Extracellular Trap (NET) Formation. J Biol Chem 290, 22174-22183.

Azzollini, J., Rovina, D., Gervasini, C., Parenti, I., Fratoni, A., Cubellis, M.V., Cerri, A., Petrogrande, L., Larizza, L. (2014). Functional characterization of a novel mutation affecting the catalytic domain of MMP2 in siblings with multicentric osteolysis, nodulosis and arthropathy. Journal of Human Genetics, 59:631-637

Bachant, J., Alcasabas, A., Blat, Y., Kleckner, N., and Elledge, S.J. (2002). The SUMO-1 isopeptidase Smt4 is linked to centromeric cohesion through SUMO-1 modification of DNA topoisomerase II. Mol Cell *9*, 1169-1182.

Bah, A., Vernon, R.M., Siddiqui, Z., Krzeminski, M., Muhandiram, R., Zhao, C., Sonenberg, N., Kay, L.E., and Forman-Kay, J.D. (2015). Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. Nature *519*, 106-109.

Banovich, N.E., Lan, X., McVicker, G., [...], Pritchard, J.K., Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS genetics, 10(9):e1004663

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., Gilad, Y. (2015). Genomic variation. Impact of regulatory variation from RNA to protein. Science, 347(6222):664-7

Bava, K.A., Gromiha, M.M., Uedaira, H., Kitajima, K., and Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic acids research, 32(Database issue):D120–1

Beadle, G.W., and Tatum, E.L. (1941). Genetic Control of Biochemical Reactions in Neurospora. Proceedings of the National Academy of Sciences of the United States of America, 27(11):499-506

Becher, I., Werner, T., Doce, C., Zaal, E.A., Togel, I., Khan, C.A., Rueger, A., Muelbaier, M., Salzer, E., Berkers, C.R.*, et al.* (2016). Thermal profiling reveals phenylalanine hydroxylase as an off-target of panobinostat. Nat Chem Biol *12*, 908-910.

Beck, M., and Hurt, E. (2017). The nuclear pore complex: understanding its function through structural insight. Nature Reviews Molecular Cell Biology, 18(2):73-89

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011). The quantitative proteome of a human cell line. Mol Syst Biol *7*, 549.

Benesch, R., and Benesch, R.E. (1967). The effect of organic phosphates from the human erythrocyte on the allosteric properties of hemoglobin. Biochemical and Biophysical Research Communications 26, 162-167.

Benetatos, L., and Vartholomatos, G. (2016). On the potential role of DNMT1 in acute myeloid leukemia and myelodysplastic syndromes: not another mutated epigenetic driver. Ann Hematol 95, 1571-1582.

Benjamini, Y. And Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J.R. Statistic Society, 57(1):289-300

Ben-Zvi, A., Miller, E.A., Morimoto, R.I. (2009). Collapse of proteostasis represents an early molecular event in Caenorhabditis elegans aging. Proceedings of the National Academy of Sciences of the United States of America,106:14914-14919

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) The Protein Data Bank. Nucleic Acids Research, 28:235-242

Beyer, A., Hollunder, J., Nasheuer, H.P., Willhelm, T. (2004). Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. Molecular Cell Proteomics, 3(11):1083-92

Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.L., Kruglyak, L. (2013). Finding the source of missing heritability in a yeast cross. Nature, 494(7436):234-237

Boisvert, F.M., Ahmad, Y., Gierlinski, M., Charriere, F., Lamont, D., Scott, M., Barton, G., and Lamond, A.I. (2012). A quantitative spatial proteomics analysis of proteome turnover in human cells. Mol Cell Proteomics *11*, M111 011429.

Boisvert, F.M., Lam, Y.W., Lamont, D., and Lamond, A.I. (2010). A quantitative proteomics analysis of subcellular proteome localization and changes induced by DNA damage. Mol Cell Proteomics *9*, 457-470.

Boopathy, S., Silvas, T.V., Tischbein, M., Jansen, S., Shandilya, S.M., Zitzewitz, J.A., Landers, J.E., Goode, B.L., Schiffer, C.A., and Bosco, D.A. (2015). Structural basis for mutation-induced destabilization of profilin 1 in ALS. Proceedings of the National Academy of Sciences of the United States of America, 112(26):7984–9

Bork, S., Pfister, S., Witt, H., Horn, P., Korn, B., Ho, A.D., and Wagner, W. (2010). DNA methylation pattern changes upon long-term culture and aging of human mesenchymal stromal cells. Aging Cell 9, 54-63.

Brehme, M., Voisine, C., Rolland, T., Wachi, S., Soper, J.H., Zhu, Y., Orton, K., Villella, A., Garza, D., Vidal, M., Ge, H., Morimoto, R.I. (2014). A chaperome subnetwork safeguards proteastasis in aging and neurodegenerative disease. Cell Rep., 9(3):1135-1150

Brooks, C.L, Onuchic, J.N., Wales, D.J. (2001). Taking a Walk on a Landscape. Science,293(5530):612-613

Bustamante, E., Pedersen, P.L. (1977). High aerobic glycolysis of rat hepatoma cells in culture: role of mitochondrial hexokinase. Proceedings of the National Academy of Sciences of the United States of America, 74(9):3735-9

Cañas, B., D. López-Ferrer, A. Ramos-Fernández, E. Camafeita, and E. Calvo, 2006, Mass spectrometry technologies for proteomics: Brief Funct Genomic Proteomic, v. 4, p. 295-320.

Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. Nature, 474(7353):609-15

Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A. […], Taylor, R., Moore, H.M. (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreservation and Biobanking, 13(5):311-319

Carmena, M., Wheelock, M., Funabiki, H., and Earnshaw, W.C. (2012). The chromosomal passenger complex (CPC): from easy rider to the godfather of mitosis. Nat Rev Mol Cell Biol *13*, 789-803.

Carter, J., Stein. (2014). Mitosis. biology.clc.uc.edu, archived from the original

Cellerino, A., and Ori, A. (2017). What have we learned on aging from omics studies? Semin Cell Dev Biolology, 70:177-189

Challen, G.A., Boles, N.C., Chambers, S.M., and Goodell, M.A. (2010). Distinct Hematopoietic Stem Cell Subtypes Are Differentially Regulated by TGF-beta 1. Cell Stem Cell 6, 265-278.

Chang, D., Gao, F., Slavney, A., Ma, L., Waldman, Y.Y., Sams, A.J., Billing-Ross, P., Madar, A., Spritz, R., Keinan, A. (2014). Accounting for eXentricities: Analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. PLoS One, 9:e113684

Cheng, Y., LeGall, T., Oldfield, C.J., Mueller, J.P., Van, Y.Y., Romero, P., Cortese, M.S., Uversky, V.N., Dunker, A.K. (2006). Rational drug desgin via intrinsically disordered protein. Trends in Biotechnology, 24(10):435-442

Chick, J.M., Kolippakkam, D., Nusinow, D.P., Zhai, B., Rad, R., Huttlin, E.L., Gygi, S.P. (2015). A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nature Biotechnology, 33(7):743-749

Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A., Gygi, S.P. (2016). Defining the consequences of genetic variation on a proteom –wide scale. Nature, 534(7608):500-5

Chin, L., Andersen, J.N., Futreal, P.A. (2011). Cancer genomics: from discovery science to personalized medicine. Nature Medicine, 17:297-303

Chou, Y.H., Bischoff, J.R., Beach, D., and Goldman, R.D. (1990). Intermediate filament reorganization during mitosis is mediated by p34cdc2 phosphorylation of vimentin. Cell *62*, 1063-1071.

Chow, A., Lucas, D., Hidalgo, A., Mendez-Ferrer, S., Hashimoto, D., Scheiermann, C., Battista, M., Leboeuf, M., Prophete, C., van Rooijen, N., *et al.* (2011). Bone marrow CD169(+) macrophages promote the retention of hematopoietic stem and progenitor cells in the mesenchymal stem cell niche. J Exp Med 208, 261-271.

Christofk, H.R., Vander Heiden, M.G., Harris, M.H., Ramanathan, A., Gerszten, R.E., Wei, R., Fleming, M.D., Schreiber, S.L., Cantley, L.C. (2008). The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumor growth. Nature, 452(7184):230-233

Collinge, J. (2001). Prion diseases of humans and animals: their causes and molecular basis. Annual review of neuroscience, 24:519–50

Collins, B.C., Hunter, C.L., Liu, Y., Schilling, B.[...], Aebersold, R. (2017). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. Nature Communications, 8(1):291

Cook, C.E., Bergman, M.T., Finn, R., [...], Apweiler, R. (2016). The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Research, 44(D1):D20-D26

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26, 1367-1372.

Crick, F. (1970). Central Dogma of Molecular Biology. Nature, 227:561-563

D'Angelo, M., Gomez-Cavazos, J.S., Mei, A., Lackner, D.H., Hetzer, M.W. (2012). A Change in Nuclear Pore Complex Composition Regulates Cell Differentiation. Developmental Cell, 22(2):446-458

Daniel, R.M., and Danson, M.J. (2013). Temperature and the catalytic activity of enzymes: a fresh understanding. FEBS letters, 587(17):2738–43

Davies, E.C., Rowe, P.H., James, S., [...], Pirmohamed, M. (2011). An Investigation of Diagreement in Causality Assessment of Adverse Drug Reactions. Pharmaceutical Medicine, 25(1):17-24

de Lichtenberg, U., Jensen, L.J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. Science *307*, 724-727.

de Lichtenberg, U., Jensen, T.S., Brunak, S., Bork, P., and Jensen, L.J. (2007). Evolution of cell cycle control: same molecular machines, different regulation. Cell Cycle *6*, 1819-1825.

De Sousa Abreu, R., Penalva, L.O., Marcotte, E.M. and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. Molecular BioSystems, 5(12):1512-26

Deberardinis, R.J., Sayed, N., Ditsworth, D., Thompson, C.B. (2008). Brick by brick: metabolism and tumor cell growth. Current Opinion Genetic Development, 18:54-61

Dechat, T., Adam, S.A., Taimen, P., Shimi, T., and Goldman, R.D. (2010). Nuclear lamins. Cold Spring Harb Perspect Biol *2*, a000547.

Defays, D. (1977). An efficient algorithm for a complete link method. Comput J 20, 364-366.

Dephoure, N., Hwang, S., O'Sullivan, C., Dodgson, S.E., Gygi, S.P., Amon, A., Torres, E.M. (2014). Quantitative proteomic analysis reveals post-translational responses to aneuploidy in yeast. Elife, 3:e03023

Dephoure, N., Zhou, C., Villen, J., Beausoleil, S.A., Bakalarski, C.E., Elledge, S.J., and Gygi, S.P. (2008). A quantitative atlas of mitotic phosphorylation. Proc Natl Acad Sci U S A *105*, 10762-10767.

Desiere, F., Deutsch, E.W., King, N.L., Nesvizhskii, A.I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S.N., Aebersold, R. (2006). The PeptideAtlas project. Nucleic Acids Research, 34: D655-D658

Desjardins, G., Meeker, C.A., Bhachech, N., Currie, S.L., Okon, M., Graves, B.J., and McIntosh, L.P. (2014). Synergy of aromatic residues and phosphoserines within the intrinsically disordered DNA-binding inhibitory elements of the Ets-1 transcription factor. Proc Natl Acad Sci U S A *111*, 11019-11024.

Di Marco, E., Gray, S.P., Jandeleit-Dahm, K. (2013). Diabetes alters activation and repression of pro- and anti-inflammatory signaling pathways in the vasculature. Frontiers in Endocrinology, 4:68

Dinkel, H., Van Roey, K., Michael, S., Kumar, M., Uyar, B., Altenberg, B. Milchevskaya, V., [...], Gibson, T.J. (2016). ELM 2016—data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Research, 44(D1):D294-300

Domon, B. and Aebersold, R. (2010). Options and considerations when selecting a quantitative proteomics strategy. Nat. Biotechnol., 28(7):710–721.

Dosztanyi, Z., Csizmok, V., Tompa, P., Simon, I. (2005) IUPRED: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics, 21(16):3433-4

Douglas, P.M., and Dillin, A. (2010). Protein homeostasis and aging in neurodegeneration. Journal of Cell Biology, 190(5):719-729

Doulatov, S., Notta, F., Eppert, K., Nguyen, L.T., Ohashi, P.S., and Dick, J.E. (2010). Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. Nat Immunol 11, 585-U552.

Drews, J. (2000). Drug discovery: a historical perspective. Science, 287:1960-1964

Drucker, E., and Krapfenbauer, K. (2013). Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalized medicine. EPMA Journal, 4(1):7

Dykstra, B., Olthof, S., Schreuder, J., Ritsema, M., and de Haan, G. (2011). Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells. J Exp Med 208, 2691-2703.

Edfors, F., Danielsson, F., Hallstroem, B.M., Kaell, L., Lundberg, E., Ponten, F., Forsstroem, B., Uhlen, M. (2016). Gene-specific correlation of RNA and protein levels in human cells and tissues. Molecular Systems Biology, 12(10):883

Ehninger, A., and Trumpp, A. (2011). The bone marrow stem cell niche grows up: mesenchymal stem cells and macrophages move in. J Exp Med 208, 421-428.

Eisenberg, E. and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends in Genetics, 29(10):569-574

Ellis, S.L., and Nilsson, S.K. (2012). The location and cellular composition of the hemopoietic stem cell niche. Cytotherapy 14, 135-143.

Erfle, H., Neumann, B., Liebel, U., Rogers, P., Held, M., Walter, T., Ellenberg, J., and Pepperkok, R. (2007). Reverse transfection on cell arrays for high content screening microscopy. Nat Protoc *2*, 392-399.

Fabregat, A., Sidiropoulos, K., [...], Hermjakob, H., D'Eustachio, P. (2016). The Reactome pathway Knowledgebase. Nucleic Acids Research, 44:481-487

Feng, Y., De Franceschi, G., Kahraman, A., Soste, M., Melnik, A., Boersema, P.J., de Laureto, P.P., Nikolaev, Y., Oliveira, A.P., and Picotti, P. (2014). Global analysis of protein structural changes in complex proteomes. Nat Biotechnol *32*, 1036-1044.

Finn, R.D., Attword, T.K., [...], Mitchell, A.L. (2017). InterPro in 2017- beyond protein family and domain annotations. Nucleic Acids Research, 45:190-199

Foisner, R., Malecz, N., Dressel, N., Stadler, C., and Wiche, G. (1996). M-phase-specific phosphorylation and structural rearrangement of the cytoplasmic cross-linking protein plectin involve p34cdc2 kinase. Mol Biol Cell *7*, 273-288.

Fortelny, N., Overall, C.M, Pavlidis, P. and Cohen Freue, G.V. (2017) Can we predict protein from mRNA levels? Nature, 547(7664):E19-E20

Franken, H., Mathieson, T., Childs, D., Sweetman, G.M., Werner, T., Togel, I., Doce, C., Gade, S., Bantscheff, M., Drewes, G., *et al.* (2015). Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. Nat Protoc *10*, 1567-1593.

Freed, E.F., Prieto, J.L., McCann, K.L., McStay, B., and Baserga, S.J. (2012). NOL11, implicated in the pathogenesis of North American Indian childhood cirrhosis, is required for pre-rRNA transcription and processing. PLoS Genet *8*, e1002892.

Frese, C.K., Mikhaylova, M., Stucchi, R., Gautier, V., Liu, Q., Mohammed, S., Heck, A.J.R., Altelaar, A.F.M., Hoogenraad, C.C. (2017). Quantitative Map of Proteome Dynamics during Neuronal Differentiation. Cell, 18(6):1527-1542

Fridman, A., Saha, A., Chan, A., Casteel, D.E., Pilz, R.B., and Boss, G.R. (2013). Cell cycle regulation of purine synthesis by phosphoribosyl pyrophosphate and inorganic phosphate. Biochem J *454*, 91-99.

Gagliardi, C., Falkenstein, K.P., Franke, D.E., Kubisch, H.M. (2010). Estimates of heritability for reproducitve traits in captive rhesus macaque females. American Journal of Primatology, 72(9):811-819

Garcia-Bermudez, J., Cuezva, J.M. (2016). The ATPase Inhibitory Factor (IF1): A master regulator of energy metabolism and of cell survival. Bioenergetics, 1857(8):1167-1182

Gavin, A.C., Aloy, P., [...] Kuster, B., Bork, P., Russell, R.B., Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature, 440(7084):631-6

Geiger, H., de Haan, G., and Florian, M.C. (2013). The ageing haematopoietic stem cell compartment. Nat Rev Immunol 13, 376-389.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics *11*, M111 014050.

Gholami, A.M., Hahne, H., Wu, Z., Auer, F.J., Meng, C., Wilhelm, M., and Kuster, B. (2013). Global proteome analysis of the NCI-60 cell line panel. Cell Rep *4*, 609-620.

Giannandrea, M., Guarinieri, F.C.,[...], Valtorta, F. (2013). Nonsense-mediated mRNA decay and loss-of-function of the protein underlie the X-linked epilepsy associated with the W356x mutation in synapsin I. PLoS one,8(6):e67724

Gillet, L.C., Navarro, P., Tate, S., Roest, H., Selevsek, N., Reiter, L., Bonner, R., Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Molecular Cell Proteomics, 11(6):O111.016717

Gillet, L.C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics *11*, O111 016717.

Goldberg, A.L, and Dice, J.F. (1974). Intracellular protein degradation in mammalian and bacterial cells. Annual Review Biochemical, 43(0):835-869

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

Graham, A.C., Kiss, D.L., and Andrulis, E.D. (2009). Core exosome-independent roles for Rrp6 in cell cycle progression. Mol Biol Cell *20*, 2242-2253.

Grasbon-Frodl, E., Lorenz, H., Mann, U., Nitsch, R.M., Windl, O., and Kretzschmar, H.A. (2004). Loss of glycosylation associated with the T183A mutation in human prion disease. Acta neuropathologica, 108(6):476–84

Greggio, C., Jha, P., Kulkarni, S.S., Lagarrigue, S., Broskey, N.T., Boutant, M., Wang, X., Conde Alonso, S., Ofori, E., Auwerx, J., Cantó, C., Amati, F. (2017). Enhanced Respiratory Chain Supercomplex Formation in Response to Exercise in Human Skeletal Muscle. Cell Metabolism, 25(2):301-311

Grob, A., Colleran, C., and McStay, B. (2014). Construction of synthetic nucleoli in human cells reveals how a major functional nuclear domain is formed and propagated through cell division. Genes Dev *28*, 220-230.

Grover, A., Sanjuan-Pla, A., Thongjuea, S., Carrelha, J., Giustacchini, A., Gambardella, A., Macaulay, I., Mancini, E., Luis, T.C., Mead, A., *et al.* (2016). Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. Nat Commun 7.

Grubert, F., Zaugg, J.B.,Kasowski, M., Ursu, O., Spacek, D.V., [...], Steinmetz, L.M., Kundaje, A., Snyder, M. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell, 162(5):1051-1065

Guergen, D., Kusch, A., Klewitz, R., Hoff, U., Catar, R., Hegner, B., Kintscher, U., Luft, F.C., Dragun, D. (2013). Sex-Specific mTOR Signaling Determines Sexual Dimorphism in Myocardical Adaptation in Normotensive DOCA-Salt Model. Hypertension, 61:730-736

Guo, T., Kouvonen, P., Koh, C.C., Gillet, L.C., Wolski, W.E., Roest, H.L., Rosenberger, G., Collins, B.C., Blum, L.C., Gillessen, S., Joerger, M., Jochum, W. and Aebersold, R. (2015). Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nature Medicine, 21(4): 407–413.

Guo, X., Wang, X., Wang, Z., Banerjee, S., Yang, J., Huang, L., and Dixon, J.E. (2016). Site-specific proteasome phosphorylation controls cell proliferation and tumorigenesis. Nat Cell Biol *18*, 202-212.

Hankin, B.L., Abramson, L.Y. (2001). Development of gender differences in depression: an elaborated cognitive vulnerability-transactional stress theory. Psychology Bulletin, 127:773-796

Hansson J1, Rafiee MR, Reiland S, Polo JM, Gehring J, Okawa S, Huber W, Hochedlinger K, Krijgsveld J. (2012). Highly coordinated proteome dynamics during reprogramming of somatic cells to pluripotency. Cell, 2(6):1579-92

Hashizume, C., Kobayashi, A., and Wong, R.W. (2013). Down-modulation of nucleoporin RanBP2/Nup358 impaired chromosomal alignment and induced mitotic catastrophe. Cell Death Dis *4*, e854.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S.*, et al.* (2012). A census of human soluble protein complexes. Cell *150*, 1068-1081.

Heiden, M.G.V., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. Science 324, 1029-1033.

Hermann, A., Goyal, R., and Jeltsch, A. (2004). The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. J Biol Chem 279, 48350-48359.

Hernandez-Verdun, D. (2011). Assembly and disassembly of the nucleolus during the cell cycle. Nucleus *2*, 189-194.

Hinnebusch, A. G. and Johnston, M. (2011). YeastBook: an encyclopedia of the reference eukaryotic cell. Genetics, 189(3):683–684.

Hochhaus, A., Larson, R.A., Guilhot, F., [...], Menssen, H.D., et al.for the IRIS Investigators. (2017). Long-Term Outcomes of Imatinib Treatment for Chronic Myeloid Leukemia. The New England Journal of Medicine, 376:917-927

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic acids research, 43(Database issue):D512–20

Houck, S.A., Singh, S., Cyr, D.M. (2012). Cellular responses to misfolded proteins and protein aggregates. Methods Molecular Biology, 832:455-461

Houten, S.M., and Wanders, R.J.A. (2010). A general introduction to the biochemistry of mitochondrial fatty acid beta-oxidation. J Inherit Metab Dis 33, 469-477.

Houtkooper, R.H., Mouchiroud, L., Ryu, D., Moullan, N., Katsyuba, E., Knott, G., Williams, R.W., Auwerx, J. (2013). Mitonuclear protein imbalance as a conserved longevity mechanism. Nature, 497(7450):451-7

Hu, J., Locasale, J.W., Bielas, J.H., O'Sullivan, J., Sheahan, K. (2013). Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nature Biotechnology, 31:522-529

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res *37*, 1-13.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Huang, S., Chaudhary, K., Garmire, L.X. (2017). More is Better: Recent Progress in Multi-Omics Data Integration Methods. Frontiers in Genetics, 8:84

Huber, K.V., Olek, K.M., Muller, A.C., Tan, C.S., Bennett, K.L., Colinge, J., and Superti-Furga, G. (2015). Proteome-wide drug and metabolite interaction mapping by thermal-stability profiling. Nat Methods *12*, 1055-1057.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics *18 Suppl 1*, S96-104.

Hughes, C.S., Foehr, S., Garfield, D.A., Furlong, E.E., Steinmetz, L.M., and Krijgsveld, J. (2014). Ultrasensitive proteome analysis using paramagnetic bead technology. Mol Syst Biol *10*, 757.

Hyndman, R.J. (1996). Computing and graphing highest density regions. Am Stat 50, 120-126.

Ignolia, N.T. (2017). Ribosome Footprint Profiling of Traslation throughout the Genome. Cell, 165(1):22-33

Irvine, G.B., El-Agnaf, O.M., Shankar, G.M., Walsh, D.M. (2008). Protein aggregation in the brain: the molecular basis for Alzheimer's and Parkinson's diseases. Molecular Medicine, 14(7-8):451-464

Iscove, N.N., and Nawa, K. (1997). Hematopoietic stem cells expand during serial transplantation in vivo without apparent exhaustion. Curr Biol 7, 805-808.

Ishikawa, K., Makanae, K., Iwasaki, S., Ingolia, N.T., Moriya, H. (2017). Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry of Protein Complexes. PLOS Genetics, e1006554.

Jamieson, C.A., Yamamoto, K.R. (2000). Crosstalk pathway for inhibition of glucocorticoid-induced apoptosis by T cell receptor signaling. Proceedings of the National Academy of Sciences of the United States of America, 97(13):7319-7324

Jensen, L.J., Jensen, T.S., de Lichtenberg, U., Brunak, S., and Bork, P. (2006). Co-evolution of transcriptional and post-translational cell-cycle regulation. Nature *443*, 594-597.

Jiang, H., Wang, S., Huang, Y., He, X., Cui, H., Zhu, X., and Zheng, Y. (2015). Phase transition of spindle-associated protein regulate spindle apparatus assembly. Cell *163*, 108-122.

Jongsma, M.L., Berlin, I., and Neefjes, J. (2015). On the move: organelle dynamics during mitosis. Trends Cell Biol *25*, 112-124.

Joseph, J., Liu, S.T., Jablonski, S.A., Yen, T.J., and Dasso, M. (2004). The RanGAP1-RanBP2 complex is essential for microtubule-kinetochore interactions in vivo. Curr Biol *14*, 611-617.

Joseph, J., Tan, S.H., Karpova, T.S., McNally, J.G., and Dasso, M. (2002). SUMO-1 targets RanGAP1 to kinetochores and mitotic spindles. J Cell Biol *156*, 595-602.

Jovanovic M, Rooney MS, Mertins P, Przybylski D, Chevrier N, Satija R, Rodriguez EH, Fields AP, Schwartz S, Raychowdhury R, Mumbach MR, Eisenhaure T, Rabani M, Gennert D, Lu D, Delorey T, Weissman JS, Carr SA, Hacohen N, Regev A (2015) Dynamic profiling of the protein life cycle in response to pathogens. Science 347: 1259038

Khan, Z., Ford, M.J., Cusanovich, D.A., Mitrano, A., Pritchard, J.K. and Gilad, Y. (2013). Primate Transcript and Protein Expression Levels Evolve under Compensatory Selection Pressures. Science, 342(6162): 1100–1104.

Kind, B., Koehler, K., Lorenz, M., Huebner, A. (2009). The nuclear pore coimplex protein ALADIN is anchored via NDC1 but not via POM121 and GP210 in the nuclear envelope. Biochemical and Biophysical Research Communications, 390(2): 205-210

Kinor, N., and Shav-Tal, Y. (2011). The dynamics of the alternatively spliced NOL7 gene products and role in nucleolar architecture. Nucleus *2*, 229-245.

Kinyua, F., Medvedeva, Y.A., Schaefer, U., Jankovic, B.R., Archer, J.A.C., Bajic, V.B. (2012). Mutations and binding sites of human transcription factors. Frontiers in genetics, 3:100

Kleiner, R.E., Hang, L.E., Molloy, K.R., Chait, B.T., Kapoor, T.M. (2017). A Chemical Proteomics Approach to Reveal Direct Protein-Protein Interactions in Living Cells. Cell Chemical Biology, 25:110-120

Klipper-Aurbach, Y., Wasserman, M., Braunspiegel-Weintrob, N., Borstein, D., Peleg, S., Assa, S., Karp, M., Benjamini, Y., Hochberg, Y., and Laron, Z. (1995). Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. Med Hypotheses *45*, 486-490.

Kressler, D., Hurt, E., and Bassler, J. (2010). Driving ribosome assembly. Biochim Biophys Acta *1803*, 673-683.

Krogan, N.J., Cagney, G., [...], Greenblatt, J.F. (2006). Global landscape of protein complexers in the yeast Saccharomyces cerevisiae. Nature, 440(7084):637-43

Kukurba, K.R., Parsana, P., [...], Battle, A., Montgomery, S.B. Impact of the X-Chromosome and sex on regulatory variation. Genome Research, 26(6):768-77

Kurasawa, Y., Earnshaw, W.C., Mochizuki, Y., Dohmae, N., and Todokoro, K. (2004). Essential roles of KIF4 and its binding partner PRC1 in organized central spindle midzone formation. EMBO J *23*, 3237-3248.

Lallemand-Breitenbach, V., and de The, H. (2010). PML nuclear bodies. Cold Spring Harb Perspect Biol *2*, a000661.

Lam, Y.W., Lamond, A.I., Mann, M., and Andersen, J.S. (2007). Analysis of nucleolar protein dynamics reveals the nuclear degradation of ribosomal proteins. Curr Biol *17*, 749-760.

Lappalainen, T., Sammeth, M., Friedlander, M., t'Hoen, P., Rivas, M.A., [...], Rosenstiel, P., Guigo, R., Gut, I., Estivill, X., Dermitzakis, E.T. (2013). Transcriptome and genome sequencing uncovers human functional variation. Nature, 501:506-511

Laurell, E., Beck, K., Krupina, K., Theerthagiri, G., Bodenmiller, B., Horvath, P., Aebersold, R., Antonin, W., and Kutay, U. (2011). Phosphorylation of Nup98 by multiple kinases is crucial for NPC disassembly during mitotic entry. Cell *144*, 539-550.

Lechertier, T., Grob, A., Hernandez-Verdun, D., and Roussel, P. (2009). Fibrillarin and Nop56 interact before being co-assembled in box C/D snoRNPs. Exp Cell Res *315*, 928-942.

Lee, E.E., Sacharidou, A., Mi, W., Salato, V.K., Nguyen, N., Jiang, Y., Pascual, J.M., North, P.E., Shaul, P.W., Mettlen, M., Wang, R.C. (2015). A Protein Kinase C Phosphorylation Motif in GLUT1 Affects Glucose Transport and is Mutated in GLUT1 Deficiency Syndrome. Molecular Cell, 58(5):845-853

Lee, I.H., and Finkel, T. (2013). Metabolic regulation of the cell cycle. Curr Opin Cell Biol *25*, 724-729.

Lee, T.Y., Huang, H.D., Hung, J.H., Huang, H.Y., Yang, Y.S., and Wang, T.H. (2006). dbPTM: an information repository of protein post-translational modification. Nucleic acids research, 34(Database issue):D622–7

Lemeer, S. and Heck, A. J. R. (2009). The phosphoproteomics data explosion. Curr. Opin. Chem. Biol., 13(4):414–420.

Lemke, E.A. (2016). The Multiple Faces of Disordered Nucleoporins. J Mol Biol *428*, 2011-2024.

Lessard, J., Wu, J.I., Ranish, J.A., Wan, M., Winslow, M.M., Staahl, B.T., Wu, H., Aebersold, R., Graef, I.A., Crabtree, G.R. (2007). An essential switch in subunit composition of a chromatin remodeling complex during neural development. Cell Neuron, 55(2):201-15.

Leuenberger, P., Ganscha, S., Kahraman, A., Cappelletti, V., Boersema, P.J., von Mering, C., Claassen, M., and Picotti, P. (2017). Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science *355*.

Ley, K., Rivera-Nieves, J., Sandborn, W.J., and Shattil, S. (2016). Integrin-based therapeutics: biological basis, clinical use and new drugs. Nat Rev Drug Discov 15, 173-183.

Li, G.W., Burkhardt, D., Gross, C., Weissman, J.S. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell, 157(3):624-635

Li, T., and Wu, Y. (2011). Paracrine molecules of mesenchymal stem cells for hematopoietic stem cell niche. Bone marrow research 2011, 353878.

Liang, Y., Van Zant, G., and Szilvassy, S.J. (2005). Effects of aging on the homing and engraftment of murine hematopoietic stem and progenitor cells. Blood 106, 1479-1487.

Liberek, K., Lewandowska, A., Zietkiewicz, S. (2008). Chaperones in control of proteind disaggregation. EMBO Journal, 27(2):328-335

Liberti, M.V., Locasale, J.W. (2016). The Warburg Effect: How does it benefit cancer cells? Trends in Biochemical Sciences (Review). 41(3): 211-218

Lin, W. and Kang, U.J. (2008). Characterization of PINK1 processing, stability, and subcellular localization. Journal of neurochemistry, 106(1):464–74

Liu, Y., Beyer A., Aebersold R. (2016). On the Dependecy of Cellular Protein Levels on mRNA Abundance. Cell, 165(3):535-50

Liu, Y., Borel, C., Li, L., Müller, T., Williams, E.G., Germain, P.L., Bulja, M., Sajic, T., Boersema, P.J., Shao, W., Faini, M., Testa, G., Beyer, A., Antonarakis, S., Aebersold, R. (2017). Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. Nature Communications, 8(1):1212

Liu., Y., Buil, A., Collins, B.C., Gillet, L.C., Blum, L.C., Cheng, L.Y., Vitek, O., Mouritsen, J., Lachance, G., Spector, T.D., Dermitzakis, E.T., Aebersold, R. (2015). Quantitative variability of 342 plasma proteins in a human twin population. Molecular Systems Biology, 11(1):786

Lo Celso, C., Fleming, H.E., Wu, J.W.W., Zhao, C.X., Miake-Lye, S., Fujisaki, J., Cote, D., Rowe, D.W., Lin, C.P., and Scadden, D.T. (2009). Live-animal tracking of individual haematopoietic stem/progenitor cells in their niche. Nature 457, 92-U96.

Locasale, J.W., Grassian, A.R., Melman, T., Lyssiotis, C.A., Mattaini, K.R., Bass, A.J., Heffron, G., Metallo, C.M., Muranen, T., Sharfi, H., *et al.* (2011). Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. Nature Genet 43, 869-U879.

Lopez-Lazaro, M. (2008). The Warburg effect: why and how do cancer cells activate glycolysis in the presence of oxygen? Anti-Cancer Agents in Medicinal Chemistry, 8(3):305-312

Louden, E., Chi, M.M., Moley, K.H. (2013). Crosstalk between the AMP-activated kinase and Insulin Signaling pathways rescues murine blastocyst cells from insulin resistance. Reproduction, 136(3):335-344

Love, M.I., Huber, W., Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15:550

Lu, P., Min, D., DiMaio, F., Wei, K.Y., Vahey, M.D., [...], Bowie, J.U., Baker, D. (2018) Accurate computational design of multipass transmembrane proteins. Science, 359(6379):1042-1046

Maddocks, O.D.K., Labuschagne, C.F., Adams, P.D., and Vousden, K.H. (2016). Serine Metabolism Supports the Methionine Cycle and DNA/RNA Methylation through De Novo ATP Synthesis in Cancer Cells. Mol Cell 61, 210-221.

Madhukar, N.S., Warmoes, M.O., Locasale, J.W. (2015). Organization of Enzyme Concentration across the Metabolic Network in Cancer Cells. PLoS One, 10(1):e0117131

Mann, M., Kulak, N.A., Nagaraj, N., and Cox, J. (2013). The coming age of complete, accurate, and ubiquitous proteomes. Mol Cell *49*, 583-590.

Marsh, J.A., Hernandez, H., Hall, Z., Ahnert, S.E., Perica, T., Robinson, C.V., Teichmann, S.A. (2013). Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell, 153:461-470

Martin, D.I., Tsai, S.F., Orkin, S.H. (1989). Increased gamma-globin expression in a non-deletion HPFH mediated by an erythroid-specific DNA-binding factor. Nature,338(6214):438-438

Martinez Molina, D., Jafari, R., Ignatushchenko, M., Seki, T., Larsson, E.A., Dan, C., Sreekumar, L., Cao, Y., Nordlund, P. (2013). Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. Science, 341: 84-87

Mathelier, A., Lefebvre, C., Zhang, A.W., Arenillas, D.J., Ding, J., Wasserman, W.W., Shah, S.P. (2015). Cis-regulatory somatic mutations and gene expression alteration in B-cell lymphomas. Genome biology,16:84

McAlister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C., Kuhn, K., Pike, I., Grothe, R. A., Blethrow, J. D., and Gygi, S. P. (2012). Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. Anal. Chem., 84(17):7469–7478.

McCarthy, M.K., and Weinberg, J.B. (2015). The immunoproteasome and viral infection: a complex regulator of inflammation. Frontiers in Microbiology, 6:21

McManus, J., Cheng, Z., Vogel, C. (2015). Next-generation analysis of gene expression regulation--comparing the roles of synthesis and degradation. Molecular Biosystems, 11:2680-2689

McShane, E., Sin, C., Zauber, H., Wells, J.N., Donnelly, N., Wang, X., Hou, J., Chen, W., Storchova, Z., Marsh, J.A., Valleriani, A., Selbach, M. (2016). Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. Cell, 167(3):803-815.e21

Melto, C., Reuter, J.A., Spacek, D.V. and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. Nature genetics,47(7):710-716

Melzer, D., Perry, J.R.B., Hernandez, D., [...], Frayling, T.M., Singleton, A., Ferrucci, L. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS genetics, 4(5):e1000072

Mendelsohn, M.E., Karas, R.H. (2005). Molecular and cellular basis of cardiovascular gender differences. Science, 308:1583-1587

Mendelson, A., and Frenette, P.S. (2014). Hematopoietic stem cell niche maintenance during homeostasis and regeneration. Nat Med 20, 833-846.

Mendez-Ferrer, S., Michurina, T.V., Ferraro, F., Mazloom, A.R., MacArthur, B.D., Lira, S.A., Scadden, D.T., Ma'ayan, A., Enikolopov, G.N., and Frenette, P.S. (2010). Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. Nature 466, 829-U859.

Mertins, P., Mani, D.R., [...], NCI CPTAC. (2016). Proteogenomics connects somatic mutations to signaling in breast cancer. Nature, 534(7605):55-62

Meszaros, B., Tompa, P., Simon, I., Dosztanyi, Z. (2007). Molecular principles of the interactions of disordered proteins. Journal of Molecular Biology, 372:549-561

Metallo, C.M, Vander Heiden, M.G. (2013). Understanding metabolic regulation and its influence on cell physiology. Molecular Cell, 49:388-398

Meyer, K., Uyar, B., Kirchner, M.L., Cheng, J., Akalin, A., Selbach, M. (2017). Mutations in disordered regions causd disease by creating endocytosis motifs. BioRxiv

Michurina, T., Krasnov, P., Balazs, A., Nakaya, N., Vasilieva, T., Kuzin, B., Khrushchov, N., Mulligan, R.C., and Enikolopov, G. (2004). Nitric oxide is a regulator of hematopoietic stem cell activity. Mol Ther 10, 241-248.

Milles, S., Huy Bui, K., Koehler, C., Eltsov, M., Beck, M., and Lemke, E.A. (2013). Facilitated aggregation of FG nucleoporins under molecular crowding conditions. EMBO Rep *14*, 178-183.

Minguez, P., Parca, L., Diella, F., Mende, D.R., Kumar, R., Helmer-Citterich, M., Gavin, A.C., van Noort, V., and Bork, P. (2012). Deciphering a global network of functionally associated post-translational modifications. Molecular systems biology, 8:599

Miwa, S., Jow, H., Baty, K., Johnson, A., Czapiewski, R., Saretzki, G., Treumann, A., von Zglinicki, T. Low abundance of the matrix arm of complex I in mitochondria predicts longevity in mice. Nature Communications, 383

Moir, R.D., Yoon, M., Khuon, S., and Goldman, R.D. (2000). Nuclear lamins A and B1: different pathways of assembly during nuclear envelope formation in living cells. J Cell Biol *151*, 1155-1168.

Morgan, D.O. (2007). The cell cycle: principles of control (New Science Press).

Morimoto, R.I. (2008). Proteotoxic stress and inducible chaperone networks in neurodegenerative disease and aging. Genes Development, 22(11):1427-138

Morrison, S.J., Wandycz, A.M., Akashi, K., Globerson, A., and Weissman, I.L. (1996). The aging of hematopoietic stem cells. Nat Med 2, 1011-1016.

Mujoo, K., Krumenacker, J.S., and Murad, F. (2011). Nitric oxide-cyclic GMP signaling in stem cell differentiation. Free Radic Biol Med 51, 2150-2157.

Murakami, H., Goto, D.B., Toda, T., Chen, E.S., Grewal, S.I., Martienssen, R.A., and Yanagida, M. (2007). Ribonuclease activity of Dis3 is required for mitotic progression and provides a possible link between heterochromatin and kinetochore function. PLoS One *2*, e317.

Muslin, A.J., Tanner, J.W., Allen, P.M., and Shaw, A.S. (1996). Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine. Cell,84(6):889-897

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol *7*, 548.

Nakamura-Ishizu, A., and Suda, T. (2014). Aging of the hematopoietic stem cells niche. International journal of hematology 100, 317-325.

Naranjo, C.A., Busto, U., Sellers, E.M., Sandor, P., Ruiz, I., Roberts, E.A., Janecek, E., Domecq, C., Greenblatt, D.J. (1981). A method for estimating the probability of adverse drug reactions. Clinical Pharmacology Therapy, 30(2):239-245

Naugler, W.E., Sakurai, T., Kim, S., Maeda, S., Kim, K., Elsharkawy, A.M., Karin, M. (2007). Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production. Science, 317:121-124

Neduva, V., Linding, R., Su-Angrand, I., Stark, A., de Massi, F., Gibson, T.J., Lewis, J., Serrano, L., Russell, R.B. (2005). Systematic discovery of peptides mediating protein interaction networks PLoS Biology, 3, e405

Nesvizhskii, A. I. (2007). Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol. Biol., 367:87–119.

Nishikawa, K., Ishino, S., Takenaka, H., Norioka, N., Hirai, T., Yao, T., and Seto, Y. (1994). Constructing a protein mutant database. Protein engineering, 7(5):733

Northtop, D.B., Simpson, F.B. (1998). Kinetics of enzymes with isomechanisms: britton induced transport catalyzed by bovine carbonic anhydrase II, measured by rapid-flow mass spectrometry. Arch. Biochem. Biophys., 352:288-292

Nott, T.J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T.D., Bazett-Jones, D.P., Pawson, T., Forman-Kay, J.D., *et al.* (2015). Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. Mol Cell *57*, 936-947.

Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., Xue, B., Dosztanyi, Z., Uversky, V.N., Obradovic, Z., Kurgan, L., *et al.* (2013). D(2)P(2): database of disordered protein predictions. Nucleic Acids Res *41*, D508-516.

Offner, F., Kerre, T., De Smedt, M., and Plum, J. (1999). Bone marrow CD34+ cells generate fewer T cells in vitro with increasing age and following chemotherapy. Br J Haematol 104, 801-808.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A., *et al.* (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Sci Signal *3*, ra3.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Molecular Cell Proteomics, 1(5):376-86.

Oostdjik, W., Idkowiak, J., Mueller, J.W., [...], Uitterlinden, A.G., Wit, J.M., Losekoot, M., Arlt, W. (2015). PAPSS2 Deficiency Causes Androgen Excess via Impaired DHEA Sulfation In Vitro and in Vivo Studies in a Family Harboring Two Novel PAPSS2 Mutations. J. Clin Endocrinol Metabol, 100(4):E672-E680

Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andres-Pons, A., Singer, S., Bork, P., and Beck, M. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol 17.

Ori, A., Toyama, B.H., Harris, M.S., Bock, T., Iskar, M., Bork, P., Ingolia, N.T., Hetzer, M.W., and Beck, M. (2015). Integrated Transcriptome and Proteome Analyses Reveal Organ-Specific Proteome Deterioration in Old Rats. Cell Syst 1, 224-237.

Ori, A., Banterle, N., Iskar, M., Andrés-Pons, A., Escher, C., Khanh Bui, H., Sparks, L., Solis-Mezarino, V., Rinner, O., Bork, P., Lemke, E.A., Beck, M. (2013). Cell type-specific nuclear pores: a

case in point for context-dependent stoichiometry of molecular machines. Molecular Systems Biology, 9:648

Otsuka, S., Bui, K.H., Schorb, M., Hossain, M.J., Politi, A.Z., Koch, B., Eltsov, M., Beck, M., and Ellenberg, J. (2016). Nuclear pore assembly proceeds by an inside-out extrusion of the nuclear envelope. Elife *5*.

Pandit, B., Sarkozy, A., Penacchio, L.A., Carta, C., Oishi, K., […], Dallapiccola, B., Tartaglia, M., Gelb, B.D. (2007). Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. Nature Genetics,39(8):1007-1012

Pang, W.W., Price, E.A., Sahoo, D., Beerman, I., Maloney, W.J., Rossi, D.J., Schrier, S.L., and Weissman, I.L. (2011). Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. Proc Natl Acad Sci U S A 108, 20012-20017.

Papayannopoulou, T., Priestley, G.V., Nakamoto, B., Zafiropoulos, V., and Scott, L.M. (2001). Molecular pathways in bone marrow homing: dominant role of alpha 4 beta 1 over beta 2-integrins and selectins. Blood 98, 2403-2411.

Park, S.H., Kukushin, Y., Gupta, R., Chen, T., Konagai, A., Hipp, M.S., Hayer-Hartl, M., Hartl, F.U. (2013). PolyQ proteins interfere with nuclear degradation of cytosolic proteins by sequestering the Sis1p chaperone. Cell,154:134-145

Pauling, L. (1949). Sickle cell anemia a molecular disease. Science, 110(2865):543-548

Payne, S.H. (2014). The utility of protein and mRNA correlation. Trends in Biochemical Sciences, 40(1):1-3

Pearson, T.A., and Manolio, T.A. (2008). How to interpret a genome-wide association study. JAMA, 299(11):1335-1344

Pelisch, F., Sonneville, R., Pourkarimi, E., Agostinho, A., Blow, J.J., Gartner, A., and Hay, R.T. (2014). Dynamic SUMO modification regulates mitotic chromosome assembly and cell cycle progression in Caenorhabditis elegans. Nat Commun *5*, 5485.

Picotti, P., Clément-Ziza, M. Lam, H., Campbell, D.S., Schmidt, A., Deutsch, E.W., Röst, H., Sun, Z., Rinner, O., Reiter, L., Shen, Q., Michaelson, J.J., Frei, A., Alberti, S., Kusebauch, U., Wollscheid, B., Moritz, R.L., Beyer, A., Aebersold, R. (2013). A Complete mass-spectromeric map of the yeast proteome applied to quantitative trait analysis. Nature, 494(7436):266-70

Pigott, T.A. (1999). Gender differences in the epidemiology and treatment of anxiety disorders. Journal of Clinical Psychiatry, 60 (Suppl18): 4-15

Pines, J. (2006). Mitosis: a matter of getting rid of the right protein at the right time. Trends Cell Biol *16*, 55-63.

Poole, W., Gibbs, D.L., Shmulevich, I., Bernard, B., and Knijnenburg, T.A. (2016). Combining dependent P-values with an empirical adaptation of Brown's method. Bioinformatics *32*, i430-i436.

Prasad, A., Bekker, P., Tsimikas, S. (2012). Advanced glycation end products and diabetic cardiovascular disease. Cardiology in Review, 20(4):177-183

Prusiner, S.B. (1991). Molecular biology of prion diseases. Science (New York, N.Y.), 252(5012):1515–22

Ramilowski, J.A., Goldberg, T., Harshbarger, J., Kloppman, E., Lizio, M., Satagopam, V.P., Itoh, M., Kawaji, H., Carninci, P., Rost, B., *et al.* (2015). A draft network of ligand-receptor-mediated multicellular signalling in human. Nat Commun 6.

Rauniyar, N. and Yates, 3rd, J. R. (2014). Isobaric labeling-based relative quantification in shotgun proteomics. J. Proteome Res., 13(12):5293–5309.

Reiland, S., Salekdeh, G.H., and Krijgsveld, J. (2011). Defining pluripotent stem cells through quantitative proteomic analysis. Expert Rev Proteomics *8*, 29-42.

Reimand, J., Wagih, O., and Bader, G.D. (2015). Evolutionary constraint and disease associations of post-translational modification sites in human genomes. PLoS genetics, 11(1):e1004919

Reimand, J., Wagih, O., Bader, G.D. (2013). The mutational landscape of phosphorylation signaling in cancer. Scientific reports,3:2651

Reinhard, F.B., Eberhard, D., Werner, T., Franken, H., Childs, D., Doce, C., Savitski, M.F., Huber, W., Bantscheff, M., Savitski, M.M., *et al.* (2015). Thermal proteome profiling monitors ligand interactions with cellular membrane proteins. Nat Methods *12*, 1129-1131.

Rinner, O., Seebacher, J., Waltzhoeni, T., Mueller, L.N., Beck, M., Schmidt, A., Mueller, M., Aebersold, R. (2008). Identification of cross-linked peptides from large sequence databases. Nature Methods, 5:315-318

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res *43*, e47.

Roel Nusse. (2008) Wnt signaling and stem cell control. Cell research, 18(5):523–7

Romanov, N., Hollenstein, D.M., Janschitz, M., Ammerer, G., Anrather, D., Reiter, W.L. (2017). Identifying protein kinase specific effectors of the osmostress response in yeast. Science Signaling, 10(469):eaag2435

Romanova, L., Kellner, S., Katoku-Kikyo, N., and Kikyo, N. (2009). Novel role of nucleostemin in the maintenance of nucleolar architecture and integrity of small nucleolar ribonucleoproteins and the telomerase complex. J Biol Chem *284*, 26685-26694.

Rossi, D.J., Bryder, D., Zahn, J.M., Ahlenius, H., Sonu, R., Wagers, A.J., and Weissman, I.L. (2005). Cell intrinsic alterations underlie hematopoietic stem cell aging. Proc Natl Acad Sci U S A 102, 9194-9199.

Röst, H., Malmstroem, L., Aebersold, R. (2015). Reproducible quantitative proteotype data matrices for systems biology. Molecular Biology of the Cell, 26(22):3926-3931

Roumeliotis, T.I., Williams, S.P., Gonçalves, E., Alsinet, C., [...], Stegle, O., Adams, D.J., Wessels, L., Saez-Rodriguez, J., McDermott, U., Choudhary, J.S. (2017). Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells. Cell, 20(9):2201-2214

Royer, C.A.. Probing protein folding and conformational transitions with fluorescence. (2006). Chemical reviews, 106(5):1769–84

Ryan, C.J., Kennedy, S., Barjami, I., Matallanas, D., Lord, C.J. (2017). A Compendium of Co-Regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events. Cell Systems, 5(4):399-409.e5.

Salis, H.M, Mirsky, E.A., Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. Nature Biotechnology, 27:946-950

Sanchez-Arago, M., Formentini, L., Martinez-Reyes, I., Garcia-Bermudez, J., Santacatterina, F., Sanchez-Cenizo, L., Willers, I.M., Aldea, M., Najera, L., Juarranz, A., Lopez, E.C., Clofent, J., Navarro, C., Espinosa, E., Cuezva, J.M. (2013). Expression regulation and clinical relevance of the ATPase inhibitory factor 1 in human cancers.

Sanchez-Cenizo, L., Formentini, L., Aldea, M., Ortega, A.D., Garcia-Huerta, P., Sanchez-Arago, M., Cuezva, J.M. (2010). Up-regulation of the ATPase inhibitory factor 1 (IF1) of the mitochondrial H+-ATP synthase in human tumors mediates the metabolic shift of cancer cells to a Warburg phenotype. J. Biol. Chem., 285:pp25308-25313

Sangith, N., Srinivasaraghavan, K. , Sahu, I., Desai, A., Medipally, S., Somavarappu, A.K., Verma, C., Venkatraman, P. Discovery of novel interacting partners of PSMD9, a proteasomal chaperone: Role of an Atypical and versatile PDZ-domain motif interaction and identification of putative functional modules. FEBS Open Biology, 4:571-83

Santaguida, S., Musacchio, A. (2009). The life and miracles of kinetochores. The EMBO Journal, 28(17):2511-2531

Sasaki, K., Kurose, A., and Ishida, Y. (1993). Flow cytometric analysis of the expression of PCNA during the cell cycle in HeLa cells and effects of the inhibition of DNA synthesis on it. Cytometry *14*, 876-882.

Satir, P., Soren, T. Christensen (2008). Structure and function of mammalian cilia. Histochemistry and Cell Biology, Springer Berlin/Heidelberg, 129(6):688

Sauer, G., Korner, R., Hanisch, A., Ries, A., Nigg, E.A., and Sillje, H.H. (2005). Proteome analysis of the human mitotic spindle. Mol Cell Proteomics *4*, 35-43.

Savitski, M.M., Mathieson, T., Zinn, N., Sweetman, G., Doce, C., Becher, I., Pachl, F., Kuster, B., and Bantscheff, M. (2013). Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. J Proteome Res *12*, 3586-3598.

Savitski, M.M., Sweetman, G., Askenazi, M., Marto, J.A., Lang, M., Zinn, N., and Bantscheff, M. (2011). Delayed fragmentation and optimized isolation width settings for improvement of protein identification and accuracy of isobaric mass tag quantification on Orbitrap-type mass spectrometers. Anal Chem *83*, 8959-8967.

Savitski, M.M., Reinhard, F.B., Franken, H., Werner, T., Savitski, M.F., Eberhard, D., Martinez Molina, D., Jafari, R., Dovega, R.B., Klaeger, S., Kuster, B., Nordlund, P., Bantscheff, M., Drewes, G. (2014). Tracking cancer drugs in living cells by thermal profiling of the proteome. Science, 346(6205):1255784

Scaglia, N., Tyekucheva, S., Zadra, G., Photopoulos, C., and Loda, M. (2014). De novo fatty acid synthesis at the mitotic exit is required to complete cellular division. Cell Cycle *13*, 859-868.

Scharaw, S., Iskar, M., Ori, A., Boncompain, G., Laketa, V., Poser, I., Lundberg, E., Perez, F., Beck, M., Bork, P., Pepperkok, R. (2016). The endosomal transcriptional regulator RNF11 integrates degradation and transport of EGFR. Journal of Cell Biology, 215(4):543-558

Schlessinger, D., and Van Zant, G. (2001). Does functional depletion of stem cells drive aging? Mech Ageing Dev 122, 1537-1553.

Schloesser, M., Arleth, S., Lenz, U., Bertele, R.M., and Reiss, J. (1991). A cystic fibrosis patient with the nonsense mutation G542X and the splice site mutation 1717-1. Journal of medical genetics, 28(12):590-603

Schofield, R. (1978). The relationship between the spleen colony-forming cell and the hematopoietic stem-cell - Hypothesis. Blood Cells 4, 7-25.

Schulze, A., Harris, A.L (2012). How cancer metabolism is tuned for proliferation and vulnerable to disruption. Nature, 491:364-373

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature 473, 337-342.

Schwarz, D.S., and Blower, M.D. (2016). The endoplasmic reticulum: structure, function and response to cellular signaling. Cell Mol Life Sci *73*, 79-94.

Schwender, J., Hebbelmann, I., Heinzel, N., Hildebrandt, T., Rogers, A., Naik, D., Klapperstueck, M., Braun, H.P., Schreiber, F., Denolf, P., Borisjuk, L., Rolletschek, H. (2015). Quantitative Multilevel

Analysis of Central Metabolism in Developing Oilseeds of Oilseed Rape during in Vitro Culture. Plant Physiology, 168(3):828

Selbach, M., Schwanhaeusser, B., Thierfelder, N., Fang, Z., Khanin, R., Rajewski, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature, 455(7209):58-63

Semba, R.D., Ferrucci, L., Sun, K., Beck, J., Dalal, M., Varadhan, R., Walston, J., Guralnik, J.M., Fried, L.P. (2009). Advanced glycation end products and their circulating receptors predict cardiovascular disease mortality in older community-dwelling women. Aging clinical and experimental research, 21(2):182-190

Sharma, A., Takata, H., Shibahara, K., Bubulya, A., and Bubulya, P.A. (2010). Son is essential for nuclear speckle organization and cell cycle progression. Mol Biol Cell *21*, 650-663.

Shu, Q., and Nair, V. (2008). Inosine monophosphate dehydrogenase (IMPDH) as a target in drug discovery. Med Res Rev *28*, 219-232.

Shyh-Chang, N., Daley, G.Q., and Cantley, L.C. (2013). Stem cell metabolism in tissue development and aging. Development 140, 2535-2547.

Siwiak, M. and Zielenkiewicz, P. (2015). Co-regulation of translation in protein complexes. Biology Direct, 10:18

Smith, B., Schafer, X.L., Ambeskovic, A., Spencer, C.M., Land, H., and Munger, J. (2016). Addiction to Coupling of the Warburg Effect with Glutamine Catabolism in Cancer Cells. Cell Reports 17, 821-836.

Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harb Perspect Biol *3*.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., [...], Ingelsson, E., Loos, R.J. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics, 42(11):937-48

Stavru, F., Huelsmann, B.B., Spang, A., Hartmann, E., Cordes, V.C., Goerlich, D. (2006) NDC1: a crucial membrane-integral nucleoporin of metazoan nuclear pore complexes. Journal of Cell Biology, 173(4):509

Stefely, J.A., Kwiecien, N.W., Freiberger, E.C., Richards A.L., Jochem, A., Rush, M.J.P., [...], Pagliarini, D.J., Coon, J.J. (2016). Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling. Nature Biotechnology, 34(11):1191-1197

Steigemann, P., Wurzenberger, C., Schmitz, M.H., Held, M., Guizetti, J., Maar, S., and Gerlich, D.W. (2009). Aurora B-mediated abscission checkpoint protects against tetraploidization. Cell *136*, 473-484.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human genetics, 133(1):1–9

Stevens, B.J. (1965). The Fine Structure of the Nucleolus during Mitosis in the Grasshopper Neuroblast cell. Journal of Cell Biology, 24(3):349-368

Stingele, S., Stoehr, G., Peplowska, K., Cox, J., Mann, M., Storchova, Z. (2012). Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. Molecular Systems Biology, 8:608

Stirewalt, D.L., Choi, Y.E., Sharpless, N.E., Pogosova-Agadjanyan, E.L., Cronk, M.R., Yukawa, M., Larson, E.B., Wood, B.L., Appelbaum, F.R., Radich, J.P., *et al.* (2009). Decreased IRF8 expression found in aging hematopoietic progenitor/stem cells. Leukemia 23, 391-393.

Strimmer, K. (2008). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics *24*, 1461-1462.

Strong, M. (2004). Protein Nanomachines. PloS Biology, 2(3):e73

Stroud, D.A., Surgenor, E.E., Formosa, L.E., Reljic, B., Frazier, A.E., Dibley, M.G., Osellame, L.D., Stait, T., Beilharz, T.H., Thorburn, D.R., Salim, A., Ryan, M.T. (2016). Accessory subunits are integral for assembly and function of human mitochondrial complex I. Nature, 538(7623):123-126

Studer, R.A., Christin, P.A., Williams, M.A., and Orengo, C.A. Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. (2014). Proceedings of the National Academy of Sciences of the United States of America, 111(6):2223–8

Sudo, K., Ema, H., Morita, Y., and Nakauchi, H. (2000). Age-associated characteristics of murine hematopoietic stem cells. J Exp Med 192, 1273-1280.

Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., [...], Peters, A., Karstenmueller, G., Gieger, C., Graumann, J. (2017). Connecting genetic risk to disease end points through the human blood plasma proteome. Nature Communications, 8(14357)

Sullivan, G.M., Feinn, R. (2012). Using Effect Size- or Why P Value is Not Enough. Journal of Graduate Medical Education, 4(3):279-282

Sun, J., Hao, Z., Luo, H., He, C., Mei, L., [...], Feng, Y. (2017). Functional analysis of a nonstop mutation in MITF gene identified in a patient with Waardenburg syndrome type 2. Journal of human genetics, 62(7):703-709

Svenningsen, S.L., Kongstad, M., Sondergaard Stenum, T., Munoz-Gomez, A.J., Sorensen, M.A. (2017). Transfer RNA is highly unstable during early amino acid starvation in Escheria coli. Nucleic Acids Research, 45(2):793-804

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., *et al.* (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45, D362-D368.

Tafur, L., Sadian, Y., Hoffmann, N.A., Jakobi, A.J., Wetzel, R., Hagen, W.J., Sachse, C., and Muller, C.W. (2016). Molecular Structures of Transcribing RNA Polymerase I. Mol Cell *64*, 1135-1143.

Teves, S.S., An, L., Hansen, A.S., Xie, L., Darzacq, X., and Tjian, R. (2016). A dynamic mode of mitotic bookmarking by transcription factors. Elife *5*.

The Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature, 487(7407):330-7

The Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. Nature, 490(7418):61-70

Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 75, 1895-1904.

Thul, P.J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Bjork, L., Breckels, L.M., *et al.* (2017). A subcellular map of the human proteome. Science *356*.

Tiku, V., Jain, C., Raz, Y., Nakamura, S., Heestand, B., Liu, W., Spath, M., Suchiman, H.E.D., Muller, R.U., Slagboom, P.E., *et al.* (2016). Small nucleoli are a cellular hallmark of longevity. Nat Commun *8*, 16083.

Trcek, T., Larson, D.R., Moldon, A., Query, C.C., Singer, R.H. (2011). Single-molecule mRNA decay measurements reveal promoter-regulated mRNA stability in yeast. Cell, 147(7):1484-1497

Trowbridge, J.J., Snow, J.W., Kim, J., and Orkin, S.H. (2009). DNA Methyltransferase 1 Is Essential for and Uniquely Regulates Hematopoietic Stem and Progenitor Cells. Cell Stem Cell 5, 442-449.

Tukiainen, T., Pirinen, M., Sarin, A.P., Ladenvall, C., Kettunen, J., Lehtimaki, T., Lokki, M.L., Perola, M., Sinisalo, J., Valchopoulo, E. (2014). Chromosome X-wide association study identifies loci for fasting insulin and height and evidence for incomplete dosage compensation. Plos Genetics, 10:e1004127

Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., *et al.* (2015). Tissue-based map of the human proteome. Science 347.

Unwin, R.D., Craven, R.A., Harnden, P., Hanrahan, S., Totty, N., Knowles, M., Eardley, I., Selby, P.J., Banks, R.E. (2003). Proteomic changes in renal cancer and co-ordinate demonstration of both the glycolytic and mitochondrial aspects of the Warburg effect. Proteomics, 3(8):1620-1632

Uribarri, J., Woodruff, S., Goodman, S., Cai, W., Chen, X., Pyzik, R. (2010). Advanced glycation end products in foods and a practical guide to their reduction in the diet. Journal Am. Diet Association, 110:911-916

van Wijk, R., and van Solinge, W.W. (2005). The energy-less red blood cell is lost: erythrocyte enzyme abnormalities of glycolysis. Blood 106, 4034-4042.

Vander Heiden, M.G., Cantley, L.C., Thompson, C.B. (2009). Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. Science, 324(5930):1029-1033

Vavouri, T., Semple, J.I., Garcia-Verdugo, R., Lehner, B. (2009). Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell, 138:198-208

Vedadi, M., Niesen, F.H., Allali-Hassani, A., Federov, O.Y., [...], Nordlung, P., Sundstrom, M., Weigelt, J., Edwards, A.M. (2006). Chemical screening methods to identify ligands that promote protein stability, protein crystalization, and structure determination. Proceedings of the National Academy of Sciences of the United States of America, 103: 15835-15840

Vilchez, D., Boyer, L., Morantte, I., Lutz, M., [...], Gage, F.H., and Dillin, A. (2012). Increased proteasome activity determines human embryonic stem cell identity. Nature, 489(7415): 304–308.

Vissher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era- concepts and miconceptions. Nature Reviews Genetics, 9(4):255-266

Vlassara, H. (2005). Advanced glycation in health and disease: role of the modern environment. Ann. N.Y. Acad. Sci., 1043:452-460

von Appen, A., Kosinski, J., Sparks, L., Ori, A., DiGuilio, A.L., Vollmer, B., Mackmull, M.T., Banterle, N., Parca, L., Kastritis, P., *et al.* (2015). In situ structural analysis of the human nuclear pore complex. Nature *526*, 140-143.

Wagih, O., Reimand, J., Bader, G.D. (2015). MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. Nature methods,12(6):531-533

Wagner, W., Bork, S., Horn, P., Krunic, D., Walenda, T., Diehlmann, A., Benes, V., Blake, J., Huber, F.X., Eckstein, V., *et al.* (2009). Aging and Replicative Senescence Have Related Effects on Human Stem and Progenitor Cells. PLoS One 4.

Wagner, W., Horn, P., Bork, S., and Ho, A.D. (2008a). Aging of hematopoietic stem cells is regulated by the stem cell niche. Exp Gerontol 43, 974-980.

Wagner, W., Horn, P., Castoldi, M., Diehlmann, A., Bork, S., Saffrich, R., Benes, V., Blake, J., Pfister, S., Eckstein, V., *et al.* (2008b). Replicative Senescence of Mesenchymal Stem Cells: A Continuous and Organized Process. PLoS One 3.

Wang, J., Ma, Z., Carr, S.A., Mertins, P., [...], Rodland, K.D., Liebler, D.C., Zhang, B. (2017). Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. Molecular Cell Proteomics, 16(1):121-134

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science, 350(6264):1096-101

Ward, M.C., Gilad, Y. (2017). Human genomics: Cracking the regulatory code. Nature, 550(7675):190-191

Webb, S. (2018). Deep learning for biology. Nature, Technology Feature

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Elrott, K., [...] (2013). The cancer genome atlas pan-cancer analysis project. Nature Genetics, 45:1113-1120

Werner, T., Becher, I., Sweetman, G., Doce, C., Savitski, M.M., and Bantscheff, M. (2012). High-resolution enabled TMT 8-plexing. Anal Chem *84*, 7188-7194.

Werner, T., Sweetman, G., Savitski, M.F., Mathieson, T., Bantscheff, M., and Savitski, M.M. (2014). Ion coalescence of neutron encoded TMT 10-plex reporter ions. Anal Chem *86*, 3594-3601.

Whitacre, CC. (2001). Sex differences in autoimmune disease. Nature Immunology, 2:777-780

Whitford, D. (2005). Proteins: structure and function. John Wiley & Sons, 66-74

Wickner, S., Maurizi, M.R., and Gottesman, S. (1999). Posttranslational quality control: folding, refolding, and degrading proteins. Science (New York, N.Y.), 286(5446):1888–93

Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A *et al* (2014) Mass-spectrometry-based draft of the human proteome. Nature 509: 582–587

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A.M., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., *et al.* (2014). Mass-spectrometry-based draft of the human proteome. Nature 509, 582-+.

Wilkie, G.S., Korfali, N., Swanson, S.K., Malik, P., Srsen, V., Batrakou, D.G., de las Heras, J., Zuleger, N., Kerr, A.R., Florens, L.*, et al.* (2011). Several novel nuclear envelope transmembrane proteins identified in skeletal muscle have cytoskeletal associations. Mol Cell Proteomics *10*, M110 003129.

Willhelm, M., Hahne, H., Savitski, M., Marx, H., Lemeer, S., Bantscheff, M., Kuster, B. (2017). Willhelm et al. Reply. Nature, Brief Communication Arising, 547:E23

Williams, E.G. Wu, Y., Jha, P., Dubuis, S., Blattmann, P., Argmann, C.A., Houten, S.M., Amariuta, T., Wolski, W., Zamboni, N., Aebersold, R., Auwerx, J. (2016) Systems proteomics of liver mitochondria function. Science, 352(6291):aad0189

Wilson, J.E. (2003). Isozymes of mammalian hexokinase: structure, subcellular localization and metabolic function. J Exp Biol 206, 2049-2057.

Winkler, I.G., Sims, N.A., Pettit, A.R., Barbier, V., Nowlan, B., Helwani, F., Poulton, I.J., van Rooijen, N., Alexander, K.A., Raggatt, L.J., *et al.* (2010). Bone marrow macrophages maintain hematopoietic stem cell (HSC) niches and their depletion mobilizes HSCs. Blood 116, 4815-4828.

Worboys, J. D., Sinclair, J., Yuan, Y., and Jørgensen, C. (2014). Systematic evaluation of quantotypic peptides for targeted analysis of the human kinome. Nat. Methods, 11(10):1041–1044.

Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., Snyder, M. (2013). Variation and genetic control of protein abundance in humans. Nature, 499(7456):79-82

Wu, W., Wu, Q., Hong, X., Zhou, L., Zhang, J., You, L., Wang, W., Wu, H., Dai, H., and Zhao, Y. (2015). Catechol-O-methyltransferase, a new target for pancreatic cancer therapy. Cancer Sci *106*, 576-583.

Wu, Y., Williams, E.G., Dubuis, S., Mottis A, Jovaisaite V, Houten SM, Argmann CA, Faridi P, Wolski W, Kutalik Z, Zamboni N, Auwerx J, Aebersold R. (2014). Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. Cell, 158(6):1415-1430

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, Sc., [...], Goddard, M.E., Visscher, P.M. (2010). Common SNPs explain a large proportion of heritability for human height. Nature Genetics, 42(7):565-569

Yao, Q., Li, H., Liu, B.Q., Huang, X.Y., and Guo, L. (2011). SUMOylation-regulated protein phosphorylation, evidence from quantitative phosphoproteomics analyses. J Biol Chem *286*, 27342-27349.

Yates, J. R., C. I. Ruse, and A. Nakorchevsky, 2009, Proteomics by mass spectrometry: approaches, advances, and applications: Annu Rev Biomed Eng, v. 11, p. 49-79.

Yeong, F.M. (2013). Multi-step down-regulation of the secretory pathway in mitosis: a fresh perspective on protein trafficking. Bioessays *35*, 462-471.

Yokoyama, H., Nakos, K., Santarella-Mellwig, R., Rybina, S., Krijgsveld, J., Koffa, M.D., and Mattaj, I.W. (2013). CHD4 is a RanGTP-dependent MAP that stabilizes microtubules and regulates bipolar spindle formation. Curr Biol *23*, 2443-2451.

Yokoyama, H., Rybina, S., Santarella-Mellwig, R., Mattaj, I.W., and Karsenti, E. (2009). ISWI is a RanGTP-dependent MAP required for chromosome segregation. J Cell Biol *187*, 813-829.

Yu, A., Shibata, Y., Shah, B., Calamini, B., Lo, D.C., Morimoto, R.I. (2014). Protein aggregation can inhibit clathrin-mediated endocytosis by chaperone competition. Proceedings of the National Academy of Sciences of the United States of America, 11(15):E1481-E1490

Zeevi, D., Korem, T., Zmora, N., Israeli, D., [...], Segal, E. (2015). Personalized Nutrition By Prediction of Glycemic Responses. Cell, 163(5):1079-1094

Zeng, S.X., Li, Y., Jin, Y., Zhang, Q., Keller, D.M., McQuaw, C.M., Barklis, E., Stone, S., Hoatlin, M., Zhao, Y., *et al.* (2010). Structure-specific recognition protein 1 facilitates microtubule growth and bundling required for mitosis. Mol Cell Biol *30*, 935-947.

Zhang, B., Wang, J., Wang, X., [...], NCI CPTAC. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature, 513(7518):382-7

Zhang, H., Liu, T., Zhang, Z., [...], CPTAC Investigators. (2016) Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell, 166(3):755-765

Zhang, H.S., Cao, E.H., and Qin, J.F. (2005). Homocysteine induces cell cycle G1 arrest in endothelial cells through the PI3K/Akt/FOXO signaling pathway. Pharmacology *74*, 57-64.

Zhao, S., Xu, W., Jiang, W., Yu, W., Lin, Y., Zhang, T., Yao, J., Zhou, L., Zeng, Y., Li, H., Li, Y., Shi, J., An, W., Hancock, S.M., He, F., Qin, L., Chin, J., Yang, P., Chen, X., Lei, Q., Xiong, Y., and Guan, K.L. (2010). Regulation of cellular metabolism by protein lysine acetylation. Science (New York, N.Y.), 327(5968):1000–4