PhD degree in Systems Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples “Federico II”

CAUSES OF ANEUPLOIDY IN TUMOURS AND THE CONSEQUENCES ON GENE EXPRESSION AND PROTEIN COMPLEX STOICHIOMETRY

Settore disciplinare: BIO/11

Gökçe Senger

Tutor: Dr. Martin Schaefer

European Institute of Oncology (IEO)

PhD Coordinator: Prof. Saverio Minucci

Anno accademico 2022-2023

Table of Contents

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Martin Schaefer, for always being there when I needed his support, reviewing my progress constantly, and guiding me through my PhD studies. From an academic perspective, I have learned, through his guidance, how to deal with obstacles along the academic path and understood my way of working efficiently.

I would like to extend my sincere thanks to my internal advisor, Assoc. Prof. Stefano Santaguida, and my external advisor, Prof. Dr. Pedro Beltrao, for their invaluable assistance and insightful guidance throughout my PhD studies.

Besides, I would like to thank my thesis examiners, Asst. Prof. Colm Ryan and Prof. Dr. Tiziana Bonaldi, for accepting to be on my thesis examination committee members in despite of their strict schedule. It should be also noted that their encouragement and challenging questions were most appreciated.

I would like to offer my special thanks to Francesca Fiore and Veronica Viscardi for efficiently organising the process and being available whenever their help needed.

I would also like to thank all my colleagues from Schaefer Lab and those with whom I had shared the office. They made this PhD journey enjoyable, and it was a pleasure to work with them.

I owe a special thanks to my family who have no idea what I have been doing abroad - thousand kilometres apart from them - but, yet, have always supported my decisions and encouraged me in finding my own way.

List of abbreviations

ACC	Adrenocortical carcinoma
APC/C	Anaphase promoting complex/cyclosome
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CCLE	Cancer Cell Line Encyclopedia
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
cLADs	Constitutive LADs
CNAs	Copy number alterations
CNVs	Copy-number variations
COAD	Colon adenocarcinoma
COREAD	Colorectal cancer
cpm	Counts per million
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CRISPR	Clustered regularly interspaced short palindromic repeats
DLBC	Diffuse large b-cell lymphoma
ED	Exponentially degraded
eGTEx	The Enhancing GTEx
ESCA	Esophageal carcinoma
FDR	False discovery rate
FPKM	Fragments per kilobase of exon per million reads mapped
GBM	Glioblastoma multiforme
GDC	Genomic Data Commons
GEO	Gene Expression Omnibus
GO	Gene ontology
GTEx	Genotype-Tissue Expression
HCC	HBV-related hepatocellular carcinoma
HGNC	HUGO Gene Nomenclature Committee
HIPPIE	Human Integrated Protein-Protein Interaction rEference
HNSC	Head and neck squamous cell carcinoma
Interactome INSIDER	INtegrated Structural Interactome and genomic Data browsER
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LADs	Lamina-associated domains
LAML	Acute myeloid leukemia
LGG	Lower grade glioma
LIHC	Liver hepatocellular carcinoma

log2FC	Log2 fold change
LUAD	lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
ML	Machine learning
MMU 16	Mouse chromosome 16
NED	Non-exponentially degraded
OGs	Oncogenes
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCA	Principal component analysis
PCPG	Pheochromocytoma and paraganglioma
PPIs	Protein-protein interactions
PRAD	Prostate adenocarcinoma
PTMs	Post-translational modifications
RCSB PDB	Protein Data BankResearch Collaboratory for Structural Bioinformatics Protein Data Bank database
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
SNP	Single nucleotide polymorphism
STAD	Stomach adenocarcinoma
TCGA	The Cancer Genome Atlas
TFs	Transcription factors
TGCT	Testicular germ cell tumours
THCA	Thyroid carcinoma
THYM	Thymoma
TMT	Tandem mass-tags
TPM	Transcripts per million
TS	Tumour suppressor
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
UVM	Uveal melanoma

Disclaimer

I hereby declare that all information in this PhD thesis has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that two main sections; Materials and Methods (except Materials and Methods subsections 13, 14, and 15) and Results - Chapter I, are mainly adapted from my own original works that I conducted during my PhD, and are published (Senger et al., 2022; Senger & Schaefer, 2021). Further, I have fully cited and referenced all the sources used in the reference section.

Name, Surname: Gökçe Senger

Signature: ____________

Abstract

Protein complexes are dynamic assemblies in which proteins bind each other in different physiological cell conditions and stoichiometries to perform cellular functions. Failures in maintaining complex stoichiometries cause proteotoxic stress, and are associated with proliferative and survival disadvantages in normal cells. Tumours are characterised by massive dysregulation of genes leading to imbalances in protein dosage and thus in protein complex stoichiometries. Aneuploidies, arm- or chromosome-level copy number aberrations, are one of the main reasons for transcriptional dysregulation, yet paradoxically they frequently occur in cancer genomes. We use aneuploid tumours, harbouring chromosome-level amplifications and deletions, as a model to understand how tumour cells compensate for aneuploidy induced transcriptional dysregulation. We observe that regulation of co-complex members in trans acts as a compensatory mechanism to deal with abundance changes on the aneuploid chromosome itself. We show that this compensation is stronger for aggregation-prone proteins of aneuploid chromosomes and those involved in a smaller number of complexes suggesting the role of protein complex organisation in modulating those compensatory mechanisms. Further, we provide evidence that this compensation in aneuploid tumours is established through post-translational regulation, and that higher degree of success in this compensation is associated with better tumour fitness, and failure results in activation of protein degradation programs.

However, it is still unclear why aneuploidies and focal copy number alterations are occurring repeatedly in cancer genomes if they cause those compensation problems through dysregulation. To address this, we ask if we can model the observed frequencies of genomic amplifications and deletions as a function of avoiding transcriptional dysregulation of co-complex members or affecting large number of genes, as an example of negative selection, probability of occurrence (e.g. distance to telomere/centromere), and amplifying oncogenes or deleting tumour suppressor genes, representing positive selection, by using machine learning models. We find a balance among these factors in explaining to a certain degree the observed genomic alteration patterns in cancer genomes.

Taken together, our findings describe the need for compensation mechanisms to deal with the imbalances in protein complex stoichiometry induced by aneuploidy, and highlight the importance of protein complex components as potential vulnerabilities for the identification of drug targets for clinical use, in addition to providing insights into understanding tumour genome evolution and factors driving frequently observed genomic alterations.

Introduction

1. The cancer aneuploidy paradox

Cancer is a system characterised by somatic molecular alterations, which could span from single nucleotide mutations to large copy number alterations (CNAs) – gain or loss in DNA copies. Aneuploidies, a special case of CNAs defined as arm- or chromosome-level alterations, have a significant impact on the expression of numerous genes most directly by providing an additional or a reduced amount of gene copies, thereby introducing imbalances at the level of transcriptome and proteome. Under normal cellular conditions, aneuploidy is detrimental, primarily due to the disruptions it causes in the balance of protein complexes which are functional units of the cell (Brennan et al., 2019; Santaguida & Amon, 2015). For example, all human monosomies, absence of one chromosome from a pair, and majority of trisomies, having an additional copy of a chromosome, are either embryonic lethal or have very limited life span (Pai et al., 2003). Paradoxically, aneuploidy is a common feature of cancer cells: Approximately 90% of solid tumours and 65% of blood cancers exhibit aneuploid karyotypes (Ben-David et al., 2019; Garribba et al., 2023; Taylor et al., 2018). Moreover, arm-level aneuploidies encompass 22.5% of cancer genome which is more than any other somatic molecular alterations, as well as they represent 92% of the most frequent CNAs based on the data from more than 10,000 The Cancer Genome Atlas (TCGA) tumours (Shih et al., 2023). However, how tumours compensate for the extensive transcriptomic and proteomic changes induced by aneuploidy, and how aneuploidy contributes to tumour evolution are still poorly understood.

2. Expression changes in aneuploid karyotypes

Transcriptome profiling in yeast model organisms revealed that aneuploidies are to a certain extent directly affecting gene expression levels. Hughes and colleagues reported that all genes located on the aneuploid chromosome showed corresponding expression changes and concluded that there is no dosage compensation for genes affected by copy number changes at the transcriptome level (Hughes et al., 2000). Together confirming these results, Pavelka and colleagues showed that copy number induced alterations at the transcriptome level are transmitted to the proteome level (Pavelka et al., 2010). On the other hand, it has been shown that there is a buffering effect at the proteome level regulating the level of proteins encoded by copy number altered genes, particularly for the ones involved in protein complexes (Dephoure et al., 2014; Stingele et al., 2012). Studies in aneuploid human cell lines revealed similar results: Correlated changes in mRNA expression levels of genes following their copy number alterations while there is a buffering at the proteome level (Schukken & Sheltzer, 2022; Stingele et al., 2012). These results were further supported by findings in human/mouse hybrid cell lines (Upender et al., 2004), in mouse (Williams et al., 2008), as well as for CNAs in human tumour samples (Gonçalves et al., 2017).

Previous studies have unveiled that aneuploidy does not only affect the expression profiles of genes on altered chromosomes but also has a substantial influence on the expression of genes located on chromosomes outside of the aneuploid ones in normal human cells (Nawata et al., 2011), as well as in cancer cells (Upender et al., 2004). In addition, Saran and colleagues showed, in aneuploid mouse models with an extra copy of mouse chromosome 16 (MMU 16), a global transcriptional dysregulation across the entire genome in addition to the effects on genes residing on MMU 16 (Saran et al., 2003), further indicating a widespread destabilisation in gene expression levels. Interestingly, for CNAs, it has been demonstrated a correlated increase between abundance of proteins encoded by copy number altered genes and abundances of their protein complex members encoded by genes outside of the copy number changed regions (Gonçalves et al., 2017), raising the question of whether this could contribute to the expression changes in aneuploid cells even on diploid chromosomes. Despite these significant insights, our comprehension of how aneuploidy impacts the expression of genes in a genome-wide manner, particularly in the context of cancer, remains incomplete.

3. Buffering the effect of copy number alterations and aneuploidy

The ability of a cell to maintain internal balance in gene expression and protein synthesis, and perform biological functions in the face of perturbations on its genome is crucial for cellular fitness and survival. Considering the global effect of aneuploidy and CNAs on cell transcriptome and proteome, it is important to understand buffering mechanisms optimising gene expression and corresponding protein abundances in aneuploid cancer cells. Previous studies on aneuploid yeast strains have demonstrated that copy numbers of genes are substantially correlated with mRNA levels but not directly with protein levels (Chino et al., 2013; Dephoure et al., 2014; Torres et al., 2007). Another study has provided insights for the underlying mechanisms of this protein-level dosage compensation. Ishikawa and colleagues demonstrated that, in aneuploid yeast, around 10% of genes are subject to dosage compensation at the proteome level and those dosage-compensated genes are enriched in protein complex subunit encoding genes (Ishikawa et al., 2017). Moreover, their experimental efforts on the ubiquitin-proteasome system and ribosome profiling led to the conclusion that the main mechanism of the dosage compensation is protein degradation by the ubiquitin-proteasome system rather than translational efficiency (Ishikawa et al., 2017). Integration of copy-number, transcriptome and proteome measurements for TCGA tumours revealed that CNAs often affect protein levels; however, 23%-33% of proteins are post-transcriptionally attenuated (Gonçalves et al., 2017). Furthermore, they experimentally showed that some complex subunits act as rate-limiting factors for the complex assembly resulting in coordinated protein levels of the members of the same complexes (Gonçalves et al., 2017). These results further support the role of post-transcriptional mechanisms in protein-level dosage compensation also for CNAs. Recently, an effort on understanding the effect of arm-level aneuploidies on proteome of human cancer cell lines has also identified post-transcriptional regulation as a dosage compensation mechanism in response to aneuploidy (Schukken & Sheltzer, 2022).

Consequently, these results are suggesting a profound role of post-transcriptional and -translational regulation in the buffering of protein abundance changes induced by CNAs and aneuploidies, as well as highlighting the role of protein complex formation in the resulting amount of protein level. In the previous paragraph, we mentioned studies that revealed the role of degradation pathways in this buffering. One of the reasons for degradation of excess amounts of proteins (Ishikawa et al., 2017) could be the need for avoiding free aggregation-prone hydrophobic interface surfaces that could further cause proteotoxicity (Young et al., 1994). On the other hand, experimental efforts to understand protein degradation kinetics have revealed proteins with different degradation profiles; non-exponentially degraded (NED) and exponentially degraded (ED) (McShane et al., 2016). Moreover, they identified the majority of NED proteins as subunits of protein complexes and showed that they are overproduced relative to other members of the same complex and then the excess amount is degraded (McShane et al., 2016). This is in contrast to proportional synthesis appearing in bacteria (G.-W. Li et al., 2014). In addition to that they characterised NED proteins with larger interface size and faster assembly time than ED proteins, they demonstrated an increase in the initial degradation of NED proteins upon amplification of their encoding genes (McShane et al., 2016). This further provides new insights on understanding the dynamics between properties of proteins and their degradation profiles. Indeed, Sousa and colleagues showed that structural properties of proteins have an effect on the degree of post-translational buffering on changes induced by CNAs in TCGA cancer patients: Proteins with larger interface size are subject to larger degree of buffering and post-translational modifications (PTMs), phosphorylation in particular, could modulate this degradation by affecting protein-protein interactions (PPIs) (Sousa et al., 2019). All together, these findings indicate that properties of proteins and the ways they organise into complexes constrain the dysregulation patterns of protein abundances and affect protein-level dosage compensation upon genomic copy number changes. However, we still lack a full characterisation of organisation of proteins into complexes and their properties.

4. Protein complex organisation

Protein complexes are not static cellular units: They assemble and disassemble dynamically. This dynamic nature adds complexity, and makes the characterisation of PPIs within complexes more challenging. It has been shown that the cellular abundance and localization of proteins, and the strength of binding interaction between proteins (binding affinity) affect the protein complex formation (Nooren & Thornton, 2003). For example, quantitative proteomics analysis on the anaphase promoting complex/cyclosome (APC/C) across different cells lines has revealed that stable (core) complex subunits are associated with higher cellular abundances and found in the complex with unique ratio to each other (stoichiometry) (Hein et al., 2015). On the other hand, the subunit KIAA1430 was defined as a transient interactor of the APC/C and can be found in the complex with different stoichiometric ratios (Hein et al., 2015). Furthermore, a spatiotemporal analysis of protein complex stoichiometries in cancer and mouse cells has demonstrated that more than half of the well-characterised protein complexes are subjected to stoichiometric changes depending on cell types, and subunits of the complexes are variable in space and time (Ori et al., 2016). More recently, an extensive characterization of the human interactome in two different cell lines has revealed that shared interactions are enriched in complexes with essential, conserved functions while cell-line-specific PPIs play a role as rewiring these core complexes (Huttlin et al., 2021). Together these results further highlight the dynamic organisation of proteins in a context-specific manner.

Since the pioneering work on understanding the principles of PPIs in complexes by Jones and Thornton in 1996 (Jones & Thornton, 1996), many studies have focused on formalising PPIs based on binding affinity, composition and stability of the complex. These efforts have resulted in three main classes: (i) homo- or hetero-oligemers based on comprising identical or non-identical proteins, respectively (Nooren & Thornton, 2003), (ii) non-obligate and obligate interactions based on the binding affinity (Acuner Ozbabacan et al., 2011; Nooren & Thornton, 2003), and (iii) transient and permanent interactions based on the stability (lifetime of the interaction) (Acuner Ozbabacan et al., 2011; Mintseris & Weng, 2003; Nooren & Thornton, 2003). In addition to the direct classification based on structural and biophysical properties, some studies have focused on indirect classification in which functional annotations, co-expression patterns, and genetic interactions are considered to define functional protein networks (Chatr-Aryamontri et al., 2013; Eisenberg et al., 2000; Franceschini et al., 2013). All these efforts together with individual experiments on characterisation of PPIs brought the need for systematic collection of the information under a platform. To address this need some important publicly available databases have been developed: CORUM is a database providing the largest and most comprehensive dataset of manually curated experimentally characterised mammalian protein complexes, which is mainly composed of human complexes (Giurgiu et al., 2019), the Research Collaboratory for Structural Bioinformatics Protein Data Bank database (RCSB PDB) storing experimentally-determined three-dimensional structures of protein complexes (RCSB.org) (Berman et al., 2000). Another example is INtegrated Structural Interactome and genomic Data browsER (Interactome INSIDER) which is an integrative source for proteome-wide human interactome together with structural information and mutation/variant data and allows users to study disease on a genomic and proteomic scale (Meyer et al., 2018).

Understanding the 3D structure of protein complexes and dynamics of PPIs within complexes is crucial for studying cancer biology, defining vulnerabilities in cancer proteome and for validating drug targets. A quantitative proteomic study on cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) explored associations between genetics and abundance of protein complex subunits and revealed complexes which are sensitive to gene knockdown and mutations (Nusinow et al., 2020). Moreover, proteogenomic profiling of lung adenocarcinoma (LUAD) tumour samples showed that protein abundance levels are more associated with patient survival than mRNA levels, and proteomic regulatory networks contribute to identify potential therapeutic vulnerabilities among subtypes of LUAD (Soltis et al., 2022). All together these findings highlight the importance of gaining a comprehensive understanding of protein complex organisation and characteristics of proteins for identifying vulnerabilities in cancer and developing effective treatment strategies.

5. Tissue-specific recurrence of copy number alterations and aneuploidies

Different cancer types acquire distinct genomic amplification and deletion patterns suggesting that cancer genomes undergo tissue-specific rearrangements due to different selective pressures or varying occurrence probabilities of these rearrangements (Ben-David & Amon, 2020; Taylor et al., 2018). Understanding which genomic alterations are subjected to selection and which are more likely to occur in regions with specific (epi)genomic characteristics is crucial, as it helps to identify the specific changes contributing to tumorigenesis. A recent work revealed that tumour mutational burden has a pivotal role in shaping the tissue-specific evolution of CNAs in cancer: Positive selection favours specific genomic regions to be amplified so that these extra copies could serve as buffers to compensate for the deleterious effects of coding mutations in mutation-rich and essential regions (Alfieri et al., 2023). Additionally, a role for density of tumour suppressors (TSs), oncogenes (OGs) and essential genes, which are affecting proliferation upon knockout, in the observed amplification and deletion patterns has been previously identified in cell line models (Davoli et al., 2013; Sack et al., 2018).

Another critical contributor to the observed alteration patterns in cancer is tissue-specific gene expression and epigenomics. Patkar and colleagues demonstrated a correlation between alteration patterns specific to cancer types and gene expression levels in the normal tissue of tumour origin (Patkar et al., 2021). Another study has unveiled the role of lamina associated domains (LADs) in influencing the probability of chromosome segregation errors, consequently affecting the occurrence of aneuploidy patterns (Klaasen et al., 2022). Furthermore, heterochromatin structure in the tissue-of-origin has been identified as a significant determinant of the position of DNA double-strand breaks and thus influencing the resulting CNA patterns in a tissue-specific manner (Cramer et al., 2016). These findings highlight the contribution of biological data from different dimensions (genome, epigenome, transcriptome, phenotype) to the observed CNA and aneuploidy patterns in cancer genomes. Collectively, a systematic analysis of the factors contributing to these alteration patterns is needed to understand their roles in tumour genome formation and evolution, and could potentially enhance our ability to predict tissue-specific responses to therapy.

6. Application of machine learning methods in cancer biology

Machine learning (ML) methods have become important in biology as they enable us to integrate data from different dimensions of biological systems, and to capture relevant information to understand complex systems like cancer (Zitnik et al., 2019). The idea behind machine learning is that the relationship between a dependent variable and several independent variables are automatically learned by identifying patterns and correlations from a subset of the data, and making predictions for the full set. This is an iterative process that enables ML algorithms to improve their performance, resulting in more accurate predictions and models. One way of using ML is to predict biologically relevant outcomes. For example, a recent study employed ML methods to predict cancer types from DNA copy number variation (CNV) data (Attique et al., 2022). Moreover, ML algorithms were used in prediction cancer dependencies by integrating different genomic profiles such as gene expression, methylation level, and CNAs (Chiu et al., 2021) as well as in prediction of cancer driver genes (Han et al., 2019; Luo et al., 2019). Another application area of ML methods is identifying important features driving prediction. Jubran and colleagues implemented ML models to estimate the relative importance of various tissue-specific features in shaping the observed arm-level aneuploidy patterns in cancer genomes (Jubran et al., 2023). Different implementations of ML algorithms allow us to capture diverse relationships in complex datasets. For instance, linear regression based methods assume a linear relationship between input features and a continuous numerical output, and are not suitable for classification tasks. In contrast, decision tree algorithms can capture complex relationships in the data, including non-linear patterns even in high-dimensional spaces (Krzywinski & Altman, 2017).

7. Overview of the thesis work

In this thesis work, we conduct an integrative study to understand to which degree the features of proteins and their interactions within complexes limit protein abundance changes induced by tumorigenesis in general and chromosome-level aneuploidies in particular with a focus on genome-wide protein level changes by integrating aneuploidy, transcriptomic, proteomic, and structural proteomics data. We show that a large number of proteins encoded on chromosomes outside of the altered (aneuploid) ones have correlated abundance changes with their complex members encoded on aneuploid chromosomes.

Furthermore, the characteristics of the aneuploid proteins constrain the degree of co-abundance regulation in which aggregation-prone proteins and those involved in smaller numbers of complexes are related to stronger degrees. We further provide evidence that post-translational mechanisms play a substantial role in this co-abundance regulation, and the degree of success in this co-abundance regulation is related to patient survival. Finally, we fit multivariable ML models in which we combine multiple tissue- and genomic-location-specific features to predict frequently observed CNA and aneuploidy patterns in cancer genomes. Taken together, our results highlight the role of protein complex organisation in mitigating stoichiometric imbalances in protein complexes in cancer and aneuploid tumours, and the need for a systematic analysis of determinants of recurrently occurred genomic alteration patterns. Overall these findings might guide our understanding of tumour genome evolution and, ultimately, therapy response.

Materials and Methods

1. Data availability

The code performing all the analyses is available at Github repository, and the corresponding links are provided in the Data availability section of our published works (Senger et al., 2022; Senger & Schaefer, 2021). Supplementary files are adapted from our published works, re-enumerated and available at github.https://github.com/SengerG/PhD_Thesis_Senger.git.

2. Data processing

Transcriptomic data, RNA-Seq fragments per kilobase of exon per million reads mapped (FPKM) values for 11,007 tumour samples analysed in TCGA, were downloaded from the NCI Genomic Data Commons (GDC) (Grossman et al., 2016). To make transcriptomic values more comparable across samples, FPKM values were converted to transcripts per million (TPM) by normalising for gene length, and then for sequencing depth. For downstream analyses, only primary tumour samples were selected (n = 9830), and Ensembl gene IDs were mapped to gene symbols based on Human genome version GRCh38.p13 downloaded from Ensembl BioMart (Howe et al., 2021). For gene symbols mapping to more than one Ensembl IDs, the mean value was taken. Then, mitochondrial genes and the ones with zero TPM values for all samples were filtered out.

HTSeq counts for TCGA tumour and matched normal samples, comprising 21 cancer types, were downloaded from GDC. Normalisation was done by using the cpm (counts per million) function from the edgeR package (Y. Chen et al., 2016; McCarthy et al., 2012; Robinson et al., 2010).

Proteomics data used in this thesis study were generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (NCI/NIH). Tandem mass-tags (TMT)-based log-transformed proteomics data for tumour and normal adjacent tissue samples (for cohorts proteomic profiling was done for both tissues, and for the ones for which the confirmatory and/or discovery study is available, only the representative study was considered) comprising colon (COAD) (Vasaikar et al., 2019) (8,067 proteins for 96 tumour and 96 matched normal samples), HBV-related hepatocellular carcinoma (HCC) (Q. Gao et al., 2019) (6,478 proteins for 159 tumour and 159 matched normal samples) and LUAD (Gillette et al., 2020) (10,699 proteins for 110 tumour and 101 matched normal samples), and proteomics data for the available TCGA cohorts: Spectral counts for colorectal cancer (COREAD) (The Cancer Genome Atlas Network, 2012; the NCI CPTAC et al., 2014) (5,561 proteins and 90 samples), relative abundances for ovarian serous cystadenocarcinoma (OV) (Cancer Genome Atlas Research Network, 2011; Zhang et al., 2016) (7,169 proteins and 174 samples) and for breast invasive carcinoma (BRCA) (Cancer Genome Atlas Network, 2012;

Mertins et al., 2016) (10,625 proteins and 105 samples), were downloaded from the CPTAC consortium. Further processing was done when it was necessary: For the COAD cohort, proteins that were quantified in less than 50% of the samples were filtered which left us 6,554 proteins out of 8,067, for the LUAD cohort, mean log-transformed TMT values were considered when more than one proteins mapped to the same gene symbol, leaving 10,316 out of 10,699 proteins. In addition, for the TCGA COREAD cohort, spectral counts were normalised to make samples more comparable by quantile normalisation followed by log2 transformation. For all the TCGA cohorts, the mean value was considered for the replicated samples, and then samples and genes shared with corresponding transcriptomic data were considered, leaving 5,353 proteins and 88 samples for COREAD, 7,062 proteins and 119 samples for OV, and 10,467 proteins and 105 samples for BRCA.

3. Detecting cancer-type-specific chromosome-level aneuploidies

Arm-level aneuploidy scores for 10,522 TCGA samples, comprising 33 cancer types; adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), BRCA, cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), cholangiocarcinoma (CHOL), COAD, diffuse large b-cell lymphoma (DLBC), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukemia (LAML), lower grade glioma (LGG), liver hepatocellular carcinoma (LIHC), LUAD, lung squamous cell carcinoma (LUSC), mesothelioma (MESO), OV, pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), sarcoma (SARC), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), testicular germ cell tumours (TGCT), thyroid carcinoma (THCA), thymoma (THYM), uterine corpus endometrial carcinoma (UCEC), uterine carcinosarcoma (UCS), uveal melanoma (UVM), were calculated by Taylor et al., 2018 (Taylor et al., 2018) from Affymetrix single nucleotide polymorphism (SNP) 6.0 arrays by using the ABSOLUTE algorithm (Carter et al., 2012). By using this data, we calculated chromosome-level aneuploidy scores as follows: For chromosomes 1-12 and 16-20, the entire chromosome was considered as amplified, deleted, or diploid if both p and q arms are amplified, deleted, or not changed, respectively. For acrocentric chromosomes, 13-15 and 21-22, q arm aneuploidy scores were considered as chromosome-level aneuploidy scores. For samples where chromosome arms have different events (amplification, deletion or no change), or arm-score is missing for one or both arms, NA values were assigned for the corresponding chromosome. COAD and READ samples were considered as one cancer type as COREAD. Chromosome level aneuploidy scores can be found in **Supplementary File 2**.

To detect frequently occurring cancer-type-specific, chromosome-level aneuploidies, we tested the occurrence of each aneuploidy event (amplification/deletion) within each cancer type against random expectation by using chi-square test. Then multiple testing correction was applied on p-values by using Holm's method, and events with adjusted p-value lower than or equal to 0.05, and chi-square standard residual equal to or higher than 2 were selected as cancer-type-specific, chromosome-level aneuploidies. At the end, we detected 203 chromosome-level aneuploidies (86 amplifications and 117 deletions; **Supplementary File 2**). Among those detected aneuploidies, 13 amplifications and 20 deletions comprise COREAD, BRCA, and OV, for which we have both transcriptomic and proteomic data available.

For each of the detected cancer-type-specific, chromosome-level aneuploidies (86 amplifications and 117 deletions), we detected frequently co-altered chromosomes (co-amplifications and co-deletions) by using chi-square test followed by Holm's multiple testing correction. Chromosome pairs with adjusted p-value lower than 0.01, and chi-square standard residual equal to or higher than 2 were considered as significantly frequently co-amplified, resulting with 305 combinations in 60 out of the 86 cancer-type-specific, chromosome-level amplifications and 672 combinations in 90 out of 117 cancer-type-specific chromosome-level deletions (**Supplementary File 2**). To systematically test the contribution of each co-amplified and co-deleted chromosome to the expression changes on other chromosomes, for each of the 60 cancer-type-specific amplifications and 90 cancer-type-specific deletions, we, first, grouped chromosomes as co-amplified/deleted and non-co-amplified/deleted. Then we calculated contribution as the percentage of the differentially expressed genes to the total number of genes on that chromosome. To end this, we compared the mean percentage of differentially expressed genes on co-amplified/deleted chromosomes to that of non-co-amplified/deleted chromosomes by using paired Wilcoxon test.

4. Detecting transcriptomic and proteomic changes

To detect differentially expressed genes and abundant proteins between different groups, we used Wilcoxon test instead of standard tools, which are designed to analyse RNA-seq data (e.g. LIMMA or DeSeq2), for two reasons. The first reason is to be consistent between transcriptomic and proteomic data since the standard bioinformatics tools are designed to perform differential expression analysis on transcriptome data rather than proteome data, and each tool has slightly different algorithms for normalisation and detection of genes obviously introducing different sensitivities and detection biases (e.g. towards more abundant transcripts). The second is to deal with false discovery rate (FDR) in our transcriptome data with large sample size. It has been shown that Wilcoxon test performs better on data with a larger sample size than the standard differential expression analysis tools as they fail to control FDR while the sample size is increasing (Y. Li et al., 2022; Soneson & Delorenzi, 2013).

To detect set of proteins showing abundance changes between tumour and matched normal samples, we used CPTAC cohorts, COAD, HCC, and LUAD, where we have available proteomic data for tumour and matched normal samples, and performed differential protein abundance analysis, separately for each cohort, by using Wilcoxon test followed by Bonferroni multiple testing correction method. Proteins with adjusted p-value less than or equal to 0.05, and absolute log2 fold change (log2FC; calculated as the median difference of log2 transformed TMT-values between tumour and normal samples) greater than 1 were considered as differentially abundant.

To detect transcriptomic and proteomic changes between aneuploid and diploid tumour samples and make them more comparable, we considered TCGA samples where aneuploidy data is available. This resulted in 9266 samples for transcriptome analysis, covering all 32 cancer types, and 298 samples for proteome analysis covering 3 cancer types (COREAD, OV, and BRCA). Then, for each of the detected cancer-type-specific, chromosome-level aneuploidies (see Materials and Methods "Detecting cancer-type-specific chromosome-level aneuploidies"), we first split TCGA samples into two groups; the ones with chromosome amplification or deletion, and the ones diploid for the respective chromosome. Since the selected samples can vary for different detected aneuploidies, for transcriptome data, we further filtered lowly expressed genes which have zero TPM in all selected samples. Then, we performed Wilcoxon test followed by the Benjamini and Hochberg multiple testing correction method. Significantly differentially expressed genes and differentially abundant proteins were selected based on the following criteria: Adjusted p-value is lower than 0.1, and uncorrected Wilcoxon p-value lower than 0.1, respectively for transcriptomic and proteomic changes. For the cases where we were left with less than 250 differentially expressed protein coding genes after adjusted p-value cutoff, the uncorrected p-value was used ($p < 0.05$) in order to have a sufficient number of genes to perform the association tests (see Materials and Methods "Statistical analyses"). Again, for the same reason, we used a relaxed statistical cutoff to select differentially abundant proteins.

To detect differentially expressed genes between tumour and matched normal samples (for 21 TCGA cohorts; BLCA, BRCA, CESC, CHOL, COREAD, ESCA, GBM, HNSC, KIRC, KIRP, LIHC, LUAD, LUSC, PAAD, PCPG, PRAD, SARC, STAD, THCA, THYM, UCEC), we used Wilcoxon test. Genes with uncorrected p-value lower than 0.05 were considered as significantly differentially expressed genes.

5. Gene ontology analysis

To understand which molecular functions frequently dysregulated genes in aneuploidy are associated with, we, first, counted how many times a gene was dysregulated across different aneuploidy cases (203 and 33 detected cancer-type-specific, chromosome-level aneuploidies, respectively, for transcriptomic and proteomic data). Then we performed gene ontology (GO) analysis on the most frequently dysregulated 150 genes in amplification and deletion cases by using WebGestalt (Liao et al., 2019). The GO analysis was performed separately for the gene sets from transcriptomic and proteomic data. All protein coding genes were used as a background.

6. Classifying proteins and protein interactions

Systematic characterisation of proteins and protein-protein interactions was done based on structural data, proteomics measurements in different cellular conditions, human interactome data, literature information, and experimentally-validated human protein complex information obtained from different public sources (each source is indicated in the related section) (**Table 1**). To integrate this classification data with cancer proteomics to perform further analysis, all proteins were kept with their corresponding gene names (ID conversion was done when it was necessary).

6.1. Calculation of stoichiometric ratio

Protein complexes are hierarchical structures in which amino acid chains fold to form individual proteins (subunits) which then interact with each other and assemble into complexes. The RCSB PDB (Berman et al., 2000) is a structural database where experimentally-determined 3D structures of protein complexes are stored and represented based on this hierarchical organisation (each protein complex is called “entry” and identified by a PDB ID, and each protein in a protein complex is called “entity”). We retrieved structural data for available protein complexes in PDB in March 2020 which covers 9,840 PDB entries. Uniprot ID of each entity within each entry was converted to gene name, and only the human protein complexes were considered for further analyses. This left us with 8,388 protein complexes comprising 3,075 proteins. To calculate stoichiometric ratio for protein pairs within the same complex, for all possible protein pairs, we took the ratio of the number of chains of one protein to that of another protein in a pair. Then, we grouped protein pairs as the ones involved in complexes with even (e.g. 1:1, 2:2, 4:4) and with uneven (e.g. 1:2, 3:1, 1:2) stoichiometric ratio (**Supplementary File 1**). We excluded protein pairs involved in complexes, where the stoichiometric ratio can vary between even and uneven, from subsequent analyses.

6.2. Calculation of co-occurrence frequency

To calculate how many times two proteins together participate in the same protein complexes, we, first, obtained experimentally-validated human protein complexes together with their subunit information (2,916 complexes comprising 3,664 proteins) from the CORUM database (CORUM 3.0 current release, September 2018) (Giurgiu et al., 2019). Then co-occurrence frequency, for each protein pair found together in at least one protein complex, was calculated as the ratio of the number of complexes in which the two proteins were found together to the number of complexes in which at least one of them is found. In this way, we aimed to address a potential bias arising from different tendencies of proteins to participate in complexes.

6.3. Context-specific and general interactions

To define proteins that are interacting with each other in an environment-dependent and -independent manner, we obtained proteome-scale, cell-line-specific human interactome data from the BioPlex Interactome (Huttlin et al., 2021) in which protein interactions were profiled via affinity-purification mass spectrometry method in HCT116 and in 293T cells. In total 167,374 protein interactions were obtained from the BioPlex Interactome for these two cell lines. Only the protein interactions that were detected by the baits targeted in both cell lines were considered to prevent possible technical biases. This left us a total number of 33,739 interactions detected in both cell lines, named as general interactions, and 89,330 interactions detected either in HCT116 or in 293T cell lines, named as context-specific interactions.

6.4. Competitive and cooperative interactions

Binding interface residues for experimentally-determined human binary interactions were calculated by using 3D structures obtained from RCSB PDB when available (where interface residues are defined as the ones with a decrease in solvent-accessible surface area equal to or larger than 1.0 Å2 upon binding), or predicted by applying ensembl-based machine learning models, and reported in Interactome INSIDER (Meyer et al., 2018). We extracted this publicly available protein interaction interface data covering 121,575 human binary interactions among 14,380 proteins. After converting Uniprot IDs to gene names, we removed binary interactions that have binding site information only for one protein, leaving us 70,355 binary interactions. To determine competitive and cooperative interactions, we first defined proteins that are interacting with at least one common protein. Then, we counted the number of intersecting residues on the corresponding binding interfaces of the common partner for two proteins. Finally, we normalised it by the total length of binding interfaces of the common partner (Jaccard index as a measure of binding similarity between two proteins) (Eq. 1). Two proteins with Jaccard index equals to or higher than 0.1 are classified as competitive, and cooperative otherwise. For the protein pairs having more than one common partner, the one with the highest Jaccard index was considered for classification.

$$Jaccard\ index\ = \frac{Number\ of\ residues\ in\ A \cap B}{Number\ of\ residues\ in\ A \cup B} \qquad \text{Eq. 1}$$

where A and B represent the corresponding interaction sites on the common partner for protein A and B.

6.5. Transient and permanent interactions

Previous efforts used PDB complexes, and characterised transient and permanent protein-protein interactions based on the stability of a protein complex and physicochemical characteristics of protein interfaces which are determined by machine learning algorithms and/or atomic contact vectors (Block et al., 2006; Mintseris & Weng, 2003). We used this publicly available data, and obtained 147 permanent and 198 transient interactions from Block et al., 2006 (Block et al., 2006), and 209 transient interactions from Minteris and Weng, 2003 (Mintseris & Weng, 2003). Only heterodimers (complexes with non-identical monomers) and human complexes were considered, and if an interaction was defined by both studies, only one of them was kept for further analyses. After this filtering, we obtained 58 transient and 9 permanent interactions.

6.6. Aggregation-prone proteins

Aggregation-prone proteins (n=300) were obtained from Määttä et al., 2020 (Määttä et al., 2020), in which they used a mass spectrometry based proteomics approach to measure solubility of a protein after heat-shock experiments.

6.7. Promiscuous and non-promiscuous proteins

Experimentally-validated human protein complexes together with their subunit information (2,916 complexes comprising 3,664 proteins -subunits-) from the CORUM database (CORUM 3.0; September 2018) (Giurgiu et al., 2019). Then, we counted the number of complexes a protein is involved in for each subunit. Promiscuity was determined based on the number of complexes a protein: If a protein is involved in more than five complexes, it was classified as promiscuous, and as non-promiscuous otherwise.

7. Statistical analyses

For the CPTAC cohorts that proteomic data is available for tumour and matched normal samples (COAD, HCC, and LUAD), a principal component analysis (PCA) was performed by using prcomp function from stats package (version 3.6.2) in R. Standard deviations in protein abundances, for each detected protein in the corresponding cohort, were calculated across tumour and normal samples, separately. Then the distribution of standard deviations in tumour and normal samples were compared by using t-test.

Cancer-type-specific protein abundance correlations, for all possible protein pairs among the detected proteins in the corresponding cohort, were calculated across tumour samples (primary tumour samples for TCGA CPTAC cohorts) for all 6 CPTAC cohorts (COAD, HCC, LUAD, COREAD, OV, and BRCA) by using Spearman method. Correlations were compared (i) between the protein pairs among the members of the same complexes and that of members from different complexes (correlations from different cohorts were pooled), (ii) between different groups of the defined protein and protein interaction classes (see Materials and Methods “Classifying proteins and protein interactions”) (separately for each cohort) by using Wilcoxon test.

In addition, to understand the effect of protein features in dysregulation patterns in aneuploid tumours, for the TCGA CPTAC cohorts in which we detected proteomic changes between aneuploid and diploid tumour groups, we only used correlations between differential abundant proteins encoding by genes found on aneuploid chromosomes and their complex members encoding by genes on other, non-aneuploid chromosomes. At the end, we obtained a unique set by pooling correlations from all three TCGA CPTAC cohorts (COREAD, OV, and BRCA). For the protein pairs for which we could compute a correlation on more than one cancer type, only the maximum absolute correlation was considered. At the end, we obtained 2772 and 3818 protein pair correlations for chromosome-level amplification and deletion cases, respectively (**Supplementary File 3**). Wilcoxon test was used to compare correlations between different groups.

To understand the association between complex proteins and (i) differentially abundant proteins in tumours compared to matched normal samples (for the cohorts COAD, HCC, and LUAD), (ii) differentially abundant proteins of other chromosomes in aneuploid tumours compared to diploid tumour samples (for 13 amplification and 20 deletion cases comprising TCGA CPTAC cohorts; COREAD, OV, BRCA), and the association between complex partners of differentially abundant proteins of aneuploid chromosomes and that of other chromosomes (again only for TCGA CPTAC cohorts), chi-square test was used. We repeated the association tests for transcriptome-level changes in aneuploid tumours (86 amplification and 117 deletion cases comprising all 32 cancer cohorts in TCGA). Since we performed a much larger number of association tests on transcriptome data, multiple testing correction was performed on p-values by using Holm’s method and associations with adjusted p-value lower than 0.01 were considered as significant.

The known human protein complexes from the CORUM database was used to assess complex membership in correlations and association tests analyses.

For each protein pair covered by the structural data obtained from the RCSB PDB (Berman et al., 2000) we tested the relationship between proteins by using a linear regression model. In the model, the dependent variable was the protein abundances of the protein with smaller copy number across tumour samples. For the protein pairs with even stoichiometric ratio, the first protein was considered as the dependent variable. Only the relationships where the coefficient of the dependent variable was significantly different from zero (p-value < 0.05, linear regression model) were considered for the comparison of slopes between protein pairs with even and uneven stoichiometric ratio. Wilcoxon test was used for the comparison.

8. Functional annotations of protein complexes

To investigate functional relevance of co-abundance regulation of complex members of aneuploid proteins in aneuploid tumours, we performed a chi-square test. To do this, we used protein abundance correlations between differentially abundant aneuploid proteins and their complex members of other chromosomes, separately for each detected cancer-type-specific aneuploidy case (13 amplifications and 20 deletions covering TCGA CPTAC cohorts) and, first, classified them into two groups: Top-correlated pairs including 20 strongest positive and negative correlations (40 in total), and a background group comprising protein pairs with correlations between -0.2 and 0.2. Then, for each group, we listed protein complexes, in which protein pairs are involved in, and their functional annotations - associated GO terms - by using the known human protein complex data in the CORUM database. To perform the chi-square test, for each GO term, we tested the number of complexes related to the corresponding term in the top correlated group against that of in the background group. An enrichment score was calculated by dividing the difference of the observed complex number and the expected one obtained from the chi-square test by the square root of the expected value. GO terms with p-value lower than 0.05 were considered as significantly associated.

We further aimed to understand if the relatively higher number of ribosomal genes and/or larger complexes biassed the functional analysis. To test this, we, first, obtained ribosomal genes from the HUGO Gene Nomenclature Committee (HGNC) database (Tweedie et al., 2021). Then, we repeated the chi-square test once by removing protein pairs including ribosomal genes, and once by removing larger complexes which were defined as the ones with more than 10 subunits.

9. Randomization tests

To further assess the statistical significance of enrichment of differentially abundant proteins of other chromosomes in complex members of differentially abundant aneuploid proteins, we performed a randomization test by using binary PPIs. To do this, we retrieved PPI data from Human Integrated Protein-Protein Interaction rEference (HIPPIE) (v2.2) (Alanis-Lobato et al., 2017), and counted the number of physical PPIs between differentially abundant aneuploid proteins and that of other chromosomes, separately for each detected cancer-type-specific aneuploidy case (13 amplifications and 20 deletions). To obtain a background distribution, we replaced the set of differentially abundant aneuploid proteins by an equal size set of proteins with the same degree distribution, and recounted the number of PPIs between the random set and differentially abundant proteins of other chromosomes. This was repeated 100 times. Then, we used background distribution to estimate the p-value by counting how often the original observed value was smaller than or equal to a randomised value.

To test at which degree differential expression of transcription factors (TFs) could explain the expression changes on other chromosomes in aneuploid tumours and overall dysregulation patterns in tumours, we performed a randomization test for 203 detected cancer-type-specific aneuploidies comprising all TCGA cohorts and for cancer types that have available transcriptome data for tumour and matched normal samples in TCGA (21 cancer types). To do this, we first retrieved ENCODE gene-TF associations (1651393 in total) detected by ChIP-Seq experiments from the Harmonizome database (ENCODE Project Consortium, 2004; Rouillard et al., 2016). Then, we counted the number of targets of differentially expressed TFs (on aneuploid chromosomes for aneuploidy cases) among differentially expressed genes (the ones located on other chromosomes were considered for aneuploidy cases). To create a background distribution, we recounted the number of targets of an equally sized random set of TFs for 100 iterations. Then p-value was calculated using the background distribution by conducting a two-tailed test.

We repeated the same randomization test to understand the degree of TF regulation on the transcript changes of complex members of aneuploid proteins in aneuploid tumours, in which the number of targets among differentially expressed complex members of aneuploid proteins on other chromosomes. To be able to compare the degree of different regulation layers at the proteome and transcriptome level, for this analysis, we used aneuploidy cases where we have proteomic data available comprising COREAD, OV, and BRCA.

10. DNA methylation analysis

To test to which degree gene promoter methylation could explain the transcriptome changes happened on other chromosomes in aneuploid tumours and overall transcriptome dysregulation in tumour, we used promoter-level methylation data calculated by a work done in my host lab (Heery & Schaefer, 2021) from probe-level methylation data in TCGA (comprising 33 cancer types, for 21 of them, methylation data is available for normal samples). For each gene, average methylation level was calculated by taking the mean of methylation levels of the most upstream promoter across samples. To test the statistical significance of differential methylation between up- and downregulated genes, we used Wilcoxon test. This was done once for differential expressed genes in aneuploid tumours (compared to diploid tumours), and once for that in tumour samples (compared to normal samples). To test if differential methylation could explain the transcript changes of complex members of aneuploid proteins on other chromosomes in aneuploid tumours, we repeated the test by considering only the genes located on other chromosomes and encoding complex members of aneuploid proteins.

11. Ubiquitination analysis

To test if post-translational regulation, ubiquitin-mediated degradation in particular, plays a role in co-abundance regulation of complex members of aneuploid proteins, we, first, obtained experimentally observed ubiquitination sites (as a proxy for ubiquitin-mediated regulation) for human proteins from PhosphoSitePlus (Hornbeck et al., 2015)**.** The unique protein pair (aneuploid protein and its complex member) correlation sets for amplification and deletions cases (see Materials and Methods "Statistical analysis"; 2772 and 3818 protein pair correlations for amplification and deletion cases, respectively, **Supplementary File 3**) were used for this analysis, and only the proteins covered by PhosphoSitePlus data were considered. For each complex member of aneuploid proteins, we counted the number of observed ubiquitination sites, and then compared the total numbers among different protein groups: (i) Strongly positively correlated complex members (abundance correlations with aneuploid protein is equal to or higher than 0.4), (ii) strongly negatively complex members (abundance correlations with aneuploid protein is lower than or equal to -0.4), (iii) all human complex members and (iv) all human proteins. Wilcoxon test was used for the statistical comparison.

12. Calculation of the stoichiometry deviation score and survival analysis

To measure the degree of co-abundance compensation for each TCGA sample, we focused on protein abundance correlations between differentially abundant aneuploid proteins and their complex members in three TCGA CPTAC cohorts for which we have available proteomic data; COREAD, OV, and BRCA. Then, for each protein pair, we used a linear regression model, in which protein abundance of complex member was dependent variable while that of aneuploid protein was independent variable, to measure tumour sample response to the change in abundance of corresponding aneuploid protein. Since residual in a linear regression model is the measure of how far the corresponding data point is to fit the regression line, we used residual as stoichiometry deviation score of a sample. To calculate an overall stoichiometry deviation score for each sample, we took the mean of absolute residuals coming from the top 30, 40, and 50 strongest tissue-specific correlation pairs (based on absolute correlations) separately. The analysis was done separately for each TCGA CPTAC cohort, and stoichiometry deviation scores calculated by using the top 30 strongest correlations were used for the survival analysis. Spearman method was used to calculate correlations among the overall stoichiometry deviation scores calculated by using different top protein pairs sets.

Clinical data for TCGA samples was downloaded from cBioPortal (Cerami et al., 2012; J. Gao et al., 2013). To group samples based on their overall stoichiometry deviation scores, we used survminer package (version 0.4.8) in R which is using the maximally selected rank statistics to determine the optimal cutpoint. At the end, we obtained two groups: Samples with high degree of co-abundance compensation (stoichiometry deviation score is equal to or lower than the cutpoint) and those with low degree of co-abundance compensation (stoichiometry deviation score is higher than the cutpoint). Then, we performed survival analysis once with overall survival and once with disease-free survival by using Kaplan Meier method in the survival package (version 3.1.8) in R.

To compare activities of protein-degradation pathways in aforementioned sample groups, we, first, defined proteins which are involved in proteasome complexes or related to ubiquitin-binding based on their related GO annotations obtained from the UniProt database (The UniProt Consortium et al., 2023). Then, the correlation between protein abundances of the selected proteins and stoichiometry deviation scores was calculated for (i) across all samples, and across samples with at least one detected cancer-type-specific chromosomal (ii) amplification, and (iii) deletion by using Spearman method, separately for TCGA CPTAC cohorts. Wilcoxon test was used to compare correlation distributions between sample groups.

13. Frequent copy number alteration and aneuploidy patterns in cancer

Cancer-type-specific amplification and deletion CNA frequencies for 27 different segment level, ranging from 1 Mbp to 50 Mbp, for TCGA tumour samples covering 23 cancer types; BRCA, COREAD, OV, GBM/LGG, LUAD, LUSC, CESC, SKCM, UCS, LIHC, HNSC, PRAD, THCA, PCPG, ESCA, STAD, KIRC, KIRP, PAAD, TGCT, MESO, SARC and BLCA), were obtained from Alfieri et al., 2023 (Alfieri et al., 2023). Arm-level aneuploidy scores for 10,522 TCGA tumour samples, covering all 33 cancer types, were obtained from the work published by Taylor and colleagues (Taylor et al., 2018). Chromosome-level aneuploidy scores were calculated from the aforementioned arm-level data (see Methods and Materials “Detecting cancer-type-specific chromosome-level aneuploidies”). To calculate cancer-type-specific arm- and chromosome-level aneuploidy frequencies, we first counted the number of samples with the corresponding arm-/chromosome-level alteration (amplification or deletion), and divided this by the total number of samples. At the end, we obtained cancer-type-specific amplification and deletion frequencies for 39 chromosome-arms, and 22 chromosomes in 31 cancer types (COAD and READ, and GBM and LGG were considered as one cancer type, COREAD and GBM/LGG respectively to be consistent with CNA data).

14. Collection and processing of feature data

For each of the 29 segment levels (27 segments covering focal CNAs ranging from 1 Mbp to 50 Mbp, and arm- and chromosome-level aneuploidies), features (except for histone marker scores, constitutive LADs, and distance to telomere/centromere) were inferred from gene-level data. Therefore, we first listed all the genes located between start and end positions of the corresponding genomic location. Since gene information was already listed for focal CNAs in the original data, we listed genes for each arm and chromosome. For this, start and end positions for human chromosomes and chromosome-arms were downloaded from The UCSC Genome Browser for the February 2009 assembly of the human genome (hg19 - GRCh37 Genome Reference Consortium Human Reference 37 (GCA_000001405.1)) (http://genome.ucsc.edu) (Nassar et al., 2023). Human genes together with their start and end positions were obtained from Ensembl Biomart (genome version hg19 - GRCh37) (Howe et al., 2021). For the genes with more than one genomic location available, the longest range was considered.

14.1. Transcriptomic and proteomic data

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund (https://commonfund.nih.gov/GTEx) of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this thesis (i) Tissue-specific median protein abundances in 32 normal human tissues (12627 proteins in total) provided by The Enhancing GTEx (eGTEx) and (ii) tissue-specific median gene-level TPM values (covering 54 normal human tissues and 56200 genes) were downloaded from the GTEx Portal (GTEx Consortium, 2013). GTEx tissue names in proteome data were mapped to TCGA cancer tissue names, and mean abundance was considered when there are more than one tissue name mapping to the same TCGA cancer tissue name. TPM values were normalised by using log2 transformation. For each segment level, for the genes located in the corresponding segment and proteins encoded by them, we calculated median expression (mRNA level) and median protein abundance (protein level), respectively. We then found protein complex members of genes located in the corresponding segment by using human known protein complex data from the CORUM database (Giurgiu et al., 2019). By only considering partners located outside of the corresponding segment, we calculated median expression of partners (partner mRNA level) and median protein abundance of partners (partner protein level).

14.2. Calculation of gene density scores

For each segment level, the total number of genes located on the corresponding segment was considered as the gene density score. Among those genes, the total number of ones encoding for protein complex subunits (based on the CORUM data) was considered as the density of subunits.

14.3. Density of tumour suppressors, oncogenes and essential genes

Pre-calculated tissue-general density of TSs and OGs for chromosome-arms and chromosomes were obtained from Davoli et al. (Davoli et al., 2013).

To calculate tissue-specific density of essential genes, we first downloaded clustered regularly interspaced short palindromic repeats (CRISPR) gene effect scores for 17931 genes across 1095 cell lines from DepMap (Dempster et al., 2019, 2021; Meyers et al., 2017; Pacini et al., 2021). After matching DepMap IDs of cell lines with the corresponding TCGA tissue name, we calculated tissue-specific gene effect scores by taking the mean of CRISPR scores in cell lines mapping to the corresponding tissue (covering 24 TCGA tissues). Finally, the first 1000 genes were considered as tissue-specific essential genes when they are ranked in an ascending order based on the gene effect scores.

14.4. Calculation of G/C content

G/C content for each gene was obtained from Human reference genome 37 (GRCh37/hg19). The overall G/C content for a bin was calculated by taking the mean of GC contents of genes located on the corresponding genomic bin.

14.5. Calculation of histone marker scores

BigWig files for tissue-specific histone ChIP-Seq data covering 111 different cell/tissue types mapping to 20 TCGA cancer types (ACC, BRCA, COREAD, DLBC, ESCA, GBM, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, OV, PAAD, SARC, SKCM, STAD, THYM) was downloaded from Roadmap Epigenomics (https://egg2.wustl.edu/roadmap/web_portal/processed_data.html#ChipSeq_DNaseSeq) (Bernstein et al., 2010; Roadmap Epigenomics Consortium et al., 2015). For each bin at each segment level, we counted the number of peaks within the range of the corresponding bin, separately for each histone mark. For this, we used the subsetByOverlaps function from the GenomicRanges package in R (Lawrence et al., 2013). The total number of peaks was considered as histone marker score.

14.6. Constitutive LADs

LaminB1-chromatin interactions, assayed by DamID, in human ESCs and human HT1080 cells were downloaded from the NCBI Gene Expression Omnibus (GEO) database (GEO accession number: GSE22428) (Meuleman et al., 2013). LaminB1 interactions that were detected in both cell lines were named as constitutive LADs (cLADs) while genomic regions in which no interaction was detected in neither of the cell lines were named as interLADs. Then, separately for chromosome-arms and chromosomes, the cLADs to interLADs ratio was calculated. For this start and end positions for human genome version hg18 was used since LaminB1-chromatin interactions data is mapped to human reference genome assembly hg18 in the original publication. Human genome version hg18 for the March 2006 GenBank freeze assembly (hg18, Build 36.1) was downloaded from The UCSC Genome Browser (http://genome.ucsc.edu) (Church et al., 2011).

14.7. Median paralog number

Paralog information for human genes was downloaded from Ensembl Biomart (genome version hg19 - GRCh37) (Howe et al., 2021). For each gene, paralogs with an identity equal to or higher than 60%, and those located outside of the corresponding gene’s segment were kept for further analysis. Tissue-specific median paralog number was calculated for each segment by taking the median of the number of paralogs - expressed in the corresponding tissue (median GTEx expression is higher than zero) - of all genes located in the corresponding segment.

14.8. Calculation of genomic distance to telomeres and centromere

For each segment level comprising focal CNAs (1 to 50 Mbp), the distances of a genomic bin to both telomeres and centromeres were determined using cytoband data from from The UCSC Genome Browser for the February 2009 assembly of the human genome (hg19 - GRCh37 Genome Reference Consortium Human Reference 37 (GCA_000001405.1)) (http://genome.ucsc.edu) (Nassar et al., 2023). Telomere distance was defined as the number of bases to the closest chromosome end. Centromere distance was defined as the number of bases to the beginning of the centromere. Bins located within the centromeric region were omitted from the calculation, as they inherently had a centromere distance of zero.

14.9. Mutation score

Mutation scores calculated based on copy-neutral segments (genomic regions which do not show any CNAs) were obtained from Alfieri et al., 2023 (Alfieri et al., 2023) for nine cancer types where statistically significant correlations were found between amplification frequency and mutation score: BRCA, COREAD, GBM/LGG, LUAD, LUSC, PAAD THCA, HNSC, and CESC.

15. Application of machine learning models

For the classification ML model, we first converted continuous cancer-type-specific amplification and deletion frequencies to binary values. For focal-level CNAs (from 1 Mbp to 50 Mbp), we calculated quantiles of amplification and deletion frequencies, separately for each cancer type across all the bins within the corresponding segment level. Then for the bins with amplification/deletion frequency equal to or higher than 4th quantile, we assigned 1 representing the alteration event, and 0 representing neutral event. For chromosome-level alterations, we used predefined frequently occurred aneuploidies (See Material and Methods “Detecting cancer-type-specific chromosome-level aneuploidies”) and for arm-level alterations, we used the same strategy and performed an association test to calculate if the likelihood of the corresponding alteration is higher than expected by chance by using chi-square test. Multiple testing correction was applied on p-values by using Holm’s method, and events with adjusted p-value lower than or equal to 0.05, and chi-square standard residual equal to or higher than 2 were selected as cancer-type-specific, arm-level aneuploidies. Finally, we assigned 1 for arm and chromosomes which are defined as frequently amplified/deleted, 0 otherwise.

All cancer types pooled for application of ML models. For this, we focused on six cancer types (BRCA, COREAD, GBM/LGG, LUAD, LUSC, and PAAD) where we had all the features available. Two ML models were constructed: an amplification model in which we compared amplified and neutral bins at different segment levels, and a deletion model in which deleted and neutral bins were compared. XGBoost was used as a classification model which is implemented using the R xgboost package (T. Chen & Guestrin, 2016). 10-fold cross validation was used, and feature importance was obtained by using xgb.importance function from the same package.

Results

Chapter I

1. Widespread transcriptome and proteome dysregulation

1.1. Proteome changes in human tumours

To investigate proteomic changes during carcinogenesis, we, first, collected TMT-based proteome quantification data for tumour and normal adjacent tissue samples available in the CPTAC consortium for the following cohorts: COAD (Vasaikar et al., 2019), HCC (Q. Gao et al., 2019) and LUAD (Gillette et al., 2020) comprising 365 tumour and 356 matched normal samples. Since it was stated in the original study that no batch effect was observed in the final processed proteome data, we did not perform batch effect correction. However, we further verified the observation in the original papers by performing PCA, and observed no obvious batch effect within the studies but the expected separation between tumour and normal samples (**Figure 1**). We, then, detected significantly differentially abundant proteins (adjusted p-value ≤ 0.05, Wilcoxon test, and absolute log2FC > 1) between tumour and normal samples within each cohort. We found a variation in the fraction of differentially abundant proteins among the quantified proteins across different cohorts; 457 out of 6,554 proteins, 481 out of 6,478 proteins for COAD and HCC (approximately 7% of all quantified proteins), respectively, and 3,971 out of 10,316 proteins (~38% of all quantified proteins) for LUAD (**Figure 2**). The majority of those differentially abundant proteins were down-regulated in tumour samples (~90% for COAD and HCC, 60% for LUAD) (**Figure 2**).

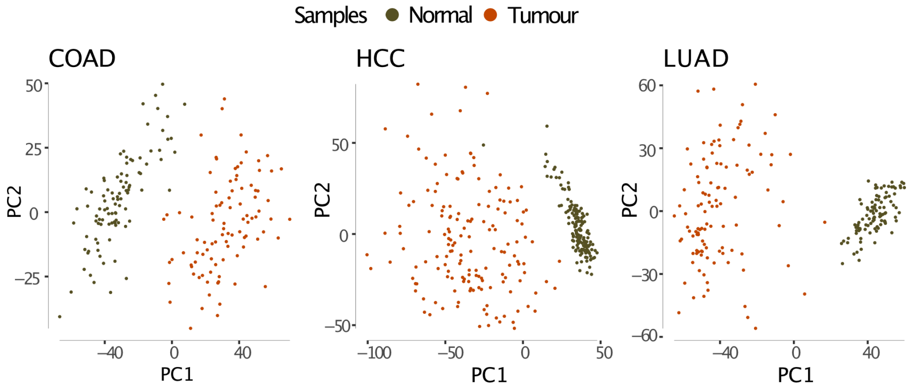


Figure 1. Principal component analysis on proteome data for the CPTAC cohorts COAD, HCC, and LUAD where proteomic measurements for tumour and matched normal samples are available.

We next aimed to understand at which degree members of protein complexes are affected by those differential abundance changes in tumour samples compared to normal samples. To do this, we performed an association test between the detected significantly differentially abundant proteins (up- and down-regulated proteins were pooled) and protein complex members obtained from the CORUM database (Giurgiu et al., 2019). We found that the number of protein complex members showing significant abundant changes is significantly less than expected value for 2 out of 3 cohorts (COAD and HCC; p-value < 0.0001, chi-square test; **Figure 3A**). Since our main source for protein complex members (CORUM database) relies on experimentally validated known protein complex information, and this could introduce size bias towards highly abundant and/or easily crystallised proteins in the comparison of complex member and non-complex member proteins, we repeated the association test by creating a background of non-complex member proteins with the same abundance distribution as complex member proteins. We reproduced our previous observation that there is a depletion of differentially abundant proteins in complex proteins in COAD and HCC cohorts. Overall, those findings suggest that protein complex members are protected from abundance changes in tumourigenesis as stoichiometric imbalances in protein complexes prevent proper functioning of complexes. In fact, Gonçalves and colleagues previously described similar observations for protein abundance changes triggered by focal genomic copy number alterations (Gonçalves et al., 2017).

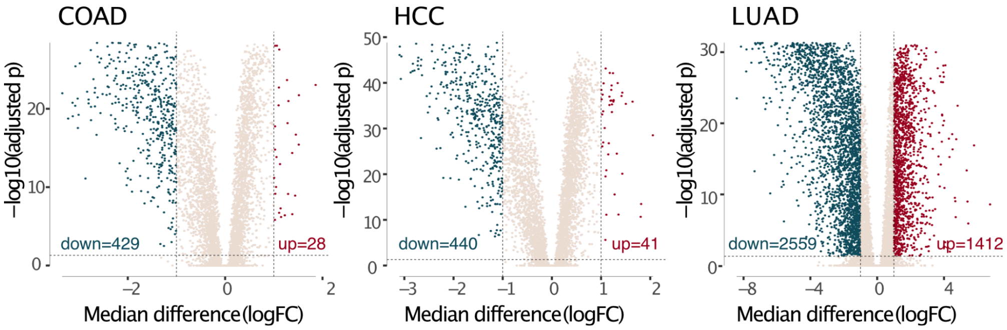


Figure 2. Protein abundance changes in cancer. Differential protein abundance changes between tumour and normal samples in COAD, HCC, and LUAD cohorts. Significantly differentially abundant proteins in tumour samples are represented by dark green and dark red colours respectively for down- and up-regulated proteins (adjusted p-value ≤ 0.05, Wilcoxon test, and absolute log2FC > 1).

Differently from COAD and HCC cohorts, where around 90% of the differentially abundant proteins were down-regulated, we observed a relatively higher proportion of up-regulated proteins in the LUAD cohort (**Figure 2**). Therefore, we performed the association test separately for up- and down-regulated proteins to assess differences in their overlap with the protein complex subunits. Consistent with the COAD and HCC cohorts, we observed a significant depletion of protein complex subunits in down-regulated proteins (p-value = 8.42e-19, chi-square test) while up-regulated proteins were significantly enriched in protein complex subunits (p-value = 6.28e-17, chi-square test) (**Figure 3B**). The overall strong depletion of protein complex members among down-regulated proteins further suggest that complexes are protected from downregulation of their members since the lack of subunits in a complex will prevent the proper functioning of the complex, which might be more detrimental for the tumours than upregulation of subunits.

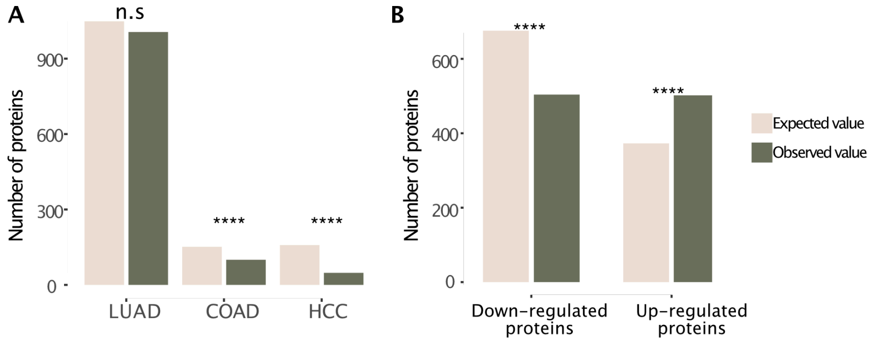


Figure 3. Depletion of protein complex members in differentially abundant proteins. (A) The expected and observed number of differentially abundant proteins in complex subunits. **(B)** The expected and observed number of complex subunits in up- and down-regulated proteins in LUAD cancer. Chi-square test was used to determine if the difference between expected and observed values is statistically significant (n.s: no statistical significance; ****: p < 0.0001).

In summary, we observed that protein complex members are, to a certain degree, protected from downregulation in cancer. However, cancer is characterised with many protein abundance changes (Guang et al., 2019). Indeed, we found that protein abundances across tumour samples showed a higher variation when compared to those in the matched normal samples in all of the 3 cohorts (p-value < 0.0001, t-test; **Figure 4**). Therefore, we further aimed to understand whether co-regulation differs between complex member proteins and non-complex member proteins. To do this, we first included the TCGA CPTAC proteome data (for which there is no matched normal samples, hence had been excluded from the differential abundance analysis) for the following cohorts; COREAD (The Cancer Genome Atlas Network, 2012; the NCI CPTAC et al., 2014), OV (Cancer Genome Atlas Research Network, 2011; Zhang et al., 2016), and BRCA (Cancer Genome Atlas Network, 2012; Mertins et al., 2016). Then, for all six cohorts, we calculated protein abundance correlations for all possible pairs of proteins covered by CORUM complexes across tumour samples by using Spearman method. We found that the distribution of correlations among proteins involved in the same complexes (co-complex members) is significantly higher than that among proteins which are members of different complexes (p-value < 2.2e-16, Wilcoxon test; **Figure 5**).

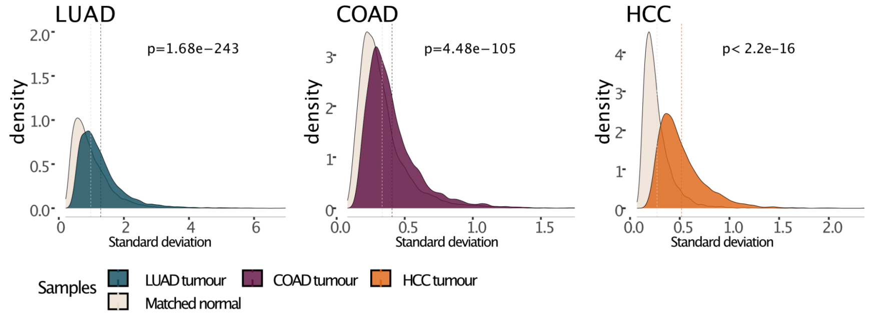


Figure 4. Overall proteome variation in cancer. The difference between the distribution of standard deviations of protein abundances in tumours and matched normal samples. Statistical significance was tested by t-test.

Together these observations suggest that the ways proteins interact within protein complexes impose limitations on dysregulation of protein abundances in cancer, and have an impact on the strength of co-abundance patterns.

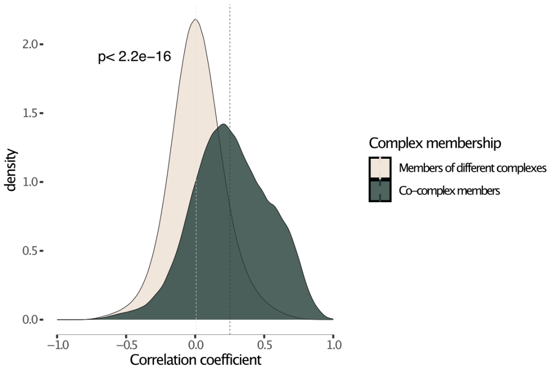


Figure 5. Abundance correlations between members of the same protein complexes. Spearman correlation coefficients of protein pairs across tumour samples. Pairwise correlations were calculated across tumour samples, separately in each cohort (LUAD, COAD, HCC, BRCA, OV, and COREAD) and then pooled. Wilcoxon test was used to test if two distributions are significantly different.

1.2. Transcriptome and proteome changes in aneuploid tumours

In the previous section, we characterised proteome-level dysregulation during tumorigenesis and provided evidence that being involved in protein complexes shapes those dysregulation patterns to avoid stoichiometric imbalances in complexes, and thus improper functioning. Considering that majority of cancer genomes are aneuploid, 90% for solid tumours, (Ben-David et al., 2019; Taylor et al., 2018), we further aimed to study the effect of chromosome-level aneuploidies on cancer transcriptomes and proteomes in aneuploid tumours. To do this, we first used publicly available cancer aneuploidy estimates (Taylor et al., 2018) for 10,522 TCGA samples, and identified cancer-type-specific, chromosome-level amplifications and deletions that occurred at significantly higher frequencies than would be expected by change (adjusted p-value <= 0.05 and standard residuals > 2, chi-square test). In total, we detected 203 cancer-type-specific, chromosome-level aneuploidies including 86 amplifications and 117 deletions comprising 32 TCGA cancer types (**Figure 6; Supplementary File 2**). Then, for each detected aneuploidy, we split the samples into two sets; those containing the respective chromosome number alteration and those diploid for the respective chromosome, and performed differential gene expression analysis to detect differentially expressed genes between aneuploid and diploid samples by using Wilcoxon test. We found that nearly half of the genes located on aneuploid chromosomes changed expression: 41% and 48% of genes, on average, located on amplified and deleted chromosomes, respectively (**Figure 7**). In addition to those intuitively expected expression changes on the aneuploid chromosomes, we observed a surprisingly large number of expression changes happening on other, typically diploid chromosomes: On average 15% and 18% of genes for amplification and deletion cases, respectively (**Figure 7**).

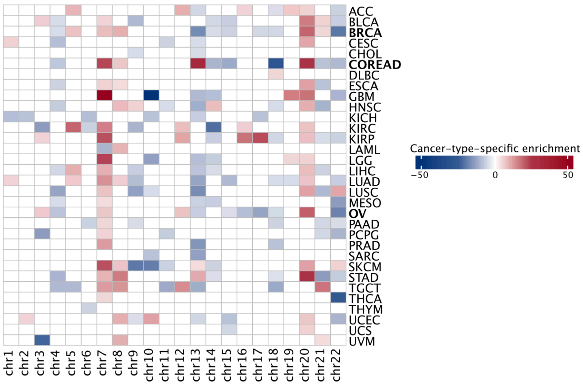


Figure 6. Cancer-type-specific, whole-chromosome-level alterations across 32 TCGA cancer types. The colour encodes the degree of their enrichment (standard residuals of the chi-square test multiplied by the alteration score: –1 in the case of deletion and 1 in the case of amplifications). Cancer types for which the proteomic data is available are indicated in bold.

We then aimed to further understand the effect of transcriptome changes in response to aneuploidy on the proteome. To do this, we obtained corresponding proteomic quantification data from CPTAC for 298 TCGA tumour samples comprising COREAD (The Cancer Genome Atlas Network, 2012; the NCI CPTAC et al., 2014), OV (Cancer Genome Atlas Research Network, 2011; Zhang et al., 2016), and BRCA (Cancer Genome Atlas Network, 2012; Mertins et al., 2016). Among the detected 203 cancer-type-specific, chromosome-level aneuploidies, 13 amplification and 20 deletions cases were found in those 3 cancer types (**Figure 6**). For those 33 aneuploidy cases, we detected protein abundance changes between aneuploid samples (with the amplified/deleted chromosome) and diploid samples. When compared to corresponding transcriptome changes, we observed that a relatively smaller number of proteins showed abundance changes (after the normalisation for the largely different gene coverage between transcriptome and proteome datasets): For aneuploid chromosomes, 24% and 33% of proteins respectively for amplified and deleted chromosomes, and for other chromosomes, 13% and 16% for amplification and deletion cases (**Figure 7**).

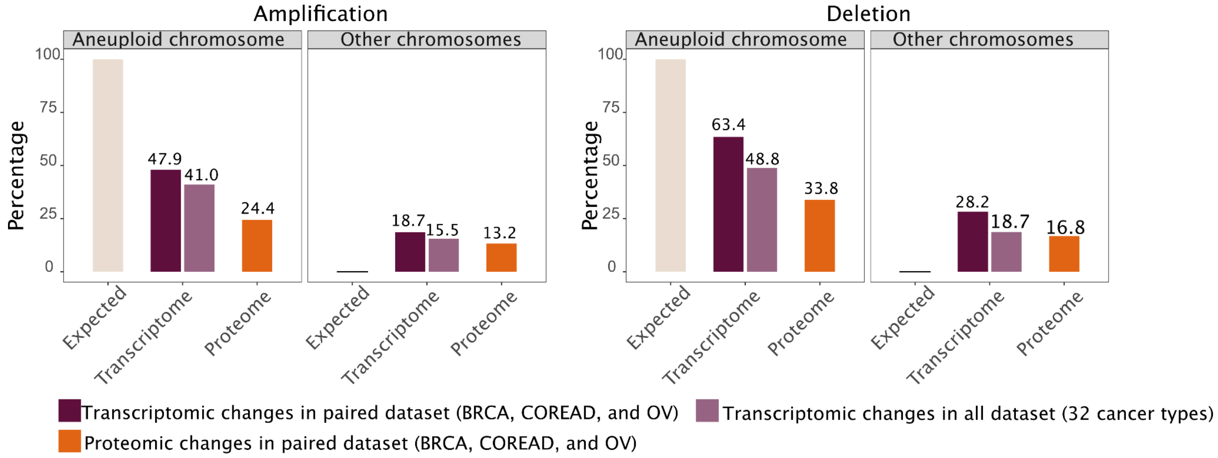


Figure 7. Transcriptomic and proteomic changes in aneuploid tumours. Average percentage of differentially expressed genes or abundant proteins on aneuploid and other, non-aneuploid chromosomes (among the detected genes on the respective chromosomes).

To better understand the crosstalk between transcriptome and proteome changes, we compared at which proportion the proteomic changes are covered by the corresponding transcriptomic changes. We found that, on average, 76% and ~90% of proteomic changes happening on amplified and deleted chromosomes, respectively, are differentially expressed on the transcriptome level (**Figure 8**). On the other hand, we observed a relatively smaller portion of proteomic changes happening on other chromosomes is covered by the corresponding transcriptomic changes (on average, 26% and 36% of genes respectively for amplification and deletion cases; **Figure 8**). These results together with overall transcriptomic and proteomic changes happening on aneuploid and other, typically diploid chromosomes, suggest the role of transcriptional control in dosage compensation for genes on the aneuploid chromosomes; however, translational or post-translational control is more prone on the regulation of genes found on other chromosomes.

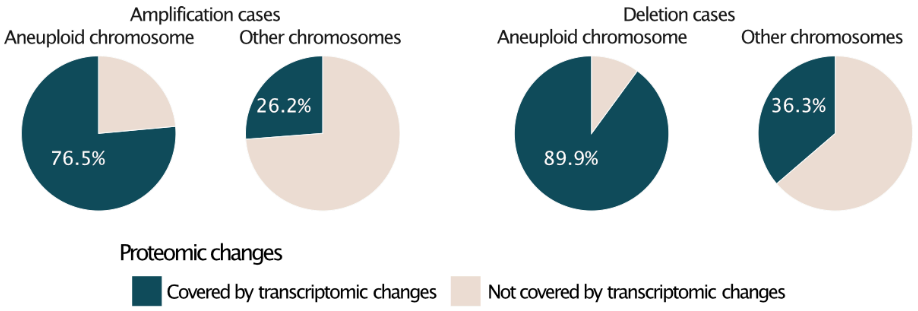


Figure 8. Crosstalk between transcriptomic and proteomic changes in aneuploid tumours. Overall proteomic changes covered by corresponding transcriptomic changes in aneuploid and other, non-aneuploid chromosomes, separately for amplification and deletion cases.

Finally, we investigated the molecular pathways associated with dysregulation patterns in aneuploid tumours separately for transcriptomic and proteomic changes. To do this, we focused on frequently differentially expressed genes and abundant proteins across all detected aneuploidy cases (13 amplification and 20 deletion cases comprising COREAD, OV, and BRCA cancer types where we have both transcriptome and proteome data available), and performed GO analysis. We found that frequently dysregulated genes and proteins in aneuploid tumours are related to GO terms belonging to cell cycle and cell cycle processes (**Supplementary File 4**) further supporting the previous findings: One of the major consequences of aneuploidy is its impact on cell proliferation (Santaguida & Amon, 2015).

Overall these observations suggest that attenuation mechanisms are in place, and we observe regulation both on transcription and translation level which prevent that all genes located on the aneuploid chromosomes are deregulated. At the same time, we observed a surprisingly large number of dysregulation events on chromosomes other than the aneuploid one raising the question of the purpose of the up- and downregulation of hundreds of genes in response to specific aneuploidies.

1.3. Contribution of co-occurrence of different aneuploidies to overall transcriptome dysregulation on other chromosomes

A systematic statistical analysis on over 15,000 cancer karyotypes previously showed a high co-occurrence rate of chromosomal gains with other gains, and losses with other losses (Ozery-Flato et al., 2011) suggesting that co-occurrence of different aneuploidies could serve as compensation mechanism to keep a balance among altered proteins and maintain cellular fitness (Ozery-Flato et al., 2011). Therefore, we aimed to test if co-occurrence patterns of chromosomal amplifications and deletions could explain the relatively high number of differentially expressed genes on other chromosomes in aneuploid tumours. To do this, for each of the 203 cancer-type-specific, chromosome-level amplifications and deletions, we first tested the statistical dependence of occurrence of other amplification/deletion events by using chi-square test. We identified 305 co-amplifications and 672 co-deletions that occurred more frequently than expected by chance covering 60 out of 86 detected amplifications and 90 out of 117 detected deletions, respectively (adjusted p-value <0.01, chi-square test; **Supplementary File 2**). Then, for each detected co-occurrence event, we quantified the contribution of the co-altered chromosome to the transcriptional dysregulation by dividing the number of differentially expressed genes located on the co-altered chromosome to the total number of genes on that chromosome. Finally, we compared the average contribution of co-altered chromosomes to that of non-co-altered chromosomes across all the detected co-amplification and co-deletion events. We found that there is no significant difference between the medians of these chromosome groups both for chromosomal amplifications and deletions (p-value = 0.4 and p-value = 0.39 respectively for amplifications and deletions, paired Wilcoxon test; **Figure 9**) suggesting that they do not substantially contribute to the overall transcriptional dysregulation on other chromosomes.

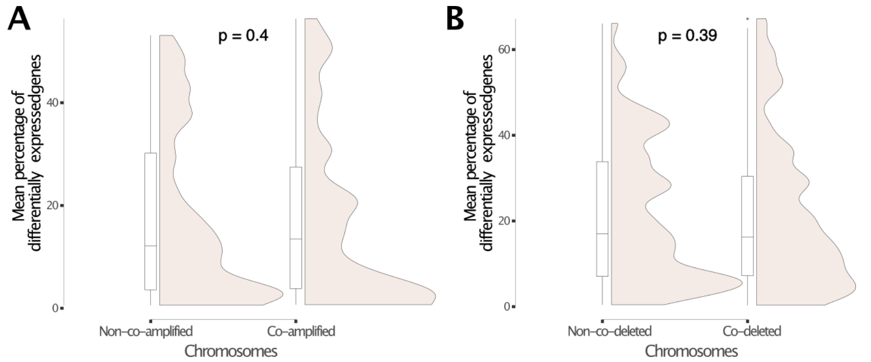


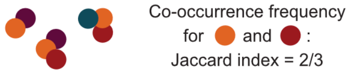
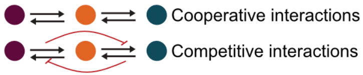
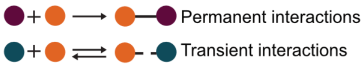
Figure 9. Contribution of co-amplification and co-deletion events to overall transcriptional dysregulation in aneuploid tumours. (A) Average percentage of differentially expressed genes on co-amplified and non co-amplified chromosomes across 60 chromosome-level amplifications. **(B)** Average percentage of differentially expressed genes on co-deleted and non co-deleted chromosomes across 90 chromosome-level deletions. Paired Wilcoxon test was used to test differences between the groups.

2. Members of the same complexes and proteins interacting through physical interactions tend to be co-deregulated

2.1. Protein-protein interactions within complexes constrains proteome dysregulation in human tumours

Analysis of proteomic changes in human tumours (see Results - Chapter I “Proteome changes in human tumours”) showed that tumour samples are characterised with relatively higher variation in protein abundances compared to normal samples; however, there is a certain level of buffering on protein complex members protecting them from this deregulation. Moreover, we showed that protein abundances of the same complex members are strongly co-regulated. Therefore, we further aimed to investigate how exactly the dynamics of protein complex organisation and the interactions between their members affect those co-abundance regulation in tumours. For this aim, we first collected protein structure data, protein quantification data in different cell lines, binding interface information and biochemical properties for PPIs from different sources (see Materials and Methods “Classifying proteins and protein interactions”), and systematically categorised different interaction types into five classes; stoichiometric ratio between proteins, co-occurrence frequency of proteins, context-specific vs. general interactions, competitive vs. cooperative interactions, and transient vs. permanent interactions (**Table 1**). Then, we calculated protein abundance correlations among all possible pairs of quantified proteins, separately for all six cohorts (COAD, HCC, LUAD, COREAD, OV, and BRCA), and tested which of the pre-defined protein interaction categories could explain differences in the strength of abundance correlations in human tumours.

Table 1: Systematic categorization of PPIs and proteins: Interaction types and protein properties, the source from which the interaction or protein information is obtained, and cartoon illustration of the corresponding category where nodes represent proteins.

Category	Source	Representation
Stoichiometric ratio (Even vs. uneven ratio)	PDB Berman et al. (2000)	Even ratio Uneven ratio
Co-occurrence frequency	CORUM Giurgiu et al. (2019)	 Co-occurrence frequency for ● and ● : Jaccard index = 2/3
Context-specificity (Context-specific vs. general interactions)	BioPlex Interactome Huttlin et al. (2021)	General interactions Context-specific interactions
Binding similarity (Competitive vs. cooperative interactions)	Interactome INSIDER Meyer et al. (2018)	
Stability of PPI (Transient vs. permanent interactions)	Block et al. (2006) Mintseris and Weng (2003)	

Proteins are involved in protein complexes with certain amounts relative to each other (Taggart et al., 2020), and keeping this stoichiometric ratio in balance is important for proper functioning of the cell. Therefore, we hypothesised that the stoichiometric ratio between protein complex members could limit the abundance dysregulation of proteins in tumours. To test this, we first calculated stoichiometric ratios among the members of protein complexes obtained from the RCSB PDB (in total, 8,388 human heteromeric protein complexes) (Berman et al., 2000). Then we compared protein abundance correlations between proteins involving in complexes with an even stoichiometric ratio (e.g. 1:1, 2:2, 4:4) to those with uneven stoichiometric ratio (e.g. 1:2, 3:1, 1:2) separately for each cohort. We found that protein pairs with even stoichiometric ratio are related to significantly stronger correlations than those participating complexes with uneven stoichiometric ratio (p-value < 0.0001, Wilcoxon test; **Figure 10**). To further understand if the stoichiometric ratio would have an impact on the strength of the abundance relationship between proteins, we performed a linear regression model for each pair and compared the steepness of the regression curve between protein pairs with an even and those with an uneven stoichiometric ratio. When considered only the significant models (coefficients are different from zero, p-value < 0.05, linear regression model), we found that positively correlated protein pairs with even stoichiometric ratio show significantly higher slopes than those with uneven stoichiometric ratio for 4 out of 6 cohorts (p-value < 0.0001, Wilcoxon test; **Figure 11A**). We observed the opposite association for negatively correlated protein pairs (significant for 2 out of 6 cohorts, p-value < 0.05, Wilcoxon test; **Figure 11B**). Overall our observations make sense since a relatively higher degree of compensation will be needed for protein complex members involved in complexes with higher number of copies relative to each other when a protein from the same complex is upregulated.

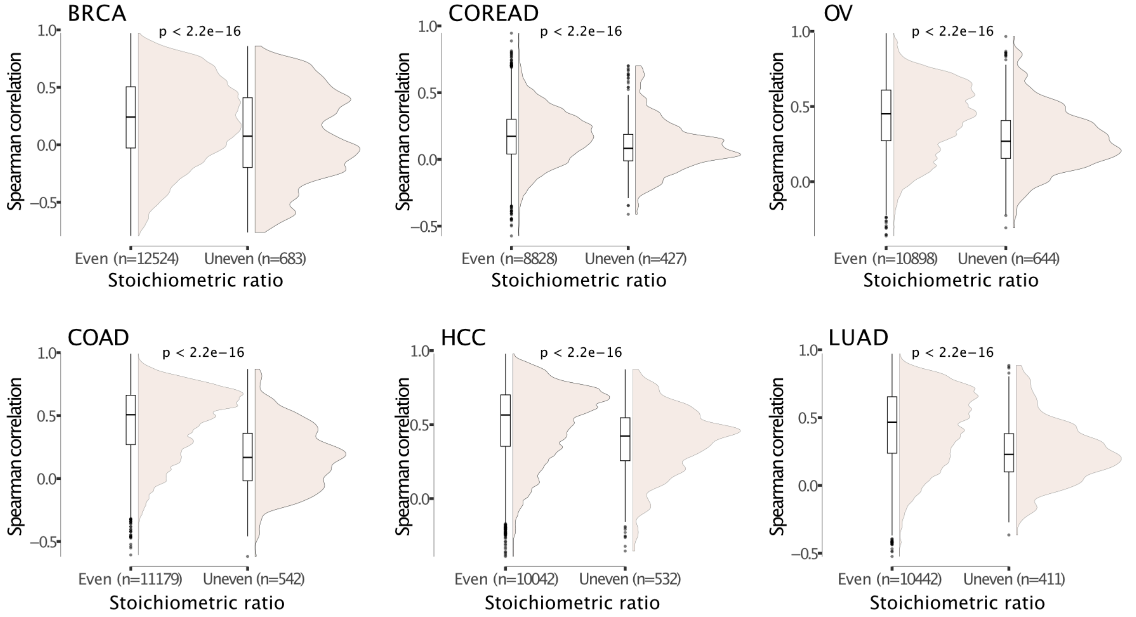


Figure 10. Effect of stoichiometric ratio between proteins to co-abundance regulation. Comparison of protein level correlations of proteins involved in complexes with an even stoichiometric ratio to those with an uneven ratio in BRCA, COREAD, OV, COAD, HCC, LUAD. Wilcoxon test was used for the statistical comparison.

Another category that we tested is co-occurrence frequency of interacting proteins in complexes as some proteins could be found together in many complexes while others only appear in a few complexes. For all possible protein pairs among the CORUM protein complex subunits, we counted how many times a protein pair co-occur in the same complex as a representative of co-occurrence frequency, and then compared it between different protein-level correlation groups separately for each of the six cohorts. We found that protein pairs showing positive abundance correlations tend to co-occur in the same complexes more than their negatively correlated counterparts, consistently for each cohort (p-value < 0.0001, Wilcoxon test; **Figure 12**). This is in line with our expectation as frequent co-membership in complexes means a larger number of complexes will depend on the proper abundance ratios between the two proteins and this will increase the need for coregulation of the protein pair.

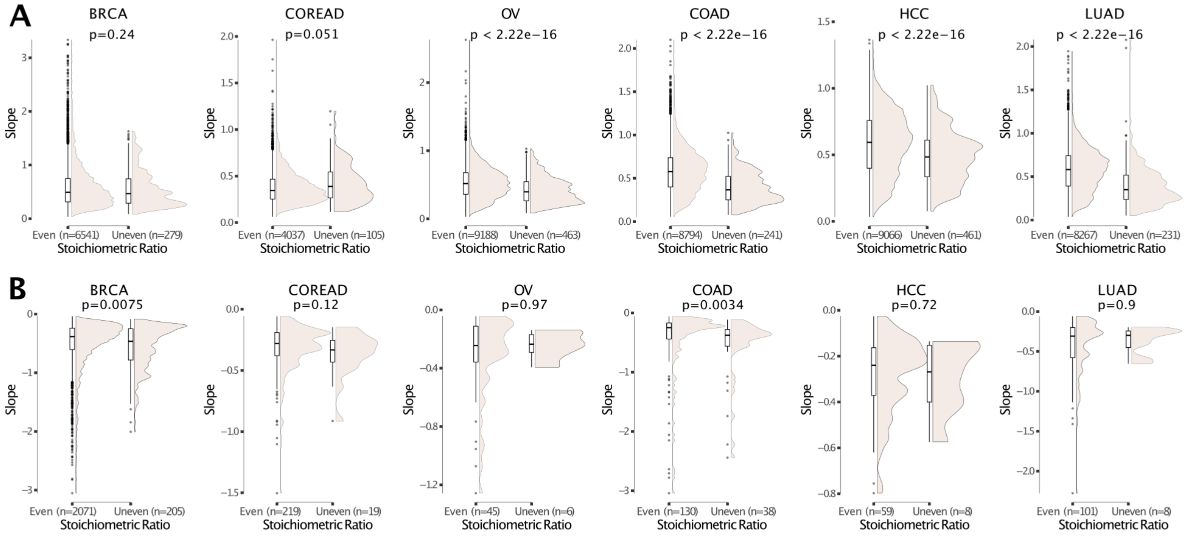


Figure 11. Linear regression analysis for protein abundances. Comparison of distributions of slopes from the linear regression analysis between protein pairs with even and uneven stoichiometric ratios for each cohort (BRCA, COREAD, OV, COAD, HCC, LUAD). **(A)** Positively correlated proteins and **(B)** negatively correlated proteins. After regression analysis, only the significant results ($p < 0.05$, linear regression model) were considered.

Even though proteins are interacting with each other in a well-defined manner, some interactions are highly dependent on cellular context (Barrera-Vilarmau et al., 2022). Therefore, we tested whether context-specificity of protein interactions impact on protein abundance correlations between protein pairs. To do this, we obtained experimentally determined cell-line specific PPIs (as a proxy for the interaction specificity) from the BioPlex Interactome (Huttlin et al., 2021), and classified protein pairs as context-specific and general interactions depending on their detection in only one of the two cell lines or in both. We found that proteins interacting in both cell lines, general interactions, are related to significantly stronger abundance correlations than those interacting in a context-specific manner in each of the six cohorts (p-value < 0.0001, Wilcoxon test; **Figure 13**). The result makes sense since general interactions tend to be less affected by different cellular contexts leading to an expectation of more control in the regulation of their co-abundance changes.

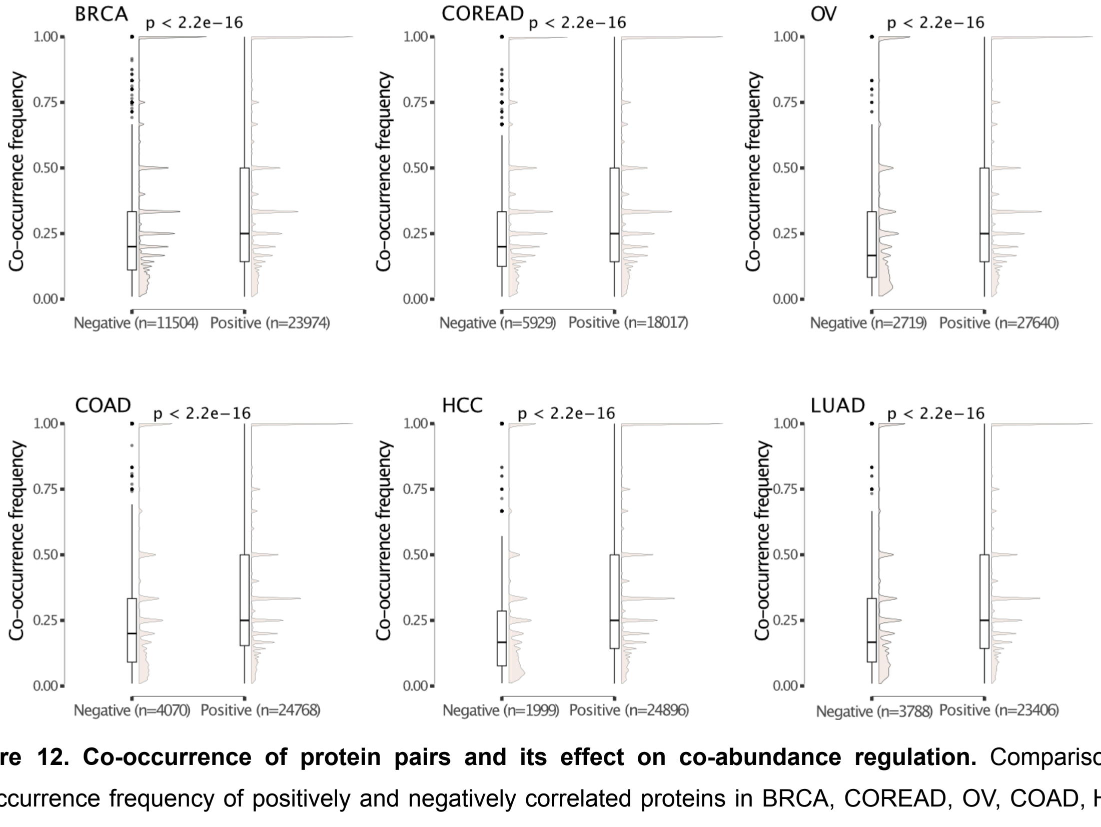


Figure 12. Co-occurrence of protein pairs and its effect on co-abundance regulation. Comparison of co-occurrence frequency of positively and negatively correlated proteins in BRCA, COREAD, OV, COAD, HCC, LUAD. Wilcoxon test was used for the statistical comparison.

While some proteins bind to few partners, some have multiple partners binding at similar (overlapping) interaction sites (Keskin & Nussinov, 2007). Therefore, we aimed to test if competing for binding affects co-abundance changes in human tumours. To identify competitive and cooperative interactions, we first obtained binding interfaces for experimentally determined human binary interactions from Interactome INSIDER (Meyer et al., 2018). Then, for protein pairs binding to another shared partner, we calculated a score to measure how similar their corresponding binding region on their shared partners (see Materials and Methods “Competitive and cooperative interactions”). Finally, we grouped protein pairs as competitive and cooperative based on this score. We observed that competitively interacting proteins have significantly higher correlations than cooperatively interacting proteins, consistently for 5 out of 6 cohorts (p-value < 0.05, Wilcoxon test; **Figure 14**). This can be robustly reproduced when comparative and cooperative interactions were grouped based on different binding similarity scores (**Figure 15**). This observation surprised us as we expected weaker or negative correlations between competitively interacting proteins as those, by definition, should not participate in the same complex at the same time, and hence an opposing expression pattern would be expected.

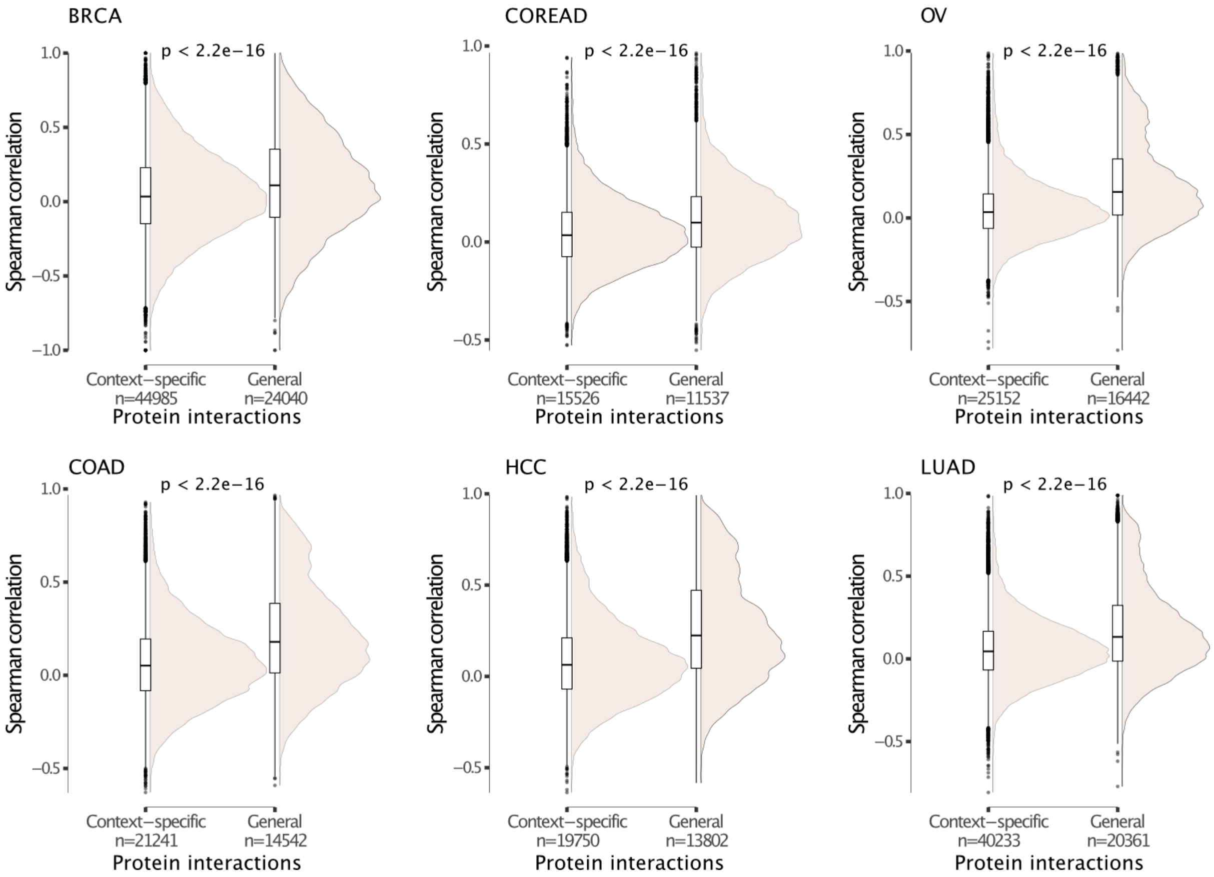


Figure 13. Proteins interacting with each other independently from the context are related to higher co-abundance regulation. Protein level correlations of proteins interacting in context-specific and general manner in BRCA, COREAD, OV, COAD, HCC, LUAD. Wilcoxon test was used for the statistical comparison.

Based on stability, PPIs are classified as permanent and transient interactions. We obtained permanent and transient interactions estimated by machine learning algorithms based on physicochemical properties of PPIs from Minteris and Weng, 2003 (Mintseris & Weng, 2003) and Block et al., 2006 (Block et al., 2006), and then compared protein level correlations between those two groups. For 3 out of 6 cohorts (COREAD, COAD, and HCC), we observed consistent trends: Permanent interactions correspond to stronger correlations while relatively weaker, transient interactions were observed between proteins whose abundances are less dependent on each other (p-value < 0.05, Wilcoxon test; **Figure 16**). This is expected as the transient interactions are more flexible for a change in binding partners during the assembly of complexes (Nooren & Thornton, 2003).

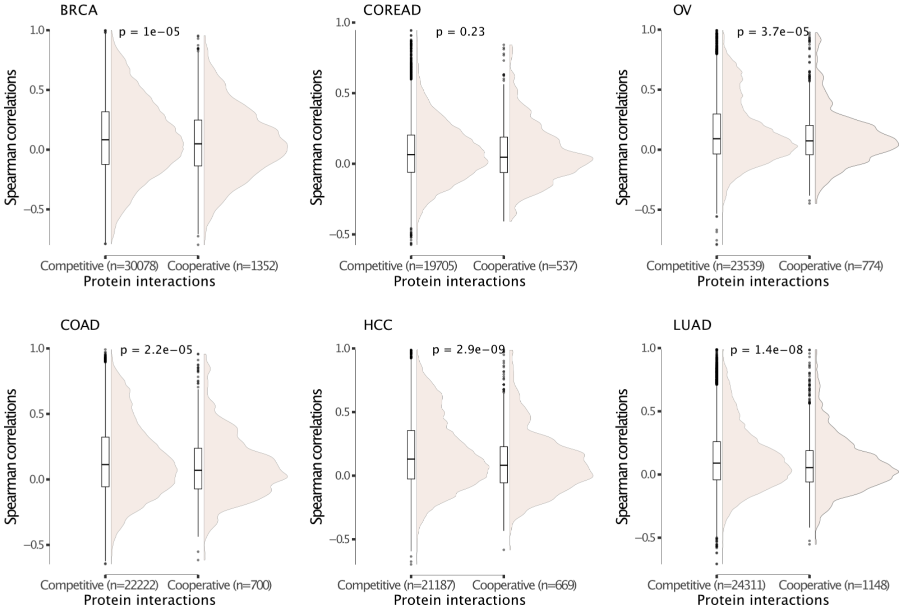


Figure 14. Competitive and cooperative PPIs and their protein level correlations in BRCA, COREAD, OV, COAD, HCC, LUAD. Wilcoxon test was performed to compare correlations between two groups in each class of PPIs.

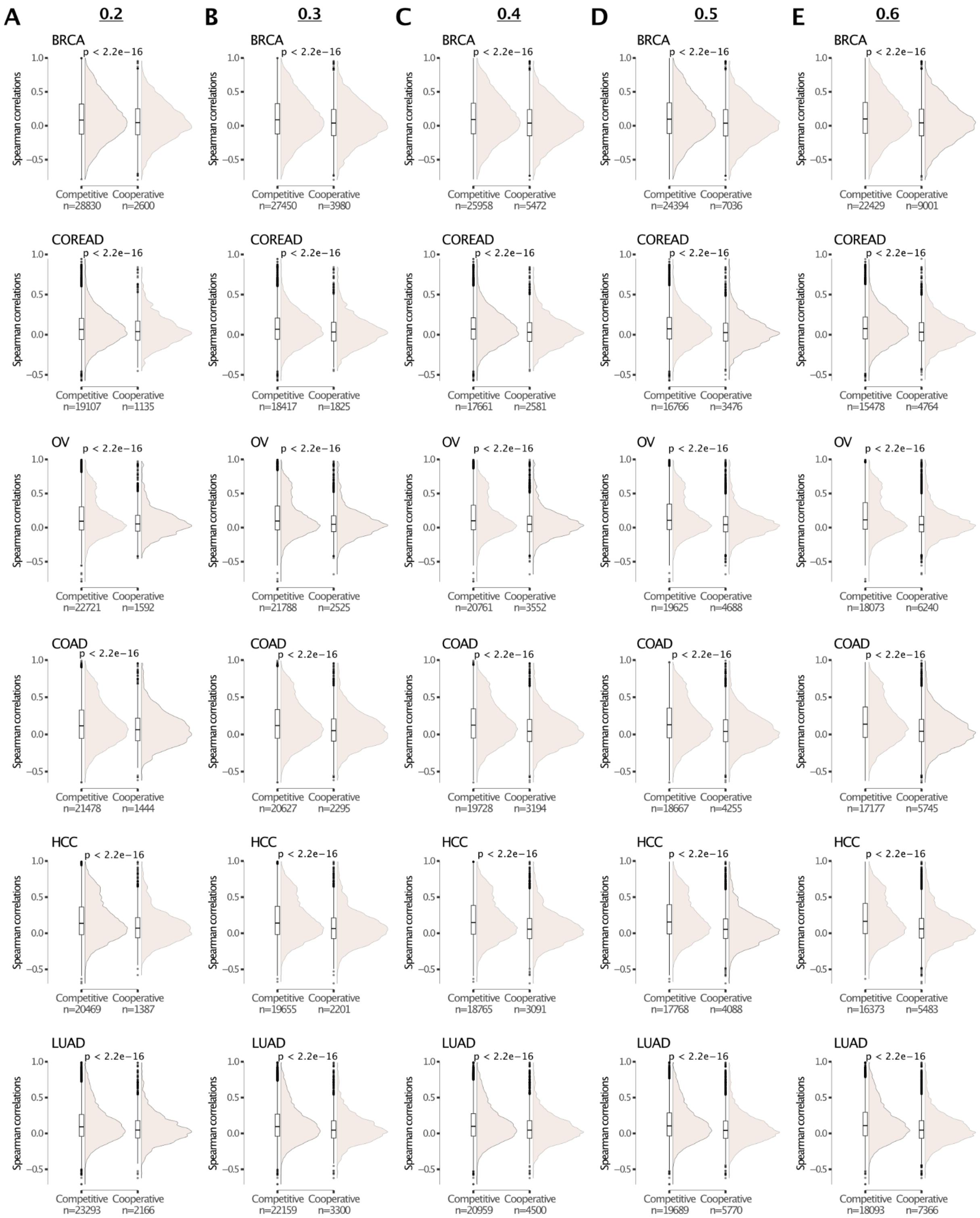


Figure 15. Comparison of protein abundance correlations between competitive and cooperative interactions when they were grouped based on different binding similarity score cut-offs. Each panel shows the comparison between groups categorised based on a certain cut-off for all cohorts (BRCA, COREAD, OV, COAD, HCC, LUAD). The interaction between two proteins is competitive, if the binding similarity score is equal to or larger than **(A)** 0.2, **(B)** 0.3, **(C)** 0.4, **(D)** 0.5, and **(E)** 0.6, otherwise cooperative.

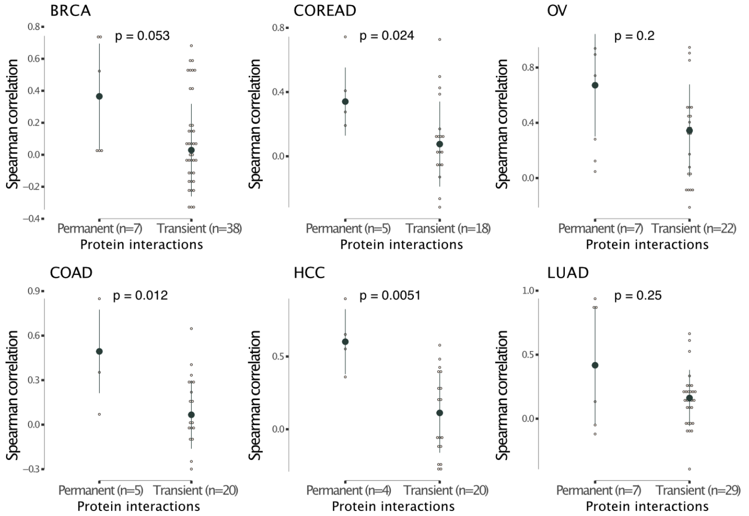


Figure 16. Permanent and transient PPIs and their protein level correlations in BRCA, COREAD, OV, COAD, HCC, LUAD. Wilcoxon test was performed to compare correlations between two groups in each class of PPIs.

2.2. Proteome-level dysregulation on non-aneuploid chromosomes could be explained by co-deregulation of members of the same complexes

Transcriptome and proteome characterization in aneuploid tumours revealed that there is a certain level of buffering on the expression levels of genes located on the altered chromosomes both at transcriptome and proteome level. Moreover, we found a surprising degree of dysregulation of genes found on other chromosomes. In the previous section, we showed that properties of proteins and the ways they organise into complexes impact on proteome-level dysregulation in human tumours. Therefore, we wondered if the same constraints could explain the abundance changes happening on other, partially diploid, chromosomes, in aneuploid tumours. To test this, we first identified co-complex members (proteins that are involved in the same protein complexes - complex partners) of differentially abundant proteins encoded on aneuploid chromosomes by using human protein complex information from the mammalian protein complex database CORUM (Giurgiu et al., 2019), and then performed an association test between those and differentially abundant proteins encoded on other chromosomes. We observed a general tendency for the differentially abundant proteins of other chromosomes to be complex partners of differentially abundant proteins from the aneuploid chromosome for both chromosome-level amplifications and deletions (p-value < 0.05, chi-square test; **Figure 17**). We found a moderate percentage (in average is 4.47%; **Supplementary File 5**) of differentially abundant proteins on other chromosomes being partners of those on aneuploid proteins. However, the coverage of proteins with CORUM complex information is rather limited (22% of proteins form part of at least one complex in CORUM). When only proteins participating in at least one complex were considered, the average fraction of partner proteins among differentially abundant proteins increased to 12.61% (**Supplementary File 5**). Moreover, comparing protein

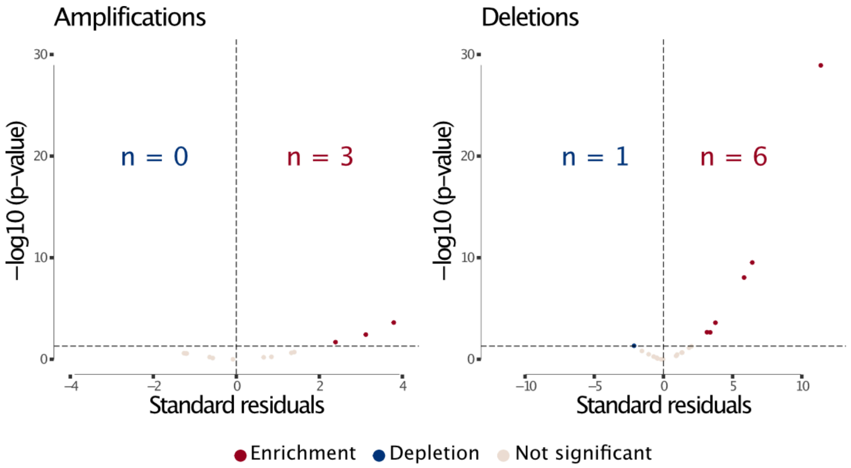


Figure 17. Enrichment of partners of aneuploid proteins in differentially abundant proteins on other chromosomes. Standard residuals and p-values for the overlap between co-complex members of differentially abundant proteins on aneuploid chromosomes and differentially abundant proteins on other chromosomes for 13 amplifications and 20 deletions.

abundance correlations between differentially abundant proteins on aneuploid chromosomes and their co-complex members with non-complex members showed significantly stronger correlations between proteins of same complexes (p-value < 2.2e-16, Wilcoxon test; **Figure 18**). Those observations are in line with previous findings claiming that complex organisation shapes protein abundance changes in response to CNA (Sousa et al., 2019).

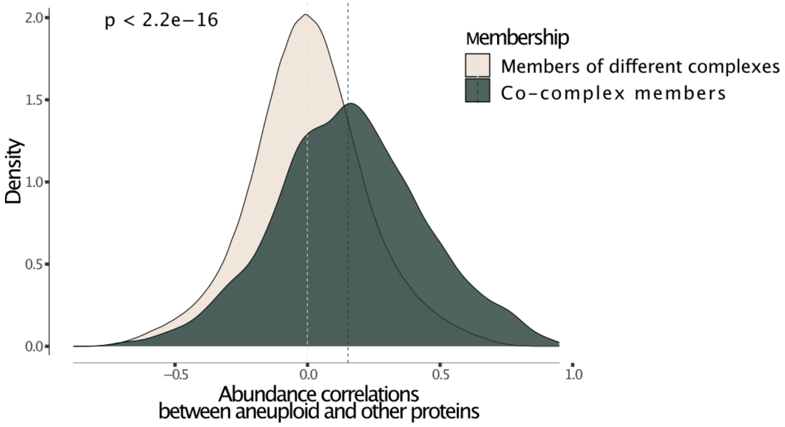


Figure 18. Co-abundance regulation of co-complex members in aneuploid tumours. Protein abundance correlations between differentially abundant proteins on aneuploid chromosomes and their co-complex and non-complex subunits. Correlations were calculated across tumour samples, separately for each cancer type, and then pooled. Wilcoxon test was used to determine whether two distributions are significantly different.

To investigate whether our observations can be generalised to binary PPIs, we used PPI data from the human protein-protein interaction database HIPPIE (v2.2) (Alanis-Lobato et al., 2017) to test if the number of interactions between differentially abundant proteins encoded on the aneuploid and those on other chromosomes is higher than expected by chance. Indeed, we found an enrichment of interactions between those protein sets in 9 out of 13 cancer type-specific amplifications and 8 of the 20 deletions (p-value < 0.05, randomization test; **Figure 19**, **Supplementary File 6**). Given the higher coverage of PPI data, we asked again which percentage of differentially abundant proteins on other chromosomes could be potentially explained by their interactions with complex members on aneuploid chromosomes. We found that on average 27.5% of the differentially abundant proteins on other chromosomes interact with those on the aneuploid chromosomes (**Supplementary File 6**). For example, for chromosome 7 in COREAD and chromosome 12 in OV, more than 40%, and for chromosome 5 in BRCA, more than 30% of the differentially abundant proteins interacted with proteins on the amplified chromosomes.

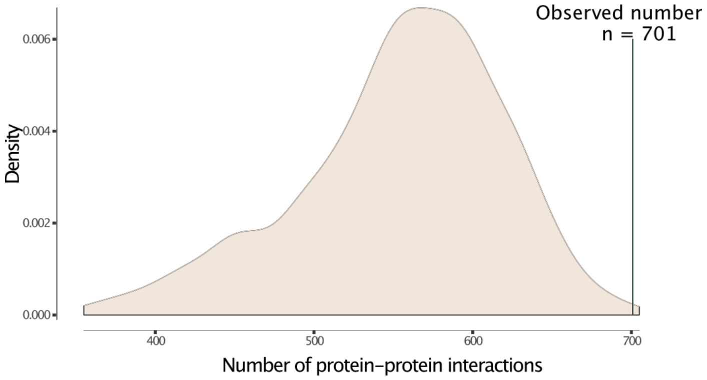


Figure 19. Enrichment of partners of aneuploid proteins in differentially abundant proteins on other chromosomes when binary interaction data is used. The number of protein-protein interactions (PPIs; n=701) between differentially abundant proteins of aneuploid chromosomes and those on other chromosomes against the background distribution for COREAD chromosome 7 amplification.

We hypothesised that these abundance changes should only affect co-complex members of differentially abundant proteins on aneuploid chromosomes, but that non-partner complex members should maintain their abundance level to prevent stoichiometric imbalances. To test this, we performed an association test for the overlap between differentially abundant proteins on other chromosomes and all known human complex members curated from CORUM. As a result of this, we found a significant depletion of complex subunits in differentially abundant proteins on other chromosomes for amplification cases (p-value < 0.05, chi-square test; **Figure 20**). This suggests that complex members overall are stably expressed to prevent disruption of complex stoichiometry upon chromosomal amplification. This effect was not observed in the case of chromosomal deletions in which differentially abundant proteins of other chromosomes are significantly enriched in complex proteins (p-value < 0.05, chi-square test; **Figure 20**).

In contrast to our observations at the proteome level where we observed a consistent pattern of enrichment for co-regulation of co-complex members, we observed both strong enrichments and depletions of co-complex members of proteins encoded by differentially expressed genes on aneuploid chromosomes in the differentially expressed genes on other chromosomes (adjusted p-value < 0.01, chi-square test; **Figure 21A**). In addition, we observed a significant enrichment of protein complex subunits among the differentially expressed genes on other chromosomes for 19 out of 28, and 20 out of 31 significant associations for amplification and deletion cases, respectively, at the transcriptome level (adjusted p-value < 0.01, chi-square test; **Figure 21B**). The lack of consistency for co-regulation of co-complex members at the transcriptome level suggests post-transcriptional compensatory mechanisms to control abundance changes induced by aneuploidy.

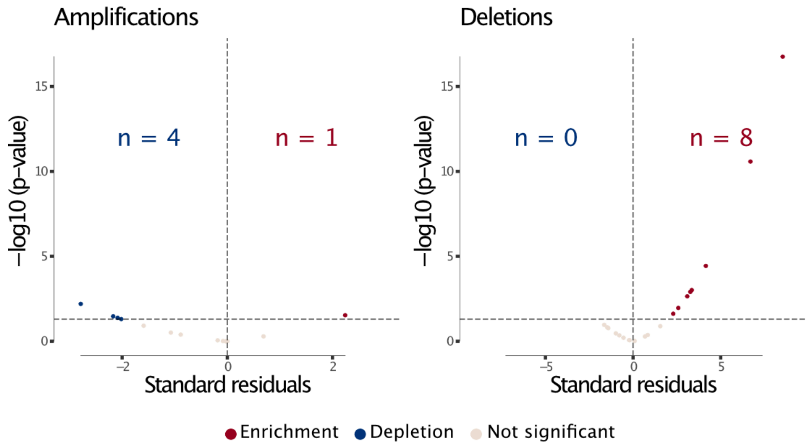


Figure 20. Buffering on abundances of protein complex members of other chromosomes in aneuploid tumours. Standard residuals and p-values for the overlap between CORUM complex subunits and differentially abundant proteins on other chromosomes for 13 amplifications and 20 deletions.

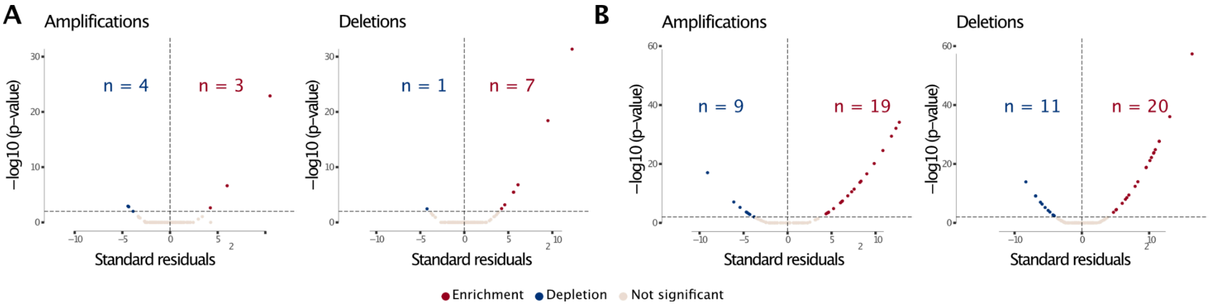


Figure 21. Transcriptome-level changes on other chromosomes. Standard residuals and adjusted p-values for the overlap between **(A)** co-complex members of differentially expressed genes on aneuploid chromosomes and differentially expressed genes on other chromosomes, for 203 detected cancer-type-specific aneuploidies; 86 amplifications and 117 deletions, and **(B)** CORUM complex subunits and differentially expressed genes on other chromosomes.

2.3. Co-deregulation of members of the same complexes as a compensatory mechanism to prevent stoichiometric imbalances in complexes and aggregation

The main expected detrimental effect of chromosomal amplifications is an excess of protein abundance of complex members leading to an aggregation of the orphan proteins (rather than a loss of function of the complex as would be expected for insufficient expression for complex assembly as a consequence of chromosome deletion) (Santaguida et al., 2015). We therefore tested if aggregation-prone proteins on the amplified chromosome show a higher tendency for strong correlations with their complex partners on other chromosomes. We grouped aneuploid proteins as aggregation-prone and non-aggregation-prone based on the data from Määttä et al (Määttä et al., 2020), and compared their protein-level abundance correlations with their complex partners. Indeed, we observed stronger correlations for aggregation-prone proteins as compared to their non-aggregating counterparts (p-value < 2.2e-16, Wilcoxon test; **Figure 22A**). This suggests that upregulating the protein expression of genes on chromosomes not affected by aneuploidy themselves serves as a compensatory mechanism to prevent proteotoxicity triggered by the aggregation of non-paired complex members located on the aneuploid chromosome.

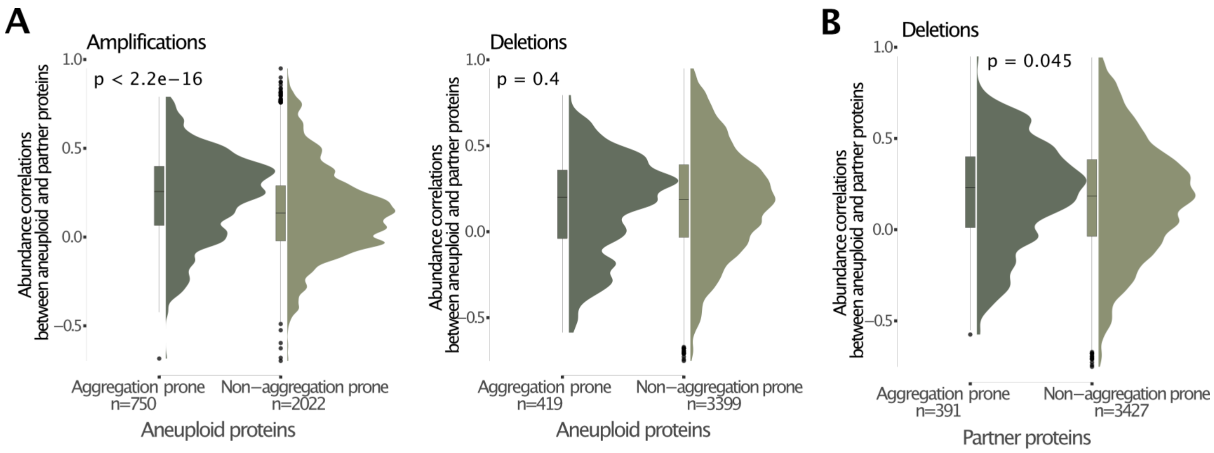


Figure 22. Compensatory mechanisms preventing aggregation of orphan subunits partly explain abundance changes on other chromosomes. **(A)** Protein abundance correlations between differentially abundant proteins on aneuploid chromosomes and their co-complex members on other chromosomes when aneuploid proteins are grouped as aggregation-prone and non-aggregation-prone. **(B)** Aggregation propensity of co-complex members of aneuploid proteins in the case of deletions. Protein abundance correlations between differentially abundant proteins on aneuploid chromosomes and their co-complex members on other chromosomes when co-complex proteins are grouped as aggregation-prone and non-aggregation-prone, in deletion cases. Wilcoxon test was used to test differences between groups.

We hypothesised that in the case of chromosomal deletions, the aggregation propensity of downregulated proteins on the aneuploid chromosome should not affect the degree of correlation with complex partners. Indeed, we observed aggregation-prone proteins to be not related to stronger correlations with their complex partners on other chromosomes when they are encoded on deleted chromosomes (**Figure 22A**). This is likely the case as downregulating those proteins would not leave them as orphan subunits and hence increase their risk of aggregation. However, one would expect that aggregation propensity of co-complex members of downregulated proteins of the deleted chromosomes should have an effect on the co-abundance correlations as this will leave them as orphan subunits. To test this, we compared the co-abundance correlation of proteins of deleted chromosomes with their aggregating co-complex members to non-aggregating ones. We found that deleted aneuploid proteins have significantly stronger correlations with their aggregating co-complex members (p-value = 0.045, Wilcoxon test; **Figure 22B**).

Assuming that the regulation of proteins on other, non-aneuploid chromosomes serves to prevent stoichiometric imbalance of protein complexes, we speculated that for proteins that are in many complexes there are more ways of being abundance-compensated by a complex partner compared to those proteins participating in few complexes and therefore each single partner should be under less stringent control for coexpression with the aneuploid protein. We therefore classified each aneuploid protein into promiscuous (participating in more than 5 complexes) and non-promiscuous (involved only in 5 or less than 5 complexes). As expected, we observed weaker correlations for promiscuous proteins of amplified chromosomes (p-value = 6e-16, Wilcoxon test; **Figure 23**) further supporting the model in which differential abundance of proteins on other chromosomes is a compensatory mechanism. We did not observe the same association in the case of chromosomal deletions (**Figure 23**).

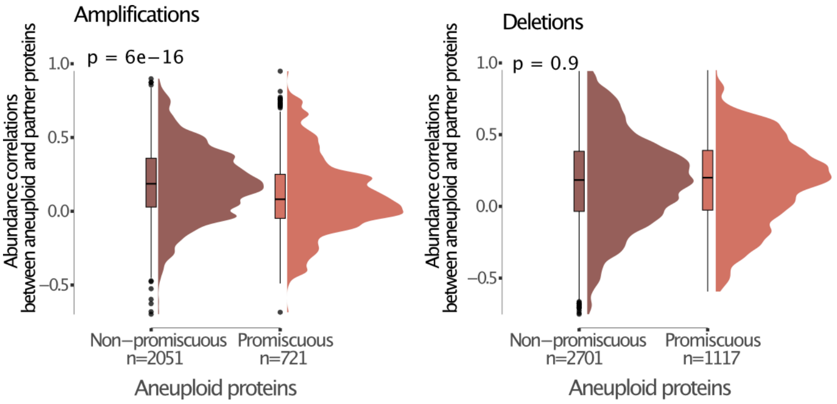


Figure 23. Compensatory mechanisms preventing imbalances in complexes partly explain abundance changes on other chromosomes. Protein abundance correlations between differentially abundant proteins on aneuploid chromosomes and their co-complex members on other chromosomes when aneuploid proteins are grouped as promiscuous and non-promiscuous. Wilcoxon test was used to test differences between groups.

Finally, we hypothesised that proteins co-occurring in many complexes should show stronger correlation as proteins found only in a few cases together in the same complex. Indeed, we found significant differences in the number of times aneuploid proteins and their positively correlated co-complex members were found together in the same complex vs their uncorrelated or negatively correlated co-complex members (p-value = 2.2e-12 and p-value = 2.8e-14 for chromosomal amplifications, p-value = 0.004 and p-value = 7.4e-05 for deletions; **Figure 24**). This, again, illustrates how complex organisation shapes the co-abundance patterns between differentially abundant proteins from the aneuploid chromosome and those located on other chromosomes.

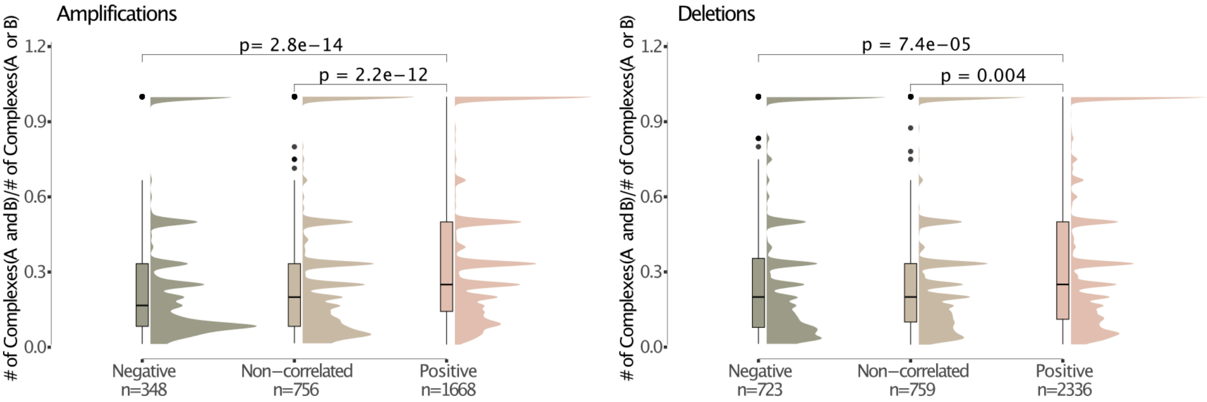


Figure 24. Proteins frequently co-occurring in many complexes with their aneuploid partner are related to stronger co-abundance regulation. Co-occurrence frequency of differentially abundant proteins on aneuploid chromosomes and their co-complex members on other chromosomes in different correlation groups, positively and negatively correlated and non-correlated co-complex members of aneuploid proteins. Wilcoxon test was used to test differences between groups.

3. Functional selection acting on keeping stoichiometry in check

In the previous sections, we proposed that co-abundance changes of protein complex partners is a compensation mechanism to prevent stoichiometric imbalance in protein complexes to avoid proteotoxicity of orphan subunits. We next wondered if besides biophysical (such as aggregation propensity) any functional properties would protect complexes and complex subunits from abundance imbalance in aneuploid human tumours. To this end, we first retrieved the most strongly correlated co-complex members of differentially abundant aneuploid proteins and identified the complexes they are involved in. Then, we obtained functional annotations of the complexes from the CORUM database. To identify functions under stronger protection from protein abundance imbalance in complexes, we computed the enrichment of these functions compared to a random set of complexes under relaxed stoichiometric protection. The analysis revealed that top correlated proteins form complexes that are frequently involved in translation (mainly driven by ribosomal proteins; see Materials and methods section “Functional annotations of protein complexes”), RNA splicing, RNA processing and protein complex assembly (**Figure 25; Supplementary File 7**). Interestingly, the functional enrichment is consistent for amplifications and deletions suggesting that not just compensatory mechanisms to prevent proteotoxicity contribute to the dysregulation of proteins on other chromosomes but also functional selection is in place, acting on important cancer-essential functions up- or down-regulating entire protein complexes while keeping their stoichiometry in check.

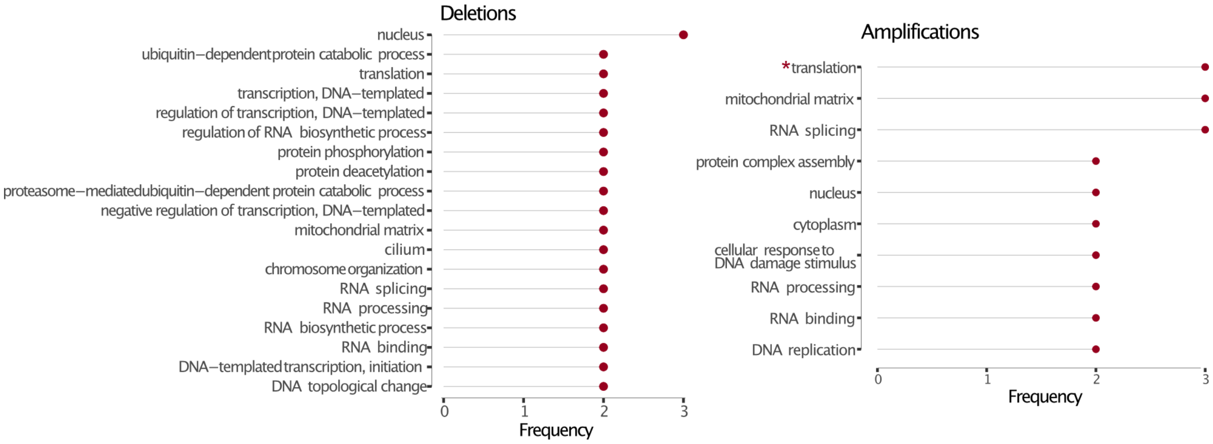


Figure 25. Enrichment of functional terms in complexes of top correlated proteins in aneuploid tumours - Part I. Most frequently enriched terms in the amplification and deletion cases. Frequency shows the number of aneuploidy cases in which the corresponding term is significantly enriched (* Enrichment is mostly driven by ribosomal genes).

To quantitatively compare the degree of enrichment between the functional terms associated with balance-protected complexes, we devised an enrichment score (see Materials and methods section “Functional annotations of protein complexes”) and compared it for the top enriched or depleted functions between amplifications and deletions. We observed that top correlations in the deletion cases are related to stronger enrichment scores when compared to their counterparts in the amplification cases (**Figure 26**).

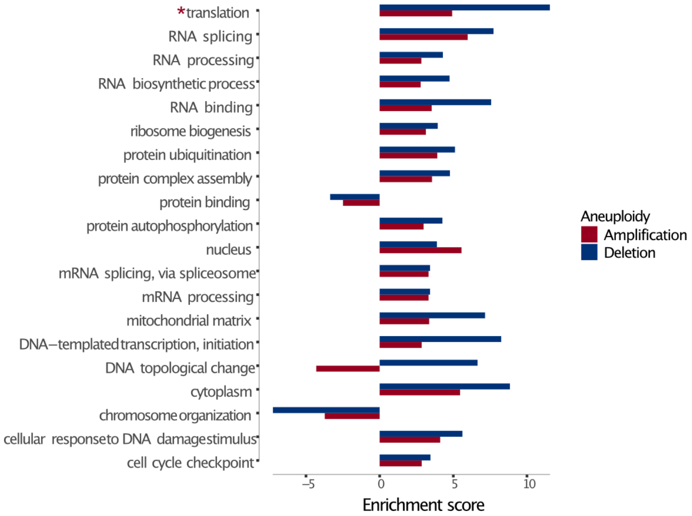


Figure 26. Enrichment of functional terms in complexes of top correlated proteins in aneuploid tumours - Part II. Enrichment scores of enriched terms both in amplification and deletion cases. For the functional term that is enriched in more than one amplification/deletion cases, the enrichment score of the ones with the lowest p-value is displayed (* Enrichment is mostly driven by ribosomal genes).

4. Levels of regulation of dysregulation on other chromosomes in aneuploid tumours

4.1. The dysregulation on other chromosomes cannot be fully explained by epigenetic and transcriptional control

Previous studies have revealed the role of epigenetic and transcriptional regulatory mechanisms in cancer: Differential DNA methylation and dysregulation of transcription factors mediate aberrant gene expression in cancer (Baylin & Herman, 2000; Bushweller, 2019; Ehrlich, 2002). Thus, we further aimed to disentangle the different regulatory layers underlying expression changes on other chromosomes induced by aneuploidy. We first tested if differential DNA methylation (epigenetic silencing) of the promoters of genes could explain the corresponding expression changes in aneuploid samples. Therefore, we focused on differentially expressed genes of other chromosomes, and compared average DNA methylation levels of those genes in aneuploid samples to diploid samples, separately for up- and downregulated genes. We found that downregulated genes are significantly related to higher methylation levels in aneuploid samples in only 6 amplification cases out of 86 (~7%) and in 5 deletion cases out of 117 (~4%) (p-value < 0.05, Wilcoxon test; **Figure 27A**). We observed significant associations between lower methylation level and upregulated genes in aneuploid samples for few cases (10 out of 86 amplification cases and 16 out of 117 deletion cases) (p-value < 0.05, Wilcoxon test; **Figure 27B**). This suggests that epigenetic regulation does not have a substantial contribution to the described genome-wide changes in gene expression induced by aneuploidy.

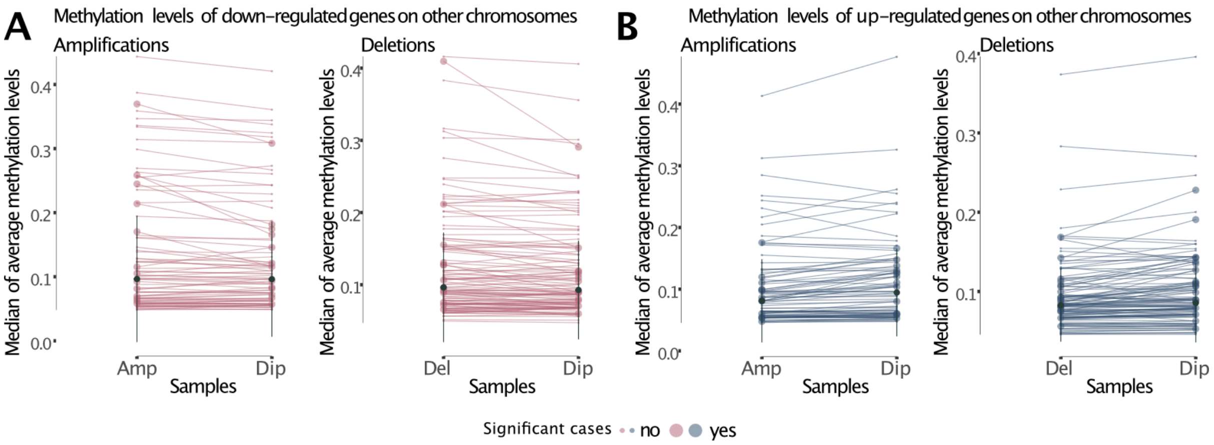


Figure 27. The contribution of promoter methylation on dysregulation patterns of other chromosomes. Average methylation level of **(A)** downregulated and **(B)** upregulated genes in aneuploid vs diploid samples. Wilcoxon test was used to test differences in methylation level of genes between aneuploid and diploid samples (*: p-value < 0.05). Each dot represents median methylation level in a cancer-type-specific aneuploidy case, and those where we observed significant differences in methylation change were highlighted with bigger point size.

We then asked if differential expression of TFs on the aneuploid chromosome could explain the large transcriptional changes on other chromosomes. We therefore tested for a large list of ENCODE gene-TF associations if there is an enrichment of targets of differentially expressed TFs of the aneuploid chromosome among differentially expressed genes on the other chromosomes. Performing a randomization test did not reveal an excess of targets for any of the tested, cancer-type-specific altered chromosomes (**Figure 28**).

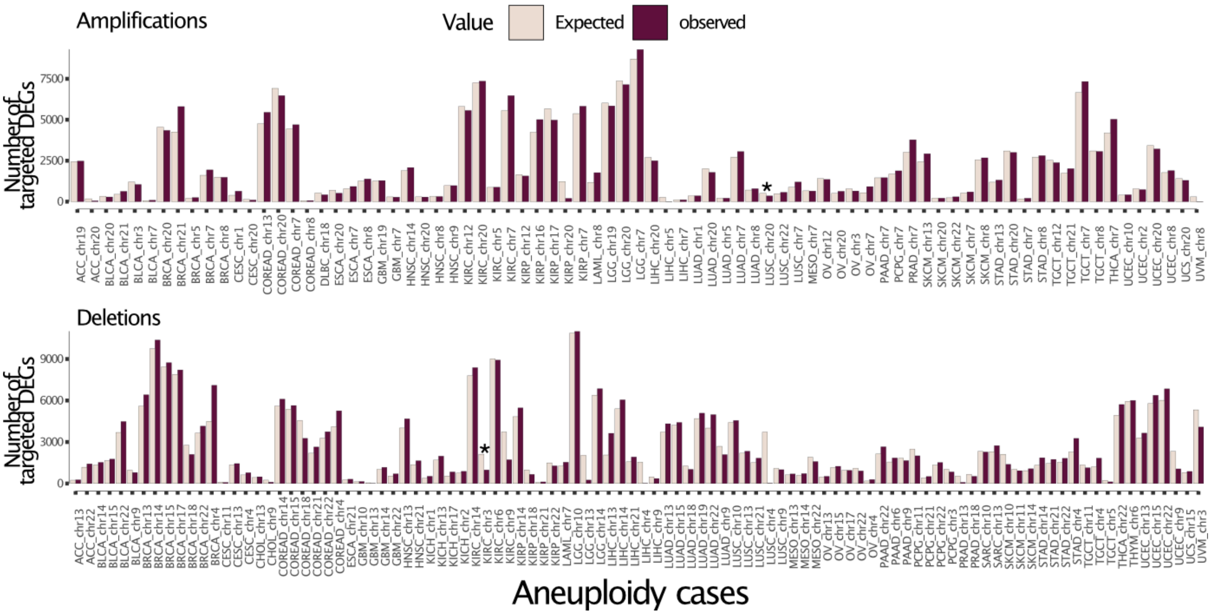


Figure 28. Transcriptional regulation of genes on other chromosomes. The number of targets of differentially abundant transcription factors on aneuploid chromosomes among differentially expressed genes on chromosomes in amplification and deletion cases. Expected value and p-value were calculated using a randomization test (*: $p < 0.05$).

As a control, we computed the differentially expressed genes between tumour and healthy samples for 21 TCGA cancer types, where we have tumour vs normal samples: In those, downregulated genes are often hypermethylated in cancer (**Figure 29**) suggesting that DNA methylation plays an important role in regulating gene expression during carcinogenesis. Even though not significant, we observed a higher number of targets of differentially expressed TFs among differentially expressed genes for 76% (16 out of 21 cancer types) of cancer types (**Figure 30**) as compared to 62% in aneuploid tumours. In addition, we observed a higher absolute difference between the number of observed and expected targets in tumour vs normal samples (353.71 and 663.34, respectively for aneuploid tumours and tumour vs normal). Lastly, we tested the regulatory impact of the differential expression of the well-known cancer-related TF MYC on the expression of its target genes. We found MYC differentially expressed in 14 out of 21 cancer types, and in those cancer types its targets are significantly enriched among differentially expressed genes (p-value < 0.05, chi-square test). Together these observations show that our

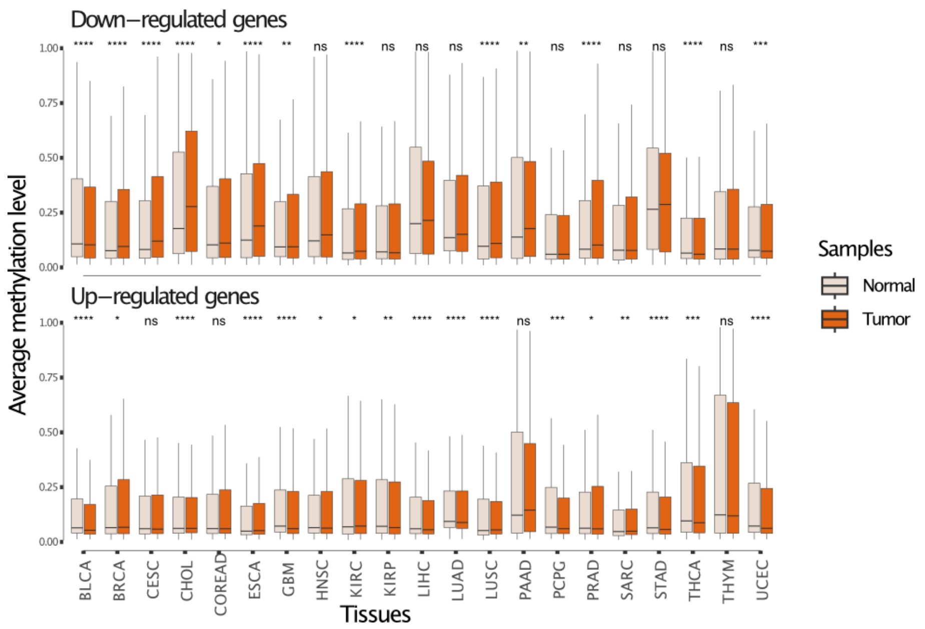


Figure 29. The contribution of promoter methylation on expression changes between tumour and normal. Average methylation level of down- and upregulated genes in tumour vs normal samples. Wilcoxon test was used to test differences in methylation levels between sample groups (n.s: Not significant, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001).

measures of transcriptional regulation can capture some regulatory activity in cancer but the absence of signals in aneuploid tumours suggests that transcriptional regulation cannot fully explain the expression changes on other chromosomes.

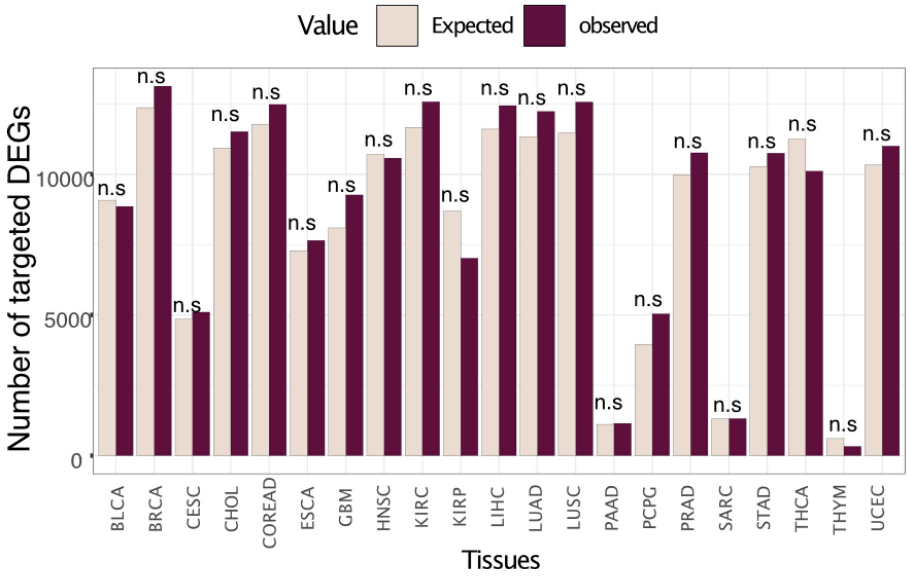


Figure 30. Transcriptional regulation on expression changes between tumour and normal. The number of targets of differentially expressed TFs in tumours vs normal among differentially expressed genes. expected value and p-value were calculated using a randomization test (n.s: Not significant, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001).

4.2. Post-translational regulation of co-complex members of aneuploid proteins

We observed a stronger association between complex partner co-regulation on proteome as compared to transcriptome level (**Figure 17 and Figure 21A**) suggesting a central role for translational or post-translational regulation in maintaining complex protein abundance balance in aneuploid cells. To further validate this, we looked at the transcript levels of co-complex members of aneuploid proteins and asked if the corresponding changes could be explained by differential methylation or differential activation of TFs encoded on aneuploid chromosomes. Indeed, we did not observe an overall significant association further supporting the role of translational or post-translational mechanisms on co-abundance regulation (**Figure 31 and Figure 32**).

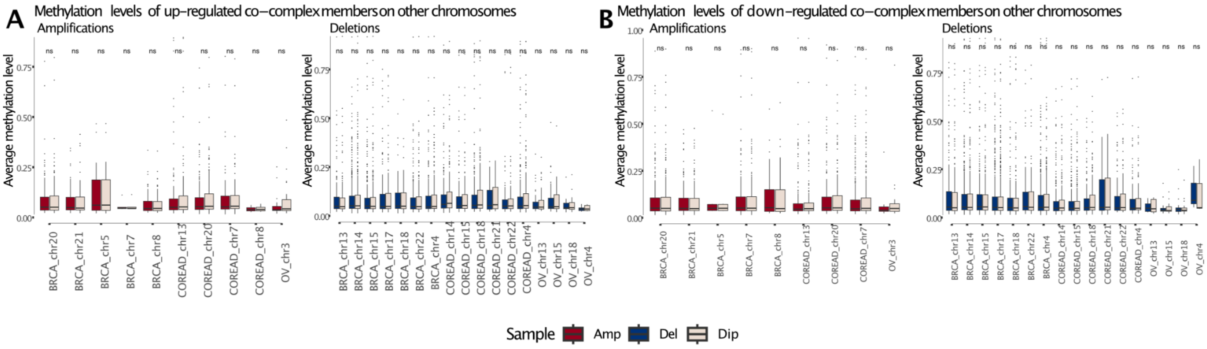


Figure 31. The effect of promoter methylation on the expression changes of co-complex members of aneuploid proteins in aneuploid tumours. Average methylation level of **(A)** up- and **(B)** downregulated co-complex members of aneuploid proteins on other chromosomes in aneuploid vs diploid samples. Wilcoxon test was used to test differences in methylation levels between sample groups (n.s: Not significant, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001).

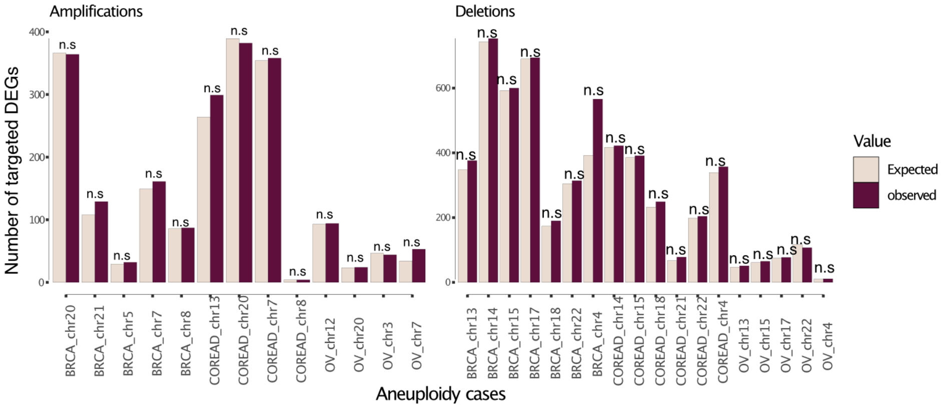


Figure 32. Transcriptional regulation on the expression changes of co-complex members of aneuploid proteins in aneuploid tumours. The number of targets of differentially expressed TFs on aneuploid chromosomes among differentially expressed co-complex members of aneuploid proteins on other chromosomes. Expected value and p-value were calculated using a randomization test (n.s: Not significant, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001).

Previous studies have suggested that ubiquitination at multiple sites is an efficient signal for degradation (Dimova et al., 2012) and further increase in the number of ubiquitination sites is related to higher binding affinity between protein and proteasome (Lu et al., 2015). We therefore hypothesised that post-translational ubiquitination of proteins could regulate co-abundance changes of partners of aneuploid proteins on other chromosomes. To test this, we retrieved ubiquitination data from PhosphoSitePlus (Hornbeck et al., 2015) and tested if top correlated co-complex members of aneuploid proteins tend to have higher number of ubiquitination sites (as a proxy to identify proteins that can be more easily targeted for degradation). Indeed we found that top correlated partners have significantly higher numbers of ubiquitination sites (p-value < 0.05, Wilcoxon test; **Figure 33**). This suggests that a primary mechanism for keeping protein complex stoichiometry in check seems to be indeed post-translational regulation (such as ubiquitin-mediated degradation).

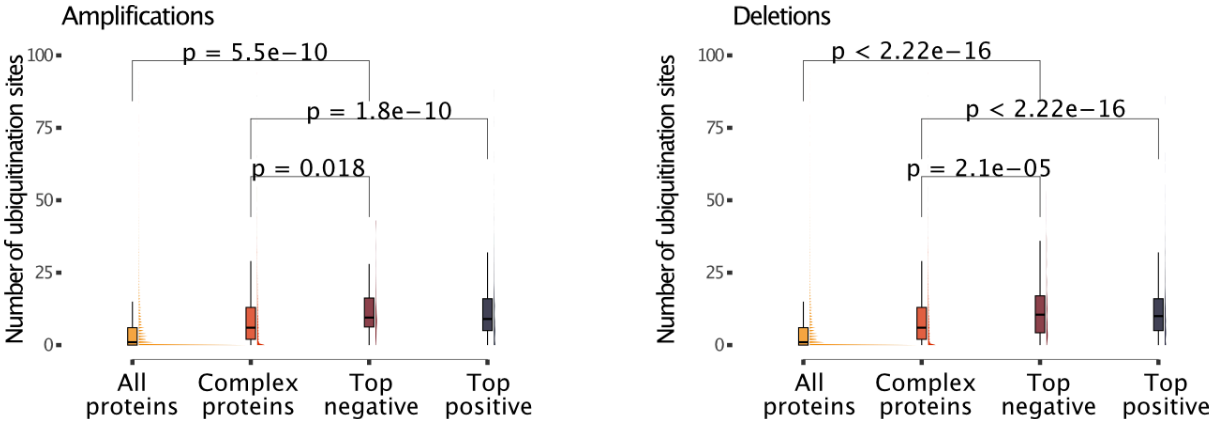
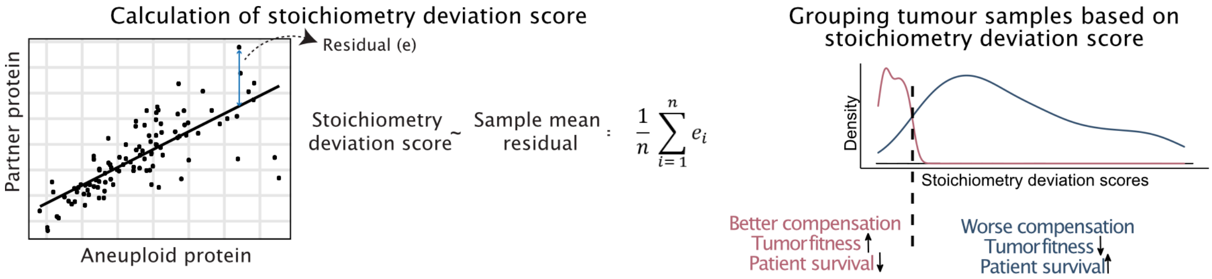


Figure 33. Post-translational regulation of co-complex members of aneuploid proteins. Number of ubiquitination sites of all, human complex, and top positively and negatively correlated proteins. Wilcoxon test was used to test differences between groups.

5. Phenotypic consequences of compensation for stoichiometric imbalances in aneuploid tumours

The previous results suggest co-regulation of co-complex members as a compensation mechanism to balance protein abundance changes caused by chromosome-level alterations, and thus to keep protein complex stoichiometry in check. We reasoned that different tumours might be able to compensate for the dysregulation of proteins on the aneuploid chromosome with a different degree of success and hypothesised that tumours that can better compensate for protein abundance changes will be associated with better survival rates while those that fail to compensate should upregulate components of the protein degradation machinery to clear the cell from the orphan complex subunits. To test this, we first aimed to measure the degree of failure to keep complex stoichiometry balance in each aneuploid tumour sample. In this regard, we focused on protein-level correlations between differentially abundant aneuploid proteins and their co-complex members of other chromosomes, of which we used the top 30 strongest correlated pairs. Finally, we performed a linear regression model in which we calculated the mean of residuals - stoichiometry deviation score - (**Figure 34**) as a degree of failure in keeping complex stoichiometry for each sample.



Calculation of stoichiometry deviation score
Residual (e)
Partner protein
Aneuploid protein
Stoichiometry deviation score ~ Sample mean residual : $\frac{1}{n}\sum_{i=1}^{n} e_i$
Grouping tumour samples based on stoichiometry deviation score
Density
Stoichiometry deviation scores
Better compensation
Tumor fitness ↑
Patient survival ↓
Worse compensation
Tumor fitness ↓
Patient survival ↑

Figure 34. Graphical representation for the calculation of sample stoichiometric deviation score and grouping samples based on this score. Pink and blue colours represent tumour samples with low and high stoichiometry deviation score, respectively (n = 30, referring to top 30 correlations).

To test if sample stoichiometry deviation score shows consistency when different numbers of strongly correlated protein pairs were considered, we re-calculated sample stoichiometry deviation score by using different sets of protein pairs, and then looked at the correlations among the scores. We, indeed, found that different sets of top-correlated protein pairs did not affect the overall stoichiometry deviation score of a sample (**Figure 35**).

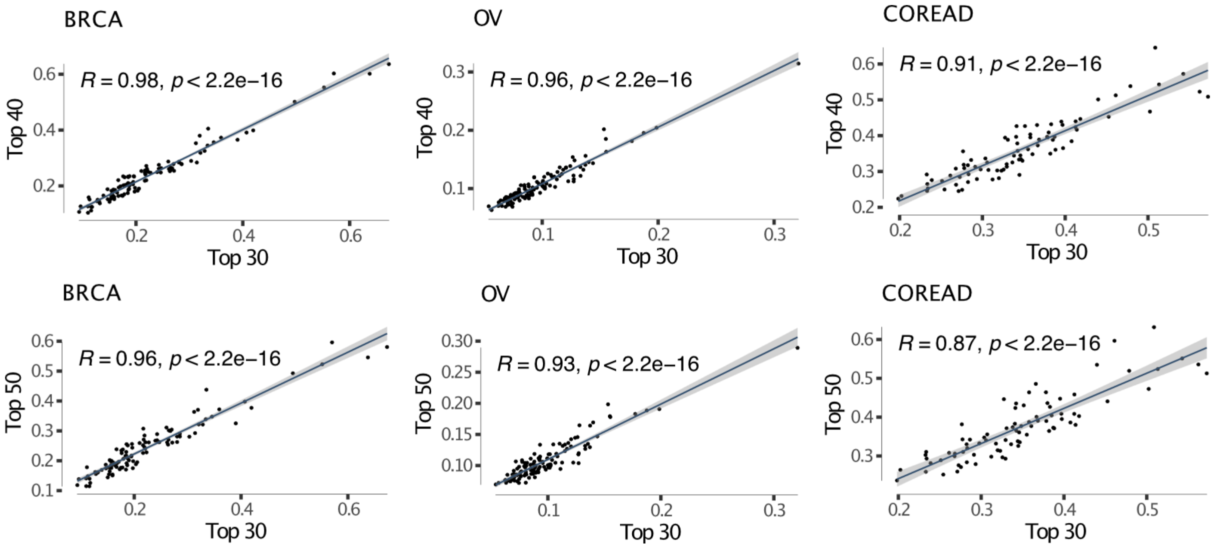


Figure 35. Correlations among different deviation scores. Correlations between the mean stoichiometric deviation scores across samples when different numbers of highly correlated protein pairs (top 30, top 40 and top 50) were considered. Spearman method was used to assess correlation and related p values.

Next, we grouped tumour samples based on their stoichiometric deviation score position relative to the overall distribution within each cancer type, and obtained two groups: Samples with high deviation score (high) and those with low deviation score (low) (**Figure 36**). Then, we performed a survival analysis once by using overall survival and once by using disease free survival. While not significant in every single case we observed a tendency that samples with low stoichiometry deviation scores are related to lower survival probabilities in all three tissue types (**Figure 37**) showing that compensation for protein abundance indeed provides a survival advantage to tumours.

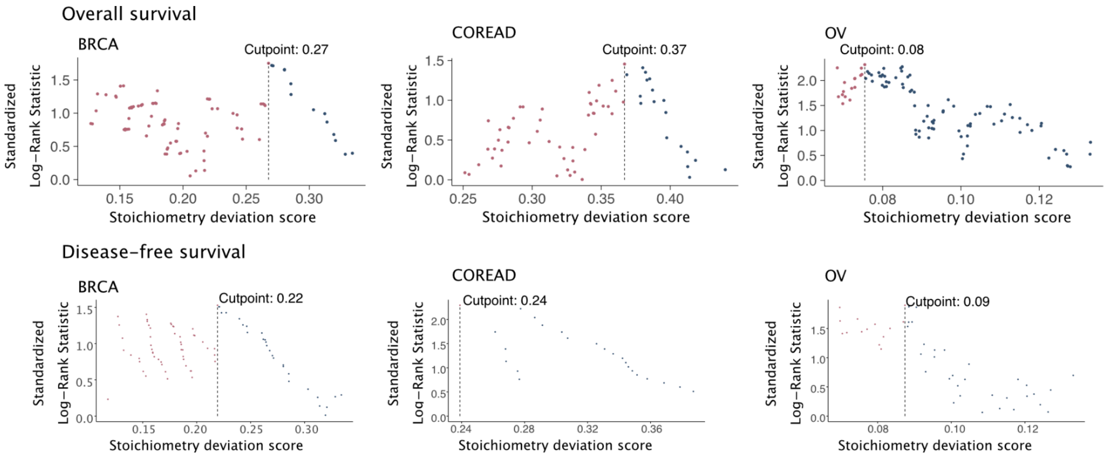


Figure 36. Classification of TCGA samples based on their stoichiometric deviation scores. Distribution of stoichiometric deviation scores and cutpoints dividing samples as high (represented in blue) and low (represented in pink) based on stoichiometry deviation scores. Cutpoints were calculated by using the maximally selected rank statistics.

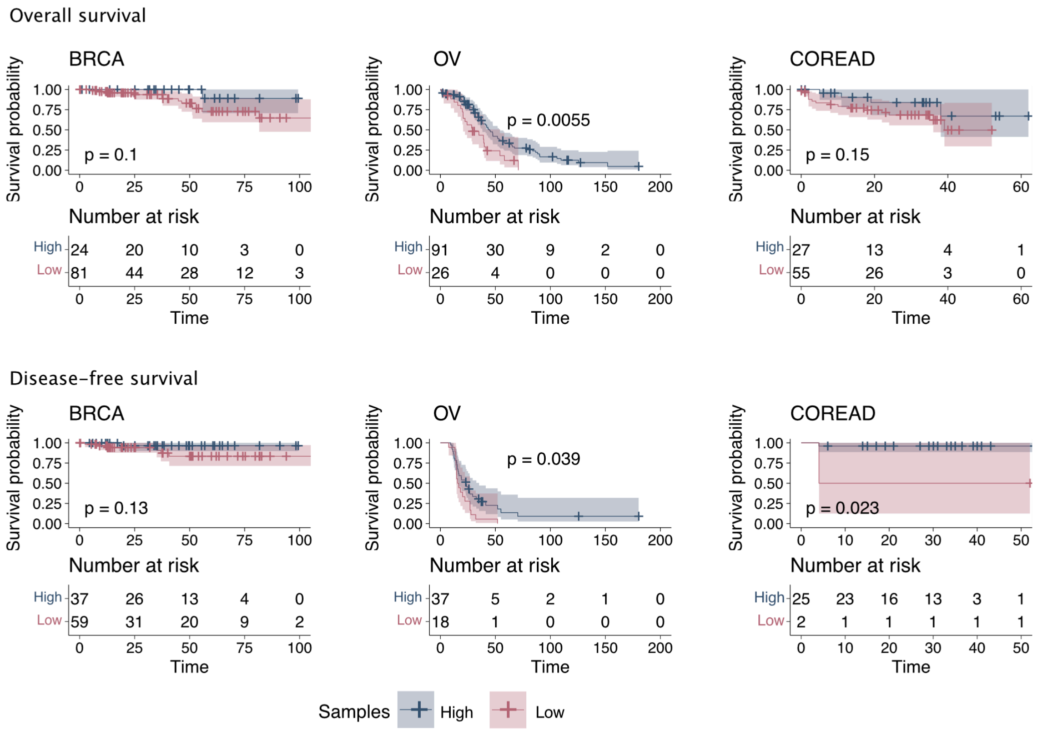


Figure 37. Consequences of stoichiometric compensation. Survival analysis results within each TCGA cancer tissue. Survival analysis was done once with overall survival and once with disease-free survival.

We further investigated if the proteins that play a role in protein degradation have higher abundances in the tumours that cannot compensate for abundance changes, and thus have to deal with the excess amount of orphan subunits. We indeed found that ubiquitin-binding proteins and components of the proteasome show significantly higher correlations between their abundances and the stoichiometry deviation scores in two out of three tissues (p-value < 0.05, Wilcoxon test; **Figure 38**), and this tendency still applies when samples are divided into amplification and deletion groups (**Figure 39**). This likely is a consequence of proteotoxic stress resulting from the inability of some tumours to keep protein complexes balanced.

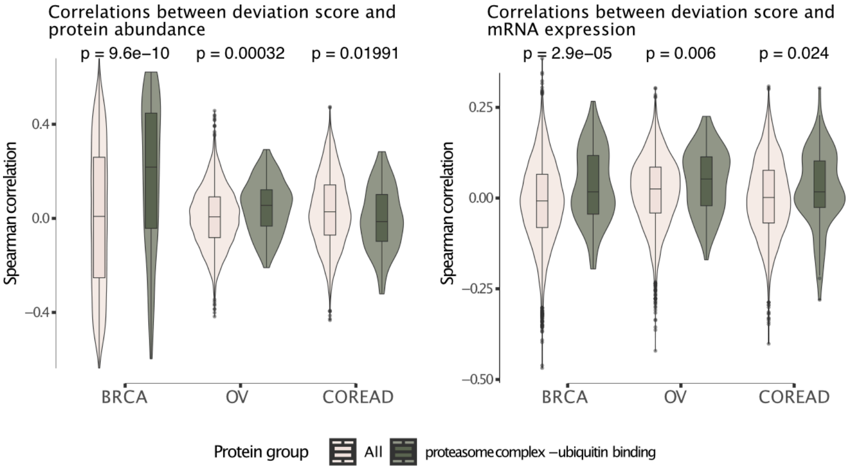


Figure 38. The effect of stoichiometric compensation on the proteome. Correlations between the stoichiometric deviation scores and protein abundance/mRNA expression of all proteins, and proteasome complex - ubiquitin binding proteins. Wilcoxon test was used to test differences between groups.

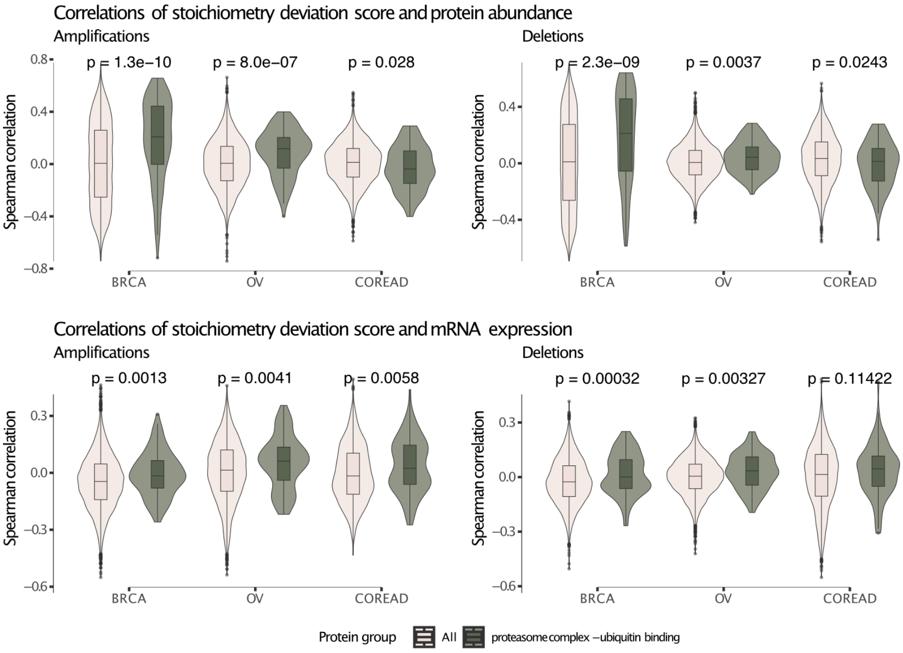


Figure 39. The effect of stoichiometric compensation on the proteome separately for amplification and deletion cases. Correlations between the stoichiometric deviation scores and protein abundance/mRNA expression of all proteins, and proteasome complex - ubiquitin binding proteins when samples were separated as amplification and deletion groups. Wilcoxon test was used to test differences between groups.

Chapter II

1. Dosage compensation in protein complexes could play a role in aneuploidy patterns in cancer genomes

In the previous chapter, we characterised transcriptome and proteome dysregulation induced by chromosome-level aneuploidies in cancer genomes, and showed the role of co-abundance regulation of protein complex members in compensation for overall dysregulation patterns to prevent proteotoxicity and maintain functional complex organisation. Moreover, we demonstrated that the degree of success in this compensation impacts the tumour fitness. We would therefore assume that negative selection acts against chromosome-level genomic alterations where an excess amount of protein product cannot be buffered. Correspondingly, Schuster-Böckler and colleagues showed that genes encoding for protein complex subunits are less prone to CNVs in human cell lines (Schuster-Böckler et al., 2010). Therefore, we aimed to understand if the same constraints could play a role in shaping the observed aneuploidy patterns in human tumours. For this, we first calculated amplification and deletion frequency for each chromosome across all tumour samples covering 32 TCGA cancer types (since certain chromosomes are repeatedly amplified or deleted across different tissues, e.g. chr20 amplification and chr13 deletion in **Figure 6**, even though there are tissue-specific patterns). Then we computed correlations between the alteration frequency of a chromosome and number of genes coding for protein complex subunit. We, indeed, found significant negative correlation between the total alteration frequency of a chromosome and the number of genes coding for complex subunits ($p < 0.05$, Spearman correlation; **Figure 40A**) suggesting that chromosomal regions encompassing a large number of complex subunit coding genes are protected from chromosome-level aneuploidies. Moreover, we found that this association is stronger for chromosomal deletions when compared to that of chromosomal amplifications (**Figure 40A**). This makes sense since the possible outcome of amplification events (e.g. excess amount of protein products) can be more easily compensated through transcriptional and/or post-transcriptional regulation (such as protein degradation machinery).

It has been previously shown that gene density is inversely correlated with the total aneuploidy score in human cell lines (Klaasen et al., 2022). Since we also observed a strong negative correlation between gene density and total alteration frequency of a chromosome in human tumours (**Figure 40B**), we further aimed to understand at which degree this correlation is driven by protein complex dynamics. To this end, we first grouped genes as those encoding for proteins involved in many complexes (promiscuous genes) and those that encode proteins involved in few complexes (non-promiscuous genes), and then computed correlations between the gene density and total aneuploidy frequency separately for two gene groups. We found a relatively stronger negative correlation for non-promiscuous genes when compared to their promiscuous counterparts (**Figure 40C & 40D**). This makes sense as there will be more ways to compensate for abundance changes of promiscuous proteins (in line with our previous finding where we showed higher level co-abundance compensation for non-promiscuous aneuploid proteins, **Figure 23**), and further suggests the role of protein complex dynamics in shaping aneuploidy patterns in human tumours.

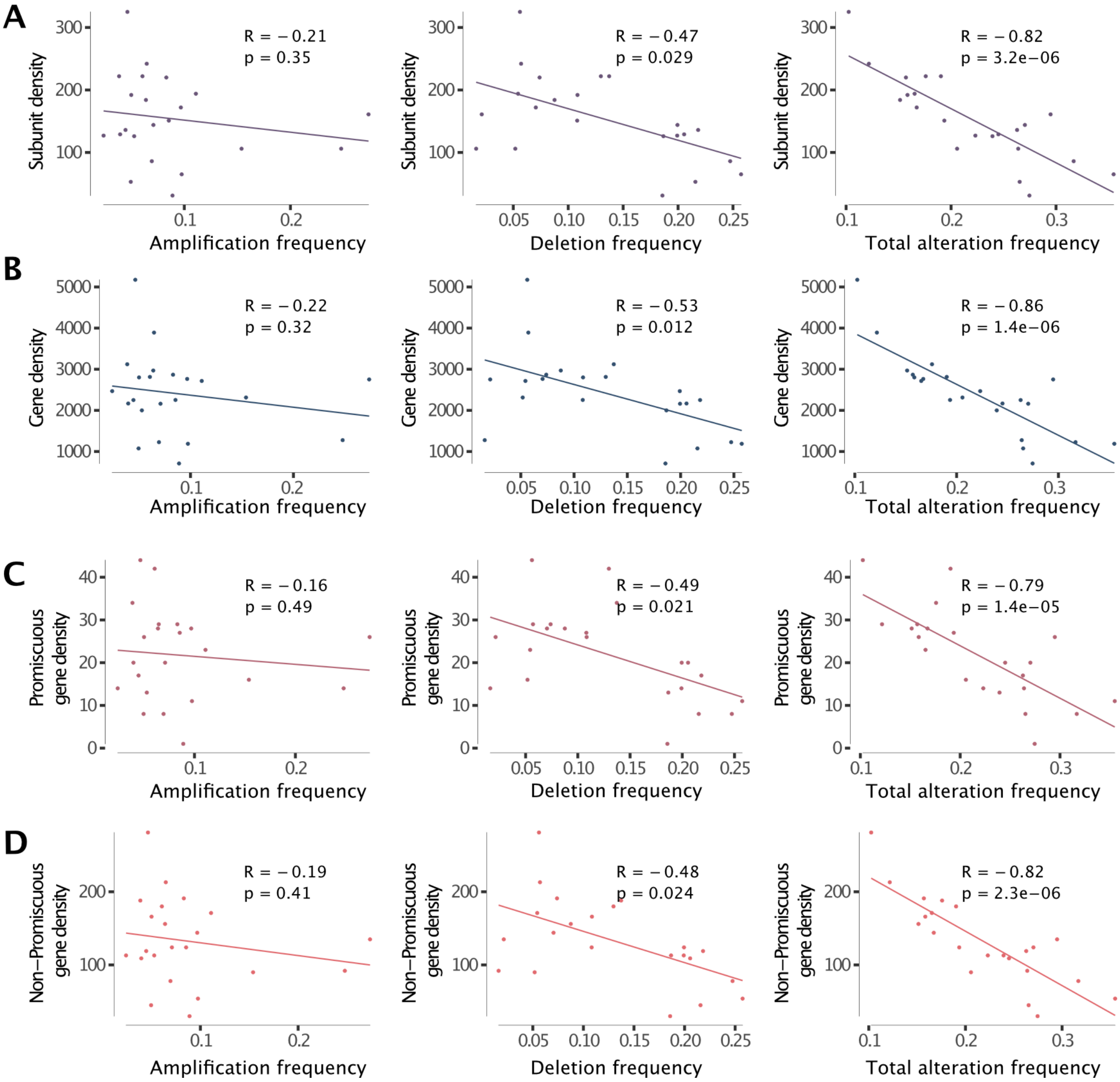


Figure 40. Correlations between alteration frequency (chromosome-level amplification, deletion, and total - amplification and deletion together-) and density of different gene groups. Each dot represents a chromosome (from chr1 to chr22). Spearman correlation was used to calculate correlation coefficient and p-value.

2. Predicting tissue-specific CNAs and aneuploidies in cancer genomes by using machine learning models

Our findings suggest that negative selection might be acting against copy number alterations in genomic regions rich in genes encoding protein complex subunits. Previous studies demonstrated the role of tissue-specific features (e.g. OGs and TS) as drivers of aneuploidy patterns in different cancer types (Davoli et al., 2013; Jubran et al., 2023). Therefore, we asked how well machine learning algorithms could predict recurrent CNAs and aneuploidies in cancer if we integrate the features that we studied in this thesis and those from other studies. In chapter I where we detected dysregulation patterns in aneuploid tumours (section 1.2. Transcriptome and proteome changes in aneuploid tumours), we calculated cancer-type-specific chromosome-level aneuploidy scores by using publicly available arm-level aneuploidy score data (Taylor et al., 2018). To better understand whether the dynamics between factors and resulting genomic alterations depend on the size of genomic alteration, we additionally considered arm-level aneuploidy scores (Taylor et al., 2018), as well as cancer-type-specific focal CNAs frequencies (Alfieri et al., 2023), ranging from 1 Mbp to 50 Mbp. This gave us 29 segment-levels with different genomic ranges (from 1 Mbp to 50 Mbp, arm- and chromosome-level) (**Figure 41**). We then calculated cancer-type-specific arm- and chromosome-level amplification and deletion frequencies by following the same calculation method used in Alfieri et al., 2023 (Alfieri et al., 2023): The number of samples with event (amplification or deletion) was divided by the total number of samples. All cancer types were pooled for the following analyses. For this, we considered the ones for which data for each feature was available, which left us with six cancer types; BRCA, COREAD, GBM/LGG, LUAD, LUSC, and PAAD.

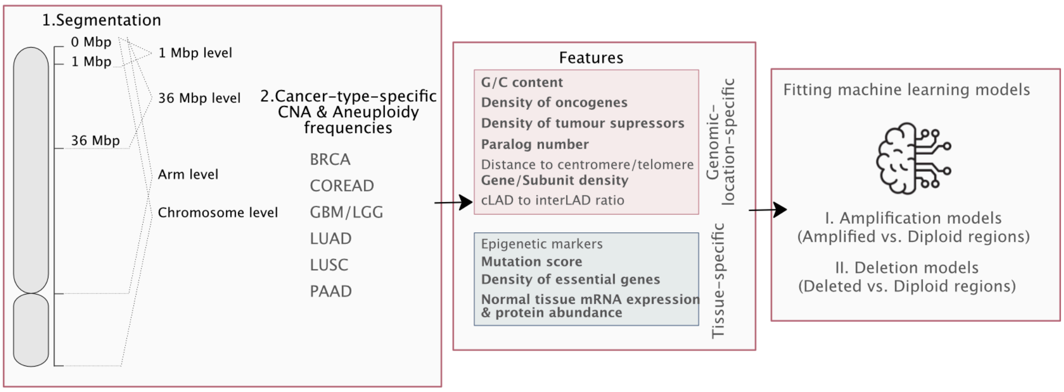

1.Segmentation
0 Mbp
1 Mbp
36 Mbp
1 Mbp level
36 Mbp level
Arm level
Chromosome level
2.Cancer-type-specific CNA & Aneuploidy frequencies
BRCA
COREAD
GBM/LGG
LUAD
LUSC
PAAD
Features
G/C content
Density of oncogenes
Density of tumour supressors
Paralog number
Distance to centromere/telomere
Gene/Subunit density
cLAD to interLAD ratio
Genomic-location-specific
Epigenetic markers
Mutation score
Density of essential genes
Normal tissue mRNA expression & protein abundance
Tissue-specific
Fitting machine learning models
I. Amplification models (Amplified vs. Diploid regions)
II. Deletion models (Deleted vs. Diploid regions)

Figure 41. Graphical representation of machine learning approach predicting CNA and aneuploidy patterns in cancer. Features modulating selection are indicated in bold. The other features are associated likelihood of occurrence.

Next, we aimed to construct a feature set by incorporating tissue- and genomic-location-specific features. To do this, we collected a comprehensive set of features (21 in total) which we classified into two groups: Factors impacting the likelihood of occurrence of amplifications/deletions and those playing a role in the selection of alteration patterns during tumour evolution. First group contains (i) histone marks representing the number of peaks inferred from histone modification profiles in normal tissue (Bernstein et al., 2010; Roadmap Epigenomics Consortium et al., 2015), (ii) cLADs to interLADs ratio representing the part of genomic regions with conserved LaminB1 interactions (Meuleman et al., 2013), (iii) absolute distances to the centromere and the closest telomere. The second group contains (i) density of TS and OGs (Davoli et al., 2013) and essential genes inferred from CRISPR gene effect score data from DepMap (Dempster et al., 2019, 2021; Meyers et al., 2017; Pacini et al., 2021), (ii) mutation score (Alfieri et al., 2023), (iii) tissue-specific median expression of genes located on the corresponding altered genomic region and that of their protein complex partners encoded by genes located outside of the corresponding genomic region (based gene expression profiles in normal tissue from GTEx) (GTEx Consortium, 2013), and (iv) their corresponding tissue-specific protein abundances calculated by using proteome profiling in normal tissue from eGTEx (GTEx Consortium, 2013), (v) G/C content, (vi) the number of genes on the corresponding altered genomic region (gene density) and (vii) the number of their paralogs located outside of the corresponding region, and finally (viii) the number of genes (on the corresponding altered genomic region) encoding for protein complex subunits (Giurgiu et al., 2019) (**Figure 41**).

To systematically test individual effects of features on amplification and deletion frequencies, we first pooled all six cancer types, and then computed Spearman correlations between features and amplification/deletion frequencies separately for each of the 29 segment-levels. We observed that the density of TS, on average, ranked as first and second among all the features based on the absolute correlation, respectively for deletion and amplification model (**Figure 42**). Moreover, the direction of the correlation, negative and positive for amplification and deletion frequency, respectively, (**Figure 42**) suggests that regions with high abundance of TS are less likely to be amplified, further supporting previous findings (Davoli et al., 2013). Another feature showing relatively higher correlations than others is the median expression levels of partner proteins (ranked as first and second for amplification and deletion models, respectively; **Figure 42**), further highlighting the role of co-regulation of members of the same complexes in shaping observed alteration patterns in cancer genomes. Recently it has been shown that amplification of particular genomic regions are under positive selection so that they could buffer the deleterious effects of coding mutations in mutation-rich and essential regions (Alfieri et al., 2023). Indeed, we found that the mutation score is strongly correlated with amplification frequency while we did not observe this trend for deletion frequencies (**Figure 42**). Interestingly, some features follow an increasing/decreasing correlation trend from 1 Mbp-level to chromosome-level (e.g. density of essential genes, OGs and TS) while some features are following stable close-to-zero correlations across focal CNAs and relatively higher (in absolute term) values for arm- and chromosome-level aneuploidies (e.g. the number of gene and protein coding genes on segment, median abundance of partner proteins; **Figure 43**) suggesting the contribution of different features to the resulted alteration patterns in cancer could vary depending on the size of genomic alterations.

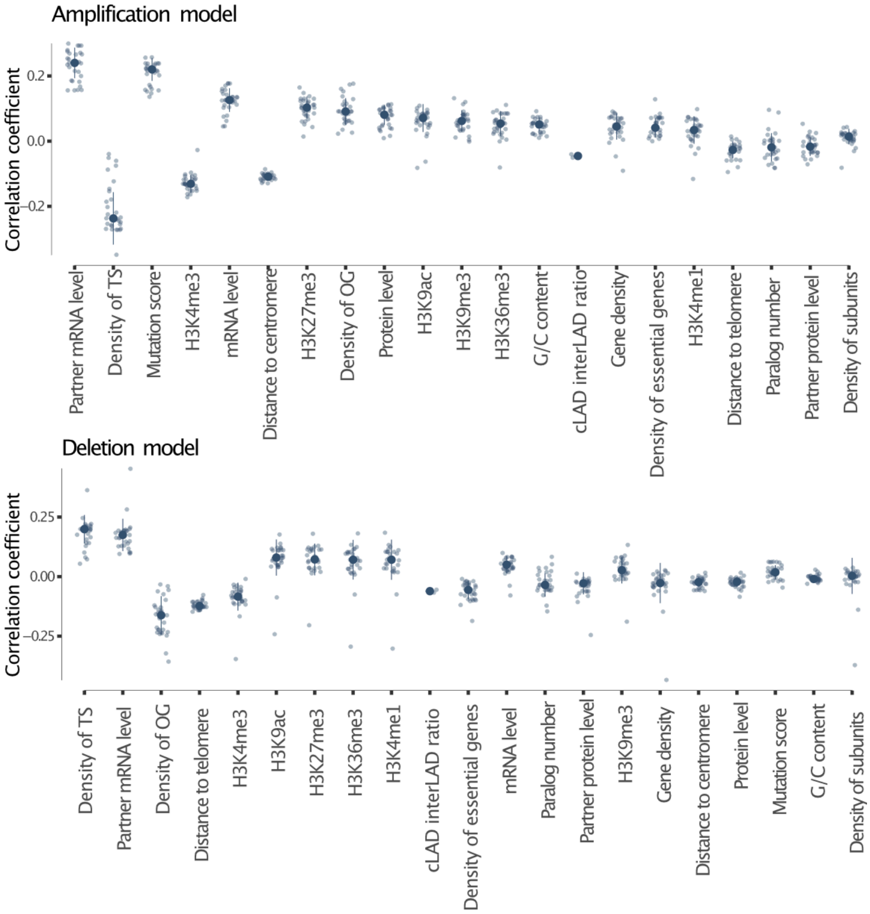


Figure 42. Spearman correlation between features and CNA and aneuploidy frequencies. Correlations were calculated across all six cancer types. Each dot represents correlation value for a specific segment level (29 levels in total).

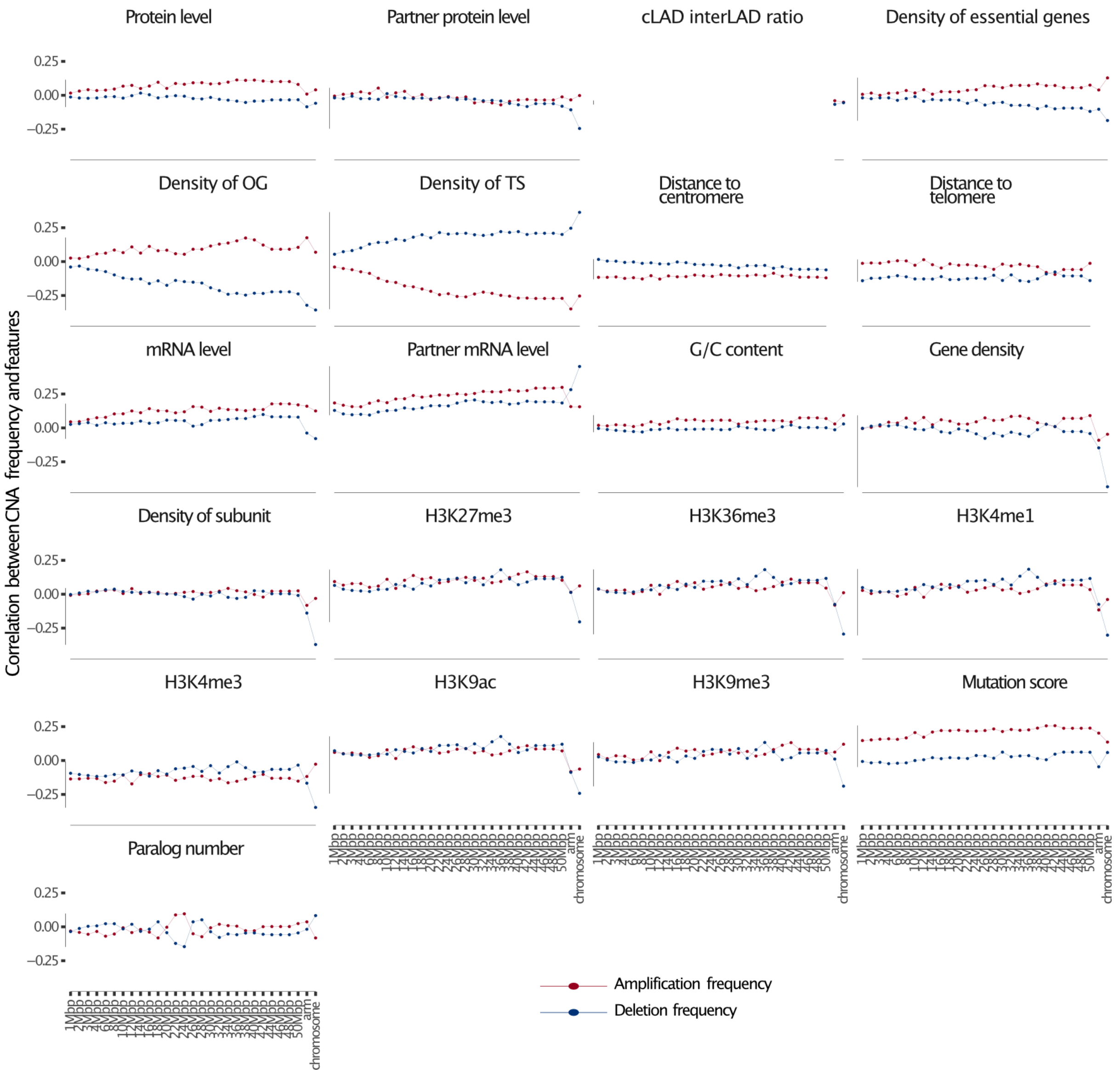


Figure 43. Segment-level correlations between genomic alteration frequency and features. For feature cLAD to interLAD ratio, correlations only at arm- and chromosome-level were shown since that feature data is only available for those levels. Spearman method was used to calculate correlations.

Next, we aimed to gain a comprehensive understanding of how effectively tissue- and genomic-location-specific features work together in explaining cancer-type-specific amplification and deletion patterns, and to determine the extent to which each feature drives the observed alteration patterns. Therefore, we employed ML models separately for amplification and deletion patterns. To this end, we, first, converted continuous values (CNAs and aneuploidy frequencies for each segment-level) to binary classes; we assigned 1 (event) to the corresponding alteration if it occurred more than expected by chance and 0 otherwise for arm- and chromosome-level aneuploidies. For focal level CNAs, we assigned 1 if the frequency of alteration was higher than the fourth quartile of the corresponding frequency distribution. To be able to capture non-linear relationships between features and observed genomic alteration patterns, we used XGBoost classifier, a decision tree method, and trained and tested our ML method on data of amplified- vs. diploid-regions (amplification model), and that of deleted- vs. diploid-regions (deletion model) separately for each segment-level. We found that our model gave better performances for amplification models (accuracy was 78% on average) than deletion models (accuracy was 67% on average) (**Figure 44A**). On the other hand, we did not observe sharp differences in the performances of models between focal CNAs and aneuploidies (**Figure 44B**).

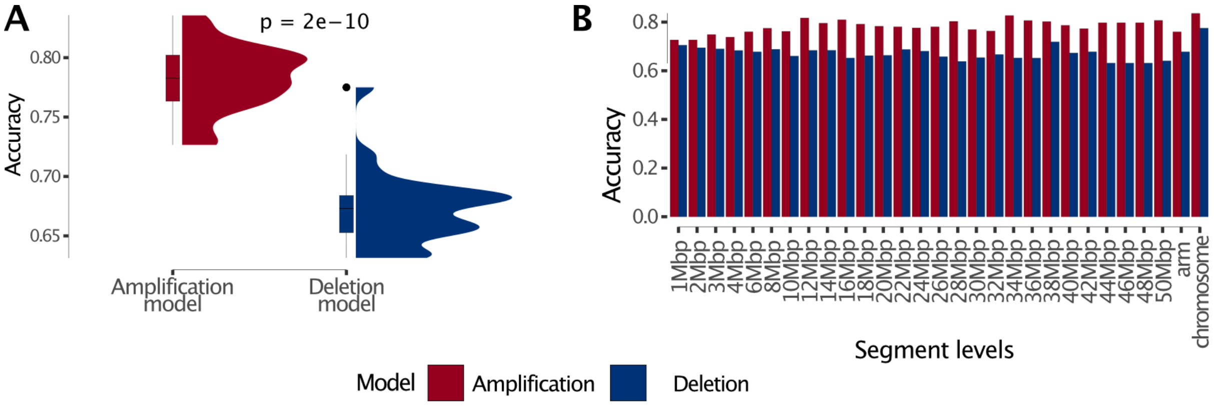


Figure 44. Performance of amplification and deletion models. (A) Overall performances of amplification and deletion models across all the segment levels. Wilcoxon test was used for the statistical comparison. **(B)** Performance (accuracy) of ML models for each segment level.

Understanding the importance of features in contribution to the model's decision-making process could shed light on the factors driving observed amplification and deletion patterns in cancer genomes. Therefore, we aimed to investigate the relative contribution of each feature to the amplification and deletion models. To this end, we calculated feature importance in our models by using xgb.importance built-in function of xgboost package in R. We found that, for amplification models, the density of TS was consistently ranked as the first or second feature based on the feature importance for relatively larger focal CNAs (equal to or more than 18 Mbp), arm- and chromosome-level aneuploidies (**Figure 45**). This is in agreement with recent findings showing the contribution of TS density to the selection of arm-level aneuploidy patterns in TCGA tumour samples and cancer cell lines (Jubran et al., 2023) but demonstrates the importance even on a smaller genomic scale. On the other hand, for amplifications in smaller focal regions ranging from 1 Mbp to 16 Mbp, distance to centromere/telomere, gene density and mRNA level were ranked among the top most important features (**Figure 45**). We found that those features were placed in the top ranks for smaller focal regions covering the same range (1-16 Mbp) for deletion models as well (**Figure 46**). Furthermore, we observed that epigenetic markers were placed very low in the rank for focal amplification models while H3K27me3 and H3K4me3 were ranked as the second and third most important feature for arm- and chromosome-level amplification models, respectively (**Figure 46**). Together these findings suggest that mechanisms driving the adaptive evolution of focal CNAs differ from those influencing alterations at larger scale (relatively larger focal CNAs and aneuploidies), and different sets of factors are in play in shaping observed genomic alterations patterns at distinct levels.

In chapter I, we showed that co-regulation of protein complex members is a way to compensate for aberrant changes induced by chromosome-level aneuploidies in aneuploid tumours. Interestingly, we found that mRNA and protein levels of partners, and density of subunits were ranked among the Top6 most important features for deletion models both for focal CNAs and aneuploidies (**Figure 45 & 46**) further supporting the role of compensation for stoichiometric imbalances in complexes in the observed genomic alteration patterns in cancer.

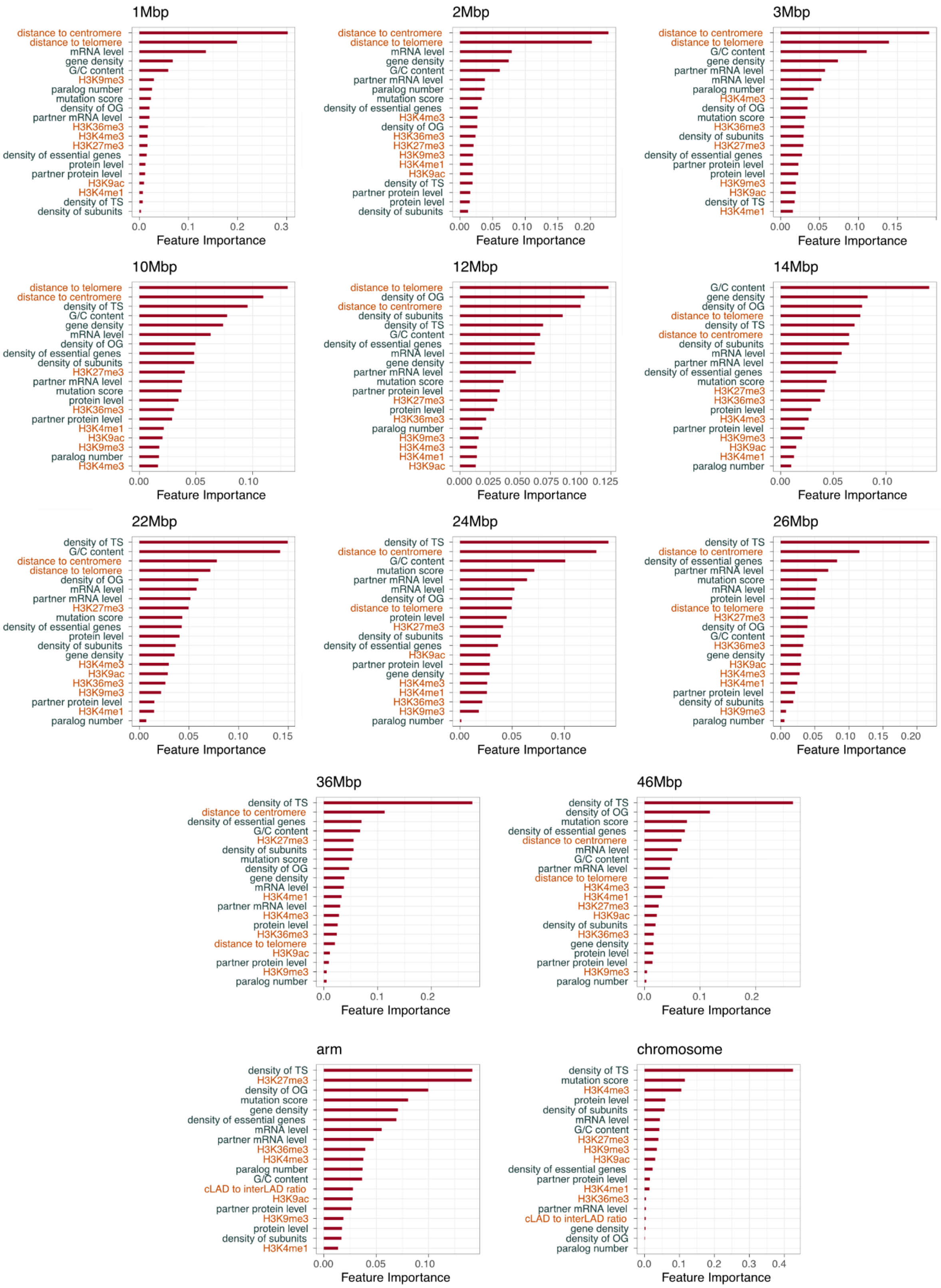


Figure 45. Feature contribution in amplification models. All the features used in the ML models were listed along the y-axis, and ranked based on their importance value within each segment level. For simplicity reasons, some segment levels were removed from the plot. Features modulating selection were represented in dark green colour while those associated with likelihood of occurrence in dark orange.

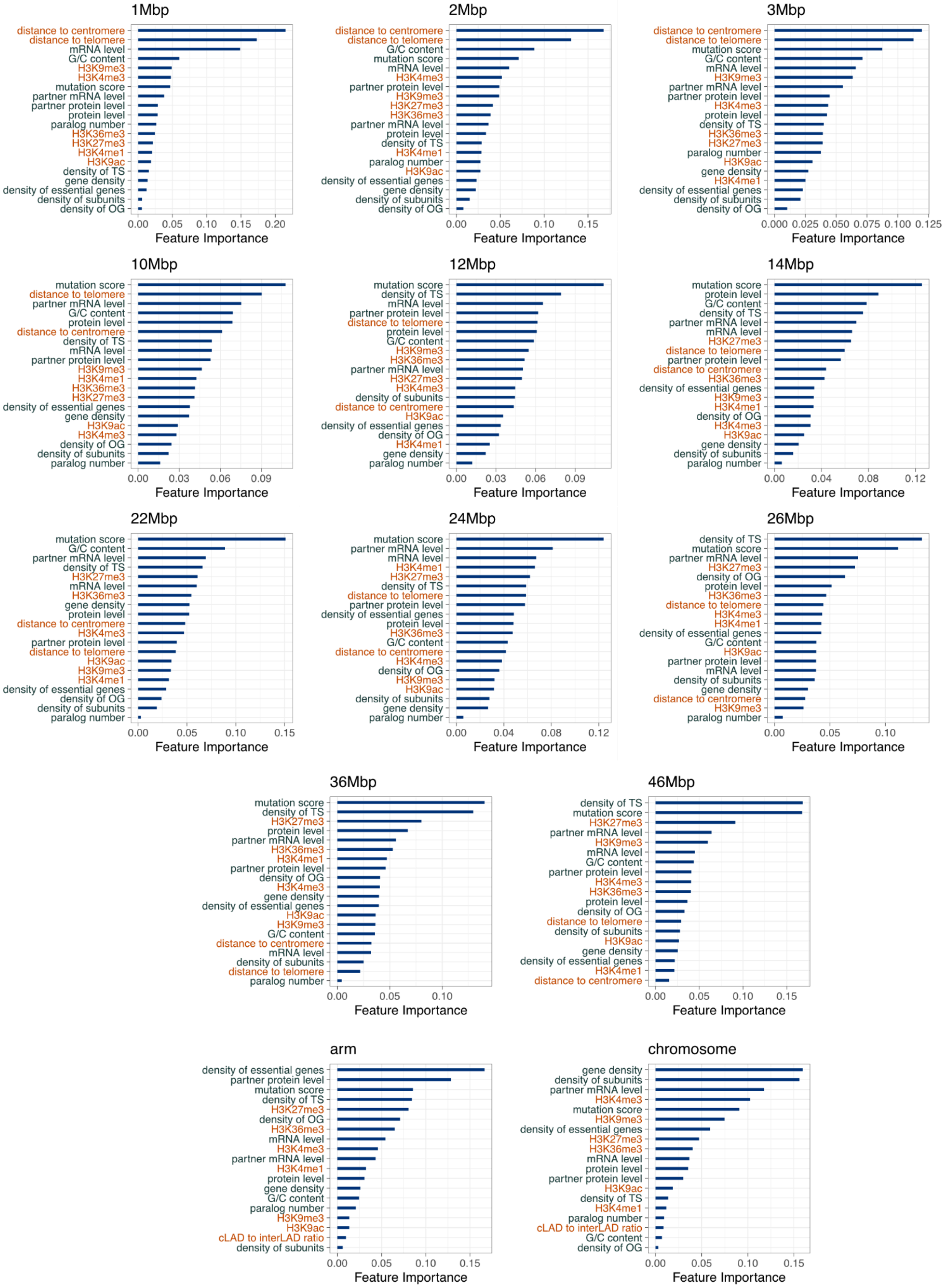


Figure 46. Feature contribution in deletion models. All the features used in the ML models were listed along the y-axis, and ranked based on their importance value within each segment level. For simplicity reasons, some segment levels were removed from the plot. Features modulating selection were represented in dark green colour while those associated with likelihood of occurrence in dark orange.

Discussion

Cancer genomes are characterised by various CNAs ranging from focal CNAs to arm- and chromosome-level aneuploidies. Among them, aneuploidies have a significant impact on protein complex stoichiometry as they lead to massive dysregulation patterns both at mRNA and protein level by introducing additional or diminished copies of hundreds to thousands of genes. Given that the inability to maintain protein complex stoichiometries leads to proteotoxic stress, and is linked to proliferative and survival disadvantages in normal cells, it is important to address these questions yet to be fully answered; (i) how cancer cells compensate for transcriptional and translational dysregulation induced by aneuploidy, and (ii) why aneuploidies and focal copy number alterations are occurring repeatedly in cancer genomes. This will help us gain a comprehensive understanding on tumour genome evolution and the factors underlying it.

Here, we present a comprehensive study integrating aneuploidy, transcriptome and proteome data for hundreds of TCGA tumour samples, harbouring chromosome-level aneuploidies, to understand how aneuploidy-induced dysregulation is compensated in cancer genomes. We conduct an extensive characterization of the transcriptome and proteome level changes, and demonstrate that 47-63% of genes on aneuploid chromosomes show expression changes at the transcriptome level while this ratio is 24-33% at the proteome level (**Figure 7**). A comparable impact of arm-level aneuploidy on gene expression has been previously observed in human cancer cells, with 50% and 25% of genes located on the altered arms exhibiting copy-number-correlated expression changes at the transcriptome and proteome levels, respectively (Schukken & Sheltzer, 2022). The effect of aneuploidy on expression changes is more pronounced in aneuploid yeast models, where up to 70-80% of genes on aneuploid chromosomes undergo transcript and protein level changes by the degree expected based on chromosome number (Dephoure et al., 2014; Gasch et al., 2016). In line with these findings, an extent summary on the effect of aneuploidy on gene expression changes conclude that degree of aneuploidy-induced expression changes depends on many factors including the cellular environment (Kojima & Cimini, 2019); however, independently from the study system, a consistent observation is that there is stronger buffering at the proteome level than transcriptome level (Dephoure et al., 2014; Stingele et al., 2012).

On the other hand, we observe a surprising degree of differential expression for genes located on chromosomes other than aneuploid ones (**Figure 7**). Furthermore, a comparative analysis of transcriptome and proteome level changes reveals that proteomic changes from aneuploid chromosomes could primarily be explained by differential expression of their corresponding genes at the transcriptome level (**Figure 8**). However, this pattern differs for proteomic changes on other, non-aneuploid chromosomes. In the case of amplification, only 26% of the differentially abundant proteins also exhibit differential expression in their corresponding transcripts, compared to 76% for aneuploid chromosomes. Similarly, for deletion cases, only 36% of the differentially abundant proteins show differential expression in their corresponding transcripts, in contrast to 89% for aneuploid chromosomes. (**Figure 8**). Together these findings suggest different levels of dosage compensation for genes located on aneuploid and non-aneuploid chromosomes. For aneuploid chromosomes, transcriptional control plays a prominent role, while translational or post-translational control appears to have a relatively stronger importance for gene regulation for other chromosomes.

We propose that a significant portion of vast proteomic changes happening on other, non-aneuploid chromosomes may function as a compensatory mechanism by binding aggregation-prone proteins upregulated due to their location on the amplified chromosomes. This correlation is also observed between differentially abundant proteins of deleted chromosomes and their differentially abundant aggregation-prone co-complex members, suggesting that correlated abundance patterns to compensate for aggregation-prone orphan proteins are detectable for both chromosome amplification and deletion cases. Furthermore, we observe that up to 40% of differentially abundant proteins of other chromosomes involve in physical interactions, either within a complex or in a binary manner, with their partners encoded by genes on aneuploid chromosomes. Based on these observations, we propose a novel compensatory mechanism in aneuploid tumours which complements the previously described dosage compensation addressing the differential expression of protein complex subunits encoded by genes directly on the aneuploid chromosomes (Schukken & Sheltzer, 2022; Stingele et al., 2012). Together, these suggested mechanisms might largely prevent the otherwise detrimental overexpression of orphan complex subunits.

Similar, interaction-mediated compensatory mechanisms have been previously identified to be induced by focal CNAs (Gonçalves et al., 2017; Sousa et al., 2019). Here, we demonstrate that protein interactions, either within complexes or in binary interactions, play a role in abundance compensation also for large genomic alterations. This genome-wide abundance compensation likely serves to prevent stoichiometric imbalances in protein complexes and proteotoxic stress triggered by aggregation of orphan subunits as suggested by the stronger correlations formed by non-promiscuous and aggregation-prone proteins (**Figure 22 & 23**). Given that around 90% of solid tumours are aneuploid, our work addresses the question of how the majority of cancer cells can tolerate the vast gene expression changes induced by alteration of large genomic regions.

While aneuploidy is very common, cancer is characterised by various additional genomic alterations introducing transcriptome and proteome dysregulation. We find similarities between compensation for aneuploidy-induced gene expression and that for the broader spectrum of gene expressions associated with tumorigenesis. By systematically categorising interaction types in protein complexes and testing each category for its impact on abundance changes induced by tumorigenesis, we provide a proof of concept that the way proteins organise into complexes affects co-abundance patterns in cancer in general.

In the majority of comparisons, the impact of each interaction category on co-abundance regulations in cancer aligns with our expectations across all cancer types tested. However, in some comparisons, the degree of the expected trend among different groups of an interaction category is not statistically strong for all cancer types: e.g., for the observation that permanent interactions are associated with higher correlations compared to transient interactions (**Figure 16**). This variation can potentially be explained by the context-specific nature of the stability of an interaction meaning that permanent interaction may become transient under certain conditions or vice versa (Nooren & Thornton, 2003). Furthermore, computational predictions of transient and permanent interactions may not fully represent how proteins interact in different local environments such as different cancer types.

Someone could expect competition among proteins binding to the same proteins through overlapping interaction sites. Therefore, we hypothesised to observe negative abundance correlations between proteins competing with each other to bind their common partners. However, our observations are surprisingly in the opposite direction of our expectation (**Figure 14**), and could be reproduced over a range of thresholds for binding site overlap definition (**Figure 15**). One possible explanation could be that alternative splicing, an intermediate regulatory process between transcription and translation, is in place and prevents competition between proteins by removing binding domains involved in competition (H. Li et al., 2015). Another reason could be potential inaccuracies in computational assessments of competitive and cooperative binding. For instance, the biophysical characteristics of protein interaction sites, such as steric hindrance, may either favour or prevent interactions between them, and therefore, classifying proteins based on the overlap in their interaction interfaces may not accurately reflect the true nature of competitive or cooperative binding. Additionally, it's important to acknowledge that these estimates are derived from *in vitro* experiments, and the proteins categorised as competing may have different localizations or expression patterns *in vivo*, preventing them from ever encountering each other in reality.

All together, our findings highlight the role of protein interactions and complex organisation as compensation mechanisms to deal with stoichiometric imbalances in protein complexes and to prevent proteotoxic stress induced by both aneuploidy in specific and tumorigenesis in general. In this aspect, the primary limitation of our study is the still incomplete understanding of the nature of the human interactome, and, in particular, the limits on the available protein complex information. For example, we classified permanent vs. transient interactions by mining the literature where we left only a few instances (58 transient and 9 permanent interactions). Additionally, we used the proteomic measurements of only two different cell lines to define context-specific vs. general interactions because of two related reasons: 1) Even though there are other studies measuring complexes in different cell lines, the 293T interactome provided by BioPlex Interactome is one of the most comprehensive data (Huttlin et al., 2021); 2) the interactome in HCT116 cell lines was determined by the same experimental pipeline which minimises technical biases in the comparison of the two interactomes. We expect that a larger number of cell-line-specific interactome data and improvements in protein complex measurements will potentially provide a more comprehensive understanding of the roles of protein interactions in dosage compensation in aneuploidy and cancer.

In addition to biophysically interacting within protein complexes, we defined specific cancer-essential functional terms that are under stronger protection from protein abundance imbalances induced by aneuploidy. This suggests the role of functional selection as a driving force in shaping the genome-wide co-abundance regulation. Given that 23% of the translation-related genes are ribosomal genes, it is not surprising that enrichment of translation is mainly driven by a relatively larger fraction of ribosomal genes in our gene sets.

Our findings describe the need for compensation mechanisms to deal with stoichiometric imbalances in protein complexes caused by aneuploidy-induced transcriptome and proteome dysregulation, and highlight the role of protein properties and complex organisation in this compensation. We provide evidence that this compensation is performed by post-translational regulation. Another interesting finding is that the higher degree of success in this compensation is associated with better tumour fitness, and tumours that fail to compensate for aneuploidy-induced stoichiometric imbalances have to deal with excess amount of subunits, and therefore activate protein degradation machinery. These results further highlight the high cost of aneuploidy on the cell - in particular, it increases the burden on energy and protein homeostasis (Sheltzer & Amon, 2011).

Given the high cost of aneuploidy on cell homeostasis, it is not surprising that recent studies have focused on understanding selection pressures and cell division dynamics underlying recurrent CNA and aneuploidy patterns in cancer genomes. While previous efforts reveal the role of both selection (positive and negative) (Jubran et al., 2023; Sack et al., 2018; Shih et al., 2023) and cell division errors, which create non-random genomic substrate for selection (Klaasen et al., 2022), the focus is either on a specific factor or small set of factors or on genomic alterations at one level (focal CNA or aneuploidy). Here, we conducted a comprehensive and comparative study investigating the impact of 21 factors - impacting the likelihood of occurrence and those playing a role in the selection - on both focal CNAs at various scales, and arm- and chromosome-level aneuploidies. We found that features associated with the likelihood of occurrence (such as distance to centromere/telomere) provide a better explanation for smaller genomic alterations (1-8 Mbp) (**Figure 45 & 46**). On the other hand, we observed that both features modulating selection (such as density of TS and mutation score) and epigenetic markers (likelihood of occurrence) were ranked among the most contributing features for larger CNAs and arm/chromosome-level aneuploidies (**Figure 45 & 46**), suggesting both selection and the likelihood of occurrence play a role in explaining larger genomic alterations. These results reveal that mechanisms driving the adaptive evolution of focal CNAs are different from those underlying alterations at larger scales, and different sets of factors are in play in shaping recurrent genomic alterations at different scales.

One interesting finding is that genomic-location-specific features are ranked as the top most contributing features in explaining cancer-type-specific CNA and aneuploidy patterns, both for amplification and deletion models. Therefore, we wondered about the relative contribution of tissue-specific features, and trained and tested our ML models only by using tissue-specific features. Despite observing a slight decrease in model performance compared to models with all features, we still obtained favourable outcomes: Model accuracies, on average, were 66% and 63% respectively for amplification and deletion models (which were 78% and 67% when all features were included, respectively for amplification and deletion models). Together these findings highlight that both tissue-specific and genomic-location-specific (non-tissue-specific) features play a significant role in shaping CNA and aneuploidy patterns in cancer.

Previous studies noted the importance of density of TS and OGs in the prediction of arm- and chromosome-level aneuploidies, and further showed negative associations between density of TS and amplifications, and density of OGs and deletions (Davoli et al., 2013; Jubran et al., 2023) highlighting the role of negative selection in shaping recurrent aneuploidy patterns in cancer. By investigating feature importance in ML models, we showed that similar selection pressures also apply at the focal level CNAs, which was further supported by correlation analyses (**Figure 43**).

Overall, our research provides insights into compensation mechanisms to cope with aneuploidy-induced stoichiometric imbalances in protein complexes, and highlight the role of ubiquitin-dependent protein degradation in keeping the complex stoichiometry and better tumour fitness (**Figure 47**). Given that targeting essential genes in aneuploid cells eventually results in proliferation defects and activation of cell death pathways (Cohen-Sharir et al., 2021), our findings might be used in identification of potential drug targets suitable for clinical use. Furthermore, our findings provide a comparative overview of tissue- and genomic-location-specific factors shaping the observed focal CNAs and aneuploidy patterns in cancer genomes, and highlight the importance of applying ML algorithms for understanding recurrent genomic alterations in cancer. Ultimately, given the high number of aneuploid tumours, studying and understanding compensatory mechanisms and the potential vulnerabilities they create in aneuploid tumours, and dynamics underlying recurrent genomic alterations will have profound implications for both basic cell biology as well as cancer biology.

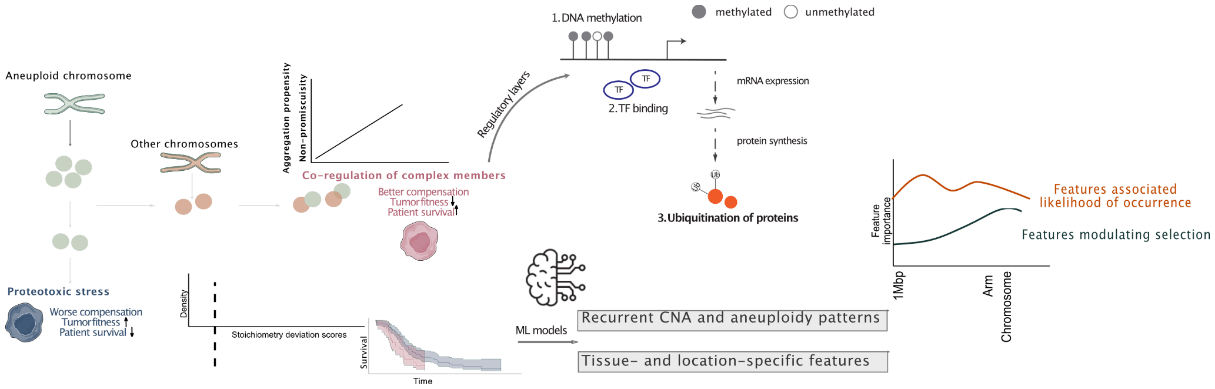


Figure 47. Overall representation of main findings. Regulation of co-complex members of aneuploid proteins is a compensatory mechanism to prevent proteotoxicity and imbalances in protein complexes. This co-abundance compensation provides fitness advantages to tumours. Post-translational regulation (ubiquitin-dependent protein degradation) plays a more substantial role in this co-abundance regulation. Different sets of factors are in play in shaping recurrent genomic alterations at different scales.

References

Acuner Ozbabacan, S. E., Engin, H. B., Gursoy, A., & Keskin, O. (2011). Transient protein-protein interactions. *Protein Engineering, Design & Selection: PEDS*, *24*(9), 635–648. https://doi.org/10.1093/protein/gzr025

Alanis-Lobato, G., Andrade-Navarro, M. A., & Schaefer, M. H. (2017). HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Research*, *45*(D1), D408–D414. https://doi.org/10.1093/nar/gkw985

Alfieri, F., Caravagna, G., & Schaefer, M. H. (2023). Cancer genomes tolerate deleterious coding mutations through somatic copy number amplifications of wild-type regions. *Nature Communications*, *14*(1), 3594. https://doi.org/10.1038/s41467-023-39313-8

Attique, H., Shah, S., Jabeen, S., Khan, F. G., Khan, A., & ELAffendi, M. (2022). Multiclass Cancer Prediction Based on Copy Number Variation Using Deep Learning. *Computational Intelligence and Neuroscience*, *2022*, 4742986. https://doi.org/10.1155/2022/4742986

Barrera-Vilarmau, S., Teixeira, J. M. C., & Fuxreiter, M. (2022). Protein interactions: Anything new? *Essays in Biochemistry*, *66*(7), 821–830. https://doi.org/10.1042/EBC20220044

Baylin, S. B., & Herman, J. G. (2000). DNA hypermethylation in tumorigenesis: Epigenetics joins genetics. *Trends in Genetics: TIG*, *16*(4), 168–174. https://doi.org/10.1016/s0168-9525(99)01971-x

Ben-David, U., & Amon, A. (2020). Context is everything: Aneuploidy in cancer. *Nature Reviews. Genetics*, *21*(1), 44–62. https://doi.org/10.1038/s41576-019-0171-x

Ben-David, U., Beroukhim, R., & Golub, T. R. (2019). Genomic evolution of cancer models: Perils and opportunities. *Nature Reviews. Cancer*, *19*(2), 97–109. https://doi.org/10.1038/s41568-018-0095-3

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*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., Farnham, P. J., Hirst, M., Lander, E. S., Mikkelsen, T. S., & Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, *28*(10), 1045–1048. https://doi.org/10.1038/nbt1010-1045

Block, P., Paern, J., Hüllermeier, E., Sanschagrin, P., Sotriffer, C. A., & Klebe, G. (2006). Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins*, *65*(3), 607–622. https://doi.org/10.1002/prot.21104

Brennan, C. M., Vaites, L. P., Wells, J. N., Santaguida, S., Paulo, J. A., Storchova, Z., Harper, J. W., Marsh, J. A., & Amon, A. (2019). Protein aggregation mediates stoichiometry of protein complexes in aneuploid cells. *Genes & Development*, *33*(15–16), 1031–1047. https://doi.org/10.1101/gad.327494.119

Bushweller, J. H. (2019). Targeting transcription factors in cancer—From undruggable to reality. *Nature Reviews. Cancer*, *19*(11), 611–624. https://doi.org/10.1038/s41568-019-0196-7

Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. https://doi.org/10.1038/nature11412

Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, *474*(7353), 609–615. https://doi.org/10.1038/nature10166

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., & Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, *30*(5), 413–421. https://doi.org/10.1038/nbt.2203

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*, *2*(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., & Tyers, M. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, *41*(Database issue), D816-823. https://doi.org/10.1093/nar/gks1158

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y., Lun, A. T. L., & Smyth, G. K. (2016). From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, *5*, 1438. https://doi.org/10.12688/f1000research.8987.2

Chino, A., Makanae, K., & Moriya, H. (2013). Relationships between cell cycle regulator gene copy numbers and protein expression levels in Schizosaccharomyces pombe. *PloS One*, *8*(9), e73319. https://doi.org/10.1371/journal.pone.0073319

Chiu, Y.-C., Zheng, S., Wang, L.-J., Iskra, B. S., Rao, M. K., Houghton, P. J., Huang, Y., & Chen, Y. (2021). Predicting and characterizing a cancer dependency map of tumors with deep learning. *Science Advances*, *7*(34), eabh1275. https://doi.org/10.1126/sciadv.abh1275

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., … Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biology*, *9*(7), e1001091. https://doi.org/10.1371/journal.pbio.1001091

Cohen-Sharir, Y., McFarland, J. M., Abdusamad, M., Marquis, C., Bernhard, S. V., Kazachkova, M., Tang, H., Ippolito, M. R., Laue, K., Zerbib, J., Malaby, H. L. H., Jones, A., Stautmeister, L.-M., Bockaj, I., Wardenaar, R., Lyons, N., Nagaraja, A., Bass, A. J., Spierings, D. C. J., … Ben-David, U. (2021). Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition. *Nature*, *590*(7846), 486–491. https://doi.org/10.1038/s41586-020-03114-6

Cramer, D., Serrano, L., & Schaefer, M. H. (2016). A network of epigenetic modifiers and DNA repair genes controls tissue-specific copy number alteration preference. *eLife*, *5*, e16519. https://doi.org/10.7554/eLife.16519

Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., & Elledge, S. J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, *155*(4), 948–962. https://doi.org/10.1016/j.cell.2013.10.011

Dempster, J. M., Boyle, I., Vazquez, F., Root, D. E., Boehm, J. S., Hahn, W. C., Tsherniak, A., & McFarland, J. M. (2021). Chronos: A cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biology*, *22*(1), 343. https://doi.org/10.1186/s13059-021-02540-7

Dempster, J. M., Rossen, J., Kazachkova, M., Pan, J., Kugener, G., Root, D. E., & Tsherniak, A. (2019). *Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines* [Preprint]. Cancer Biology. https://doi.org/10.1101/720243

Dephoure, N., Hwang, S., O'Sullivan, C., Dodgson, S. E., Gygi, S. P., Amon, A., & Torres, E. M. (2014). Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *eLife*, *3*, e03023. https://doi.org/10.7554/eLife.03023

Dimova, N. V., Hathaway, N. A., Lee, B.-H., Kirkpatrick, D. S., Berkowitz, M. L., Gygi, S. P., Finley, D., & King, R. W. (2012). APC/C-mediated multiple monoubiquitylation provides an alternative degradation signal for cyclin B1. *Nature Cell Biology*, *14*(2), 168–176. https://doi.org/10.1038/ncb2425

Ehrlich, M. (2002). DNA methylation in cancer: Too much, but also too little. *Oncogene*, *21*(35), 5400–5413. https://doi.org/10.1038/sj.onc.1205651

Eisenberg, D., Marcotte, E. M., Xenarios, I., & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, *405*(6788), 823–826. https://doi.org/10.1038/35015694

ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, *306*(5696), 636–640. https://doi.org/10.1126/science.1105136

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., & Jensen, L. J. (2013). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, *41*(Database issue), D808-815. https://doi.org/10.1093/nar/gks1094

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., & Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, *6*(269), pl1. https://doi.org/10.1126/scisignal.2004088

Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., Song, Z., Huang, C., Li, J., Dong, X., Zhou, Y., Liu, Q., Ma, L., Wang, X., Zhou, J., Liu, Y., Boja, E., Robles, A. I., Ma, W., Wang, P., … Fan, J. (2019). Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. *Cell*, *179*(2), 561-577.e22. https://doi.org/10.1016/j.cell.2019.08.052

Garribba, L., De Feudis, G., Martis, V., Galli, M., Dumont, M., Eliezer, Y., Wardenaar, R., Ippolito, M. R., Iyer, D. R., Tijhuis, A. E., Spierings, D. C. J., Schubert, M., Taglietti, S., Soriani, C., Gemble, S., Basto, R., Rhind, N., Foijer, F., Ben-David, U., … Santaguida, S. (2023). Short-term molecular consequences of chromosome mis-segregation for genome stability. *Nature Communications*, *14*(1), 1353. https://doi.org/10.1038/s41467-023-37095-7

Gasch, A. P., Hose, J., Newton, M. A., Sardi, M., Yong, M., & Wang, Z. (2016). Further support for aneuploidy tolerance in wild yeast and effects of dosage compensation on gene copy-number evolution. *eLife*, *5*, e14409. https://doi.org/10.7554/eLife.14409

Gillette, M. A., Satpathy, S., Cao, S., Dhanasekaran, S. M., Vasaikar, S. V., Krug, K., Petralia, F., Li, Y., Liang, W.-W., Reva, B., Krek, A., Ji, J., Song, X., Liu, W., Hong, R., Yao, L., Blumenberg, L., Savage, S. R., Wendl, M. C., … Shi, Z. (2020). Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell*, *182*(1), 200-225.e35. https://doi.org/10.1016/j.cell.2020.06.013

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., & Ruepp, A. (2019). CORUM: The comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Research*, *47*(D1), D559–D563. https://doi.org/10.1093/nar/gky973

Gonçalves, 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.e4. https://doi.org/10.1016/j.cels.2017.08.013

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *The New England Journal of Medicine*, *375*(12), 1109–1112. https://doi.org/10.1056/NEJMp1607591

GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

Guang, M. H. Z., Kavanagh, E. L., Dunne, L. P., Dowling, P., Zhang, L., Lindsay, S., Bazou, D., Goh, C. Y., Hanley, C., Bianchi, G., Anderson, K. C., O’Gorman, P., & McCann, A. (2019). Targeting Proteotoxic Stress in Cancer: A Review of the Role that Protein Quality Control Pathways Play in Oncogenesis. *Cancers*, *11*(1), 66. https://doi.org/10.3390/cancers11010066

Han, Y., Yang, J., Qian, X., Cheng, W.-C., Liu, S.-H., Hua, X., Zhou, L., Yang, Y., Wu, Q., Liu, P., & Lu, Y. (2019). DriverML: A machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Research*, *47*(8), e45. https://doi.org/10.1093/nar/gkz096

Heery, R., & Schaefer, M. H. (2021). DNA methylation variation along the cancer epigenome and the identification of novel epigenetic driver events. *Nucleic Acids Research*, *49*(22), 12692–12705. https://doi.org/10.1093/nar/gkab1167

Hein, M. Y., Hubner, N. C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I. A., Weisswange, I., Mansfeld, J., Buchholz, F., Hyman, A. A., & Mann, M. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, *163*(3), 712–723. https://doi.org/10.1016/j.cell.2015.09.053

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., & Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research*, *43*(Database issue), D512-520. https://doi.org/10.1093/nar/gku1267

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Hughes, T. R., Roberts, C. J., Dai, H., Jones, A. R., Meyer, M. R., Slade, D., Burchard, J., Dow, S., Ward, T. R., Kidd, M. J., Friend, S. H., & Marton, M. J. (2000). Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature Genetics*, *25*(3), 333–337. https://doi.org/10.1038/77116

Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., Gygi, M. P., Thornock, A., Zarraga, G., Tam, S., Szpyt, J., Gassaway, B. M., Panov, A., Parzen, H., Fu, S., Golbazi, A., Maenpaa, E., Stricker, K., Guha Thakurta, S., … Gygi, S. P. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, *184*(11), 3022-3040.e28. https://doi.org/10.1016/j.cell.2021.04.011

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*, *13*(1), e1006554. https://doi.org/10.1371/journal.pgen.1006554

Jones, S., & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(1), 13–20. https://doi.org/10.1073/pnas.93.1.13

Jubran, J., Slutsky, R., Rozenblum, N., Rokach, L., Ben-David, U., & Yeger-Lotem, E. (2023). *Machine-learning analysis of factors that shape cancer aneuploidy landscapes reveals an important role for negative selection* [Preprint]. Cancer

Biology. https://doi.org/10.1101/2023.07.05.547626

Keskin, O., & Nussinov, R. (2007). Similar binding sites and different partners: Implications to shared proteins in cellular pathways. *Structure (London, England: 1993)*, *15*(3), 341–354. https://doi.org/10.1016/j.str.2007.01.007

Klaasen, S. J., Truong, M. A., van Jaarsveld, R. H., Koprivec, I., Štimac, V., de Vries, S. G., Risteski, P., Kodba, S., Vukušić, K., de Luca, K. L., Marques, J. F., Gerrits, E. M., Bakker, B., Foijer, F., Kind, J., Tolić, I. M., Lens, S. M. A., & Kops, G. J. P. L. (2022). Nuclear chromosome locations dictate segregation error frequencies. *Nature*, *607*(7919), 604–609. https://doi.org/10.1038/s41586-022-04938-0

Kojima, S., & Cimini, D. (2019). Aneuploidy and gene expression: Is there dosage compensation? *Epigenomics*, *11*(16), 1827–1837. https://doi.org/10.2217/epi-2019-0135

Krzywinski, M., & Altman, N. (2017). Classification and regression trees. *Nature Methods*, *14*(8), 757–758. https://doi.org/10.1038/nmeth.4370

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

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. https://doi.org/10.1016/j.cell.2014.02.033

Li, H., Zhou, Y., & Zhang, Z. (2015). Competition-cooperation relationship networks characterize the competition and cooperation between proteins. *Scientific Reports*, *5*, 11619. https://doi.org/10.1038/srep11619

Li, Y., Ge, X., Peng, F., Li, W., & Li, J. J. (2022). Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biology*, *23*(1), 79. https://doi.org/10.1186/s13059-022-02648-4

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. (2019). WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, *47*(W1), W199–W205. https://doi.org/10.1093/nar/gkz401

Lu, Y., Lee, B., King, R. W., Finley, D., & Kirschner, M. W. (2015). Substrate degradation by the proteasome: A single-molecule kinetic analysis. *Science (New York, N.Y.)*, *348*(6231), 1250834. https://doi.org/10.1126/science.1250834

Luo, P., Ding, Y., Lei, X., & Wu, F.-X. (2019). deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks. *Frontiers in Genetics*, *10*, 13. https://doi.org/10.3389/fgene.2019.00013

Määttä, T. A., Rettel, M., Sridharan, S., Helm, D., Kurzawa, N., Stein, F., & Savitski, M. M. (2020). Aggregation and disaggregation features of the human proteome. *Molecular Systems Biology*, *16*(10), e9500. https://doi.org/10.15252/msb.20209500

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297. https://doi.org/10.1093/nar/gks042

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. https://doi.org/10.1016/j.cell.2016.09.015

Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., Gatza, M. L., Wilkerson, M., Perou, C. M., Yellapantula, V., … NCI CPTAC. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, *534*(7605), 55–62. https://doi.org/10.1038/nature18003

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.-B., Pagie, L., Kellis, M., Reinders, M., Wessels, L., & van Steensel, B. (2013). Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. *Genome Research*, *23*(2), 270–280. https://doi.org/10.1101/gr.141028.112

Meyer, M. J., Beltrán, J. F., Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., & Yu, H. (2018). Interactome INSIDER: A structural interactome browser for genomic studies. *Nature Methods*, *15*(2), 107–114. https://doi.org/10.1038/nmeth.4540

Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Dharia, N. V., Montgomery, P. G., Cowley, G. S., Pantel, S., Goodale, A., Lee, Y., Ali, L. D., Jiang, G., Lubonja, R., Harrington, W. F., Strickland, M., Wu, T., Hawes, D. C., … Tsherniak, A. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nature Genetics*, *49*(12), 1779–1784. https://doi.org/10.1038/ng.3984

Mintseris, J., & Weng, Z. (2003). Atomic contact vectors in protein-protein recognition. *Proteins*, *53*(3), 629–639. https://doi.org/10.1002/prot.10432

Nassar, L. R., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, B. T., Lee, C. M., Muthuraman, P., Nguy, B., Pereira, T., Nejad, P., Perez, G., Raney, B. J., Schmelter, D., Speir, M. L., … Kent, W. J. (2023). The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research*, *51*(D1), D1188–D1195. https://doi.org/10.1093/nar/gkac1072

Nawata, H., Kashino, G., Tano, K., Daino, K., Shimada, Y., Kugoh, H., Oshimura, M., & Watanabe, M. (2011). Dysregulation of gene expression in the artificial human trisomy cells of chromosome 8 associated with transformed cell phenotypes. *PloS One*, *6*(9), e25319. https://doi.org/10.1371/journal.pone.0025319

Nooren, I. M. A., & Thornton, J. M. (2003). Diversity of protein-protein interactions. *The EMBO Journal*, *22*(14), 3486–3492. https://doi.org/10.1093/emboj/cdg359

Nusinow, D. P., Szpyt, J., Ghandi, M., Rose, C. M., McDonald, E. R., Kalocsay, M., Jané-Valbuena, J., Gelfand, E., Schweppe, D. K., Jedrychowski, M., Golji, J., Porter, D. A., Rejtar, T., Wang, Y. K., Kryukov, G. V., Stegmeier, F., Erickson, B. K., Garraway, L. A., Sellers, W. R., & Gygi, S. P. (2020). Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell*, *180*(2), 387-402.e16. https://doi.org/10.1016/j.cell.2019.12.023

Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andrés-Pons, A., Singer, S., Bork, P., & Beck, M. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. *Genome Biology*, *17*, 47. https://doi.org/10.1186/s13059-016-0912-5

Ozery-Flato, M., Linhart, C., Trakhtenbrot, L., Izraeli, S., & Shamir, R. (2011). Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy. *Genome Biology*, *12*(6), R61. https://doi.org/10.1186/gb-2011-12-6-r61

Pacini, C., Dempster, J. M., Boyle, I., Gonçalves, E., Najgebauer, H., Karakoc, E., van der Meer, D., Barthorpe, A., Lightfoot, H., Jaaks, P., McFarland, J. M., Garnett, M. J., Tsherniak, A., & Iorio, F. (2021). Integrated cross-study datasets of genetic dependencies in cancer. *Nature Communications*, *12*(1), 1661. https://doi.org/10.1038/s41467-021-21898-7

Pai, G. S., Lewandowski, R. C., & Borgaonkar, D. S. (2003). *Handbook of chromosomal syndromes*. Wiley-Liss.

Patkar, S., Heselmeyer-Haddad, K., Auslander, N., Hirsch, D., Camps, J., Bronder, D., Brown, M., Chen, W.-D., Lokanga, R., Wangsa, D., Wangsa, D., Hu, Y., Lischka, A., Braun, R., Emons, G., Ghadimi, B. M., Gaedcke, J., Grade, M., Montagna, C., … Ried, T. (2021). Hard wiring of normal tissue-specific chromosome-wide gene expression levels is an additional factor driving cancer type-specific aneuploidies. *Genome Medicine*, *13*(1), 93. https://doi.org/10.1186/s13073-021-00905-y

Pavelka, N., Rancati, G., Zhu, J., Bradford, W. D., Saraf, A., Florens, L., Sanderson, B. W., Hattem, G. L., & Li, R. (2010). Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature*, *468*(7321), 321–325. https://doi.org/10.1038/nature09529

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. https://doi.org/10.1038/nature14248

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G., & Ma’ayan, A. (2016). The harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database: The Journal of Biological Databases and Curation*, *2016*, baw100. https://doi.org/10.1093/database/baw100

Sack, L. M., Davoli, T., Li, M. Z., Li, Y., Xu, Q., Naxerova, K., Wooten, E. C., Bernardi, R. J., Martin, T. D., Chen, T., Leng, Y., Liang, A. C., Scorsone, K. A., Westbrook, T. F., Wong, K.-K., & Elledge, S. J. (2018). Profound Tissue Specificity in Proliferation Control Underlies Cancer Drivers and Aneuploidy Patterns. *Cell*, *173*(2), 499-514.e23. https://doi.org/10.1016/j.cell.2018.02.037

Santaguida, S., & Amon, A. (2015). Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nature Reviews. Molecular Cell Biology*, *16*(8), 473–485. https://doi.org/10.1038/nrm4025

Santaguida, S., Vasile, E., White, E., & Amon, A. (2015). Aneuploidy-induced cellular stresses limit autophagic degradation. *Genes & Development*, *29*(19), 2010–2021. https://doi.org/10.1101/gad.269118.115

Saran, N. G., Pletcher, M. T., Natale, J. E., Cheng, Y., & Reeves, R. H. (2003). Global disruption of the cerebellar transcriptome in a Down syndrome mouse model. *Human Molecular Genetics*, *12*(16), 2013–2019. https://doi.org/10.1093/hmg/ddg217

Schukken, K. M., & Sheltzer, J. M. (2022). Extensive protein dosage compensation in aneuploid human cancers. *Genome Research*, *32*(7), 1254–1270. https://doi.org/10.1101/gr.276378.121

Schuster-Böckler, B., Conrad, D., & Bateman, A. (2010). Dosage Sensitivity Shapes the Evolution of Copy-Number Varied Regions. *PLoS ONE*, *5*(3), e9474. https://doi.org/10.1371/journal.pone.0009474

Senger, G., Santaguida, S., & Schaefer, M. H. (2022). Regulation of protein complex partners as a compensatory mechanism in aneuploid tumors. *eLife*, *11*, e75526. https://doi.org/10.7554/eLife.75526

Senger, G., & Schaefer, M. H. (2021). Protein Complex Organization Imposes Constraints on Proteome Dysregulation in Cancer. *Frontiers in Bioinformatics*, *1*, 723482. https://doi.org/10.3389/fbinf.2021.723482

Sheltzer, J. M., & Amon, A. (2011). The aneuploidy paradox: Costs and benefits of an incorrect karyotype. *Trends in Genetics: TIG*, *27*(11), 446–453. https://doi.org/10.1016/j.tig.2011.07.003

Shih, J., Sarmashghi, S., Zhakula-Kostadinova, N., Zhang, S., Georgis, Y., Hoyt, S. H., Cuoco, M. S., Gao, G. F., Spurr, L. F., Berger, A. C., Ha, G., Rendo, V., Shen, H., Meyerson, M., Cherniack, A. D., Taylor, A. M., & Beroukhim, R. (2023). Cancer aneuploidies are shaped primarily by effects on tumour fitness. *Nature*, *619*(7971), 793–800. https://doi.org/10.1038/s41586-023-06266-3

Soltis, A. R., Bateman, N. W., Liu, J., Nguyen, T., Franks, T. J., Zhang, X., Dalgard, C. L., Viollet, C., Somiari, S., Yan, C., Zeman, K., Skinner, W. J., Lee, J. S. H., Pollard, H. B., Turner, C., Petricoin, E. F., Meerzaman, D., Conrads, T. P., Hu, H., … Wilkerson, M. D. (2022). Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways. *Cell Reports. Medicine*, *3*(11), 100819. https://doi.org/10.1016/j.xcrm.2022.100819

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, *14*, 91. https://doi.org/10.1186/1471-2105-14-91

Sousa, A., Gonçalves, E., Mirauta, B., Ochoa, D., Stegle, O., & Beltrao, P. (2019). Multi-omics Characterization of Interaction-mediated Control of Human Protein Abundance levels. *Molecular & Cellular Proteomics: MCP*, *18*(8 suppl 1), S114–S125. https://doi.org/10.1074/mcp.RA118.001280

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. https://doi.org/10.1038/msb.2012.40

Taggart, J. C., Zauber, H., Selbach, M., Li, G.-W., & McShane, E. (2020). Keeping the Proportions of Protein Complex Components in Check. *Cell Systems*, *10*(2), 125–132. https://doi.org/10.1016/j.cels.2020.01.004

Taylor, A. M., Shih, J., Ha, G., Gao, G. F., Zhang, X., Berger, A. C., Schumacher, S. E., Wang, C., Hu, H., Liu, J., Lazar, A. J., Cancer Genome Atlas Research Network, Cherniack, A. D., Beroukhim, R., & Meyerson, M. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell*, *33*(4), 676-689.e3. https://doi.org/10.1016/j.ccell.2018.03.007

The Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337. https://doi.org/10.1038/nature11252

the NCI CPTAC, Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J. C., … Liebler, D. C. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature*, *513*(7518), 382–387. https://doi.org/10.1038/nature13438

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., … Zhang, J. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Torres, E. M., Sokolsky, T., Tucker, C. M., Chan, L. Y., Boselli, M., Dunham, M. J., & Amon, A. (2007). Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science (New York, N.Y.)*, *317*(5840), 916–924. https://doi.org/10.1126/science.1142210

Tweedie, S., Braschi, B., Gray, K., Jones, T. E. M., Seal, R. L., Yates, B., & Bruford, E. A. (2021). Genenames.org: The HGNC and VGNC resources in 2021. *Nucleic Acids Research*, *49*(D1), D939–D946. https://doi.org/10.1093/nar/gkaa980

Upender, M. B., Habermann, J. K., McShane, L. M., Korn, E. L., Barrett, J. C., Difilippantonio, M. J., & Ried, T. (2004). Chromosome transfer induced aneuploidy results in complex dysregulation of the cellular transcriptome in immortalized and cancer cells. *Cancer Research*, *64*(19), 6941–6949.

https://doi.org/10.1158/0008-5472.CAN-04-0474

Vasaikar, S., Huang, C., Wang, X., Petyuk, V. A., Savage, S. R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O. A., Gritsenko, M. A., Zimmerman, L. J., McDermott, J. E., Clauss, T. R., Moore, R. J., Zhao, R., Monroe, M. E., Wang, Y.-T., Chambers, M. C., … Zhang, Z. (2019). Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell*, *177*(4), 1035-1049.e19. https://doi.org/10.1016/j.cell.2019.03.030

Williams, B. R., Prabhu, V. R., Hunter, K. E., Glazier, C. M., Whittaker, C. A., Housman, D. E., & Amon, A. (2008). Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. *Science (New York, N.Y.)*, *322*(5902), 703–709. https://doi.org/10.1126/science.1160058

Young, L., Jernigan, R. L., & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Science: A Publication of the Protein Society*, *3*(5), 717–729. https://doi.org/10.1002/pro.5560030501

Zhang, H., Liu, T., Zhang, Z., Payne, S. H., Zhang, B., McDermott, J. E., Zhou, J.-Y., Petyuk, V. A., Chen, L., Ray, D., Sun, S., Yang, F., Chen, L., Wang, J., Shah, P., Cha, S. W., Aiyetan, P., Woo, S., Tian, Y., … CPTAC Investigators. (2016). Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell*, *166*(3), 755–765. https://doi.org/10.1016/j.cell.2016.05.069

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *An International Journal on Information Fusion*, *50*, 71–91. https://doi.org/10.1016/j.inffus.2018.09.012