TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Multi-omics analysis as a tool to investigate causes and consequences of impaired genome integrity

vom Fachbereich Biologie der Technischen Universität Kaiserslautern
zur Verleihung des akademischen Grades
Doktor der Naturwissenschaften (*Dr. rer. nat.)*
genehmigte Dissertation

von **Paul Robert Menges**

Wissenschaftliche Aussprache: Kaiserslautern, 13.01.2022

Referent: Prof. Dr. Zuzana Storchova

Korreferent: Prof. Dr. Michael Schroda

Vorsitz: Prof. Dr. Johannes M. Herrmann

Kaiserslautern, 2022 - D 386

Eidesstaatliche Erklärung

Ich erkläre hiermit, daß die eingereichte Dissertation mit dem Titel *„Multi-omics analysis as a tool to investigate causes and consequences of impaired genome integrity“* von mir selbständig unter Anleitung von Prof. Dr. Zuzana Storchova verfaßt wurde.

Die für die Arbeit benutzten Hilfsmittel wurden genannt und die Ergebnisse beteiligter Mitarbeiter sowie anderer Autoren klar gekennzeichnet.

Die Dissertation dient einzig und allein der Erlangung des akademischen Grades „Doktor der Naturwissenschaften“ (Dr. rer. nat.) an der Technischen Universität Kaiserslautern. Die Dissertation oder Teile daraus wurden nicht als Prüfungsarbeit bei einem anderen Fachbereich oder einer anderen Fakultät eingereicht. Es wurde bisher bei keiner anderen Hochschule ein Promotionsverfahren beantragt.

Die Promotionsordnung des Fachbereichs Biologie der Universität Kaiserslautern ist mir in ihrer derzeit gültigen Fassung bekannt.

Kaiserslautern, 09.11.2021

Paul Robert Menges

List of publications

"*Systems approaches identify the consequences of monosomy in somatic human cells*" (2021) Narendra Kumar Chunduri, Paul Menges, Vincent Leon Gotsmann, Xiaoxiao Zhang, Balca R. Mardin, Christopher Buccitelli, Jan O. Korbel, Felix Willmund, Maik Kschischo, Markus Raeschle, Zuzana Storchova

Nature Communications (2021), doi: 10.1038/s41467-021-25288-x

"*Scaling of cellular proteome with ploidy*" (2021). Galal Yahya, Paul Menges, Devi Anggraini Ngandiri, Daniel Schulz, Andreas Wallek, Nils Kulak, Matthias Mann, Patrick Cramer, Van Savage, Markus Raeschle, Zuzana Storchova

BioRxivs (2021), doi: 10.1101/2021.05.06.442919

Summary

Impaired genome integrity has severe consequences for the viability of any cell. Unrepaired DNA lesions can lead to genomically unstable cells, which will often become predisposed for malignant growth and tumorigenesis, where genomic instability turns into a driving factor through the selection of more aggressive clones. Aneuploidy and polyploidy are both poorly tolerated in somatic cells, but frequently observed hallmarks of cancer. Keeping the genome intact requires the concentrated action of cellular metabolism, cell cycle and DNA damage response.

This study presents multi-omics analysis as a versatile tool to understand the various causes and consequences of impaired genome integrity. The possible computational approaches are demonstrated on three different datasets. First, an analysis of a collection of DNA repair experiments is shown, which features the creation of a high-fidelity dataset for the identification and characterization of DNA damage factors. Additionally, a web-application is presented that allows scientists without a computational background to interrogate this dataset. Further, the consequences of chromosome loss in human cells are analyzed by an integrated analysis of TMT labeled mass spectrometry and sequencing data. This analysis revealed heterogeneous cellular responses to chromosome losses that differ from chromosome gains. My analysis further revealed that cells possess both transcriptional and post-transcriptional mechanisms that compensate for the loss of genes encoded on a monosomic chromosome to alleviate the detrimental consequences of reduced gene expression. In my final project, I present a multi-omics analysis of data obtained from SILAC labeled mass spectrometry and dynamic transcriptome analysis of yeast cells of different ploidy, from haploidy to tetraploid. This analysis revealed that unlike cell volume, the proteome of a cell does not scale linearly with increasing ploidy. While the expression of most proteins followed this scaling, several proteins showed ploidy-dependent regulation that could not be explained by transcriptome expression. Hence, this ploidy-dependent regulation occurs mostly on a post-transcriptional level. The analysis uncovered that ribosomal and translation related proteins are downregulated with increasing ploidy, emphasizing a remodeling of the cellular proteome in response to increasing ploidy to ensure survival of cells after whole genome doubling. Altogether this study intends to show how state-of-the-art multi-omics analysis can uncover cellular responses to impaired genome integrity in a highly diverse field of research.

Zusammenfassung

Beeinträchtigte Genomintegrität hat drastische Folgen für die Überlebensfähigkeit jedweder Zelle. Unreparierte DNA-Läsionen können zu genomisch instabilen Zellen führen, welche häufig für unkontrolliertes Wachstum und Tumorentstehung prädisponiert sind. Die genomische Instabilität wird so durch die Selektion aggressiverer Klone zu einem treibenden Faktor in Krebs. Aneuploidie und Polyploidie werden in somatischen Zellen schwer toleriert, sind aber häufig beobachtete Krebsmerkmale. Um das Genom intakt zu halten, ist ein Zusammenspiel von Zellstoffwechsel, Zellzyklus und DNA-Reparatur erforderlich. In dieser Studie wird Multi-Omics-Analyse als Werkzeug zur Untersuchung diverser Ursachen und Folgen beeinträchtigter Genomintegrität vorgestellt. Mögliche Analysen werden anhand von drei verschiedenen Datensätzen demonstriert. Zunächst wird eine Analyse einer Sammlung von DNA-Reparaturexperimenten präsentiert, welche die Entstehung eines Datensatzes zur präzisen Identifizierung und Charakterisierung von DNA-Schadensfaktoren ermöglichte. Darüber hinaus wird eine Web-App vorgestellt, die es Biowissenschaftlern ermöglicht, diesen Datensatz zu untersuchen. Weiterhin werden die Folgen des Chromosomenverlusts in menschlichen Zellen durch eine Analyse von TMT-Massenspektrometrie und Sequenzierungsdaten untersucht. Diese Analyse zeigte, daß sich die heterogene zelluläre Reaktion auf Chromosomenverlust von der auf Chromosomengewinn unterscheidet. Meine Analyse ergab, daß Zellen sowohl transkriptionelle als auch post-transkriptionelle Mechanismen besitzen, die den Verlust von Genen auf einem monosomischen Chromosom kompensieren, um die Folgen der reduzierten Genexpression zu mildern. In meiner dritten Studie präsentiere ich eine Multi-Omics Analyse, in welcher ich Daten aus SILAC-Massenspektrometrie und einer dynamischen Transkriptomanalyse von Hefezellen verschiedener Ploidie, von haploid bis tetraploid, untersuche. Diese Analyse ergab, daß das Proteom einer Zelle im Gegensatz zum Zellvolumen nicht-linear mit zunehmender Ploidie skaliert. Während die Expression der meisten Proteine dieser Skalierung folgte, zeigten mehrere Proteine eine ploidie-abhängige Regulierung, die nicht durch die Expression des Transkriptoms erklärt werden konnte und folglich post-transkriptionell stattfindet. Die Analyse ergab, daß ribosomale und translationsbezogene Proteine mit steigender Ploidie herunterreguliert werden, was auf einen Umbau des Proteoms als Reaktion auf höhere Ploidie deutet, um das Überleben der Zellen nach einer Genomverdopplung sicherzustellen. Insgesamt soll diese Studie zeigen, wie moderne Multi-Omics-

Analysen zelluläre Reaktionen auf beeinträchtige genomische Instabilität in einem vielfältigen Forschungsbereich aufdecken können.

1. Introduction

1.1 Advances of proteomics and computational biology

1.1.1 Advances of proteomics approaches in systems biology

In the cells of any living organism the information about their appearance, function and purpose is encoded as a nucleotide sequence in their DNA. According to the simplified central dogma of molecular biology, the information stored this way is transcribed from gene to RNA, processed to mRNA and translated to a protein [1]. Yet, the process of protein biogenesis as a whole is affected by a multitude of factors. While the genome of a cell is largely static, the functional proteome, or a set of all expressed proteins at a given time, is dynamic in response to environmental factors, disease or impaired genome stability and reacts steadily to external and internal perturbations. The precise abundance of each protein is tightly regulated by biogenesis and degradation to control the proteome composition [2]. As proteins are the biochemical actors in all cellular as well as many disease related processes, understanding and associating them to a biological function is playing an increasingly crucial role in modern biological research.

The precise function of proteins is influenced by post-translational regulation and the information about post-translational modifications (PTMs) is determined by the amino acid sequence. Modifications regulate the activity, degradation or localization of a protein, hence play an important role in its biological processes [3]. Together with alternative splicing of RNA, polymorphisms and translation efficiency, there are numerous other factors that shape the highly complex proteome and cannot be studied by DNA or RNA sequencing alone (Figure 1).

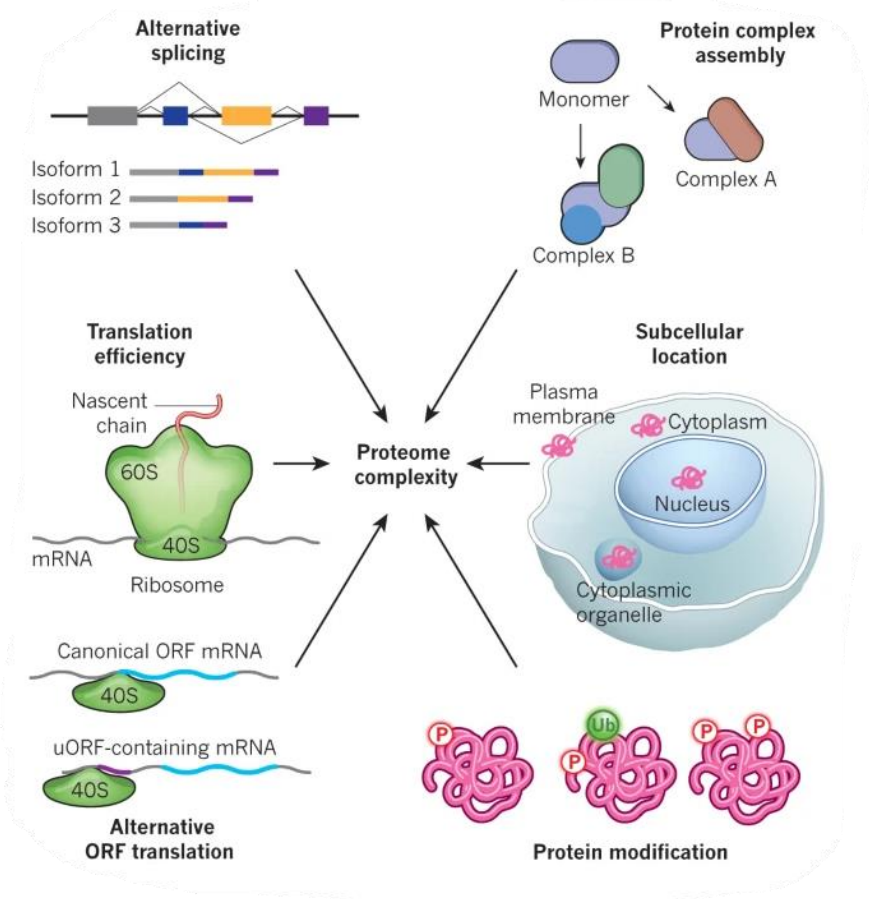


Figure 1 | Proteome complexity. The complexity of the proteome is influenced by various factors. Post-translational modifications are involved in regulating various protein mechanisms, from conformational changes that lead to function activation of the protein, to degradation by polyubiquitin fusion or dynamic translocation of proteins between cellular compartments. Translation efficiency and protein degradation steadily adapt and regulate the levels of proteins. Alternative splicing and protein complex assembly produce different isoforms and influence protein dynamic and stoichiometry (adapted from Rev. [4]).

Since the completion of the human genome project in 2004 [5], it became increasingly clear that to gain a systematic insight to the physiological function of the proteome of a cell, genomics and transcriptomics alone were not sufficient. As the proteome is highly dynamic, established methods for analyzing individual proteins, such as the direct analysis of the amino acid sequence going back to Edman degradation [6] to staining of gel separated or antibody-based methods, do not provide sufficient information about the entirety of all expressed proteins in a given state.

“Proteomics” therefore has gained traction as a rapidly growing, post-genomic field in the recent years, which focuses on large scale, quantitative study of all expressed

proteins in a cell. The most common modern proteomics approaches are based on mass spectrometry, a tool that allows scientists to reliably identify and quantify the total protein complement of a biological system.

In summary, the function of a protein is determined by more than only its underlying nucleotide sequence. Due to a non-static biogenesis and degradation of proteins as well as PTMs the composition of the proteome is highly dynamic. To gain meaningful systematic insights about the physiological function of a cell, modern mass spectrometry is gaining traction as reliable method for quantitative proteomics analysis.

Mass spectrometry-based protein analysis

Any mass spectrometer consists of an ion source, a mass analyzer and a detector where the measurement of mass-to-charge ratios (m/z) is carried out in the gas phase of ionized analytes. The two most established techniques to ionize peptides from an ion source, or in certain measurement approaches proteins directly, are matrix assisted laser desorption and ionization (MALDI) and electrospray ionization (ESI). MALDI ionizes analytes from a solid phase, crystallized matrix utilizing UV laser pulses. ESI ionizes analytes by creating a high voltage aerosol from a solution. It is therefore compatible with liquid-based chromatography and fractionation methods for analysis of samples with higher complexity [7,8].

Many mass spectrometers have been established using hybrid designs with a combination of multiple analyzers and separators in the recent years. MALDI based ionization is generally, although not exclusively, used with time-of-flight (TOF) analyzers in combination with one or multiple quadrupoles (tandem or triple-quadrupole/TOF). The quadrupole mass analyzer allows the separation of ionized analytes based on their m/z ratio. Oscillating electric fields are applied to four parallel metal rods. Since the stability of a the ion trajectory while passing through the quadrupole is relative to its m/z ratio, the system functions as a dynamic mass filter [9]. In most modern tandem mass spectrometers, a quadrupole is connected to a collision chamber, in which ions are broken into fragments by collision-induced dissociation (CID) with molecules of an inert gas before being passed to a time-of-flight detector. The time of flight reflector orthogonally accelerates the ion fragments by an electric

field towards a detector that amplifies and counts the signal of arriving ions. Therefore, the impulse injection pattern produced by MALDI is better suited for time-of-flight detectors [10,11].

Developments in modern mass spectrometry strived to increase the confidence in the identification of analytes by using sophisticated mass analyzers with increased accuracy, resolution and resolving power to get a better understanding of samples with higher complexity. Therefore, ESI ionization is used in many mass spectrometry facilities as it allows the analysis for more complex, soluble samples in combination with previous liquid-phase chromatography (LC-MS) and state of the art systems utilizing orbitrap mass analyzers. In those mass spectrometers a quadrupole is used to filter ions based on their m/z ratio and direct them into a high energy collision chamber (HCD), in which analytes are fractioned with high kinetic energies in the kilovolt range, which produce smaller fragments. These are collected in a C-trap that passes them pulse-wise to the orbitrap. Inside this specialized ion trap, the ion populations harmonically oscillate around the length of a central spindle-like electrode in a vacuum, where the frequency and intensity of the oscillation are measured and resolved via Fourier transformation. This process, while computationally intensive for heterogeneous ion populations, enables commercially available instruments with an resolving power at full-width-half-maximum (FWHM) of up to 280,000 at a peak with an m/z ratio of 200 [12]. This is sufficient to distinguish ion footprints of most analytes, still in non-commercial, experimentally tuned orbitrap systems the resolving power in recent studies could exemplarily be increased to 2,000,000, which allows isotopic resolution of intact monoclonal antibodies [13].

In summary, the most commonly used approach for the ionization of peptides is ESI, as it can be easily integrated in many “bottom-up” approaches for complex samples and combined with many modern LC-MS systems, such as cutting edge orbitrap mass analyzers that facilitate high accuracy spectra identification.

Bottom-up mass spectrometry.

The “bottom-up” approach is the most widespread proteomics workflow for the quantitative identification of proteins in complex mixtures. This technique relies on a sample preparation stage, in which proteins are digested to peptides by sequence-specific enzymes, such as the pancreatic trypsin that cuts peptide chains at the carboxyl side of the amino acids lysine or arginine, or the bacterial endoproteinase Lys-C or Lys-N for the carboxy- or amino- end of lysine. Peptides are then commonly separated by fractionation and reversed-phase high-performance liquid chromatography (HP-LC) to increase the identification rate for individual peptides [14]. Three common approaches for bottom-up proteomics workflows are depicted in Figure 2 A, data-dependent acquisition (DDA), data independent acquisition (DIA) and targeted proteomics.

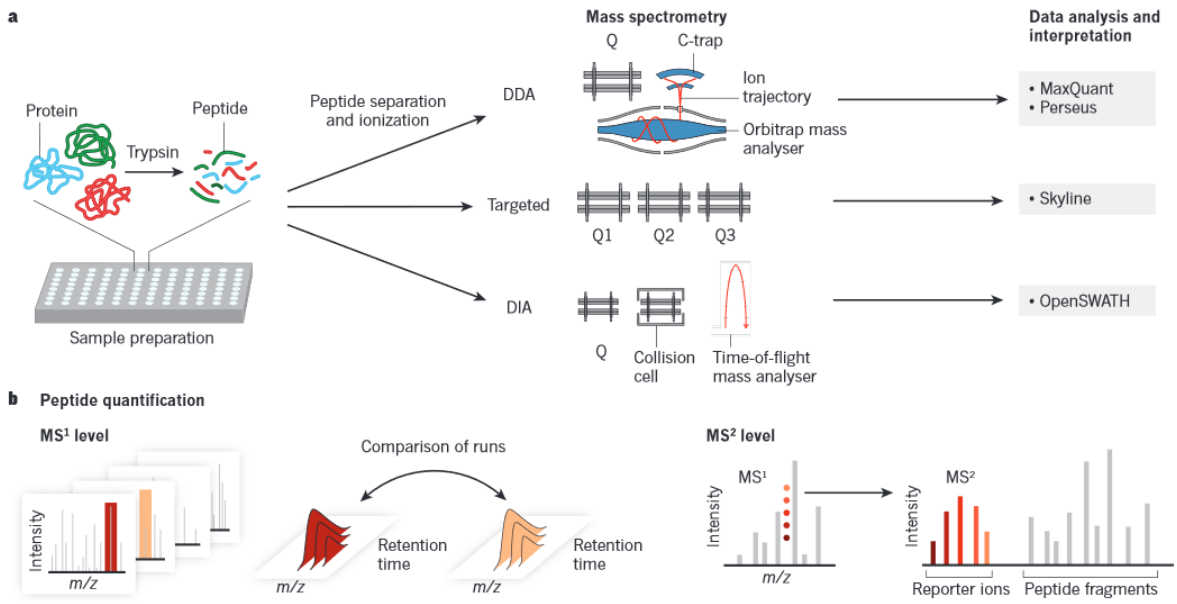


Figure 2 | Bottom-Up Mass Spectrometry. The figure shows data acquisition and peptide quantification approaches commonly used in bottom-up mass spectrometry. (A) highlights the data acquisition technique, from left to right, starting with a sample preparation stage including enzymatic digest of the protein to peptides, to peptide ionization with an ion source. The mass spectrometry approaches are shown in the middle as three methodical classes. DDA, data dependent acquisition or “shotgun” proteomics, which uses a qPol-Orbitrap setup to select and fragment individual precursor ions from a MS^1. Targeted proteomics, utilizing a triple quadrupole mass spectrometer setup in which a known peptide is selected based on its m/z ratio in the first quadrupole Q1, fragmented in Q2 and lastly the fragments are monitored over time in Q3. Data independent acquisition (DIA), which uses a mass-to-charge window to consecutively fragment and measure the intensities of all passing analytes in a qpol-TOF setup. On the right the computational end of the pipeline is depicted, showing exemplarily

software platforms to process the acquired spectrum data. (B) shows different peptide quantification methods. (left) shows the integration of multiple MS^1 spectra utilizing alignment procedures, such as the matching of peptide masses to spectra with similar retention times, where the ion has been fragmented in one but not the other. (right) shows the selection of unique reporter ion peaks being selected for quantification at the MS^2 level (from Rev. [15]).

The first and most broadly used approach is the data-dependent acquisition (DDA) or "shotgun" proteomics for the unbiased identification of proteins. Here, all peptides that co-elute from the reversed-phase liquid chromatography are ionized by electrospray ionization at the same time. The full range of detectable precursor ions is identified in a "full scanned" mass spectrum (MS^1). The mass spectrometer alternates between acquiring novel precursor ions and choosing ions resembling the *top N* peaks, whereas N determines the number of peaks with the highest intensities in the current MS^1 for fragmentation and consecutive measurement of MS/MS (MS^2) fragment-ion spectra in the mass analyzer. DDA approaches are popular as they allow scientists to create hypotheses by quantifying all proteins in a sample without previous knowledge, while still maximizing the amount of detected peptide features based on unbiased precursor characterization [16]. Furthermore, not all fragment ions need to be identified. Instead, empirical "fingerprints" of individual peptide spectra are matched and evaluated based on probability scoring against publicly available reference databases, such as UniProtKB [17] or NCBI RefSeq [18], which contain amino acid sequences of different organisms.

Multiple search engines have been established for data dependent approaches, such as Sequest [19], Mascot [20], or Andromeda [21]. A variety of both commercial and non-commercial software applications incorporate those data-dependent search engines and automate many of the required downstream processes. OpenMS [22] or Proteome discoverer [23], utilizing different search engines, and MaxQuant [21], relying on the in-house developed search engine Andromeda, are among of the most sophisticated solutions. For the calculation of protein group intensities of the commonly used label-free quantification approach, MaxQuant employs mass and retention time alignment between measurements to transfer peptide identifications from fragmented to un-fragmented peptides to increase the total amount of peptide ratios across samples (Figure 2 B). The absolute quantities of unique "reporter" MS^2 ion intensities are summed up to determine a protein group intensity. The problem of peptide ion signals

differing across multiple measurement runs in unlabeled samples, which would require individual normalization coefficients, is avoided by utilizing “delayed normalization”. This technique relies on first summing up the individual intensities with normalization factors as free variables. Their quantities are determined later via a global optimization procedure based on determining the least overall proteome variation. This process, termed “MaxLFQ” [24], relies more than other measurement approaches on highly accurate peptide identification, as it correlates with the number of data points for pairing of the corresponding peptides.

Recent advances using deep learning techniques aimed at increasing the overall peptide identification rate across all intensity ranges by predicting the peptide-spectrum matches (PSM). By constructing regression models, exemplarily either by bidirectional recurrent neural networks (RNN) or sliding-window-based approaches, the achieved MS/MS spectrum prediction was nearly as accurate as the limit of technical reproducibility. While predictive techniques still lag behind in certain areas, such as the prediction of ion spectra with higher charges or with modifications, these shortcomings will decrease with the steadily increasing amount of training data. Machine learning techniques will continuously increase the efficiency of both data dependent and data independent acquisition for existing peptide search engines either by supplementary spectral prediction or the generation of in silico–based spectral libraries [25].

In contrast, data independent acquisition (DIA) is used as an alternative approach to reduce dimensionality compared to “shotgun” mass spectrometry. Figure 2 A depicts a common mass spectrometer setup with a qpol-TOF setup, which facilitates consecutive, time-resolved data acquisition. Instead of recording a full MS^1 at a set timepoint, the DIA approaches, such as Sequential Window Acquisition of All Theoretical Mass Spectra (SWATH-MS), rely on shifting small, sequential mass windows of 25 m/z. All acquired precursor ions are concurrently fragmented and the resulting fragment ions recorded in a time-of-flight analyzer. This produces a two-dimensional record of fragment ion signals that are measured in continuous time and fragment-ion intensity [26]. Multiple software applications largely automate the computational pipeline, such as Spectronaut [27] or OpenSWATH [28]. The detection rate of this approach is generally limited by the accuracy as well as by the size of the mass window, which is used for precursor ion acquisition. This is dependent on the

increasing complexity due to multiplexing, as the fragment spectra contains ions of more precursors with larger mass windows, while the link between a specific precursor and the related fragment ions is unknown. The collected spectra can be interpreted by database searching [29], or *de novo* identification of mass spectra.

De novo identification does not rely on commonly used databases, but instead identifies the precursor ion mass depending on a composition of the peaks of the fragmented MS^2 ion spectra, hence inferring the mass of a precursor based on its related MS^2 spectra, mass and charges. The principle of *de novo* MS relies on a stepwise calculation of the precursor mass as combination of the m/z ratios of the individual peaks in the MS^2 spectra, representing b- or y-ion fragments, from the amino or carboxyl terminus of the amino acid sequence, of the precursor. The initial graph theory based approach was originally described in 1990 by Bartels [30]. A spectrum graph is calculated, in which nodes represent the m/z ratios of the peaks in the MS^2 spectrum and edges between them are inferred if the m/z matches an amino acid, combination or modification of amino acids. The longest, connected path of edges that has a mass resembling the precursor mass contains the information about its amino acid sequence, and is present in a spectrum for b- and y-ions, respectively. Through the steady rise of computational biology, the *de novo* mass spectrometry has seen performance and probabilistic improvements in the recent years, and allows scientists to identify novel proteins or modifications that have not been described yet and therefore do not exist in publicly available databases. The shifting focus to machine and deep learning techniques is facilitating neural network based *de novo* peptide-sequencing methods for both data dependent [31] and data independent approaches [32]. This new techniques play an increasing role in personalized immunotherapy in cancer through the identification of individual neoantigens [33].

The third approach to bottom-up mass spectrometry is targeted proteomics, serving as a completement to “shotgun” proteomics. Instead of an unbiased measurement of the complete proteome, targeted proteomics allows high accuracy quantification of predetermined proteins, a technique that has gained traction in biomedical research for the identification of drugs or small molecules. Figure 2 depicts the setup of a triple quadrupole mass spectrometer (QQQ), which is commonly coupled with liquid chromatography and enzymatic digest. As the fragmentation properties of the analyzed protein are known, the quadrupoles act as a filter to selectively monitor only the protein

of interest as well as its fragments [34]. The measurement of a fragment-ion pair is called single reaction monitoring (SRM) and is commonly multiplexed for multiple fragments in a multiple reaction monitoring (MRM). Skyline is a popular bioinformatics platform for the downstream processing of targeted proteomics, which relies on the scientists providing the protein sequence information as well as the list of peptides, precursors and expected product ion transitions lists to robustly identify the protein of interest or quantify it based on labeled internal standards [35]. With the development of high accuracy orbitrap mass analyzers, the third quadrupole in the instrument setup could be replaced to perform a full scan of each transition to scan all product ions of the fragmentation together with the precursor in parallel (PRM, parallel reaction monitoring), which adds a novel qualitative ability to the measurements. This workflow has been recently established to identify SARS-CoV-2 proteins directly from nasopharyngeal and oropharyngeal swabs to potentially facilitate large-scale, clinical population screening [36].

In summary, bottom-up mass spectrometry is the most wide-spread proteomics approach. The three techniques, data dependent acquisition, data independent acquisition and targeted proteomics facilitate reliable identification and quantification of proteins by utilizing specialized mass spectrometer setups and software to cover a wide range of studies.

Label-based mass spectrometry

To further increase the quantitative performance of bottom-up mass spectrometry, many metabolic or chemical labeling approaches have been developed in parallel to the technical and computational advances. Early methods relied on stable isotope dilution techniques, earliest described in experiments measuring the protein expression in yeast *Saccaromyces cerevisiae* as Isotope Coded Affinity Tags (ICAT) [37]. This approach uses eight heavy deuterium atoms, incorporated in the labeled peptides. This weight shift of 8 Dalton (Da) allows the separation of labeled and unlabeled peptides and a quantification based on relative signal intensity in those peptide pairs. This inspired the creation of isobaric tagging methods, like iTRAQ [38] and TMT [39]. These methods rely on reagents with functionally designed groups (Figure 3

top), an amine reactive group that facilitate binding to the peptide, a mass normalizer region, and a mass reporter. The mass normalizer and reporter together have the same structure and mass, yet are substituted by isotopes at various positions (Figure 3 top, red marked star). Due to their identical total molecular weight, labeled molecules are indistinguishable during LC and in MS^1 spectra, yet produce distinguishable mass reporter ions for quantification in MS^2 after fragmentation, differing in weight by 1 Da from 126 to 131 for TMT sixplex (Figure 3 bottom). Recently this multiplexing strategy could be expanded up to a 16-plex by combining isotopic substitution at both 13C and 15N positions [40]. A disadvantage to this technique stems from the coelution of peptides within set isolation windows; while this effect can be diminished by mass spectrometers with higher accuracies and setting smaller isolation windows, it still leads to a systematic underestimating of peptide/protein ratios that has to be considered [41]. More recent advances in isobaric tagging, such as Easily Abstractable Sulfoxide-based Isobaric-tag (EASI) [42], have been shown to maintain the advantages of isobaric tagging, while preventing ratio compression effects. This is facilitated by using tags that fragment at a lower collision energy than the peptide backbone, and measuring with asymmetric ^{12}C-specific isolation windows to minimize the isolation of co-eluting peptides.

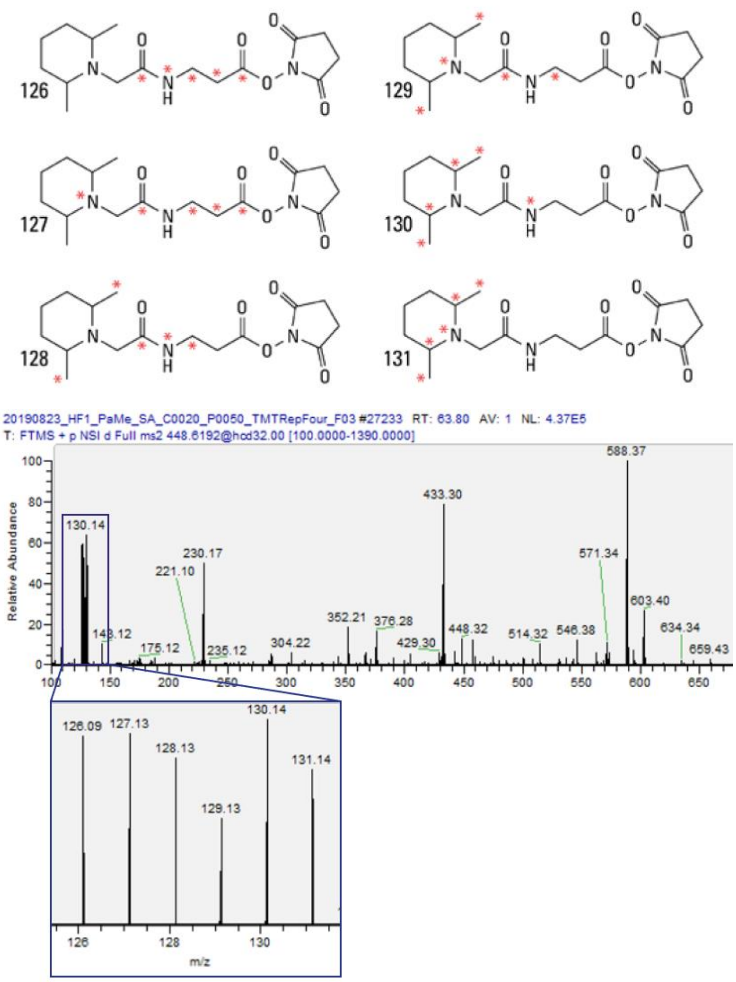


Figure 3 | TMT Labeling. The figure shows the schematic of a TMT sixplex (top: adapted from thermofischer.com) and the TMT signal in an MS^2 spectra (bottom: in-house measurement, screenshot from the software Qual Browser).

Alternative labeling methods have been proven to work without the need for specifically designed tags, such as stable isotope labeling by amino acids in cell culture (SILAC). Here, different populations of cells are grown in the presence of isotopes ^{12}C, ^{14}N or ^{13}C, ^{15}N, respectively. The incorporation of isotopes after multiple cell doublings produces a “heavy” population, which can be mixed with unlabeled “light” populations of cells and treated together for a single LC-MS workflow [43]. Due to the incorporation of isotopes, the measurement produces ion pairs, which can be quantified by the calculation of ratios between the “heavy” and “light” ion signal. This technique can readily be expanded by the creation of a mixed “Super”-SILAC standard, in which multiple labeled populations are mixed to create an internal standard for the computation of light-to-heavy ratios of individual populations. This method has been proven to be efficient highly efficient for the comparison of cellular states or the study of intracellular signal transduction [44]. It is a powerful technique applied in many experiments in different organisms, from plant studies in *Arabidopsis thaliana* [45] to the quantification of multiple human tumor tissue proteomes [46].

In summary, there have been numerous advances in proteomics research across different approaches, from “bottom-up” data acquisition to computational downstream pipelines. This facilitates a steady increase in the analytical performance of mass spectrometry, with an increase in resolving power, resolution and sensitivity, and plays an important role in the identification of novel factors for clinical research, different proteoforms of individual proteins as well as the understanding of the proteome to provide a novel perspective of the landscape of an organism in addition to its genome and transcriptome.

1.1.2 Big data analysis in modern biology

Mass spectrometry and other omics approaches in general produce massive amounts of data. The number of proteins detected in a single cell type encoded by over 10.000 protein-coding genes can go beyond 10.000 proteins, which in the recent years could be readily identified in a single run when using orbitrap systems and SWATH-DIA measurement approaches [47, 48]. This provides a highly dynamic range of protein expression in any given system, yet independent of the data acquisition method the downstream biological analysis to create meaningful hypotheses remains a bottleneck in omics experiments to this day. The processing of mass spectrometry data is in most cases performed in functional or object-oriented programming languages, such as C#, Python or R, using scripts tailored to the individual analysis, hence requiring significant computational expertise. While there are several accessible computational platforms or applications for the data processing of omics data, far less exist for the data analysis and interpretation of them. This highlights a gap between “dry lab” bioinformatics and “wet lab” biological research that further increases over time with the development of more sophisticated analysis methods in both fields. As data analysis is a highly varied field, applications that attempt to facilitate the generation of hypotheses from omics data are usually specialized and either limited in functionality or scope. For example Perseus [49], as supplementary application for MaxQuant uses C# functionality to facilitate explorative data analysis for scientists without a background in bioinformatics. Due to this it is limited in functionality to the curated processing steps the developers implemented for proteomics analysis. “OmicsVolcano”, as recently released platform utilizes web-server-based structure for the visualization and exploration of omics data, but is limited to only the visualization, not the analysis of data [50]. The “DNA Repair Atlas”, presented in Results Chapter 1, allows visualization and mining of *Xenopus laevis* proteomics data focused on DNA repair, yet is currently limited in scope to an expansive, manually curated dataset. Many more similar applications exist, yet the problematic of closing the gap between dry and wet lab persists.

In summary, recent advances in data acquisition lead to a drastic increase in the amount of generated data. Downstream data processing and hypothesis generation from the data is a bottleneck in MS analysis to this day. Improving the general accessibility of this increasingly complex field for scientists without a computational background is an important ongoing topic.

Normalization as a critical step of data analysis

The hypotheses generation from proteomics data is based on the experimental design, while the most common approaches revolve around the identification of novel protein interactions, its functions, or the comparison of the proteome between two different physiological states of a system. Comparative, quantitative approaches therefore can be categorized as systematical or “top-down”, to investigate the changes of the global proteome in response to a perturbation and identify novel cellular and functional mechanisms [51], and specific or “bottom-up”, to bait or tag and pull-down individual proteins of interest to identify novel protein-protein interactions. Affinity purification-mass spectrometry (AP-MS) has long been established as the standard “bottom-up” technique to investigate the dynamic interaction changes of a protein of interest in response to perturbed conditions by examining a baited subset of proteins [52]. A schematic of a downstream computational analysis workflow is depicted in Figure 4.

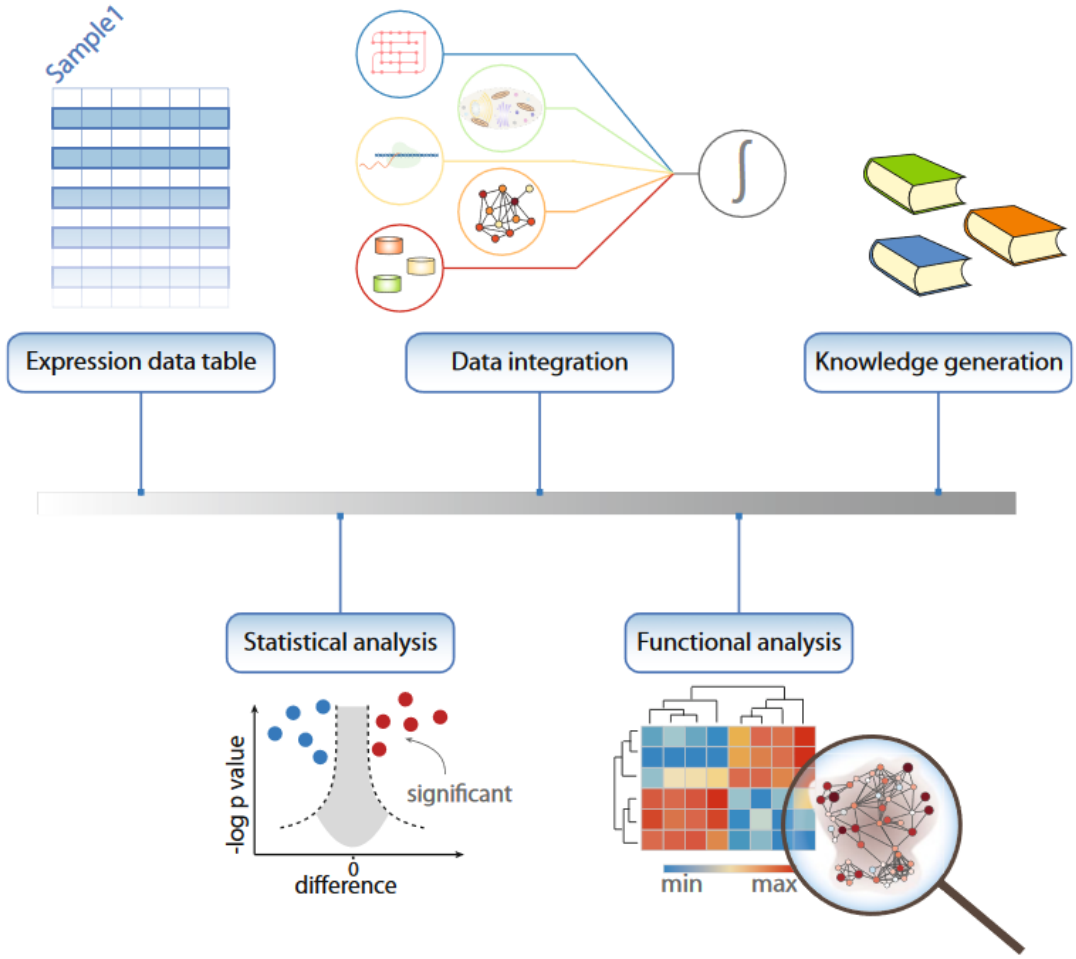


Figure 4 | Converting data to knowledge. The figure shows a schematic for the generation of information from omics data, starting from the statistical analysis to identify outliers or significantly deregulated values. This is followed by data integration and annotation by external resources and databases. From there a functional analysis is shown, highlighting meaningful relations between measured values that lastly produce information about biological processes or molecular functions (from [53]).

While there are numerous approaches for the analysis of proteomics data, the entry point is commonly an expression data table containing expression data in the form of protein group or label intensities or spectral counts. Before any statistical analysis, this set needs to be adjusted for a systematic bias that commonly is introduced by nonbiological factors, such as differences in preparation of samples, labeling efficiency or deviations in the experimental conditions during mass spectrometric measurement, so called “batch effects”. To handle these issues, several normalization approaches have been developed, as careful normalization to remove or at least down-weight the systematic variation in between measured samples and replicates is crucial. This is only amplified by the introduced technical bias generally being unknown at the time of data analysis. While certain computational platforms for the processing of mass spectrometry data already have integrated normalization methods, such as MaxLFQ [24] in MaxQuant that applies “delayed normalization” of label-free measurements, developments over the last years resulted in multiple standalone alternatives.

Older and straightforward normalization methods stem from the analysis of cDNA microarray data, but can be readily applied for other omics datasets, assuming they follow a normal distribution. A straightforward normalization is for example median shifting. The data is transformed to a $\log_2$ scale and the individual measurements scaled to have a difference in median of 0 [54]. This assumes that the individual samples differ only by a systematic factor, such as the labelling efficiency of an isobaric tag. An alternative computation to achieve a comparable effect is described as quantile normalization, where each value in a sample is transformed to the mean of the corresponding quantile to make the individual distributions identical in their statistical properties [55].

Later normalization methods managed to incorporate further factors that could explain systematic bias. Linear regression normalization can adjust a bias that is linearly

dependent on the peptide abundances. In LC-MS experiments, this can result from leftover peptides on the analytical column in between measurements that elute and lead to an overlap in detected peaks in consecutive measurements. This normalization is performed by applying the least squares estimations, a linear regression model that fits data by minimizing the sum of squared residuals, which describe the difference between the measured observation and the fit obtained by the trained model [56]. For a nonlinear relationship between bias and peptide abundances, the closely related local regression normalization can be used. A nonlinear bias can potentially be included by background noise during measurements, the peptide signals reaching mass detector saturation or other ion suppression factors [57]. The most commonly used local regression approach is based on the locally estimated scatterplot smoothing (LOESS), that is incorporated in the R/Bioconductor software package *limma* [58] (see below).

Another approach to normalize data for systematic experimental factors, such as labelling efficiency and detector sensitivity, is variance stabilization normalization (VSN). This model-based method applies a two-step procedure including an affine transformation, which scales data in between measurements to adjust for previous calibration or normalization steps, and a variance stabilization step, which uses maximum likelihood estimation to transform variances in a measurement to be independent of the mean intensity. As this method only adjusts to this specific dependence, it does not consider other factors influencing the variance aside the mean intensity, such as differing transcriptional regulation between sample conditions or gene-inherent properties [59].

Nevertheless, in recent comparative studies it was shown that VSN consistently and successfully reduced the intragroup variation between technical replicates in both spike-in and label-free data compared to ten other normalization methods, including local and linear regression models as well as median and quantile normalization [60]. Interestingly, in previous studies the regression models outperformed VSN, which in turn was ranked only average [61]. This was explained by the authors of the recent study by the stronger difference and variation between their spike-in datasets, resembling different instrumentation or protocols used. It is plausible that this reflects “real world” settings better, where VSN adjusts better for unknown systematic biases introduced during sample preparation and measurement.

In summary, measurement of omics data always introduces a technical or experimental bias to a dataset. For ideal comparison and generation of hypotheses it is crucial to adjust this bias to minimize variance between measurements. Several algorithms, from MaxLFQ to linear regression-based models and variance stabilization have been established to approach this in the recent years.

Identification of differentially expressed factors

The next step in the generation of information from normalized omics data is commonly the identification of outliers or differently expressed genes between measurements of samples in different conditions, treatments or untreated controls. The identification of outliers in this approach is commonly performed either via either hard-threshold cutoff based methods, such as $\log_2$ estimation of at least n-fold up or downregulation of fold changes, or statistical inference usually employing two or multi sample t-tests or model-based methods. While standard two sample t-tests have long been established for omics data, determining significance of individual values in datasets with usually smaller sample sizes due to a low number of overall runs, but in turn often high variance, is an ongoing topic.

Adapted from the field of genomics, the Significance Analysis of Microarrays (SAM) has found its use for proteo- and transcriptomics data as well. This approach identifies statistically significant genes by computing the strength of the relationship of each individual gene expression to a set response variable, such as treatment, condition or pairing, by assimilating a set of gene-specific t-tests. To determine whether the gene relation is significant, SAM employs a tuning parameter as a cutoff based on the false discovery rate (FDR), which is estimated from nonsense genes identified from permutations of the measurements [62].

An alternative solution are empirical Bayesian methods, inference methods based on the shrinking of sample variance via the estimation of the standard error towards a common mean for calculated fold changes between set groups. By shrinking proteins with larger abundancy and fold change, that generally tend to also have larger variability after logarithmic transformation due to the smaller sample size, become

more significant as compared to regular t-tests. in which proteins with large fold changes and variance often are missed due to being deemed not significant.

Proteins with extreme abundancy and fold changes tend to also have larger variability after logarithmic transformation due to smaller sample size. By shrinking the variance of those proteins towards a common mean they show a higher significance in statistical tests, in which they otherwise could be missed due to high variance. This methodology has been described in the highly cited “linear models for microarray data” (LIMMA) [63] and is included in the R/Bioconductor package with the same name [58]. In agreement with the high impact of the method in the field of omics research, several studies confirmed its efficiency over regular t-statistics for proteomics [64] as well as for the identification of differentially expressed peptides [65].

A further issue in the identification of significantly deregulated proteins are the so-called missing values in measurements. They occur either due to a low abundance, peptide-spectrum mismatches, or imperfect ionization of peptides. Those missing values severely impact the statistical inference. Several methods to offset or impute those values have been established over the years, most notably single and multivariate imputation.

The most straightforward approach is the limit of detection (LOD) imputation, which replaces missing values by either a constant or randomly selected values from a normal distribution near the detection limit and is most suited for the imputation of missing values of bottom-up approaches, such as “shotgun” proteomics. More complex multivariate imputations are based on either distance-based methods, such as the K-nearest neighbors (KNN) imputation utilized in SAM, or regression based, such as local least square (LSS) or least-squares adaptive (LSA) [66]. Recent in-depth assessments of imputation for differentially expressed proteins in “shotgun” proteomics highlighted that SAM and empirical Bayesian methods show better statistical evaluation of datasets compared to classical t-testing, and imputation to some extent enhance this performance, yet it should be noted that the imputation efficiency generally decreases with the number of missing values [67].

In summary, statistically significant identification of differentially expressed genes is crucial to draw conclusions about any given dataset. To achieve this, complex permutations of the dataset to decrease overall variance of outliers or replace missing values have been established to facilitate a reliable identification of meaningful genes.

Annotation and pathway analysis of proteins

As proteins rarely function alone and rather operate as components in macromolecular complexes to achieve their biological function [68], annotation and categorization of differentially expressed genes or proteins is essential for a functional analysis. Specifically, interpretation of big data requires careful integration of data to gather information about cellular functions and pathways. There are numerous databases, from general to highly specialized, which facilitate categorical annotation of genes.

The Gene Ontology Consortium (GOC) provides one of the most comprehensive resources available to compute general gene related annotations. They hierarchically categorize genes in three independent ontologies: Biological Process, Cellular Compartment and Molecular Function. Each term in the ontology annotates a set of genes, has a numerical identifier and is linked by relation in a layered graph structure to more universal parental terms up to the root term and with increasing specificity to child terms. The edges in the graph represent the type of the relation [69]. The Gene Ontology Consortium dynamically expands the ontology by novel terms, proteins and functionalities [70].

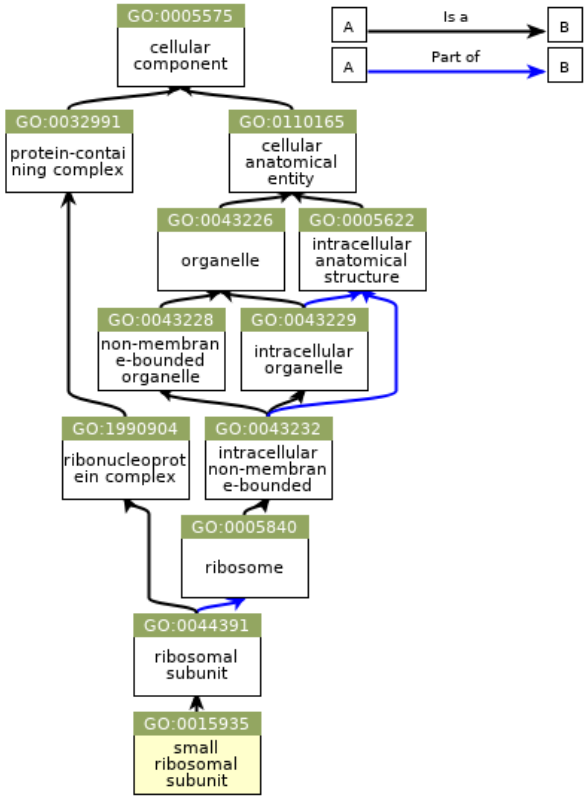


Figure 5 | The GO graph. The figure shows the hierarchical gene ontology "cellular compartment" graph for the term GO:0015935 (small ribosomal subunit) including its related parent terms.

(Adapted from: https://www.ebi.ac.uk/QuickGO)

Further commonly used databases in similar structure are the Kyoto encyclopedia of genes and genomes (KEGG) for systematic annotation of high level gene functions. It is a manually curated database resource to integrate genomic, chemical and systemic function information to a dataset [71]. CORUM is often used as a more specialized database to specifically annotate mammalian protein complexes [72].

A long-established approach to analyze differentially expressed genes in order to extract novel insights about their collective function is the gene set analysis. The most common approach is the overrepresentation analysis (ORA), which is based on the identification of differentially expressed genes that are annotated by biologically relevant terms, such as from GO, KEGG or Corum. Generally, this method applies the Fisher Exact test assuming hypergeometric distribution for small, or binomial distribution for large sets, as well as an independence of the values. Under the null hypothesis that there is no association between differential expression and annotated term, a p-value is calculated by testing whether a gene being part of an annotated term is by chance via random sampling [73]. Many additional approaches expanded on this idea, such as the web-application EnrichR, which computes an additional combined score as product of p-value from the Fisher Exact test and z-score from the deviation from a calculated expected rank for better evaluation of the results and interactively visualizes them for different annotated gene sets [74].

Further methods include the expression table directly, such as the Gene Set Enrichment Analysis (GSEA), an univariate functional class scoring method that calculates an enrichment score by Kolmogorov-Smirnov statistic from a ranked list of input genes through a set input parameter, such as a p-value, fold changes or gene expression measure [75]. While broadly used, a disadvantage of this method is the overreliance on the absolute size of the term, over scoring terms with genes centered in the ranked list, which are often not significant for the studied phenotype. Multiple attempts to address this shortcoming were established, such as using a weighted Kolmogorov-Smirnov statistic and FDR base adjustment for multiple comparisons [76] or unsupervised, competitive testing as “spectral” gene set enrichment (SGSE) [77].

The use of GSEA as a comparative tool was implemented in Perseus by one or two-dimensional annotation enrichment. This ranks independent sets of genes or gene products by fold change, expression or p-value and computes an enrichment score by multivariate analysis of variance (MANOVA), which is corrected by a Benjamini-Hochberg FDR [78]. A positive score shows upregulation, a negative downregulation of the related gene set. The benefit of a strictly rank based computation of enrichment scores is the ability to integrate sets originating from different measurement or experimental backgrounds, as shown in Figure 6.

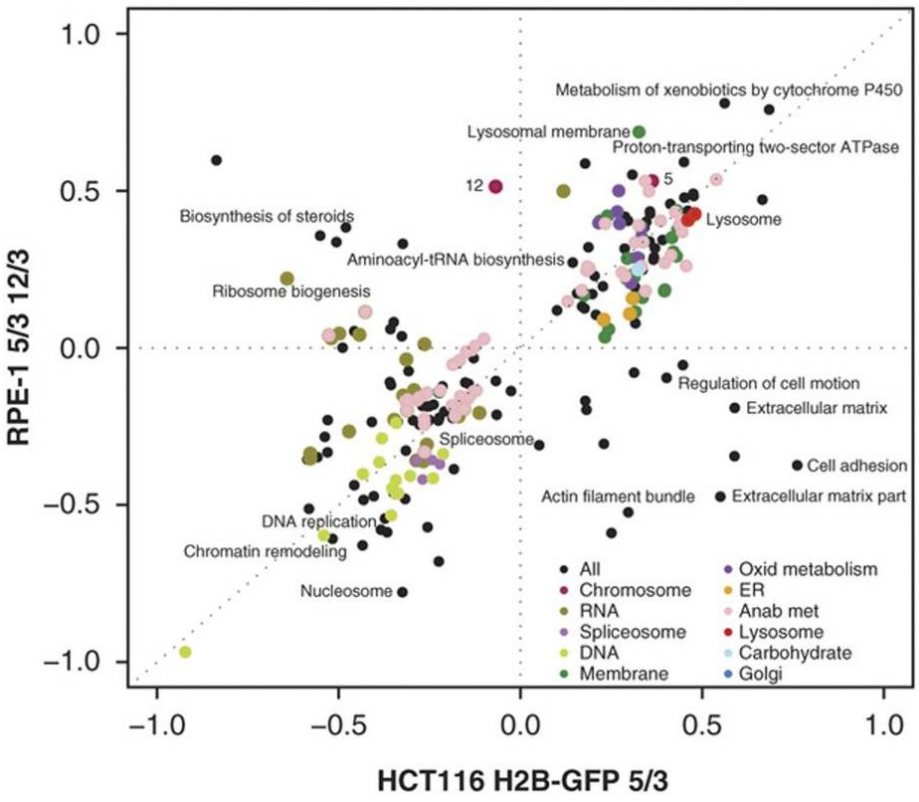


Figure 6 | 2D Annotation Enrichment. The figure shows the result of a two dimensional annotation enrichment of the proteome of the aneuploid **H**uman **C**olon **T**umor (HCT) cell HCT116 H2B-GFP 5/3 and the **R**etinal **P**igment **E**pithelium (RPE) cell lines RPE 5/3 12/3. Shown are differentially regulated pathways in the cell line with chromosome 12 and 5 trisomy compared to a cell line with chromosome 5 trisomy (Benjamini-Hochberg FDR threshold 0.02) [79].

In this analysis, the results of proteomics measurement of cell lines with a trisomy in chromosome 12 and 5 and trisomy 5 were annotated with the gene sets from KEGG and GO. The enrichment scores of both measurements are scaled from -1 to 1 and plotted as scatter plot with term-specific coordinates in a Cartesian coordinate system. The position in the plot determines correlation or anticorrelation of a term in the independent datasets, as well as up- or downregulation (top right/bottom left: correlating region, top left/bottom right: anticorrelating region) (Figure 6).

The term *Lysosomal membrane* (top-right) exemplarily shows an upregulation and correlation in both cell lines. The terms for *DNA Replication* (bottom left) a downregulation in both cell lines and a correlation between both. *Ribosome biogenesis* (top left) shows slight upregulation in the HCT cell line, but downregulation in RPE, hence anticorrelation between the sets.

In summary, identification of the expression of a single deregulated gene usually has only limited value for the analysis of biological processes. Therefore, many specialized annotation databases exist as resource to categorize sets of genes by function or cellular compartment. The analysis of the deregulation of entire gene sets by enrichment analyses facilitates more meaningful, global hypothesis generation.

Network analysis approaches

Pathway analysis has an obvious limitation, as it relies on curated gene sets and therefore can only create information about genes that are assigned to one of those sets. This inherently prevents this analysis from gaining insights about novel relationships between genes and results in a bias towards well-studied pathways that are more comprehensively annotated. Network analysis is therefore a rapidly developing area of research, as it can create hypotheses on a systematic, unbiased level and predict novel interactions between proteins, gene regulatory mechanisms or biological processes. Specifically, in the identification of disease mechanisms or the creation and mining of protein-protein interactomes, network-based approaches haven been shown to be exceedingly powerful in the development of therapeutic targets [80]. Figure 7 shows the schematic workflow for the creation and analysis of a co-expression network.

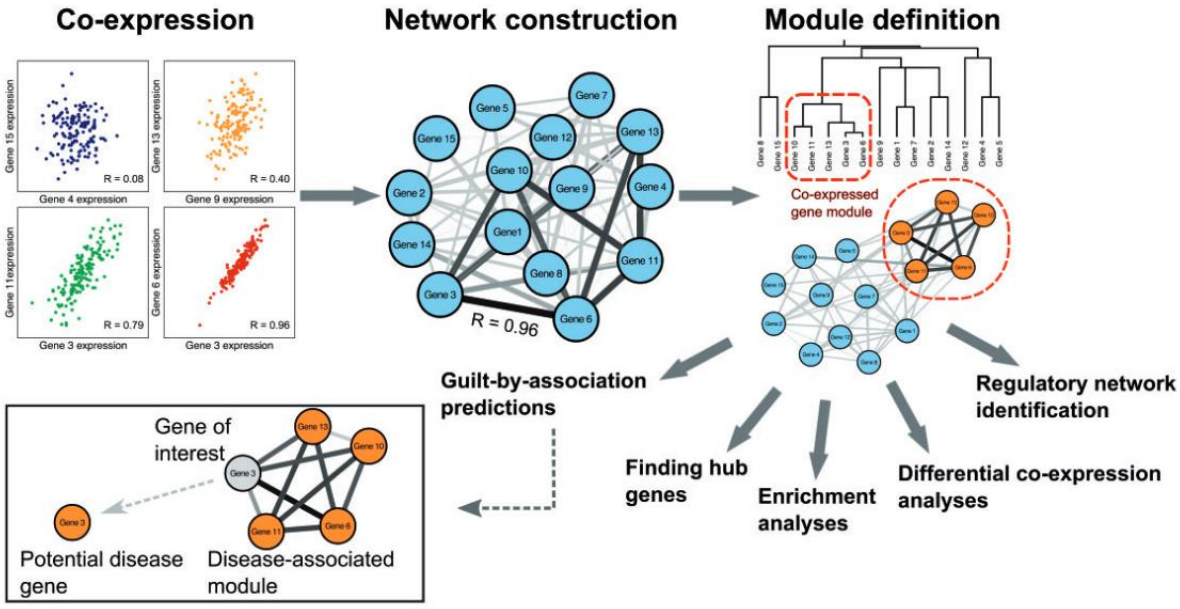


Figure 7 | Workflow of a network analysis. The figure shows a schematic workflow of a co-expression network analysis. In a first step an expression table is correlated and a network graph inferred. Next, modules in the network are identified that contain information about the relationship of the represented genes or gene products. From those modules the network can be characterized or information about novel interactors can be predicted by guilt by association, exemplarily shown for a disease-associated cluster of genes [81].

Co-expression network representations are usually based on pair-wise correlation to describe the similarity of a coordinated expression of genes across multiple samples. For similarity measures, a straightforward approach is to apply the Pearson correlation, or for a rank-based approach the Spearman's correlation. Although these approaches are generally applicably, their disadvantage is that every value has a non-zero correlation with every other expressed gene in the created network, leading to high redundancy. Generally, permutation-based filtering cutoffs are set to reduce said redundancy, which brings the risk to remove meaningful correlation just below the threshold [82]. An alternative is the use of least absolute error regression [83] or Bayesian networks [84].

The most widely used data representation are weighted, directed or undirected graphs created from a correlation matrix. A graph G is described as a paired list of nodes and edges $G = (V, E)$, where V is a set of nodes, and $E = \{(u, v) | \, u, v \in V\}$ the edges which connect them. In an undirected graph, the edge is usually described as a node pair of start and end note. The weight of this edge, which generally presents the information about the strength of the pairwise correlation, is either stored directly as a third parameter with the edge, considering object-oriented programming, or as a supplementary function, which maps the weight of the edge [85]. A widely used resource based on undirected graphs is the STRING protein-protein interaction network, which shows protein-protein interactions in both experimentally validated and computationally predicted modules [86] .

Those protein-protein interaction networks have been experimentally analyzed in yeast with two-hybrid assays. It was shown that proteins of similar function and cellular localization cluster together, and thus potentially novel functions of proteins could be assigned by associating them to close interaction partners [87]. This approach can be transferred from yeast experiments to the general functional enrichment of genes in a module and is commonly referred to as "guilt-by-association" [88]. Yet, this approach is

not without its limitations. The accurate identification of clusters or modules in networks depends on several important factors, from the quality of the measurement of the data and sample size, to the correlation, to the inference of the network and modules. As each of these steps is error-prone in its own way, association to a cluster is often susceptible to false-positive identification [89].

To remedy this issue, several graph-based clustering or module inference methods have been established to better group genes with similar expression patterns in multiple samples. Hierarchical clustering has long been used as a robust method for building a hierarchy of clusters. This clustering methods works either as a agglomerative, meaning as a “bottom-up” method considering the entire network as individual list of entries, where each observation starts its own clusters and adds fitting members to it, or as divisive, the “top-down” method, where the network first is considered as one cluster that is split into sub-clusters based on the observations. “Neighborhood joining” is a popular agglomerative approach [90], whereas divisive clustering could be combined with maximum likelihood estimation, which has proven to be a highly accurate, yet computationally expensive approach for hierarchical clustering [91]. The results of hierarchical clustering are commonly visualized in a tree-like, so called dendrogram structure. The R package weighted gene correlation network analysis (WGCNA) is the most widely used clustering tool for co-expression analysis and visualization [92].

Alternative to hierarchical clustering, density, partition or spectral based clustering has been established. Density based clustering methods, such as the Markov Clustering algorithm (MCL) are specifically designed for weighted graphs. This algorithm applies an iterative random walk that identifies clusters in the network based on the correlation of nodes. This algorithms requires dense networks, as sparse connectivity lead frequently to loops in a random walk algorithm [93].

Partition based clustering, such as the k-means algorithm, have long been established for gene expression analysis. Related to the divisive hierarchical clustering, they divide the graph into optimal partitions by clustering iteratively towards the nearest mean, hence creating *k* clusters from *n* observations. This method counts as supervised clustering as it requires a predefined end point, or number of final clusters [94].

As those algorithms are generally vulnerable to noise and higher dimensionality, which is common in gene expression data, semi-supervised spectral clustering (SSCC) has been developed. This semi-supervised approach combines an unsupervised first-step clustering (consensus clustering) with the graph generation method hybrid bipartite graph formulation (HBGF), and a spectral clustering approach that reduces dimensionality of the graph. This combinatory approach together with the dimension reduction increases robustness of the resulting clusters [95].

For the direct visualization, many of these methods have been integrated in Cytoscape, an open source modularized network analysis platform that both integrates clustering as well as direct network analysis methods and the visualization of the resulting modules [96].

Due to the previously mentioned limitations of guilt-by-association, direct network-based methods have recently gained traction. Direct methods have been shown more efficient to determine the function of individual genes or gene products instead of the overall structure of a biological network. This includes neighborhood, path algorithms as well as network topological analyses. Topological analyses haven been proven to be powerful for protein-protein networks. The degree of a node here determines the number of edges to other nodes, representing the amount of physical interaction partners. This can be used to determine nodes that play a significant role in the studied phenotype, as well as hubs - highly connected nodes with multi-interacting or regulating functions. Recently, pathway analysis of nodes with a higher degree enabled identification of causal genes associated with disease-loci in type 1 and 2 diabetes [97].

Shortest path analysis algorithms are used to determine and rank-direct interaction partners based on weighted networks. Dijkstra's-algorithm, the original shortest path algorithm with the lowest complexity can be applied to calculate the shortest distance between two gene or gene products of interest by a stepwise comparison of the weight of the connected edges. Dijkstra's shortest path algorithm has been successfully adapted for biological networks, such as the STRING database, by ranking nodes between two proteins of interested based on their betweenness, a metric that determines the amount of shortest path the node is part of. This approach has been used to identify colorectal cancer-related genes by a shortest path analysis between

genes identified by maximum relevance minimum redundancy (mRMR) [98]. While this shortest path algorithm has some limitations, such as gaps in the path due to error-prone data, false positive edges and the small world problem, emphasizing that in dense networks all nodes can be reach by all other nodes in a small number of steps, the method outperformed identification based on gene expression profiles alone. Bellmann-Ford, a further development of Dijkstra, can be used to find shortest path in potentially imperfect networks that contain missing nodes. It is a single-source shortest path algorithm that starts at a set entry point and can determine all shortest paths to all other nodes by iteratively comparing all combinations of edges and dismissing those leading to a longer distance. These methods can be employed to find missing interaction partners, such as for pathways of the osmotic stress response in *S. cerevisiae* [99].

Network propagation has been introduced as an approach that can overcome the limitations of module based or shortest path algorithms, as it considers all possible paths between nodes in the network at the same time. This weights down the impact of nodes that have a strong correlation with a low degree. Many different network propagation or score diffusion algorithms have been described over the years in different fields of research. Popular examples include heat diffusion or random walks with restarts, as variants to Google's PageRank algorithm for the ranking of websites. A set score is determined for each node in the network, being either empirical as 1 for known and 0 for unknown nodes, or a metric of choice. This information is propagated iteratively through the edges to nearby nodes, until the diffused score reaches a steady-state. For this scoring, either random walkers with a set restart probability at each step and a maximum target of steps, or heat diffusion models are used [100]. The resulting score is influenced by the degree of nodes, which has been shown to recognize hubs as well as to amplify lower signals in a network, where nodes are interconnected with each other [101]. Furthermore, this method recently was expanded to not only identify individual clusters or associate genes to pathways for function prediction, but rank subnetworks over a range of biological scales in highly mutated biological networks from The Cancer Genome Atlas (TCGA) to identify a range of novel candidate cancer genes [102].

In summary, there have been numerous advances in modern computational biology. From normalization to the diverse approaches to generate hypotheses from omics data, every field has received a significant number of novel tools and improved methods. The general workflow is still largely unchanged, starting with normalization, to the identification of differentially expressed genes or gene products, to data integration and functional analysis to generate knowledge from the data. The functional analysis consists of two different principles, resembling a “bottom-up” or “top-down” approach, which depends on the analysis aiming at the investigation of changes to the global proteome in response to a perturbation, or originates from the level of individual differentially regulated proteins to find interaction partners.

1.1.3 Integrative -omics landscape of cancer cells

Advances in all fields of big data -omics analyses have been enabled by the rapid development of various ways to distribute data to the scientific community and cross-reference related datasets. Over the years, various public databases have been established for storing and rapid integration of -omics datasets. For genomic and transcriptomic sequencing datasets, the gene expression Omnibus from NCBI has proved to be a valuable resource [103]. The database PRIDE [104] from the ProteomeXchange consortium was established to facilitate the easy publishing of proteomics datasets. The combined effort of public databases facilitates both the integration, as well as the analysis of related databases that were previously inaccessible due to a missing or unclear documentation of data. For example, the cancer genome atlas (TCGA) [105], curated by the U.S National Cancer Institute reported the genetic profiling of 10,000 different human tumors by a collaborative effort of 16 nations that resulted in identification of nearly 10 million cancer-related mutations already in 2015 [106]. While this achievement is remarkable, it has highlighted a higher complexity of the genomic heterogeneity in cancer than anticipated. Copy number alterations (CNA) of transcription factors have been analyzed together with expression changes in individual DNA repair genes and summarized at pathway level. This study lead to the novel insights in DNA repair dysregulation based on transcriptional changes, leading to genomic instability in breast cancer [107]. More recent studies have highlighted the capability of implementing deep-learning (DL) based models in multi-omics datasets. This has been shown in hepatocellular carcinoma (HCC) studies that integrated RNA seq, miRNA seq and methylation data from TCGA. The model is able to predict heterogeneity in HCC etiology, leading to better prognostication and improving risk-adapted therapy. This is the first study that robustly employed DL techniques to identify multi-omics features linked to differential survival of patients [108].

Furthermore, current studies of proteomic cancer datasets from the PRIDE database have shown a strong correlation of baseline gene expression across different tumor types and cell lines, indicating a robust comparability between proteomics sets of similar cellular origin. By integration of related mRNA data, it was shown that variation in mRNA levels alone is a poor predictor of changes to protein abundance [109]. As a first meta-analysis of proteomics datasets, this study highlights a step towards a broader annotation and analysis of public datasets. In the future, multi-omics studies

will likely likely gain more popularity thanks to the continuous growth of publicly accessible datasets of all omics types.

In summary, advances in high-throughput sequencing, MS and data acquisition have led to a rapid increase in the number of available datasets across all omics types. The integrative analysis thereof is gaining momentum, indicating a trend towards more investigative studies that focus on the interaction of multiple layers of -omics data in diverse biological system. Specifically for fields focused on highly complex mechanisms, such as the identification of driving forces behind cancer and genomic instability this will likely facilitate more novel insights.

1.2 Causes and consequences of impaired genome integrity

1.2.1 From impaired genome integrity to cancer

Genetic alterations, ranging from a single nucleotide mutation to whole chromosome rearrangements can predispose cells towards malignancy and uncontrolled autonomous expansion, commonly categorized as cancer. In an influential review of Hanahan and Weinberg, the authors organized the many complexities of how cancer clones achieve their detrimental behavior in a series of hallmarks: Self-sufficiency in growth signaling, insensitivity to anti-growth signals, evading apoptosis, limitless replicative potential, sustain angiogenesis and tissue invasion [110]. A decade later the same authors added two more hallmarks, modification of energy metabolism as well as evasion from the innate immune response, to this list. They concluded that this is mostly facilitated by two enabling traits, shared between different cancer types: inflammation, which promotes tumorigenesis and, even more prominently, genome instability [111].

Impaired maintenance of genome integrity leads to genomically unstable tumor cells. This facilitates a higher rate of mutations and genomic alterations, which in turn enable a selection of more aggressive clones with faster proliferation and advantages materializing in the described hallmarks. A key factor for the disruption of genome integrity is DNA damage. This damage can be categorized in two main classes: endogenous and exogenous. It is estimated that each human cell is a subject to approximately 70,000 endogenous DNA lesions daily as a consequence of reactive metabolites and hydrolysis, including single-strand DNA (ssDNA) breaks that can lead to more severe double-strand breaks (DSBs) [112]. Exogenous DNA damage, on the other hand, occurs when physical, chemical or environmental agents damage the DNA. This includes exposure to radiation, such as UV or ionizing radiation, alkylating or crosslinking agents [113].

Inactivation of DNA repair pathways or genotoxic stress induced from replication promotes genomic instability and drives the tumor development through an increased amount of spontaneous mutations [114]. Loss-of-function mutations in tumor suppressor or anti-oncogenes, that help in the prevention of unrestrained cellular growth and promote DNA repair and cell cycle checkpoint activation are frequently found in tumors. The activation of those genes counteracts gain-of-function mutations of so-called

oncogenes, which stimulate cell growth and division and induce replication stress. Loss-of-function of tumor suppressors, such as *TP53* and the homologous recombination (HR) repair genes *BRCA1/2* are strongly linked to an increased risk of prostate, ovarian, pancreatic and breast cancer [115-117]. In this cancer type gain-of-function mutations in oncogenes *ErbB2*, *PI3KCA*, *MYC*, and *CCND1* were shown to play a promoting role [118].

The most prominent tumor suppressor is *TP53*, also titled the “guardian of the genome”. It is responsible for a complex signal transduction network, commonly referred to as the p53 pathway. This pathway regulates diverse cellular responses to oncogenic stresses and maintains genomic integrity. It dynamically induces either apoptosis in case of severe, sustained stress signaling, or cell cycle arrest to facilitate DNA damage repair and maintain genomic stability in case of transient stress signals. The cellular response thereby is mediated by posttranslational modifications of two regulatory domains of p53 that determine its activity and stability: phosphorylation on Ser-15/20 induces cell cycle arrest, while on Ser-46 triggers apoptosis [119]. This protects the cell from propagating detrimental mutations and reduces the tumorigenic consequences and maintain genomic integrity. *TP53* restricts cell proliferation in response to a variety of stress signals, from replicative stress and DNA damage [120], to oxidative stress [121], to proteotoxic stress from mis-segregation in cells [122].

Studies of aggregated whole-genome sequencing data from 38 tumor types have shown different signatures for structural variations resulting from various mutational events in tumor suppressors, emphasizing how heterogenic genomic instability can restructure the genome [123]. A large proportion of tumors show chromosomal instability (CIN) as a drastic form of genomic instability, described as an increased rate of loss or gain of a whole or a part of chromosomes by mitotic errors or structural changes due to breakage and repair of DNA, as depicted in Figure 8 [124].

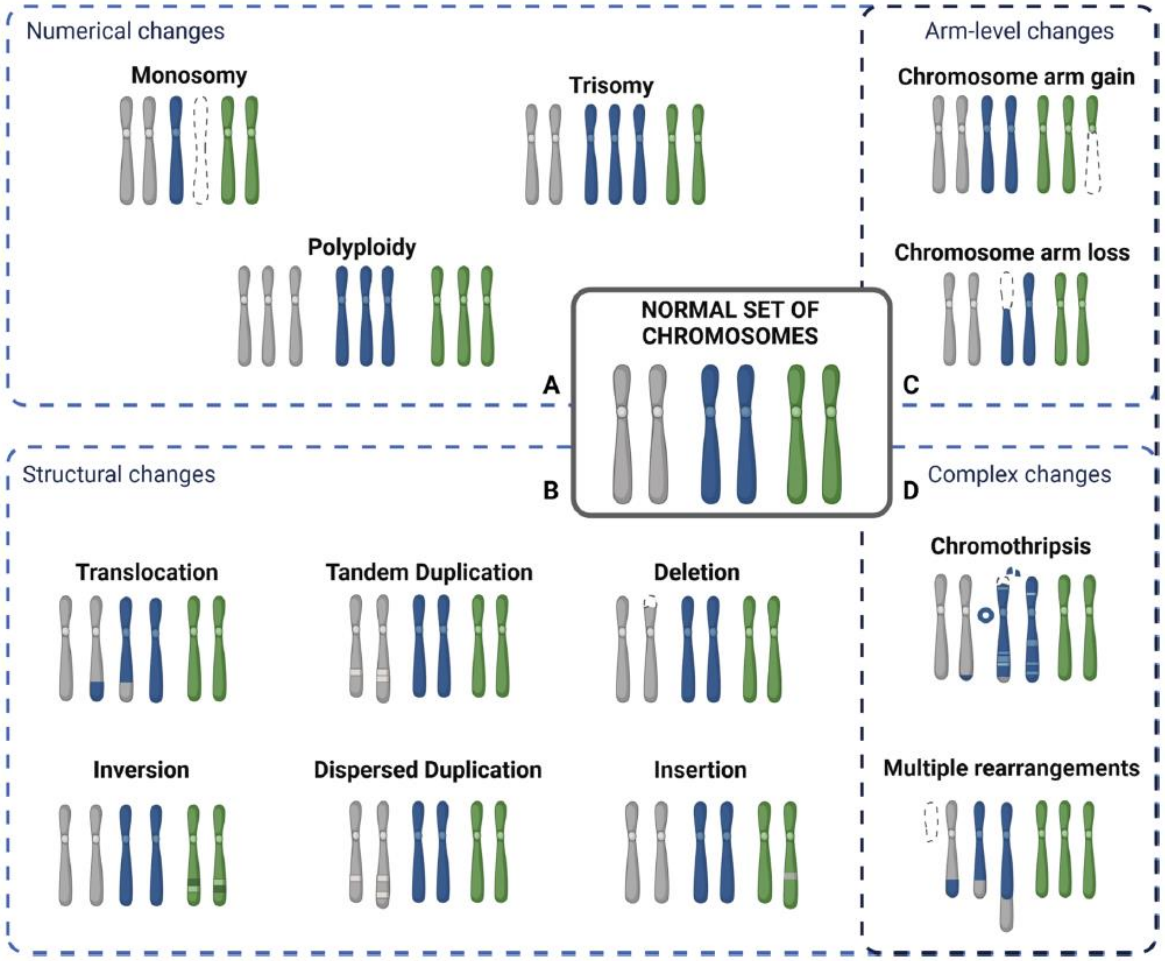


Figure 8 | Numerical and structural chromosomal changes. The figure shows the diverse numerical (top) and structural changes (bottom) to chromosomes. Numerical changes, either by mitotic or meiotic errors during chromosome segregation, can affect (A) whole chromosomes through, gain of chromosome (trisomy), loss of chromosome (monosomy) or gain of whole set of chromosomes (polyploidy). (C) Alternatively, they can arise as partial, or arm-level changes. (B) Structural changes affect part of the chromosome and lead to balanced or unbalanced karyotypes. (D) Complex changes are a combination of numerical and structural changes and include multiple rearrangements as well as chromothripsis, the catastrophic fragmentation and reassembly of genetic material [125].

While CIN is a commonly observed from of genomic instability and a characteristic in sporadic cancers, other forms genomic instabilities have been described in hereditary cancers. Microsatellite instability (MSI), a process in which the number of oligonucleotide repeats present in DNA sequences fluctuates, has been described in hereditary colon cancer [126]. Another example is an increased frequency of base-pair mutations that result in increased G•C to T•A transversion frequencies due to a mutation in the DNA glycosylase MYH [127].

In summary, maintenance of genomic integrity plays a central role in the prevention of cancer. Genomic instability, a general hallmark of cancer, is facilitated by impaired DNA damage response. It is suggested that oncogene-induced DNA replication stress is responsible for genomic instability in sporadic cancers. Its consequences include both numerical and structural chromosomal changes, often lead to unbalanced karyotypes, copy number changes and aneuploidy. Maintenance of genome integrity therefore is shown to be crucial in the prevention of uncontrolled cell growth, division and tumorigenesis and its consequences present an important field of research.

1.2.2 Maintaining genome integrity: DNA damage and response

Since the consequences of genomic instability are highly detrimental, cells have evolved an intricate toolbox of mechanisms to maintain genomic integrity during all phases of the cell cycle. Impaired genomic integrity leads to both numerical changes in chromosome or DNA copy number and structural instability. Whereas aneuploidy is the direct consequence of chromosome segregation errors, structural changes are caused largely by DNA damage and impaired DNA replication machinery, as shown in Figure 8 (top: numerical, bottom: structural changes). However, the causes and consequences of both structural and numerical changes are interconnected. Chromosome segregation errors can result in structural chromosome aberrations, such as translocations induced by chromosome breakage during cytokinesis, which triggers a double-strand-break response involving ATM, CHK2 and p53 [128]. Strikingly, DNA damage that occurs during mitosis induces errors that lead to mis-segregation of whole chromosomes. This was shown by inhibition of the DNA damage response (DDR) proteins ATM or CHK2, which abolished a stabilizing effect of kinetochore-microtubule attachments by Aurora-A and PLK1 kinases, which in turn induced the chromosome segregation errors [129]. Additionally, the tumor suppressor BRCA2 was shown to suppress replication stress by homologous recombination. This was shown by CRISPR-Cas9 induced depletion of BRCA2, that lead to an increase in replication stress, G1 arrest as well as 53BP1 nuclear body formation [130]. Defective DNA damage repair can lead to catastrophic events that result in complex genomic rearrangement affecting one or a few chromosomes, such as chromothripsis and chromoanasynthesis, which have been identified in numerous tumors. Those events occur as consequence of multiple double strand breaks, which lead to a fragmentation of a chromosome and re-ligation by error-prone repair processes. Chromothripsis triggers multiple genomic changes in one event, potently accelerating tumorigenesis by generating oncogenic changes, such as amplification of oncogenes, as was demonstrated in case of the MYC/MYCN [131].

Thus, given the far-reaching consequence of impaired maintenance of genome integrity for cell, the study of the various DNA repair mechanisms as well as the identification of involved proteins is essential for the understanding and treatment of cancer.

Various causes of DNA damage

The genomic integrity of a cell is continually challenged by a variety of endogenous and exogenous factors. Endogenous factors are normal cellular metabolism byproducts, including reactive oxygen (ROS) or nitrogen species (RNS), lipid peroxidation products and reactive metabolites that give rise to spontaneous DNA damage. This damage mostly leads to single strand breaks (SSB) and interstrand crosslinks (ICL) by oxidizing DNA bases, but also includes spontaneous reactions such as depyrimidination, depurination and deamination by hydrolysis. Similarly, errors in DNA replication caused by deoxyribonucleoside 5´-triphosphate (dNTP) misincorporation can potentially lead to mutagenic damages [132,133]. Thus, endogenous DNA damage results in ongoing potentially mutagenic burden in cells. This effect has to be minimized by efficient DNA repair processes to facilitate genome integrity.

DNA defects can arise also due to external damage. Exogenous physical sources are for example various forms of radiation. Ionizing radiation (IR) or the ultraviolet (UV) component of sunlight cause up to $1x10^5$ DNA lesions daily in each cell resulting in single or double strand breaks in the DNA helix, which may facilitate mutations and structural rearrangements when incorrectly repaired [134]. Chemical exogenous sources are highly variable. Many of these agents are commonly used in laboratories to induce DNA damage or arise spontaneously during metabolic cellular processes, such as methyl methanesulfonate (MMS), N-methyl-N′-nitro-N-nitrosoguanidine (MNNG) and methylnitrosourea (MNU) [135] or the highly genotoxic, DNA protein crosslink inducing (DPC) formaldehyde [136]. They are also found in cigarette smoke, and DNA adduct levels as well as various cancer types have been correlated with smoking [137]. DNA damaging agents are also used in chemotherapy to attach alkyl groups to the DNA or as crosslinking agents, inducing covalent intrastrand or interstrand crosslinks between bases of the DNA strand, such as cyclophosphamide, cis-platinum, melphalan or mitomycin C [138]. Unrepaired DNA damage from both endogenous and exogenous sources can lead to various disease phenotypes, ranging from cancer predispositions or impaired development phenotypes, to chromosomal and genomic instability or mutations, which induce premature aging. Additionally, some damage, such as intrastrand crosslinks or DSBs, stall replication forks, which leads to an inability to complete the replication of the chromosome and subsequently to cell death [114]. An

overview of different DNA lesions, as well as their sources and response mechanisms are shown in Figure 9.

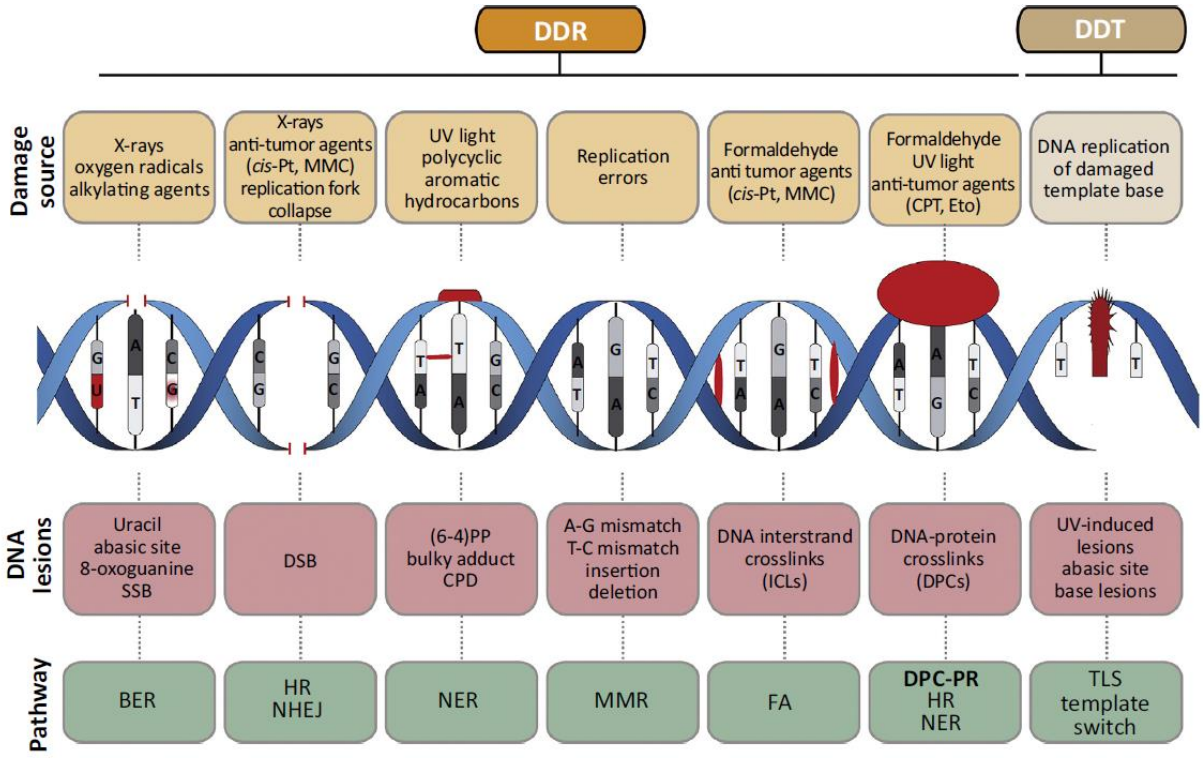


Figure 9 | DNA repair mechanisms maintain genomic stability. This figure shows an overview of different DNA lesions, from single and double strand breaks, to adducts, base mismatches, crosslinks to base lesions. Each DNA lesion triggers a defined DNA damage response (BER: Base excision repair; HR: Homologous recombination; NHEJ: Non-homologous end joining; NER Nucleotide Excision repair; MMR: Mismatch repair; FA: Fanconi Anemia; DPC-PR: DNA-protein Crosslink proteolysis repair; TLS: Translesion synthesis) with a set of key factors, that identify the lesion and facilitate the responses to repair or tolerate the DNA damage, as in the case of TLS (adapted from [139]).

Specialized DNA damage responses

To counteract the various DNA lesions, cells developed complex, interwoven DNA damage responses that are essential to maintain the genomic integrity. Recognition and signaling of DNA lesions plays an important role in the repair and replication of DNA. Key components of the DDR signaling are the protein kinases Ataxia telangiectasia and RAD3-related protein (ATR), which is activated by RPA coated single-strand DNA regions arising from the uncoupling between the MCM helicase and the DNA polymerase, and Ataxia-telangiectasia mutated kinase (ATM), which is mainly activated by the presence of DSBs. Facilitated by the recognition of phosphorylated

sites, the assembly of DNA repair complexes is promoted. The DDR additionally decides whether the cells undergo a cell cycle arrest or apoptosis [140,141]. Below, key DNA damage repair pathways are briefly summarized.

Less severe lesion that infer with the structural properties of DNA bases and their pairing properties, arising from the reaction to oxidative agents, alkylation or metabolic byproducts as well as single strand breaks can be repaired by different enzymes during base excision repair (BER) [142]. DNA glycosylases recognize and remove the damaged bases, which could otherwise lead to mutations by mispairing, or to DNA breaks. The injured strand is incised at the abasic site by endonucleases, resynthesized by polymerases and ligated by DNA ligase III. The process is coordinated by temporal and spatial protein-protein interactions of multiple repair factors with XRCC1 [143], while PARP1 and PARP2 stabilize the replication forks through FBH1-dependent regulation of RAD51 [144].

Bulkier single strand DNA lesions, such as formed UV light or environmental mutagens are repaired by the nucleotide excision repair (NER). This process can be split in two sub-pathways: Global genome NER (GG-NER), that can occur anywhere in the genome, and transcription-coupled NER (TC-NER), essential for an accelerated repair of lesions in the transcribed strand of DNA [145]. TC-NER is initiated by RNA polymerase stalled at a DNA lesion, facilitated by the factors CSA, CSB, and XAB2, whereas GG-NER detects disrupted base pairing and relies on recognition by XPC-RAD23B. The core of the repair pathway functions otherwise similarly, starting with a dual incision at the flanking regions of the lesion via the recruitment of ERCC1-XPF by interaction with XPA at the NER complex, followed by polymerase facilitated re-filling and sealing of the nick by DNA ligase III [146].

Mismatch repair deals with base pair mismatches and insertion/deletion loops that form during recombination of by polymerase errors and distort the helical structure. Those sites are recognized by the mismatch-recognition factors MUTSα and MUTSβ. After ATP-dependent recruitment of MUTL, an exonuclease-mediated degradation of the error containing region of the strand is initiated, followed by newly synthesized DNA encompassing at the mismatch site and resynthesis of the DNA [147,148].

Repair of double-strand-breaks, which are more severe lesions, rely on two major pathways. Non-homologous end joining (NHEJ) and homologous recombination (HR).

In the process of NHEJ, which occurs throughout the cell cycle, DSBs are recognized by the Ku70/80 heterodimer. It binds to the DNA and activates the protein kinase DNA-PKcs, DNA ligase IV and the polymerases μ and λ. This complex first removes damaged nucleotides at the ends of the DNA lesions, and in a second step the polymerases μ and λ fill in the gaps. After the end processing, the DNA ligase IV ligates the DSB ends together to finish the repair. However, this process can cause mutation or small DNA deletion around the DSB site and therefore acts mutagenically [149,150].

Homologous recombination is restricted to S- and G2-phase, as it requires a complementary sister-chromatid as a template. After the formation of a DSB, the DNA ends are resected by the MRE11-RAD50-NBS1 complex (MRN) to yield 3' single-strand DNA overhangs. Facilitated by RAD51 and the breast-cancer susceptibility proteins BRCA1 and BRCA2, the single-stranded DNA then invades the undamaged sister chromatin, where it anneals to its complementary strand. The damage is repaired via synthesis of new DNA, which uses the sister-chromatid sequence as template, followed by ligation [151,152].

Importantly, these "core" repair responses do not work alone. To repair complex DNA damages, such as DNA-protein crosslinks or interstrand crosslinks multiple repair mechanisms interact.

Interstrand crosslinks are highly toxic DNA lesions that prevent transcription and replication of DNA by covalently linking the DNA strands. Once a replication fork encounters an ICL during S phase, it is stalled and the resulting structure is recognized by the Fanconi Anemia Group M (FANCM) protein, which contains an interaction site for the rest of the Fanconi Anemia complex (FA) [153]. The DNA is cleaved on both sides of the ICL in a process called "unhooking" by the Fanconi-associated nuclease activity and the endonuclease XPF/ERCC1. This results in a DSB on one of the strands of the sister chromatids. This unhooked lesion can be bypassed by translesion DNA polymerases ζ (Pol ζ) and Rev1, which extend the 3' end past the remaining adduct [154]. RAD51 nucleoprotein together with other DNA repair factors promote the repair of the double-strand by homologous recombination. The leftover adduct is removed by NER. In a final step, the deubiquitylating enzyme ubiquitin-specific peptidase 1 (USP1) together with the cofactor UAF1 deubiquitinates the heterodimer FANCI-FANCD2,

which acted as a “clamp” for the DNA and stalled replication in a monoubiquitinated state [155].

DNA-protein crosslinks are poorly understood DNA lesions, that are created if DNA is covalently linked to proteins through radiation or chemical agents. Similarly, to ICL, they interfere with DNA replication, transcription and therefore impair genome integrity. Smaller DPCS, at less than 11 kDA, can be repaired in eukaryotes by nucleotide excision repair and homologous recombination [156]. It is not yet fully understood how larger lesions are removed. DPCs can be categorized in type I DPCs, which are linked to uninterrupted Duplex DNA, type II DPCs, which are flanked on one side by a single-strand break or as type III DPCs by a double strand break [157]. Initial studies in mammalian cells suggested that the proteasome plays a role in the removal or DPCs, as its inhibition prevented removal of topoisomerases and DNA Polβ as well as increased sensitivity to formaldehyde treatment [158]. Novel factors for this repair pathway have recently have been identified, as described below in more detail.

Genomic integrity is facilitated by a coordinated effort of DNA repair responses and cell cycle checkpoints, that safeguard the transition of the DNA through the cell cycle (reviewed in [159]). Yet, identification of new members of these complex mechanisms is demanding. Facilitated by the development of more powerful proteomics techniques, new approaches could be established that allow scientists to gain a better understanding of the repair of complex DNA lesion and the members of the involved DNA repair mechanisms.

Identification of DDR proteins for complex DNA lesions

Together with technical advances in mass spectrometry, several mass-spectrometry based methods to study and analyze the recruitment of factors involved in the DNA replication and repair were developed, such as chromatin mass spectrometry (CHROMASS) or plasmid-pulldown mass spectrometry (PP-MS). These approaches are often based on the egg extract system of the model organism *Xenopus laevis*. This system has offered several advantages for the study of DNA damage. Its replication and repair mechanisms are very similar to mammalian cells, and various extracts can

be prepared from *Xenopus* eggs, that cycle highly synchronously through S and M phase (Figure 10) [160].

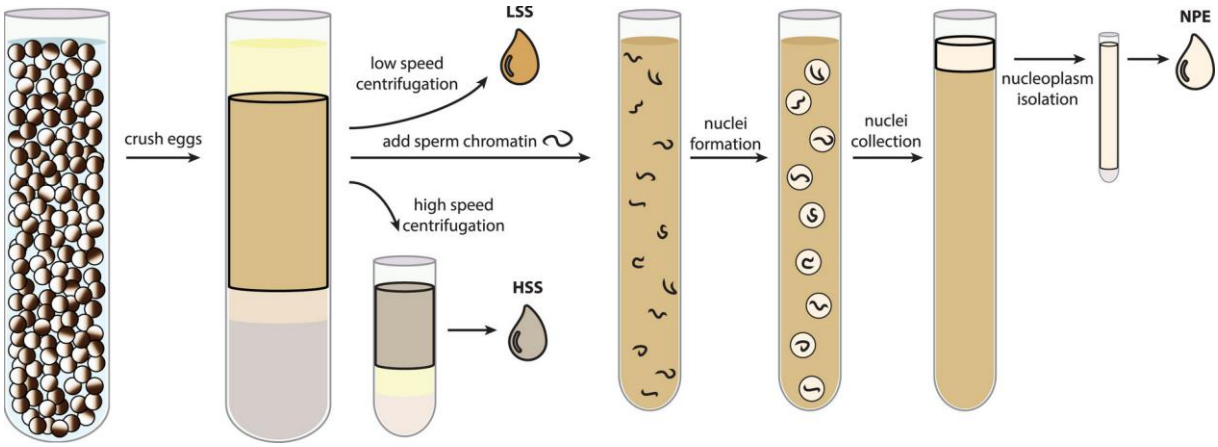


Figure 10 | Preparation of *Xenopus* egg extract. The schematic shows the preparation of diverse *Xenopus* egg extracts. After crushing unfertilized eggs, the cytoplasmic faction is collected. Membrane-containing low-speed supernatant (LSS) or membrane-free high-speed supernatant (HSS) can be collected through different centrifugation steps (100,000xg for LSS; 260,000xg for HSS). Nucleoplasmic egg extract can be collected after adding sperm chromatin to facilitate nuclei formation, that can trigger replication if added to the extracts [161].

This allows extensive study of the recruitment of proteins to DNA, carrying various induced lesions and the repair mechanisms of DNA damages at stalled replication forks. At specific time intervals, either sperm chromatin or a plasmid with an induced DNA lesion is isolated from the extract and the factors recruited to the DNA lesion are quantified by label-free mass spectrometry to provide an unbiased view on recruitment profiles of DNA repair factors in a time-resolved manner [162,163].

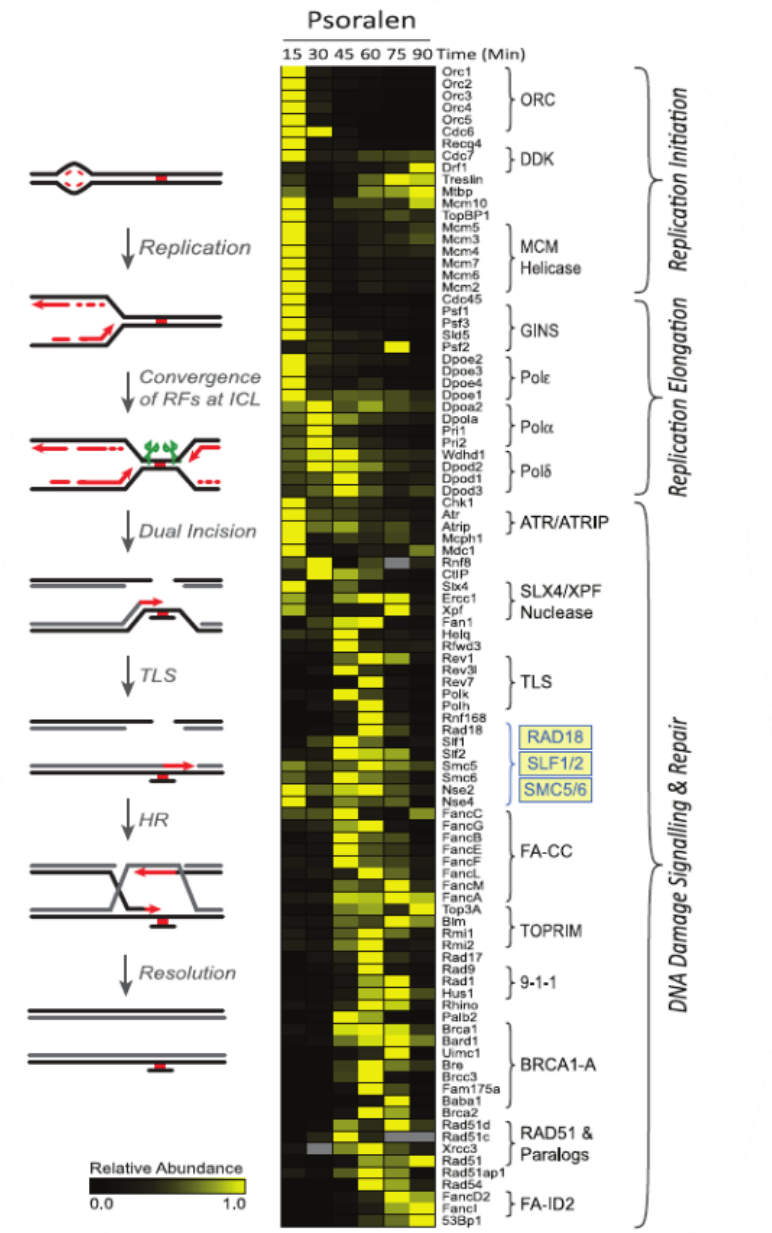


Figure 11 | Relative abundance of replication and repair factors at psoralen cross linked DNA. The figure shows a schematic of the steps involved in the repair of psoralen cross-linked DNA in *Xenopus* egg extract. The right side shows the recruitment profiles of the involved factors as relative abundance in comparison to an undamaged control (adapted from [162]).

CHROMASS, as shown in Figure 11, proved to be extremely efficient technique for unraveling new factors of complex DNA repair pathways, For example, using interstrand crosslinks as a model DNA lesion, it could be shown that SLF1/2 physically link RAD18 to the SMC5/6 complex, thereby targeting it to the damaged DNA [162]. SMC5/6 has a known function in the structural maintenance of chromosome (SMC) and genomic stability by regulating sister chromatid resolution and genomic location-dependent promotion or suppression of homologous recombination, and it's importance for ICL was unexpected [164]. CHROMASS can be used also for analysis of other DNA repair pathways, for example, in a more recent study, this method allowed identification of ETAA1, a novel ATR activator, by surveilling the recruitment of proteins to RPA coated single-strand DNA. ETAA1 binds and activates the ATR/ATRIP kinase responsible for DNA damage checkpoint pathway [165].

In further experiments in *Xenopus* egg extracts it was shown that DPCs encountered by the replisome are proteolytically degraded to short peptide adducts, what facilitated bypass by translesion synthesis polymerase complex REV1-Polζ [166]. Recent PP-MS studies identified the metalloprotease SPRTN, which is recruited with the proteasome to pDPC, as shown in Figure 12.

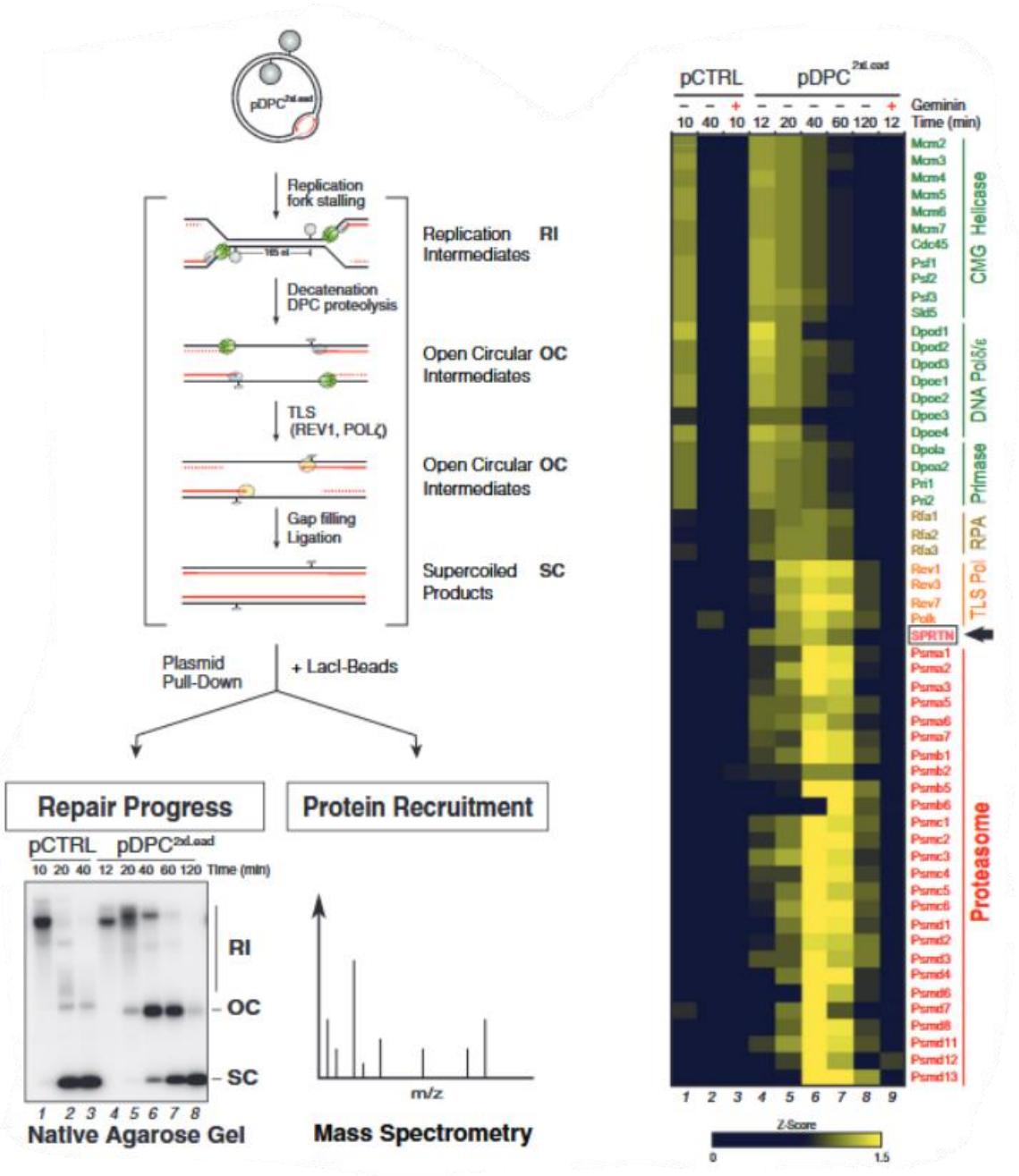


Figure 12 | Recruitment of SPRTN and the proteasome to a pDPC during replication. The figure shows a schematic of the replication and repair in the presence of a pDPC as well as the analysis by PP-MS. The replication intermediates were analyzed by agarose gel electrophoresis and bound proteins pulled down and analyzed by label-free quantified mass spectrometry. The right side shows relative intensities of the proteins recruited to the pDPC at multiple timepoints, including the proteasome and SPRTN, in comparison to an undamaged control (adapted from [163]).

SPRTN is the functionally similar homologue of the yeast protease WSS1, that was shown to remove trapped topoisomerase I complexes from the DNA and confers resistance to formaldehyde [167]. In humans, the loss of function of SPRTN causes Ruijs-Aalfs syndrome (RJALS), that is characterized by hepatocellular carcinoma and genetic instability [168]. SPRTN is regulated by multiple factors to safeguard from its potentially proteotoxic activity. It is deubiquitylated in response to a DPC, which facilitates its localization to the chromatin requiring the ubiquitin ligase activity of TRAIP [163]. Its proteolytic activity is stimulated upon DNA binding to degrade non-ubiquitylated DPCs until the DPC is degraded enough to be bypassed by TLS, then it degrades itself [169,170].

The identification of individual key factors for complex DNA repair processes shows the need for precision identification of function workflows. CHROMASS and PP-MS have been proven to be strikingly powerful for the investigation of DNA damage repair and the identification of novel factors that facilitate genome stability. State of the art proteomics provides an excellent opportunity to analyze cellular responses to defects in maintenance of genome integrity.

1.2.3 Aneuploidy as consequence of impaired genomic stability

Aneuploidy is often associated with chromosomal instability and commonly described as copy number changes of whole chromosomes or large chromosomal segments. While chromosomal instability, which facilitates formation of various aneuploid karyotypes, has been the focus of numerous studies in regards to targeted cancer therapies [171], the role of aneuploidy in tumorigenesis and how it affects eukaryotic cells is not fully understood. The challenge of aneuploidy research in cancer stems from its highly diverse consequences on a cellular level. Experimentally induced aneuploidy is detrimental for the development, proliferation and viability of the cell in most circumstances [172]. However, it appears to be well tolerated in cancers. In a large-scale study across over 10,500 cancer genomes from the TCGA it was shown that whole chromosome or chromosome arm-level changes are detectable in 88% of all investigated tumor types, ranging up to 99 % in glioblastoma. Further, it revealed a characteristic tumor-specific aneuploidy pattern, with different arms or whole chromosomes altered at different frequencies. This shows not only a tolerance of aneuploidy in cancer, but also indicates that aneuploidy is a promoting or facilitating factor of tumorigenesis [173]. This overall paradoxical effect to cell proliferation is referred to as “aneuploidy paradox” [174].

Causes of aneuploidy

The causes of whole chromosome aneuploidy are mostly chromosome segregation errors during mitosis or meiosis. The chromosome segregation machinery in humans is imperfect and the natural error rate is increased by chromosomal instability. *In vitro* it is estimated that diploid cells mis-segregate chromosomes once every hundred cell divisions, and the rate of chromosome gains or losses is increased by impaired cellular mechanisms leading to further chromosomal instability [175,176]. There are multiple ways how erroneous mis-segregation can lead to aneuploidy.

After chromosome duplication, the sister chromatids are held together by a multi-protein cohesion complex until their separation during anaphase. It consists of the subunits SCC1, SMC1 and SMC3, which form a tripartite ring structure around sister chromatids until bipolar spindle attachment to the kinetochores is completed [177]. The

shugoshin proteins SGO1 and SGO2 protect this complex from prematurely releasing the chromatids at the centromere by recruiting the phosphatase PP2A-B56 to the centromere, that antagonizes mitotic kinase activity [178]. Once the spindle checkpoint is silenced, the separase activation and a redistribution of SGO1 from centromeres to kinetochores [179] allows the cleavage of cohesin and the transition to anaphase can start [180]. Inactivating of the SGO1-mediated protection leads to premature sister-chromatid separation and SAC facilitated mitotic arrest. It was shown that this response is not robust enough to halt cell division fully, resulting in abnormal mitotic exit and aneuploidy (Figure 13 D) [181].

Further, erroneous kinetochore attachment can lead to aneuploidy. The kinetochore is a large proteinaceous complex assembled at the centromeric region of each chromosome [182]. Microtubules connect centrosomes to chromosomes via the kinetochore. Once the chromosomes are stably attached cell division progresses from metaphase to anaphase. Potential errors are recognized by the SAC, that holds anaphase until tense connections are established. Errors are asymmetric attachments, such as monotelic, where only a single kinetochore attaches to the microtubules, or syntelic and merotelic, with both kinetochores connected to the same or only one pole. The merotelic conformation creates a connection, which is under sufficient tension and is therefore not recognized by the SAC, and can therefore lead to lagging chromosomes and aneuploidy if the cell division continues [183] (Figure 13 A).

Supernumerary centrosomes that lead to multipolar mitosis, have also been shown as a cause of aneuploidy. Centrosomes are the microtubule organizing centers, which are duplicated during S-Phase. Centrosome amplification has been frequently observed in cancer, is known to promote tumorigenesis, and is associated with CIN [184]. The tumor suppressor *TP53* has been implicated in the duplication control and thereby helps to maintain the correct amount of centrosomes [185]. The presence of additional centrosomes can be mitigated by centrosome clustering and the formation of pseudo-bipolar spindles (Figure 13, B). This allows the cell to survive, but leads to a higher frequency in merotelic attachments of kinetochores and chromosome mis-segregation [186].

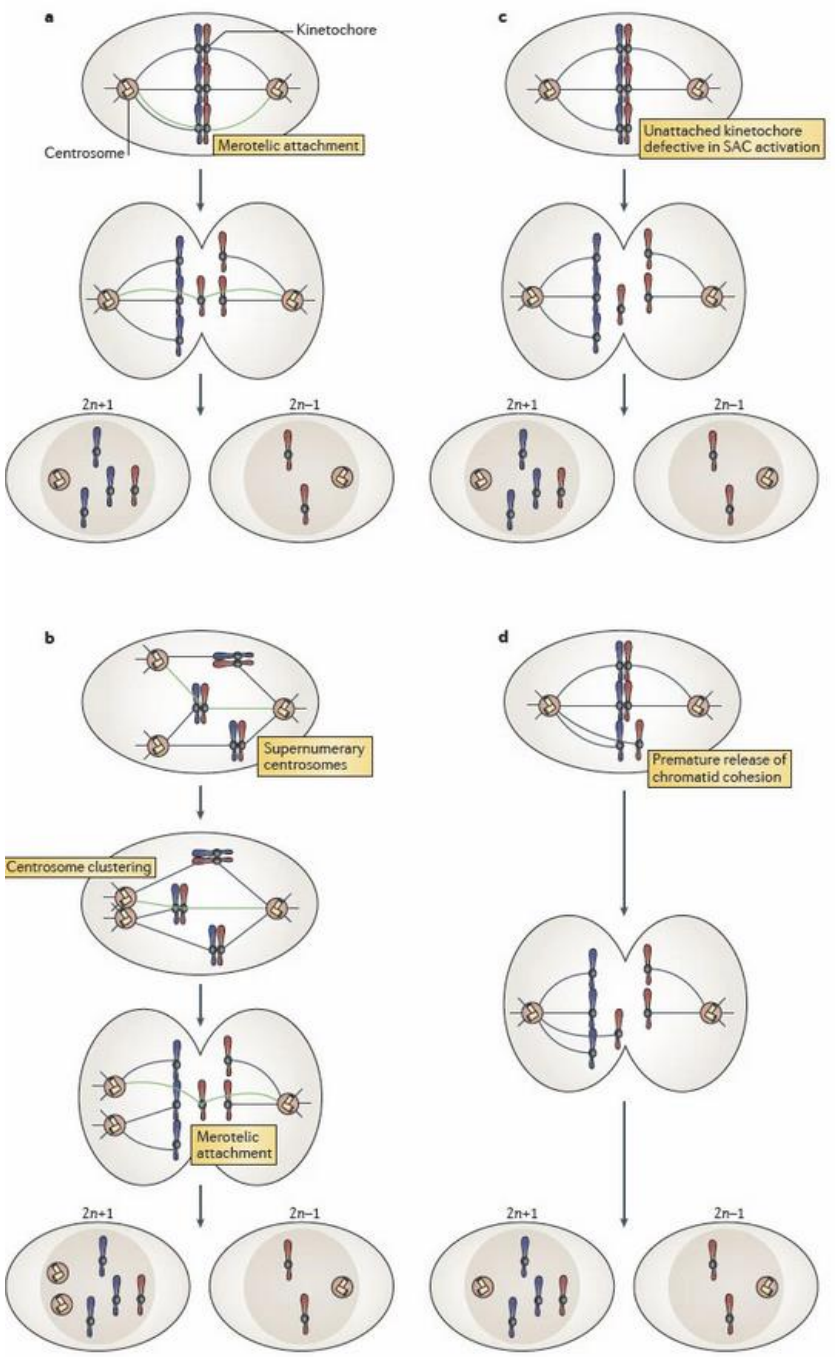


Figure 13 | Whole-chromosome aneuploidy. The figure shows mis-segregations that lead to aneuploidy. (A) Merotelic kinetochore attachment. A kinetochore is attached to microtubules of both poles. Left uncorrected chromatids mis-segregate towards the same pole in anaphase. (B) Supernumerary centrosomes lead to multipolar mitosis. By creation of a pseudo-bipolar spindle merotelic attachment are created that mis-segregate towards one pole. (C) A defective spindle assembly checkpoint does not recognize unattached kinetochores, that enter anaphase and end in the same daughter cell. (D) Impaired sister chromatid cohesion leads to mis-segregation before microtubules are attached to kinetochores (adapted from [187]).

A lack of an attachment, or insufficient tension on an attachment is recognized by the spindle assembly checkpoint. During mitosis this maintains genome stability by formation of the mitotic checkpoint complex (MCC), which is composed of BUBR1, BUB3, MAD2 and CDC20. This complex inhibits the transition to anaphase by inducing a “wait anaphase” signal that prevents the activation of the anaphase-promoting complex/cyclosome (APC/C) until the microtubule attachments are corrected [188]. In case of a defective SAC, the cell division proceeds to anaphase despite unattached kinetochores, leading to either apoptosis after premature mitotic exit or mis-segregation of chromosomes by mitotic slippage (Figure 13 C) [189].

In summary, the segregation of chromosomes during mitosis a tightly regulated process. Impairment of any segregation step can lead to mis-segregation and consequential, the loss or gain of a whole chromosome. The resulting aneuploidy karyotype can be observed frequently in tumors, and has highly diverse consequences.

Consequences of aneuploidy on gene expression

Whole chromosome aneuploidy is generally poorly tolerated in humans. Stable, non-cancer karyotypes, such as the trisomy syndromes of chromosome 13 (Patau -), 18 (Edwards -) and 21 (Down’s Syndrome) are associated with severe pathological consequences and lower viability [190,191].

Aneuploidy alters the relative dosage of genes on the affected chromosome, which leads to various cellular consequences. Studies that used chromosomal transfer strategies to generate aneuploid yeast strains and investigated the effect on aneuploidy on the transcriptome by gene expression mRNA microarrays have shown that the overall gene expression scales proportionally to the gene dosage [192]. Similar observations were drawn from engineered human, murine and patient-derived tissue cell lines [193-195]. This lead to the conclusion that there is no gene dosage compensation on the mRNA level, which would allow the aneuploid cell to minimize the effects of gene copy number changes from the aneuploid chromosome. In contrast, dosage compensation has been observed for additional copies of X chromosomes in human, that are in healthy women silenced by long non-coding RNA XIST [196]. Strikingly,

autosomal gene dosage compensation was observed by some research groups in naturally occurring aneuploid yeast strains, although re-evaluation of this data has shown again tight correlation between gene copy number and expression in both laboratory and wild strains [197,198]. Overall this highlights both that transcriptome levels scale with gene copy number but also that the observable effect presents considerable difficulties for the measurement and statistical evaluation of data.

With the advances in mass spectrometry also proteome levels of aneuploid cells could be investigated with higher accuracy, facilitating insights about the scaling of proteins with copy number. For example, in a landmark publication from Stingele et al. [79], DNA, mRNA and proteome levels of tri- and tetrasomic human colon tumor cells (HCT116) and non-cancerous cells (RPE1) have been analyzed. As expected, the presence of the additional chromosome revealed a significant adverse effect of extra chromosomes on cellular growth in the G1 and S phase. The comparison of the individual mRNA and proteome levels in this study furthermore revealed a lower abundancy than expected in ~25 % of proteins coded on the extra chromosome compared to the mRNA level, which increased according to chromosome copy number changes. Further studies in yeast have confirmed that while the levels of genes encoded on the extra chromosomes largely scale with the gene copy number, part of the protein do not. Enrichment analyses in both yeast and human aneuploidy cell lines have shown that the compensated part of the proteome is predominately associated with subunits of multi-molecular complexes as annotated in CORUM database [199,200].

In summary, aneuploidy alters gene copy number and thereby unbalances gene expression. Studies have shown that proteome levels do not exactly correspond to mRNA expression and increased copy number. While this effect could be linked to multi-molecular complexes the exact mechanisms responsible for this dosage compensation effect are still to be investigated.

Global consequences of aneuploidy

Overall, the increased gene dosage induced by aneuploidy strongly impacts the phenotype of the affected cell, both on the level of global gene expression, as well as by highly variable chromosome-specific effects and induces an “aneuploidy specific

stress response", as shown in Figure 14. Aside the mentioned downregulation of nucleic acid metabolism, shown in the finding of Stingele et al. [79] this aneuploid stress also involves specifically the downregulation of ribosomal subunits and biogenesis and cell cycle regulation together with an upregulation of lysosomal pathways, membrane metabolism, glycolysis and regulation of autophagy, as well as the response to type I interferons (IFN). Importantly, the data showed a strong downregulation of several pathways linked to RNA and DNA metabolism. Intriguingly, this global changes in pathway regulation were similar in cells with different abnormal karyotypes, indicating that aneuploidy triggers a general cellular response and might be used as a potential cancer treatment target. [201,202].

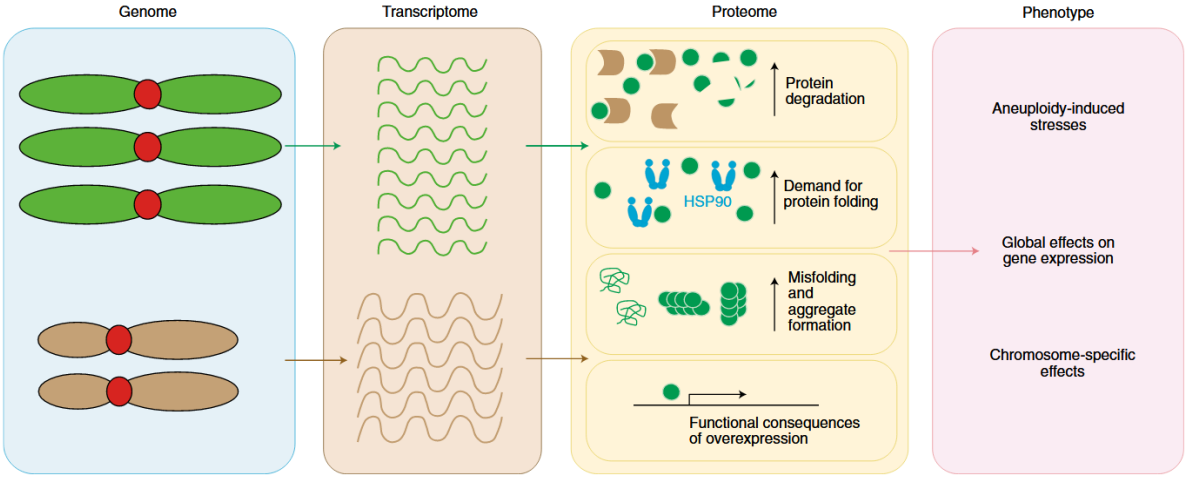


Figure 14 | Gene expression changes and consequences of aneuploidy. Additional gene dosage leads to higher amounts of expressed mRNA and protein. This leads to an unbalancing of the protein homeostasis and an increased demand for protein degradation, folding and the formation of misfolded aggregates if these systems are overwhelmed. Additionally, chromosome specific functional consequences are to be expected. Altogether this leads to a strongly affected phenotype, including a specific aneuploid stress response, as well as global effects on gene expression and effects specific for the genes expressed on the aneuploid chromosome (adapted from Rev. [203]).

A higher gene copy number leads to more expressed mRNA and proteins of genes located on the aneuploidy chromosome. This unbalancing of gene expression is facilitated by alterations in protein stoichiometry. This also leads to the overburdening of cellular mechanisms that maintain protein homeostasis and folding. This is critical, as the function of multimolecular complexes relies on a tightly balanced expression of the individual protein components [204]. Unbalancing of this homeostasis varies strongly depending on the affected complex or aggregated proteins. For example, the gain of

chromosome 6 in yeast cells, which carries beta-tubulin-coding genes, leads to lethality. Yet this phenotype can be rescued by a gain of chromosome 13, which contains the gene encoding for alpha-tubulin, hence restoring the stoichiometry of the alpha/beta-tubulin dimer [205]. By carrying one or more additional chromosomes, an overproduction of proteins overwhelms the chaperone systems needed for protein folding, as well as the degradation systems that in turn remove the misfolded protein aggregates [206]. These protein aggregates accumulate in the cytoplasm due to reduced heat shock protein (HSP) folding capacity. Unlike euploid cells, aneuploid cells show no change in HSP90 and HSP70 expression in response to heat shocks, likely due to already existing proteotoxic stress [207]. In agreement with this, aneuploid cells are sensitive to HSP90 inhibition. Additionally, overexpression of the transcription factor heat shock factor 1 (HSF1), which regulates the expression of heat shock proteins rescues the defects in protein folding [208]. Aneuploid cells in yeast with extra chromosomes 4, 12, 13 and 16 were also shown to be sensitive to inhibition of the proteasome inhibitor MG132. Yeast disomic strains are also sensitive to the presence of a HSP90 inhibitor and to higher temperature, and show an increased amount of protein aggregates and altered kinetics of heat adaptation [192]. Altogether these findings suggest the presence of misfolded proteins due to unbalanced gene expression as consequence of aneuploidy, that overwhelm the folding capacity and therefore are detrimental to cell viability.

A loss of function mutation in the deubiquitinating enzyme UBP6, regulating proteasome-mediated degradation and ubiquitin recycling was found in a bottom-up study that screened for fitness-increasing mutations in aneuploid yeast [209]. This enzyme allows substrates of the proteasome to escape degradation by cleaving the ubiquitin recognition motif [210]. Loss of the deubiquitinase function has a strong negative effect on the viability of aneuploid cells. A loss of the conserved human homolog of UBP3, USP10 is also detrimental for the fitness of human cells upon chromosome mis-segregation. This effect is specifically accompanied by autophagy inhibition [211].

Lysosome-mediated autophagy was detected from a global analysis of human aneuploids [79] and accumulation of the autophagy marker LC3 was observed immediately after chromosome mis-segregation in mammalian cells. This study revealed that aggregated proteins were encapsulated within autophagosomes, yet not subjected to lysosomal degradation. Strikingly, lysosomal stress does not manifest

immediately after the aneuploidy inducing mis-segregation but requires 2-3 cell divisions. This indicates that aneuploidy continuously burdens the degradative capacity of cells and induces a lysosome stress response to facilitate the production of further autophagosomes and lysosomes [212]. Considering the upregulation of lysosomal pathways in constitutive aneuploid cells, highlighted in the paper from Stingele et al. [79], this suggests that aneuploid cells can adapt to impaired protein homeostasis by activation of autophagy.

DNA replication was another pathway identified to be globally affected by aneuploidy. Accurate eukaryotic DNA replication requires precise stoichiometric balance of protein complexes. The replication of eukaryotic DNA starts from multiple genomic sites, or origins of replication (ORI), in a process involving the licensing and firing of the ORI. This relies on an origin recognition complex (ORC), which together with the proteins CDC6 and CDT1 load the replicative helicase consisting of the proteins MCM2-7. After transition to the S-Phase, the replication forks form bidirectionally, originating from the origin recognition complex [213].

Aneuploidy, which impairs protein stoichiometry, can interfere with the functionality of those complexes and lead to replication stress. This facilitates slowing or stalling of replication forks, which potentially allows the repair or breaking of the DNA double-strand. Thereby, replication stress delays or arrests the cell cycle. Induction of aneuploidy in human retinal pigment epithelial (RPE-1) cells by an MPS1 inhibitor has shown replication fork stalling and reduction in replication fork rate [214]. Similarly, slower replication rates were observed in human triploid and tetraploid RPE1 and HCT116 cell lines, in which the replication stress and DNA damage was also increased and resulted in novel chromosomal rearrangements and overall genomic instability [215]. Replication defects can be explained by an imbalance in the replication helicase subunits MCM2-7, which is consistent with the general downregulation of proteins involved in DNA replication found by Stingele et al. as consequence of aneuploidy.

It is not yet fully understood what causes this reduced expression. Speculatively, the impaired protein folding, degradation and proteotoxic stress in aneuploid cells could lead to a deregulation of replication protein expression, which was observed following the inhibition of HSF1 [208]. Also, a p53-mediated transcriptional cell cycle repression includes proliferation factors and the helicase subunits [216]. Generally, an increase in

DNA damage, specifically double-strand breaks or an increased mutational load accompanies replication stress in aneuploid cells, likely due to the elongated exposure of single-stranded DNA due to stalled or slowed replication fork. In budding yeast this increase of DNA damage was demonstrated by the accumulation of RAD52-GFP foci during S-Phase, indicating the replication defects as a cause [217].

In agreement with these findings, similar 53BP1 foci accumulation was also shown in human aneuploid RPE-1 cells [214]. Furthermore, next generation sequencing and SNP analysis of human trisomic and tetrasomic cells revealed break point junctions suggestive of replication defects. Together with the reduced expression of MCM helicase levels in response to chromosome gain, this highlights how gain of chromosome causes genomic instability due to replication stress and likely promotes to tumorigenesis [215].

Aneuploid cells affected by replication stress suffer from multiple different phenotypes next to stalling or slowing of replication forks. For example, aneuploid human pluripotent stem cells (hPSCs) that undergo replication stress have shown defective chromosome condensation and segregation [218]. The replication stress is also induced at telomeres in murine and human aneuploid cells, leading to telomeric DNA damage and p53 activation, and this lead to premature senescence and hematopoietic cell depletion [219]. Additionally, an inflammatory response due to DNA damage and genomic instability can be triggered in response to increased levels of cytoplasmic DNA in aneuploids [220], leading to an activation of type I interferon through the cGAS-STING pathway of innate immunity [221].

In summary, aneuploidy has a highly diverse set of consequences. First, it affects gene copy number and therefore gene expression. This effect can not be fully compensated on protein level by the cell, leading to an unbalanced protein expression. Aneuploidy further leads to overburdening of the cellular protein folding and degradation machinery, proteotoxic and replication stress as well as an increase in genomic instability. Multiple chromosome specific consequences, as well as a general “aneuploidy specific stress response” could be identified. Therefore, aneuploidy presents a potentially attractive target for cancer treatment. Yet, most of the current knowledge about aneuploidy is derived from gain of chromosome cell lines.

Monosomy

Next to the gain of one or more chromosomes, there are also other observed aneuploid karyotypes commonly occurring in cancer and likely playing a role in tumorigenesis. Monosomy, the loss of a chromosome or chromosomal arm is highly detrimental as embryos die during early stages of embryonic development. Exception is the Turner syndrome with a loss of the chromosome X in females, which is better tolerated due to the inherent silencing of one copy of the two sex-chromosome X [222]. Microdeletions, a copy number variant (CNV) affecting less than 5 mb, which can be diagnostically detected by modern karyotype analysis and fluorescence in situ hybridization (FISH), are viable, yet associated with syndromic forms of intellectual disability and developmental delay [223] such as 15q11-q13 deletion (Prader-Willi and Angelman syndrome [224]) or 17p11 deletion (Smith-Magenis syndrome [225]). Little is known about the specific consequences of chromosome loss, mostly due to the lack of a defined model system to study.

It is hypothesized that the lethality is caused by haploinsufficiency, where a single copy of a gene is not capable of supporting the wild type phenotype. Additionally, loss of the second gene copy may lead to the unmasking of recessive mutations. A recessive mutation is masked in the presence of a second, functional allele on the partner chromosome. In case of biallelic mutation, or in the case of monosomy, this previously masked phenotype is expressed [226-228]. Similar to the previously described gain-of-chromosome, the loss of a chromosome is paradoxically also frequently observed in cancer cells. Both whole chromosome as well as losses of chromosomal arms have been detected in large scale TCGA analyses across numerous tumors, highlighting specific aneuploidy patterns similar to those of gain-of-chromosomes and likely representing a tumor promoting factor [173,229].

This can be potentially explained by haploinsufficiency in tumor suppressors, which is observed in the frequent deletions of the p-arm of chromosome 17 in various tumors that is associated with loss of function of *TP53*. In mice with an artificial deletion of the corresponding chromosome, several other tumor suppressor genes have been identified that contributed to more aggressive lymphoma and leukemia development than *TP53* deletion [230]. In a recent study in our lab, the consequences of the loss of chromosome in somatic human cells have been further investigated, and stable

monosomic cell lines have been established [231]. This study will be explored in detail in Results Chapter 2.

In summary, compared to gain of chromosome the loss of chromosome aneuploidy is a lot less understood. This is due to the lack of generally usable model system and the highly detrimental phenotype of chromosome loss. Likely this strong phenotype is caused by haploinsufficiency of genes or unmasked genetic mutations. Paradoxically, monosomy is frequently observed in cancer cells.

Whole genome doubling

Opposed to the loss and gain of individual chromosomes, polyploidy (>2N) occurs commonly in different eukaryotic organisms. Here polyploidy plays an important role in evolutionary speciation [232]. Additionally, it can facilitate resistance against environmental stresses that otherwise would not be tolerated by diploids and increase the adaptive potential on the cost of potentially disrupting effects. This is facilitated by nuclear and cell enlargement and higher levels of epigenetic and genomic instability [233-235]. On the other hand polyploid cell divisions can also lead to aneuploidy [236] and Whole genome doubling (WGD) is commonly observed in human cancer cells. Similar to other types of chromosomal aberrations, polyploidy in tumors can occur from mitotic defects. TCGA analyses that characterized somatic copy number alterations (SCNA) in nearly 5,000 cancers identified WGD in 37 % of them. It has been shown that while polyploidy generally impairs genome stability, it also increases adaptability of the cell. Hence, it is considered a driving force for evolution and tumorigenesis [237,238]. By sequencing data from around 10,000 primary human cancers it was shown that WGD facilitates both tumor-promoting genetic traits, but also common vulnerabilities that could serve as potential cancer therapeutic targets [239]. For example, KIF18A, which encodes a mitotic kinesin protein, was described as a potentially interesting therapeutic target as its loss specifically decreased viability of cells that underwent whole genome doubling by inducing mitotic errors [240].

Using yeast as a model, it was calculated that a consequence of increased ploidy is a linear scaling of cell and nuclear volume together with a non-linear scaling of two-dimensional structures (1.58 fold) and linear structures (1.26 fold). It was shown that

this is accompanied by slightly reduced growth rate, aberrant cell cycle regulation and response to nutrition, decrease in fitness and impaired genome stability [234,241-244]. Transcriptome analysis in yeast highlighted that only a few genes change in response to increasing ploidy, whereas the majority stays in balance with gene dosage. The few deregulated genes encode structural proteins for membrane and cell wall, indicating that despite an overall maintenance of homeostasis of gene expression with ploidy, the cell may need to adapt to lower surface-to-volume ratios in larger polyploids [245]. In a recent study in our lab, proteome of yeast cells of different ploidy was analyzed for the first time, to investigate scaling of proteome content with ploidy and ploidy-dependent regulation that occurs post-translationally [246]. This study will be explored in detail in Results Chapter 3.

In summary, whole chromosome aneuploidy is a multi-faceted state poorly tolerated in somatic cells, that arises from errors in chromosome segregation. A gain of chromosome is commonly associated with pathological syndromes and on cellular level induces a broad array of responses, from replication- to proteotoxic stress to global and chromosome specific gene expression changes and genomic and chromosomal instability. Strikingly, many types of aneuploidy as well as whole genome doubling events are commonly found in different tumors where slower proliferation rates and selection for impaired tumor suppressors drive tumorigenesis and lead to more aggressive clones. This paradoxical state of the cell presents an attractive field of research to understand its still unclear contribution to cancer.

2. Aims of this study

Maintenance of genome integrity is crucial to ensure faithful replication of the genome in each cell of the body. Keeping the genome intact requires a concentrated action of cellular metabolism, cell cycle and DNA damage response. The consequences of impaired genome integrity are highly diverse. While this can lead to beneficial mutations that are an evolutionary drive, there are also various detrimental phenotypes. Genomically unstable cells are often predisposed for malignant growth and tumorigenesis, where genomic instability turns into a driving factor through the selection of more aggressive clones. Aneuploidy and polyploidy are both poorly tolerated in somatic cells, but frequently observed hallmarks of cancer that represent attractive fields of study for potential cancer treatments. This study presents multi-omics analysis as a versatile tool to investigate the various causes and consequences of impaired genome integrity by analysis of three different datasets.

1. To uncover potential causes of impaired genome integrity, I focused on an analysis of a collection of DNA repair proteomics experiments and describe the creation of a high-fidelity dataset for the identification and characterization of different DNA damage responses. To this end, I created the DNA Repair Atlas, a web-application that allows scientists without a computational background to interact, mine and visualize this dataset.
2. To uncover how chromosome loss affects the balance of gene expression in human cells, I analyzed an integrative omics dataset of TMT labeled mass spectrometry and sequencing data. In this analysis, I investigate how consequences of chromosome loss manifest on a cellular level and compare them to chromosome gain. I will also identify the effect of chromosome loss on the global expression of genes by pathway analysis.
3. Whole genome duplication is another detrimental outcome of failed cell division. Using a multi-omics analysis of SILAC labeled mass spectrometry and transcriptomics data, I will evaluate a ploidy series of yeast cells from haploid to tetraploid. In this analysis I will investigate how the proteome of a cell scales with increasing ploidy. Further, I will identify how genes are differentially expressed in response to this scaling with increasing ploidy.

Altogether, this study intends to show how powerful state-of-the-art multi-omics analysis is to uncover cellular responses to impaired genome integrity in a highly diverse field of research.

3.1 The DNA Repair Atlas, a web resource for mining and visualization of proteomics data

3.1.1 Mass spectrometry data collection and combined analysis

To investigate the cellular responses to different DNA lesions, we collected proteomics data from *in vitro* DNA repair experiments performed in *Xenopus laevis* egg extracts (Introduction Figure 10) in a combined dataset. The data originated from a total of 35 chromatin mass spectrometry (CHROMASS) or plasmid-pulldown mass spectrometry (PP-MS) DNA repair experiment series evaluating 60 different treatments with DNA damaging agents or inhibitors. In these approaches, sperm chromatin or a plasmid with an induced DNA lesion is isolated from *Xenopus* egg extract at specific time intervals and the factors recruited to the DNA lesion are quantified by label-free mass spectrometry. The unique experimental framework of CHROMASS and PP-MS is based on the evaluation of cell extracts under overall conditions of mild physiological stress in multiple measured timepoints. This allows for a highly robust temporal profiling and identification of DNA repair factors recruited to stalled replication forks or double-strand breaks [162,163,247] (Figure 11 & 12). A CHROMASS workflow highlighting the data acquisition strategy is shown schematically in Figure 15.

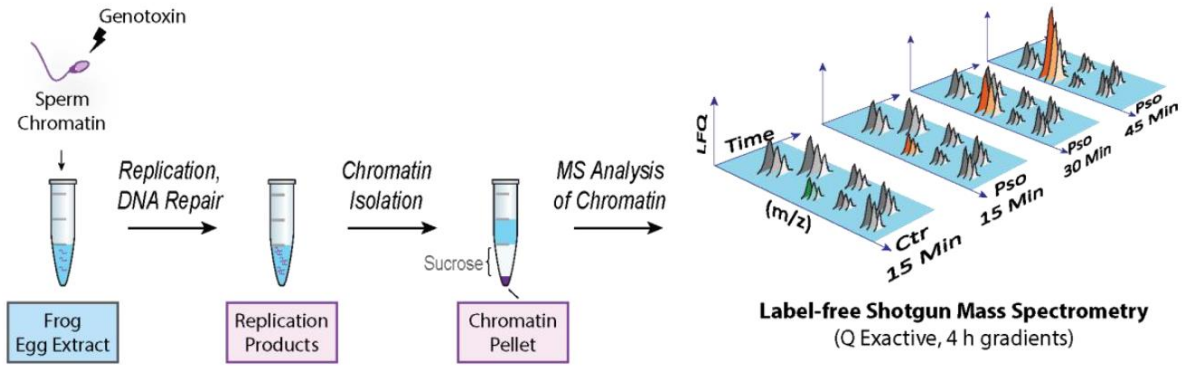


Figure 15 | Schematic of the data generation of the DNA repair experiments. The schematic highlights the workflow of a CHROMASS experiment. Sperm chromatin or plasmid substrates are incubated in frog egg extracts. DNA is recovered under mild conditions, the bound proteins digested with trypsin and quantified using label free mass spectrometry at multiple timepoints (with Markus Räschle).

The combined dataset includes time-resolved recruitment profiles of DNA repair proteins for different lesions of both published and unpublished data, such as interstrand crosslinks (ICL), double-strand breaks (DSB), DNA-protein crosslinks (DPC) and replication fork collapse (FC). A full list of included experiment series can be found in Supplementary Table 1. The collected experiments include a wide range of treatments, including presence and absence of various inhibitors, such as geminin to restrict replication or ATRi to inhibit the DNA damage signaling, and were performed in at least three replicates with matched, appropriate controls. All treatments and the corresponding abbreviations can be found in Supplementary Table 2. Figure 16 gives an overview about all measurements included in the combined dataset.

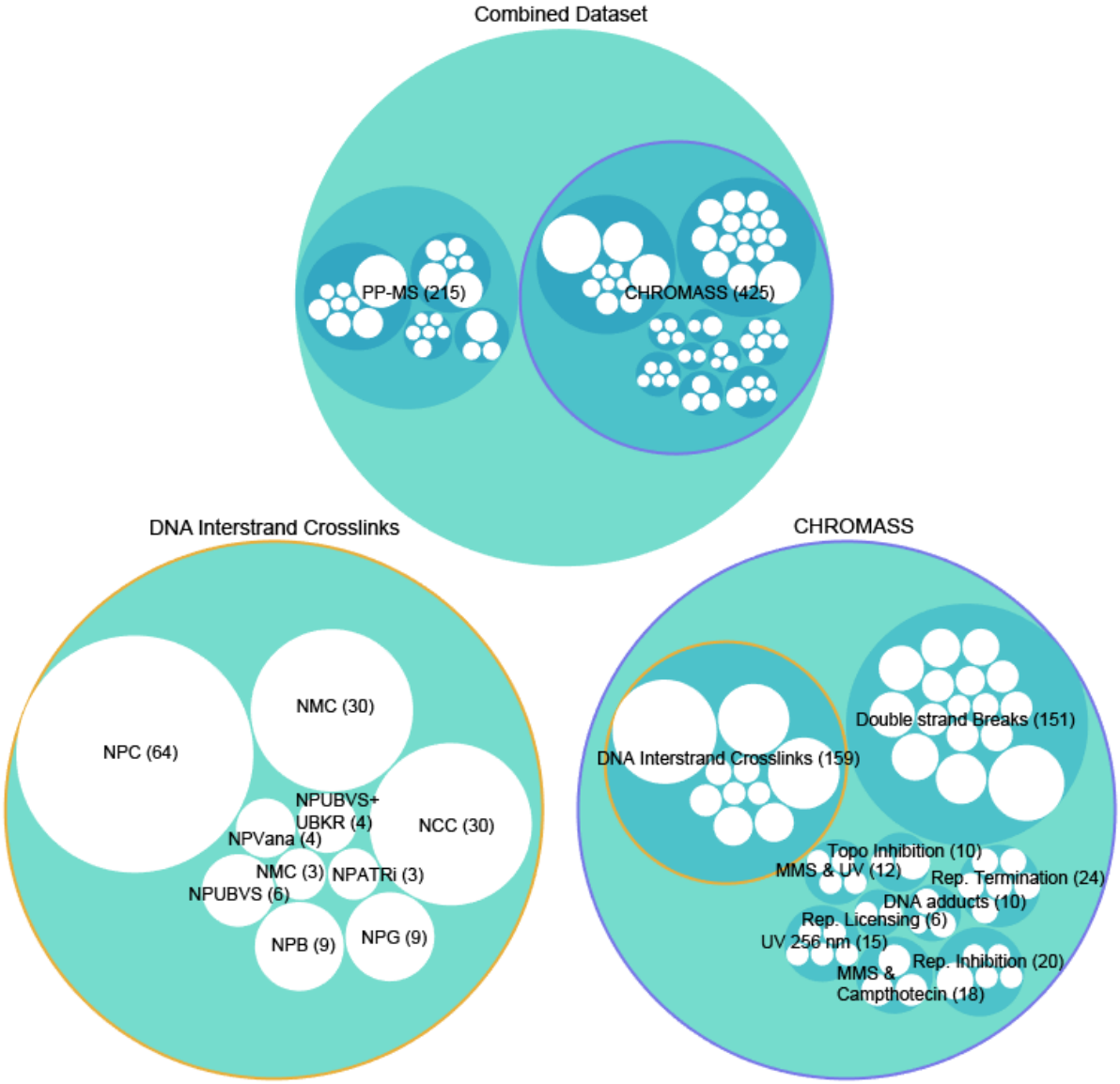


Figure 16 | Overview of the combined dataset. The hierarchical circle plot shows all collected measurements grouped by experimental method (top), experiment type (bottom right) and treatment (bottom left). Each circle shows a layer in the hierarchy. An excerpt for the CHROMASS experiment series for DNA interstrand crosslinks is shown, which is indicated by the colored outline of the circles. The size of a circle correlates with the number of related measurements, which is shown in brackets. Abbreviations: **Rep**, Replication. **MMS**, Methylmethane sulfonate. **Topo**: Topoisomerase. **HSS/NPE**: High-speed supernatant/ Nucleoplasmic extract. **NPC**: HSS/NPE - Psoralen-treated Chromatin. **NMC**: HSS/NPE - Mock (no Chromatin or DNA). **NCC**: HSS/NPE - Untreated treated Chromatin. **NPG**: HSS/NPE - Psoralen-treated Chromatin + Geminin. **NPB**: HSS/NPE - Psoralen-treated Chromatin + BRC4. **NPUBVS**: HSS/NPE - Psoralen-treated Chromatin + UB-VS. NPVANA: HSS/NPE - Psoralen-treated Chromatin + Vanadate. **NPUBVS**+**UBKR**: HSS/NPE - Psoralen-treated Chromatin + UBVS + Ubiquitin K63R. **NPATRi**: HSS/NPE - Psoralen-treated Chromatin + ATRi.

All files from in total 664 mass spectrometry measurements were processed together in a single run in MaxQuant (1.6.17) [21] on a Hyper-V server cluster, as described in the Methods. This resulted in a set of 5790 protein groups for further analysis.

Due to the dataset resulting from a high number of experiments, treatments and two different data acquisition methods, it is highly complex. Therefore, to give an overview about the investigated data, the t-distributed stochastic neighbor embedding (t-SNE) algorithm was used for a visualization in a low-dimensional space [248]. To facilitate this, individual experiments were shifted to a median of 0 and missing values were imputed from a normal distribution separately for each measurement, with a downshift of 1.8 standard deviations and a width of 0.3 (Figure 17).

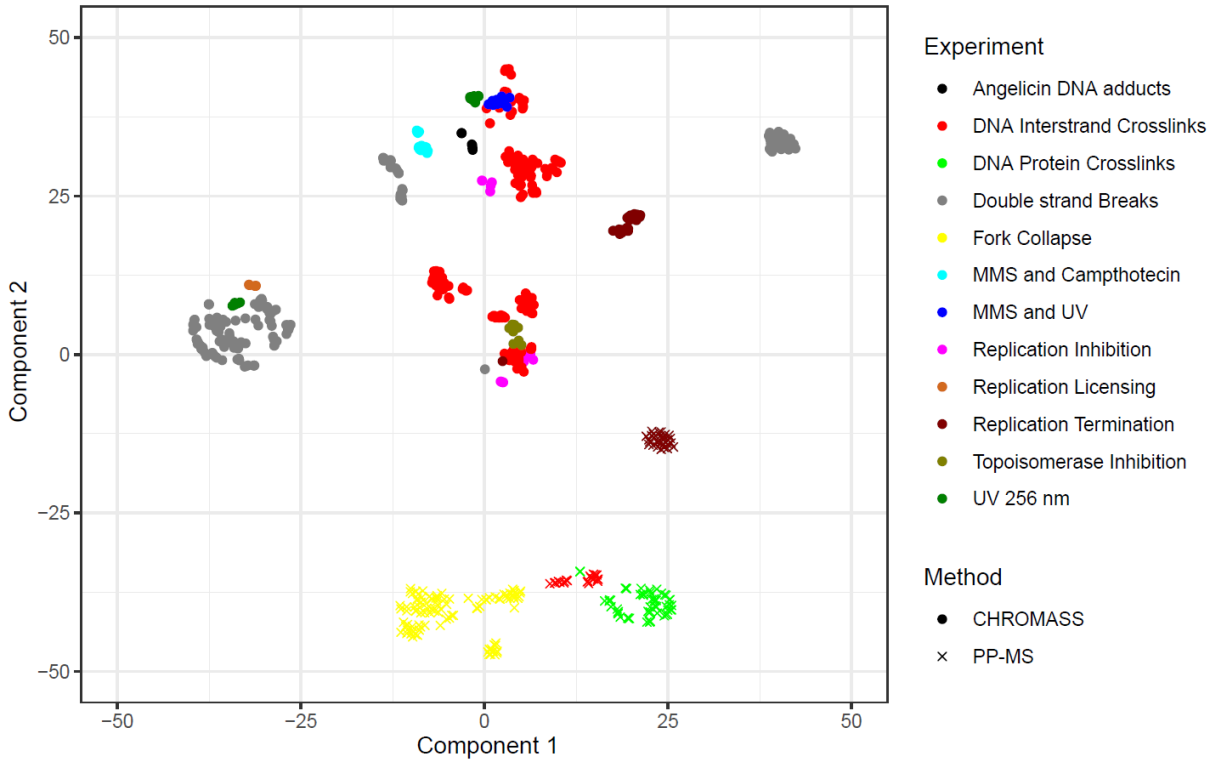


Figure 17 | Two-dimensional t-distributed stochastic neighbor embedding (t-SNE) of all sets included in the DRA. The figure shows the t-SNE dimensionality reduction to two components of all investigated datasets. Maximum iterations: 1000, perplexity 30. Theta: 0. Experiment types are grouped by color. PP-MS and CHROMASS are distinguished by shape.

Application of dimensionality reduction approaches revealed tight clustering of individual replicates. To some degree, data also clustered according to the experimental conditions, whereas the type of MS instrument or batch effects of individual measurements had only a marginal effect. Remarkably, data from CHROMASS experiments could be clearly separated from PP-MS data in this analysis

as indicated by the clusters in the top and lower half of the plot. This is likely explained by different sets of proteins co-isolated non-specifically during sedimentation of the chromatin templates compared to the proteins co-purified with the bead-mediated capture of plasmid substrates. Overall, we present an expansive proteomics dataset of *in vitro* DNA repair experiments that contains broad information about multiple DNA damage repair mechanisms due to a range of different treatments. For a computational analysis those experiments were processed with MaxQuant resulting in little technical bias.

3.1.2 A combined statistical analysis across multiple DNA lesions highlights recruitment of characteristic DNA repair factors

To identify DNA repair factors, we used the R package “SamR, significance analysis of microarray data” [62] , which determines significant enrichment in the presence of various induced DNA lesions, as described in the Methods. We defined series of two-sample t-test consisting of 139 different comparisons of replicates against appropriate controls, in which we included experiments across all investigated DNA lesions and treatments of multiple timepoints, to cover a wide range of recruitment of DNA repair factors. For each identified protein the total number of tests was calculated, in which it was significantly enriched compared to the appropriated control (Supplementary Figure 1 & Table 3). The distribution of how often proteins score significant is shown in Figure 18.

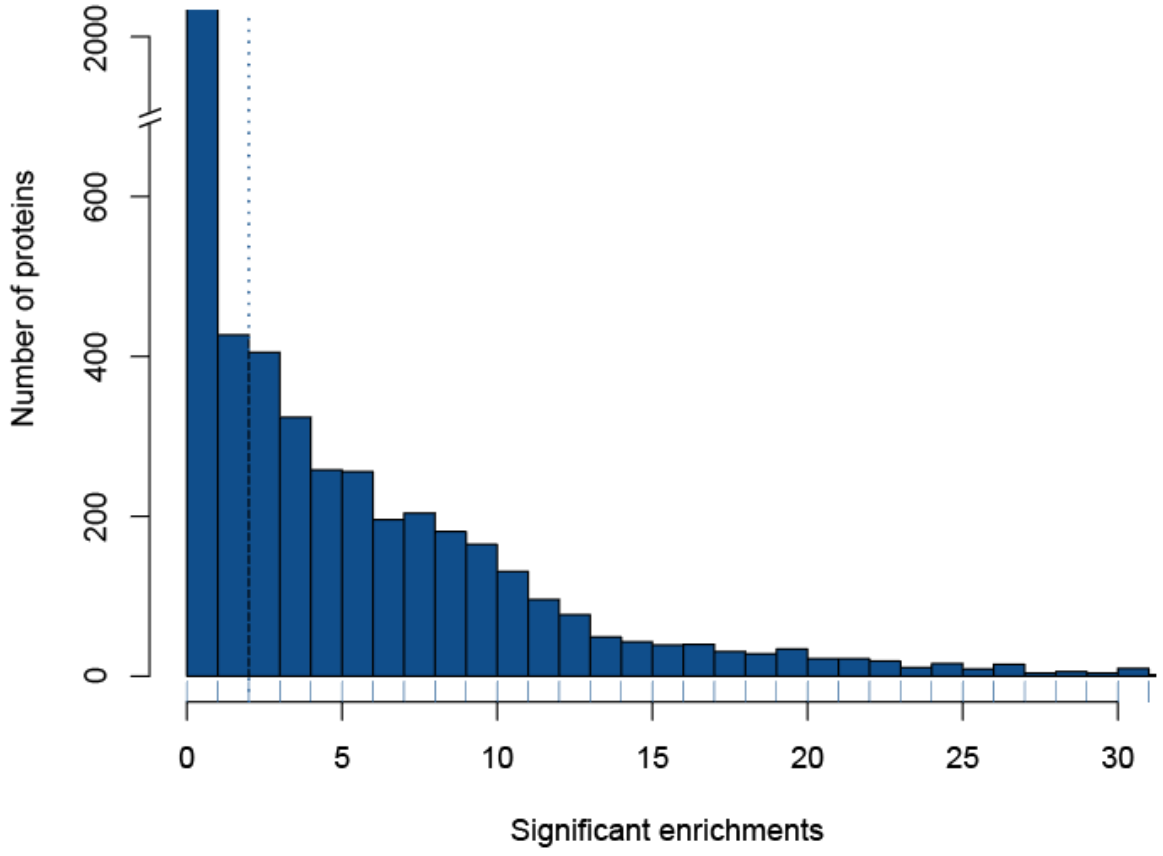


Figure 18 | Significance analysis. The histogram shows the distribution of significant enrichments in 139 SamR comparisons across all investigated DNA damages for 5790 proteins. Significant enrichments are defined as the total count of experiments the respective protein was significantly enriched in the treatment in a t-test against an appropriate control. The dotted line indicates the global median of significant enrichments at two comparisons. Supplementary table 3 shows the exact number of proteins as well as the percentages per significant enrichment.

As the dataset originates from a high number of different conditions, which lead to the identification of many chromatin associated proteins, many factors that aren't necessarily associated with DNA damage repair were identified. Therefore, the majority of proteins are significant in a single or no t-test. In total 2989 proteins (52%) enriched significantly at or below the global median of two comparisons (Supplementary Table 3). The number of proteins decreases monotonously with the number of significant enrichments. The highest numbers of significant enrichments include a set of 103 proteins with 30 to 65 significances. This indicates the identification of DNA damage markers and repair factors that play an essential role in multiple DNA repair mechanisms. The list of the most significant proteins confirms this, as it includes many factors with a known and well-described function in DNA replication, in the maintenance of DNA integrity, and DNA damage signaling and repair (Table 1). It is led by the ATP-dependent annealing helicase SMARCAL1 (65 times significant) and also contains its regulating key partners, the single strand DNA binding factors RPA1 (59 times significant), RPA2 (55 times significant) and RPA3 (51 times significant) [249]. Further it shows the MCM2-7 complex (54-56 times significant), a DNA replication helicase and a key factor of DNA replication initiation and progression [250], and the interacting factors FANCD2 (64 times significant) and FANCI (65 times significant), which are essential for the stabilization of stalled replication forks and recruitment of DNA repair factors [251].

Table 1 | Top 25 proteins with most significant enrichments. The table shows the 25 proteins that were shown to be significant in the highest number of t-tests. XB gene name is derived from XenBase. The suffix _BL indicates that the name originates from a BLAST search to human. Duplicate names (rpa1, rpa2) stem from the allotetraploid sub-genome of *X. laevis* (see below).

XB Gene Name	Significance Count
smarcal1	65
fanci	65
fancd2	64
rpa1_BL	59
mcm3	56
scai	56
mcm6	56
mcm4	56
mcm5	56
prim1	56
mcm7	55
rpa1_BL	55
rpa2	55
brca2	55
atrip	54
mcm2	54
pola2	52
prim2	51
bard1	51
pola1_BL	51
atrip	51
rpa3	51
rfwd3	50
rpa2	50
atr	49

In a next step we investigated DNA lesion specific enrichment to visualize commonly recruited factors from our combined dataset. Therefore, we grouped the dataset in four major groups of investigated DNA lesions: interstrand cross links, double-strand breaks, DNA-protein crosslinks, and fork collapse, and two minor groups: Replication - various perturbations, Plasmid - various perturbations. To investigate these DNA lesions in detail, we calculated lesion specific enrichment scores for each protein, as described in the Methods, and plotted them as “combined volcano plots” against the maximum $\log_2$ fold change of each recruited protein in all related treatment against control groups. Of note, the organism *Xenopus laevis* has an allotetraploid genome with two sub-genomes, denominated small and large, resulting from a hypothesized hybridization of two species [252]. A further novelty of our combined dataset is that we can investigate the recruitment of DNA repair factors while distinguishing between the two sub-genomic origins of most genes, as indicated by gene names with the suffix “.L” (large sub-genome) and “.S” (small sub-genome) in the shown plots.

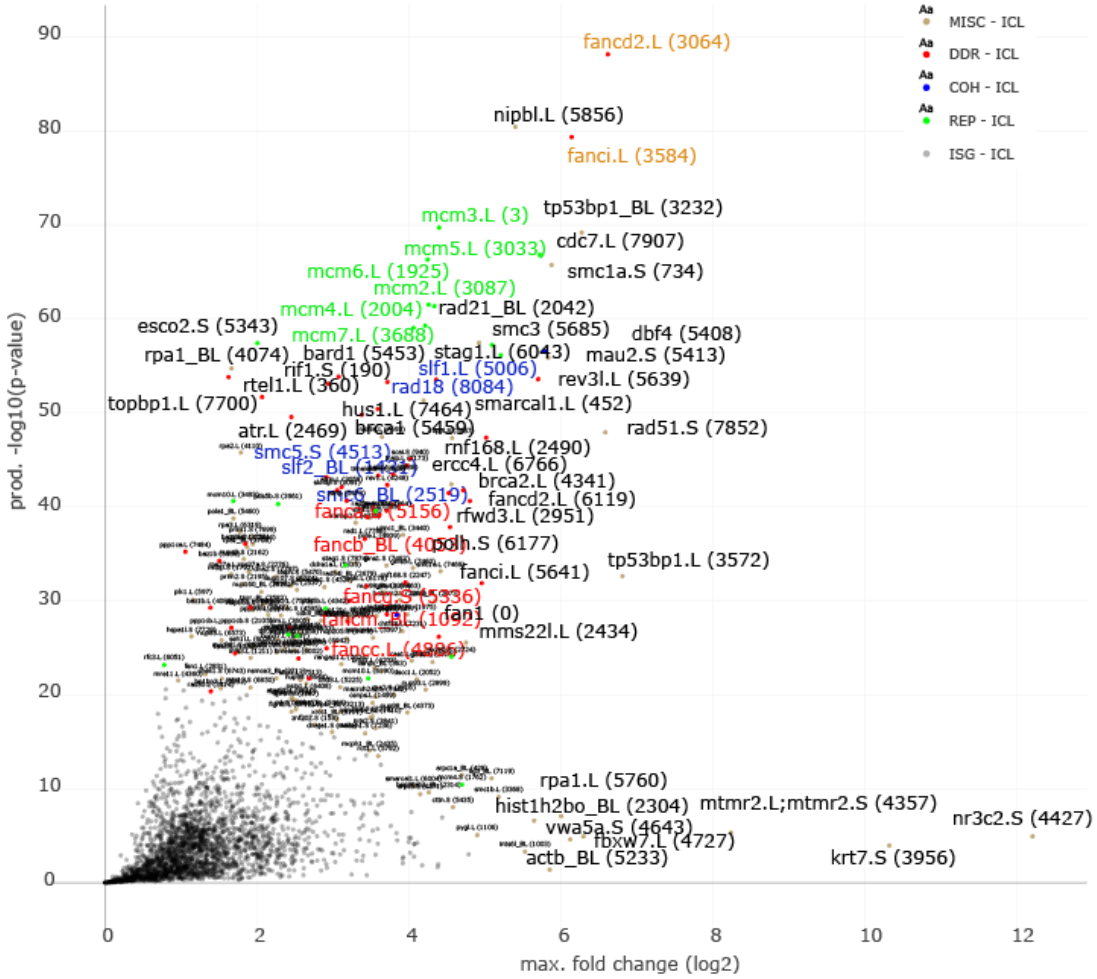


Figure 19 | Interstrand crosslink repair. The figure highlights the highest enriched factors in all investigated interstrand crosslink repair experiments. The x-axis shows the maximum fold change per proteins in all experiments, the y-axis the combined score calculated as product of all related p-values on a -$\log_{10}$ scale. The color coding shows a manually curated grouping of proteins based on their known function in replication (REP, green), cohesion (COH, blue), DNA damage response (DDR, red) or unassigned/miscellaneous (MISC, brown). Labels show gene symbol and MaxQuant id in brackets for unique identification. Several highly enriched proteins are color coded manually. The plot can be interactively visualized in the module *"Volcano Plots"* at http://dnarepairatlas.bio.uni-kl.de/.

The combined volcano plot for DNA interstrand crosslinks originates from multiple CHROMASS experiments of time courses of psoralen treated chromatin with several inhibitors, such as the replication inhibiting factor geminin or direct inhibitors for BRCA4 or ATR (Figure 19). Recruited factors contain a distinct overlap with many previously published findings in close proximity to each other. Among the highest enriched factors are FANCI/FANCD2, which regulate incision and translesion synthesis during ICL repair [253], the replication initiation and elongation proteins MCM2-7, the Fanconi Anemia complex and the RAD18-SLF1-SLF2 complex, which recruits SMC5/6 to stalled replication forks [162]. Multiple proteins are recruited in several copies to the DNA lesion and therefore show a higher enrichment than others. Those proteins include RAD51, the three RPA subunits and according to recent publications FANCI and FANCD2, which could be shown in CHIP-seq experiments and directly observed by cryo-electron microscopy [254,255].

Similar as for ICL, the combined enrichment shows a high fidelity for the identification of previously described factors involved in the repair of covalent DNA-protein crosslinks (Figure 20).

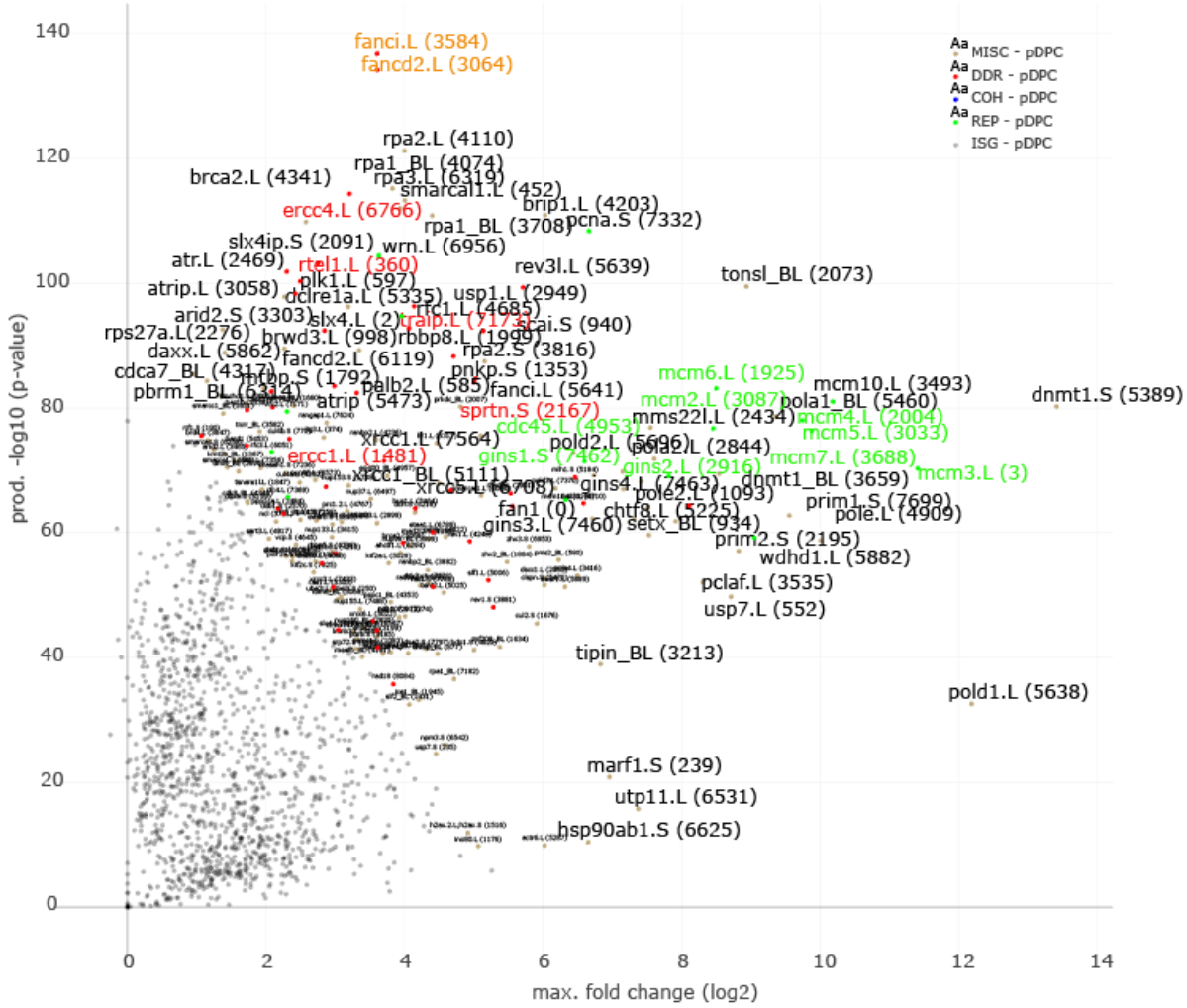


Figure 20 | DNA-protein crosslink repair. The figure highlights the highest enriched factors in all investigated pDPC repair experiments. The x-axis shows the maximum fold change per proteins in all experiments, the y-axis the combined score calculated as product of all related p-values on a -$\log_{10}$ scale. The color coding of the points shows a manually curated grouping of proteins based on their known function in replication (REP, green), cohesion (COH, blue), DNA damage response (DDR, red) or unassigned/miscellaneous (MISC, brown). Labels show gene symbol and MaxQuant id in brackets for unique identification. Several highly enriched proteins are color coded manually. The plot can be interactively visualized in the module *"Volcano Plots"* at http://dnarepairatlas.bio.uni-kl.de/.

The figure contains experiments from multiple PP-MS measurements monitoring the recruitment of factors triggered by the collision of replication forks with a covalently attached protein roadblock, with several different inhibitors, including geminin and Ubiquitin Vinyl Sulfone (UB-VS). This inhibitor was used to block proteasomal

degradation in the related experiment that lead to the identification of the degradation factors TRAIP and SPRTN, which mediate DNA-protein adduct degradation in a replication-dependent manner [163]. These factors can be found as strongly enriched together with ERCC1 and ERCC4 and the DPC bypass facilitating helicase RTEL1, which unwinds the DNA past the DPC for the CMG complex (CDC45, MCM2-7, GINS) [247]. The results show similar reliability for proteins recruited in the presence of the other major investigated experiment groups for fork collapse and double-strand break repair (Supplementary Figures 2 & 3, DNA Repair Atlas module: "*Volcano Plots"*). Overall, this highlights a high fidelity of the dataset for the identification of significant key players in all groups of investigated DNA lesions.

Repair factors in X. laevis are expressed from both sub-genomes.

The two sub-genomes of the allotetraploid model organism *Xenopus laevis* have evolved asymmetrically. 56% of all genes are retained in two homoeologous copies. One chromosome set is more often preserved in the ancestral state while the other has experienced frequent genetic alterations, such as gene loss by deletion, intrachromosomal rearrangement and overall reduced gene expression, as described in [256].

To investigate whether the repair factors are preferably expressed from one of the two genomes, I filtered for all proteins encoded on either the small or large sub-genome. This resulted in two subsets with 2035 (.S) and 2836 (.L) uniquely assignable hits. (Figure 21, A). To investigate a potential selective bias for DNA repair factors, I created a second subset that consists of proteins with both a detected .S and .L variant and at least one significant enrichment in the dataset. This resulted in a set of 765 significantly "co-expressed" proteins (Figure 21, B).

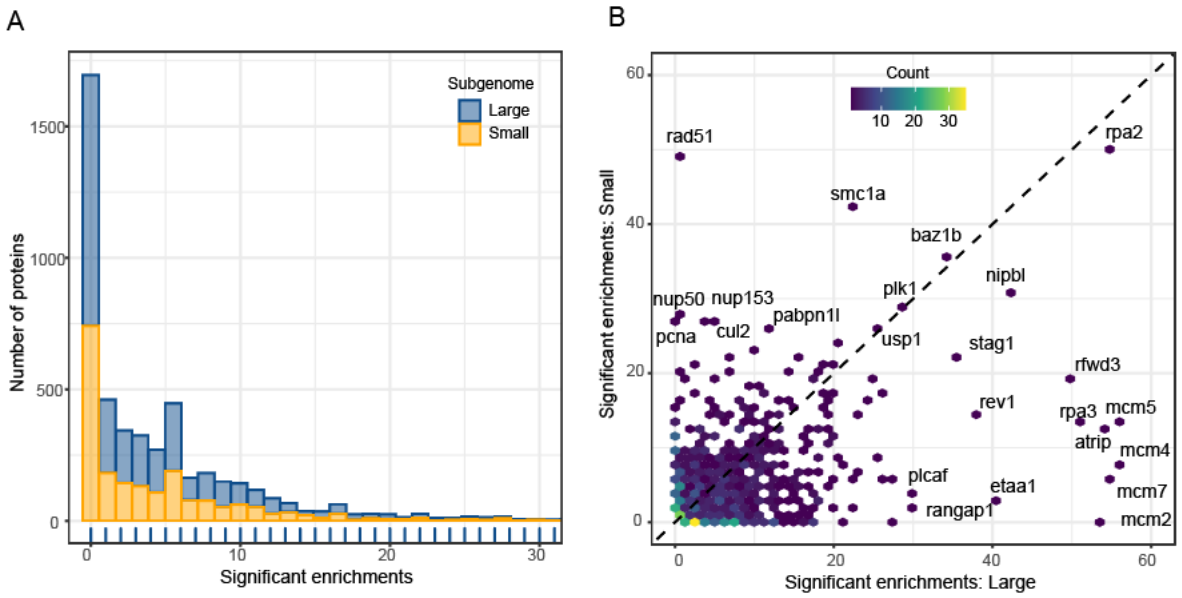


Figure 21 | Distribution of sub-genomic expression in *X. laevis*. (A) Stacked histogram indicating the total count of significant enrichments, calculated as described in Figure 18, for proteins originating from the large (blue) or small (yellow) sub-genomes. (B) 2D-Histogram of the significance count for proteins expressed on both sub-genomes. The color gradient indicates increased number of enrichments.

This analysis showed no bias for the overall expression of proteins from either sub-genome (Figure 21, A). For the subset of proteins significantly expressed from both sub-genomes, only a slight bias towards the large (.L) can be observed, specifically for the proteins with the highest number of total significances (Figure 21, B). Strikingly, this included four proteins of the MCM complex. Taken together, we conclude that while there is a slight bias towards a set of outliers, DNA repair factors overall are expressed from both sub-genomes of *X. laevis*.

3.1.3 Creation of the DNA Repair Atlas – A web resource for mining and visualization of mass spectrometry data

To further investigate the dataset and facilitate an interactive visualization of the results, we created the DNA Repair Atlas. This web-application is a server-based resource for scientists to mine and visualize our combined dataset. Figure 22 schematically highlights the used filtering and normalization strategy.

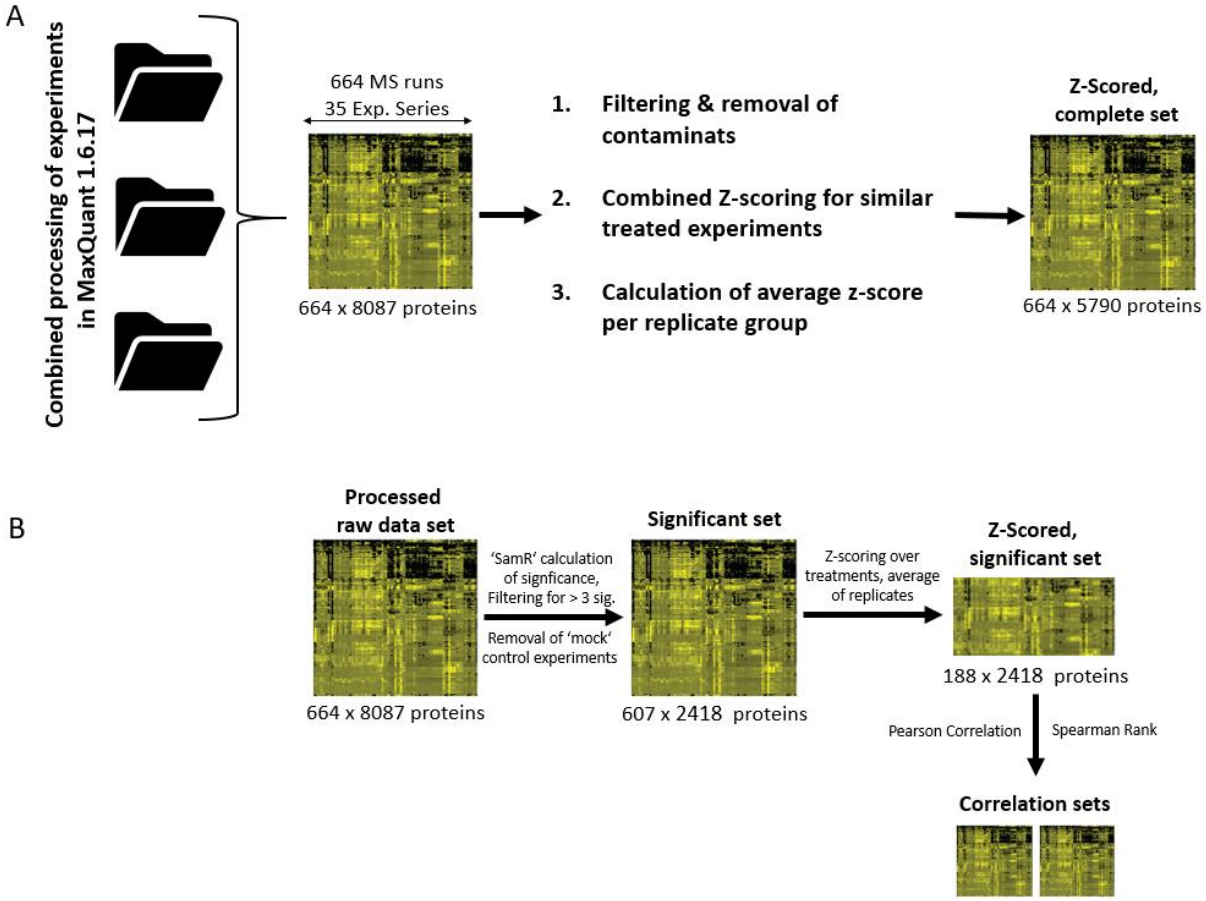


Figure 22 | Schematic for the DRA data processing. The figure highlights the processing strategy for the DNA Repair Atlas. (A) 664 MS measurements resembling 35 experiment series were pooled, filtered and normalized. (B) Significance analysis of microarray (SAM), has been used to test for significance across all experiments to create a subset of significantly enriched repair factors.

In brief, the full dataset, containing 8087 protein groups, was transformed to $\log_2$ scale and filtered to remove contaminants, reverse hits and proteins only identified by site. The resulting, “complete” set containing 5790 proteins was normalized by z-scoring across similar treatments and averaged per replicate group to create comparable values (Figure 22 A).

Next, the result of the SamR analysis was used to determine the median of two significant enrichments across all comparisons (Figure 18). A subset with significant proteins was created by filtering for proteins that scored at least four times as significantly enriched in any of the performed t-test. This resulted in a set with 2418 significantly enriched proteins (Figure 22 B). To facilitate further analysis, we used Spearman Rank and Pearson correlation to create undirected network graphs. Each protein from the set of significantly enriched proteins is represented by a node and each connection between a node pair is represented by a weighted edge containing the correlation value. Performing those preprocessing steps yields robust statistical relationships between differentially recruited repair proteins, which serve as a foundation for all visualization and function prediction approaches in the application.

A modularized resource facilitates mining and visualization of DNA repair factors

The DNA Repair Atlas functions as a collection of independent modules (Figure 23), which serve the purpose to generate hypotheses for bottom-up analysis approaches. It facilitates an unbiased analysis starting from a single protein or a list of proteins of interest to gather information about their shared function, or by yielding information about specific DNA repair pathways for a scientist, to learn more about individual proteins of interest.

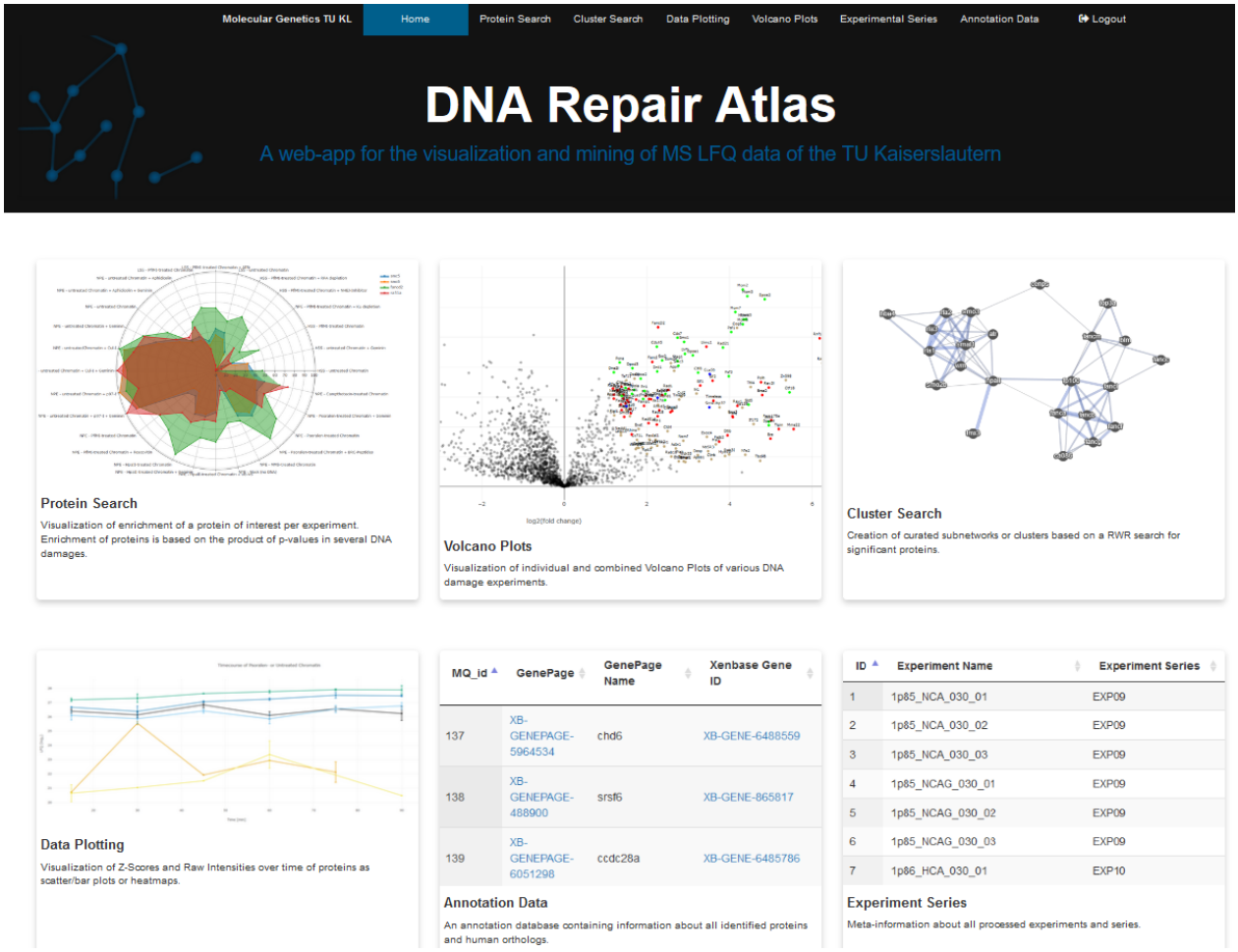


Figure 23 | The home page of the DNA Repair Atlas. The screenshot shows the hub page of the DNA Repair Atlas in http://dnarepairatlas.bio.uni-kl.de/. The hub consists of six individual modules that facilitate the analysis of the DNA repair data. The login credentials are Name: "**Guest**". Password: "**drs18**"

Six different modules were created to facilitate data mining and visualization, which I will describe in brief in the following section.

Module "Protein Search"

The module "*Protein Search"* visualizes relative enrichment of a protein as radar plot. Enrichment scores, calculated as described in the Methods, represent the product of p-values obtained for the enrichment of the proteins in all experiments belonging to one of the six grouped DNA repair pathways shown on the radar plot. To ease the comparison, enrichment scores were scaled to an interval from 0 to 1 (Figure 24).

Figure 24 | A radar plot visualized by the module "*Protein Search*". The figure shows the enrichment scores of FANCI and FANCD2. Each radial axis represents a particular perturbation corresponding to different DNA lesions: Interstrand crosslinks (ICL), fork collapse (FC), double strand breaks (DSB) and DNA-protein crosslinks (DPC) as well as the two minor groups replication - various perturbations (RV) and plasmid – various perturbations (PV). The table on the right side shows the t-tests in which the proteins were significantly enriched, as well as the related p-value and fold change. Visualization is performed with plotly.js at: http://dnarepairatlas.bio.uni-kl.de/Summary

The table on the right side lists all experiments, in which the selected protein was found to be significantly enriched with a false discovery rate of >5% and a minimal fold change of at least 1.5. Further information about the compared experimental conditions can be found under the tab *"Experiment Series"*. Using the (+) button, additional proteins can be selected. This facilitates the analysis of the pathway-specific enrichment of DNA repair complexes. As an example, data for the FANCI/FANCD2 heterodimer is shown. In agreement with the significance analysis (Table 1, Figure 18) both subunits of the complex show a high recruitment profiles across the different experiments. The control module in the top left initializes the plot via comma or

semicolon separated lists of genes names. Further explanations of the interactivity can be found by pressing the help button (?) in all modules.

Module "Data Plotting"

The module *"Data Plotting"* provides easily accessible supplementary information about the investigated experiments. Z-scores and LFQ intensities can be visualized in form of line plots, bar plots or heatmaps. To process the data for the individual plots a query list of proteins is used to extract the associated values from the dataset, which is chosen in a linked drop-down menu (Supplementary Figure 4). When initialized the application dynamically recognizes the chosen experiment, treatment and plot type and visualizes the list of input proteins accordingly (Figure 25).

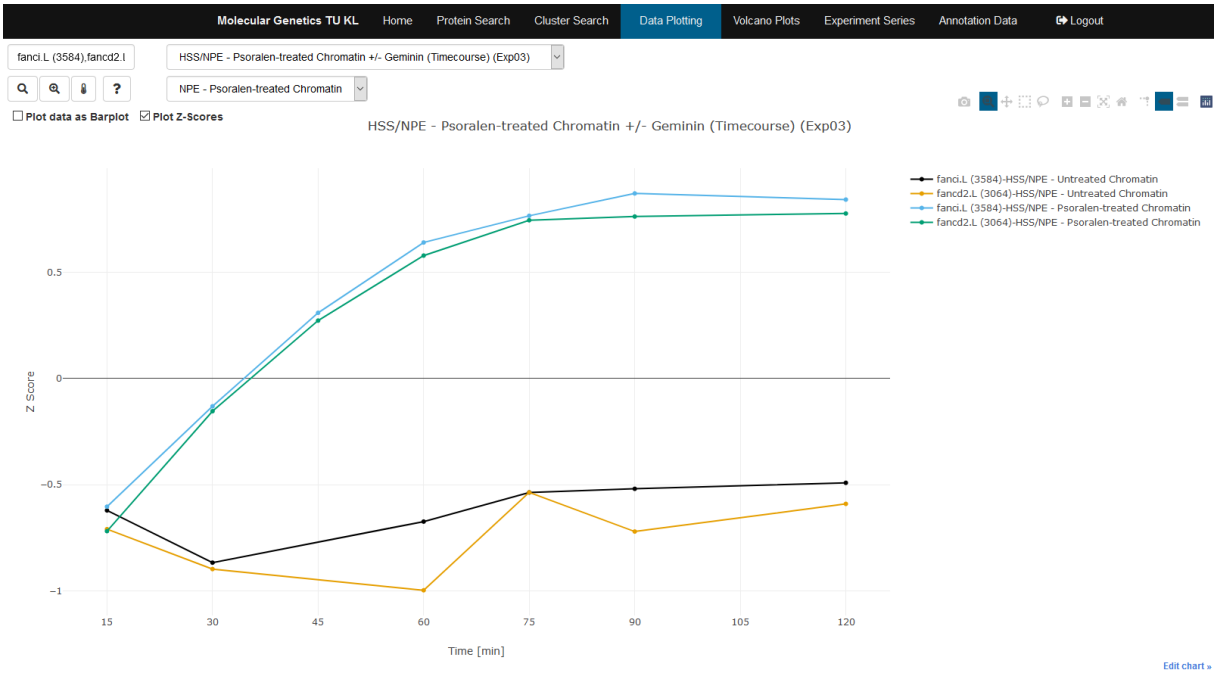


Figure 25 | Z-scored abundancy of FANCI and FANCD2 visualized by the module *"Data plotting"*. The screenshot shows the z-scored log_2 LFQ intensities of the proteins FANCI and FANCD2 in the interstrand crosslink time-course of psoralen treated and untreated chromatin. The control panel in the top left is used to initialize the plot from a list of input proteins either as line plot, bar plot or heatmap. The visualization is performed by plotly.js.at: http://dnarepairatlas.bio.uni-kl.de/ZScore

Using the check boxes, the application can dynamically switch between a chosen visualization type. As an example, z-scored recruitment profiles of FANCI and FANCD2 in the time course of psoralen treated and untreated chromatin are shown as line plot. The shown plot can be further edited in the plotly chart studio by clicking **>>edit chart** or exported as scalable vector graphic (.svg) via the save symbol.

Module "Volcano Plots"

The module *"Volcano Plots"* visualizes the enrichment of individual proteins together with the associated p-value as calculated by "SamR" based two-sample t-testing (Figure 26). To keep the computational load of the web application minimal, the volcano plots have been pre-calculated for all closely matched conditions within individual experiments.

Figure 26 | A volcano plot visualized by the module "*Volcano Plots*". The screenshot shows a volcano plot of psoralen induced interstrand crosslink repair at 30 minutes. The factors FANCD2 and FANCI are highlighted by the search function. The proteins are color coded by manually curated groups for DNA damage repair (DDR), cohesion (COH), replication (REP) and miscellaneous (MISC). Insignificant hits (ISG) are shown with lower opacity and without label. The top left shows the control panel, that allows the initialization and comparison of multiple plots next to each other. Interactivity, as on-click events and labels shown on hover are shared between multiple plots for easier tracking of proteins. Visualization is performed by plotly.js at: http://dnarepairatlas.bio.uni-kl.de/Volcano

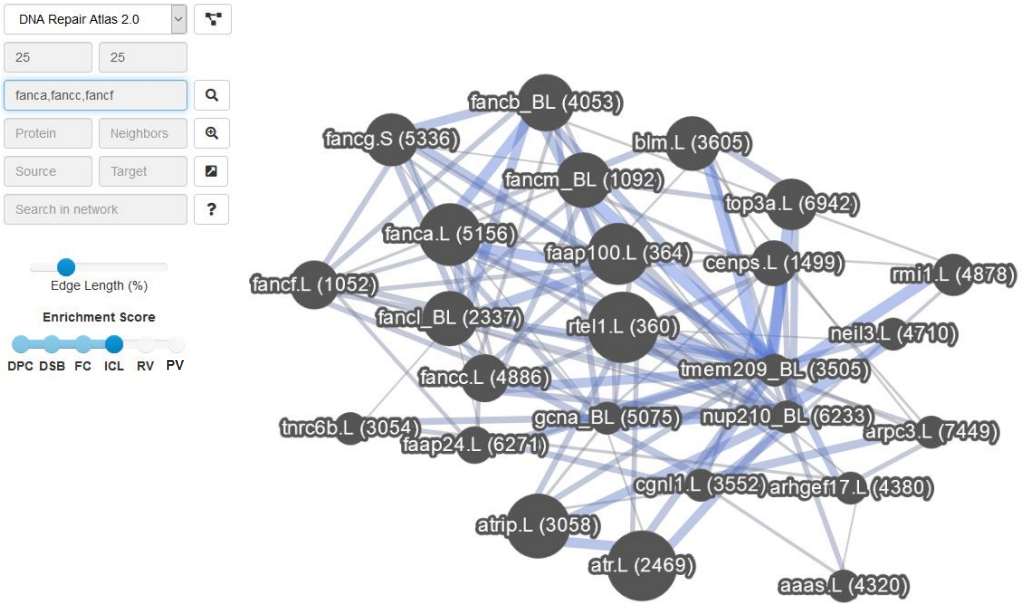
In additions to these individual plots, so-called "combined volcano plots" are available for each of the four experimental series that have been conducted to characterize the time-resolved repair of DNA interstrand crosslinks, DNA protein crosslinks, collapsed replication forks as well as restriction enzyme induced double strand beaks in *Xenopus*

laevis egg extracts (Figure 19 & 20, Supplemental Figure S2 & S3). In these “combined volcano plots”, the maximum fold change within the experiment series is plotted against the product of all p-values obtained for all comparisons within the grouped, experimental series.

To ease comparisons between different experimental series or between different time points, two volcano plots can be shown side-by-side, while specifically marked proteins of interest, such as for example FANCI and FANCD2 in Figure 26, can be highlighted by a search function. Hoovering over a data point highlights the proteins automatically in both plots. All data points are colored by manually curated categories: DNA-damage response (DDR), cohesion (COH), DNA replication (REP), miscellaneous (MISC) or, if the value is lower than a set significance, as insignificant (ISG). Alternatively, by selecting another experiment to be shown on the y-axis it is possible to plot the fold change of any two conditions against each other (choosing hide will reset the y-axis to show the associated p-value). To facilitate data mining, each data point is hyperlinked to the corresponding web-page in XenBase, Uniprot or NCBI, to which links are shown when a data point is clicked. Similar to the modules *“Protein Search”* and *“Data Plotting”*, the shown plot can be further edited in the plotly chart studio by clicking **>>edit chart** or exported as scalable vector graphic (.svg) via the save symbol.

Module "Cluster Search"

To visualize clusters of correlating proteins and potentially identify new DNA replication and repair factors, an algorithm was implemented to cluster the temporal recruitment profiles. This modified network diffusion algorithm is based on a random walk with restart (RWR) [257]. Starting from user-defined sets of proteins, an empirical score is diffused in a network of significantly enriched repair factors, created as described in the Methods (Figure 27).



protein	score
fanca.L (5156)	0.111345
fancc.L (4886)	0.110734
fancf.L (1052)	0.109541
tmem209_BL (3505)	0.048865
nup210_BL (6233)	0.039955
faap100.L (364)	0.033911
fancb_BL (4053)	0.033777
fancl_BL (2337)	0.033415
fancm_BL (1092)	0.031207
fancg.S (5336)	0.030041
gcna_BL (5075)	0.025378
faap24.L (6271)	0.016594
cenps.L (1499)	0.011521

Figure 27 | A clustering of the Fanconi Anemia core complex visualized by the module "*Cluster Search*". The screenshot shows a clustering performed for the genes FANCA, FANCC and FANCF. Visualized are the top 25 highest scoring hits according to the score diffusion of the random walk with restart. The higher the Pearson correlation between two nodes, the wider the edge and the stronger the saturation of blue color. The control panel in the top left allows the initialization of the clustering and modification of the parameters reset probability and size of the drawn cluster. Enrichment scores can be toggled via the slider below the panel, modulating the size of the nodes relative to the related enrichment score in the chosen DNA lesion. The table on the right shows the final score of the clustering for the top 100 highest scoring proteins. The clusters are visualized by cytoscape.js at: http://dnarepairatlas.bio.uni-kl.de/Clustering

In brief, the network represents all significantly enriched proteins in the dataset as nodes and the pairwise correlations between them correspond to weights of connecting edges. This algorithm accepts a single or a list of proteins of interest as input. Starting from this list it stepwise diffuses a score by visiting directly connected

neighbors, while preferring edges with a higher correlation. The user can set a restart probability that at each step determines the chance that the algorithm jumps back to a random member of the list of start nodes instead of taking a further step away from it. Thereby, the proximity of the score diffusion can be modulated and narrower or broader clusters identified. After taking 100.000 steps to diffuse the initial score to a steady state, the algorithm returns a final score for each node based on how often it was visited. A cutoff for the number of visualized proteins allows users to only draw clusters of a set size and reduce redundancy. For the visualization of the resulting cluster the top scoring proteins are passed to an interactive visualization with cytoscape.js and a table is shown, which shows the top 100 highest scoring nodes. To reduce the overall network density only edges representing biologically meaningful correlations are shown (corr. >= 0.7).

The additional functions of the control panel allow drawing of edges of interest between two shown nodes or add proteins and a set number of direct neighbors directly to the cluster. The visualization of the cluster highlights information about the protein relations: The higher the correlation between two nodes, the wider the edge and the stronger the saturation of blue color of the drawn edge. Enrichment scores can be toggled via the slider below the panel, modulating the size of the nodes relative to the related enrichment score in the chosen DNA lesion. The control panel in the top left allows the initialization of the clustering, modification of the reset probability and size of the drawn cluster. By clicking on a node in the table of highest scoring proteins its name is transferred to the input field to facilitate a consecutive analysis with an expanded selection of input proteins. All drawn clusters can be downloaded as scalable vector graphics (.svg) by clicking on the save-button.

Figure 27 for example shows a cluster of the top 25 scoring hits for the proteins FANCA, FANCC and FANCF with a reset probability of 25%. The algorithm identified the rest of the Fanconi Anemia core complex (FANCA, -B, -C, -E, -F, -G, -L and -M) with the exception of FANCE based on three members of it as input. FANCE is missing a full annotation in the UniProt FASTA of *Xenopus laevis.* Only two FANCE domain containing proteins are identified that show a lower correlation to the rest of the core complex. Further identified factors include known biologically relevant interactors, such as the bloom helicase BLM together with TOP3A and RMI1 [258] or the ATR-ATRIP kinase complex, which triggers activation of the Fanconi Anemia repair pathway [259].

Highly comparable results can be shown if only single subunits of a complex are used as input. For example, the same proteins can be identified by starting the clustering from FANCA or FANCB (Supplementary Figure S5 A, B). Alternatively, the algorithm can identify other modules with a high precision, such as the NHEJ factors (NHEJ1, LIG4, PAXX, XRCC4) and KU proteins (XRCC5, XRCC6) or proteasomal subunits based on a single input protein (Supplementary Figure S5 C, D). The fidelity of the search increases with the number of input proteins of the same module, as the score can be diffused more evenly if the algorithm has multiple start nodes of a biological module to choose from at each reset. Overall, this demonstrates how an unbiased cluster search algorithm can be used to identify direct interaction partners or proteins that share a common function in DNA repair, which are recruited together under similar conditions and timepoints and therefore are connected in the network.

Module "Experiment Series"

This module provides a list of all MS measurements comprised in the DNA Repair Atlas to give a comprehensive overview across all processed data (Figure 28). The interactive table summarizes the experimental details for each measurement, such as the used *Xenopus* egg extract system, the type of the used DNA template, inhibitor, as well as the time, at which the chromatin was isolated. For replicating extracts, the time is relative to the addition of nucleoplasmic extract (NPE), which triggers the assembly of active replisomes on the added DNA [161]. For double-strand break repair assays in absence of replication, the time is indicated relative to the addition of the PflMI restriction enzyme. Measurements belonging to the same experimental series have been prepared and measured side-by-side to ease comparison. Finally, all raw file names are provided to unambiguously retrieve them from the public repository *ProteomXchange [104].*

ID	Experiment Name	Experiment Series	Time	DNA Template	Treatment	Experiment Title	Raw File Name
89	1p170_NCC_015_01	EXP03	15	Sperm Chromatin	HSS/NPE - Untreated Chromatin	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) III	20120427_EXQ2_MaRa_SA_20cm_mr1p170_C15_01
90	1p170_NCC_015_02	EXP03	15	Sperm Chromatin	HSS/NPE - Untreated Chromatin	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) III	20120427_EXQ2_MaRa_SA_20cm_mr1p170_C15_02
91	1p170_NCC_015_03	EXP03	15	Sperm Chromatin	HSS/NPE - Untreated Chromatin	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) III	20120427_EXQ2_MaRa_SA_20cm_mr1p170_C15_03

Figure 28 | Experiment series database. The figure shows a section of the experiment series database for 664 processed measurements shown in the DNA Repair Atlas. Columns from left to right show the file ID, experiment name, series, timepoint after replication initiation in minutes, used DNA template, treatment, experiment title and the associated .raw file. The table is interactively built by the package datatables.js at http://dnarepairatlas.bio.uni-kl.de/Meta

Module "Annotation Data"

In this module, a database containing protein centric information is provided in an interactive table. Processing data in MaxQuant assigns a numerical MaxQuant identifier (MQ_id) and a majority protein ID derived from the used UniProtKB FASTA file. The MQ_id serves as common identifier for proteins in all processing steps for internal matching, as it is unique for any identified protein group and easily expandable. Gene names were assigned from the combined UniProtKB, Swiss-Prot and TrEMBL database derived from the processing in MaxQuant. The used model organism organism *Xenopus laevis* has an allotetraploid genome, as described above. For example, the two proteins associated with mcm5 have two genetic origins mcm5.S (small sub-genome) and mcm5.L (large sub-genome) with different measured abundance rates in the datasets and annotations in the database (Figure 29).

A special annotation derived from XenBase is utilized to distinguish between the allotetraploid origins. Hence, we expanded the protein annotation database by mapping from UniProtKB (majority protein id) to XenBase, an online resource collecting various genomic, genotype and phenotype data from Xenopus research [260].

MQ_id	GenePage	GenePage Name	XenBase Gene ID	Gene Symbol	Majority protein IDs	Protein names	Human Entrez Gene ID	Human UniProtKB Gene Name	Human Gene stable ID	Category	Significance Count
2724	XB-GENEPAGE-985664	mcm5	XB-GENE-985671	mcm5.S	A0A1L8GGS4;Q6PCI7	DNA replication licensing factor mcm5-B	4174	MCM5	ENSG00000100297	REP	14
3033	XB-GENEPAGE-985664	mcm5	XB-GENE-6256510	mcm5.L	A0A1L8GNF3;P55862	DNA replication licensing factor mcm5-A	4174	MCM5	ENSG00000100297	REP	56

Figure 29 | Protein annotation database. The figure shows a part of the protein annotation database of the DNA Repair Atlas for MCM5.L and MCM.S. Each entry is assigned (from left to right) a MQ_id, a XenBase GenePage identifier, Name and Gene ID. The gene symbol includes allotetraploid origin and, for proteins that could not be annotated, a BLASTed gene name with the suffix _BL. Majority protein ID and names are derived from MaxQuant (UniProtKB, SwissProt, TrEmbl). For further information about the human ortholog from ENSG (Id, name) have been added. A manually curated category group was assigned to broadly highlight the function of each protein (REP: replication, COH: cohesin complex, DDR: DNA damage response, MISC: miscellaneous/unassigned). Significance count shows the count of significant enrichments in all tested experiments. Each entry in the database interactively links back to the corresponding entry in the related annotation databases XenBase, UniProt, NCBI and Ensembl. The table is interactively built by the package datatables.js at http://dnarepairatlas.bio.uni-kl.de/Anno

The gene symbols of the .S and .L sub-genome have individual gene names and XenBase gene ids, but point to the same GenePage id and GenePage name (Figure 29). This allows the application to dynamically deconstruct the users' input by pattern matching from a GenePage name to the gene names of both sub-genomes, for example by plotting both mcm5.L and mcm5.S when a user searches for mcm5. To further annotate the identified protein groups, a blast search was carried out to identify the most closely related human homolog using the NCBI BLAST software (local, 2.12.0) [261] with the human UniProtKB FASTA 2021_03 (including both Swiss-Prot and TrEMBL, *Homo sapiens*, TaxID: 9606). The highest scoring hits based on E-Values, which determine the expected number of false positives or background noise, were used to fill the missing gene names and marked with a suffix "_BL" to indicate the BLAST background. Further human ortholog information from NCBI and ENSG (Entrez ID, gene name, stable ID) have been mapped to the protein groups via BiomaRt [262]. Lastly, a manually curated category group was assigned to broadly highlight the function of each protein (REP: replication, COH: cohesin complex, DDR: DNA damage response, MISC: miscellaneous/unassigned) as well as the total count of significant t-tests of each protein. Taken together, this annotation database serves as a comprehensive resource for the identified *Xenopus laevis* proteins by combining and matching several independent databases. Both annotation databases together with the identified protein groups from the data processing and the complete list of SamR comparisons and results are stored in a separate module "*Supplementary Data*" in the application and can be downloaded.

In summary, the DNA Repair Atlas presents a comprehensive resource for the analysis and visualization of DNA repair data for scientists without a background in computational biology. All modules in the DRA require minimal input or tuning of parameters to be user friendly and come with convenient help pages that explain the functionality in more detail. Furthermore, it provides tools to compare time-resolved protein recruitment across different DNA repair pathways. This facilitates visualization and analysis without the need for advanced computational experience and facilitates the generation of new hypotheses about individual DNA repair factors or mechanisms for further experimental validation.

3.1.4 The DNA Repair Atlas facilitates iterative bottom-up analysis of DNA repair factors

To further assess the reliability of the generated dataset on a global scale, we used the DNA Repair Atlas for a network characterization of significantly enriched repair factors in all DNA lesions. As the complete network of proteins is too dense to be visualized in full, the implemented cluster search algorithm described above can be used to create subnetworks starting from a small set of known DNA repair factors. In essence, members of the resulting clusters are chosen based on the correlation of their recruitment profiles to those of the initial seed proteins

Here, we will showcase an analysis to identify modules composed of functional DNA repair complexes implicated in the repair of DNA interstrand crosslinks. As a starting set, we chose 10 well-known DNA repair factors, which all scored at least 20 significant enrichments on damaged DNA templates compared to appropriated controls (Supplementary figure 6). These included FANCA (26), FANCB (24), RTEL1 (48), BARD1 (51), TOP3A (22), RAD21 (25), BRCA2 (31), PALB2 (37), REV3L (43) and RAD51 (49). The number in brackets indicates the total number of significant enrichments. A subnetwork of the 100 highest ranking proteins identified by the RWR algorithm was visualized by cytoscape.js in Figure 30.

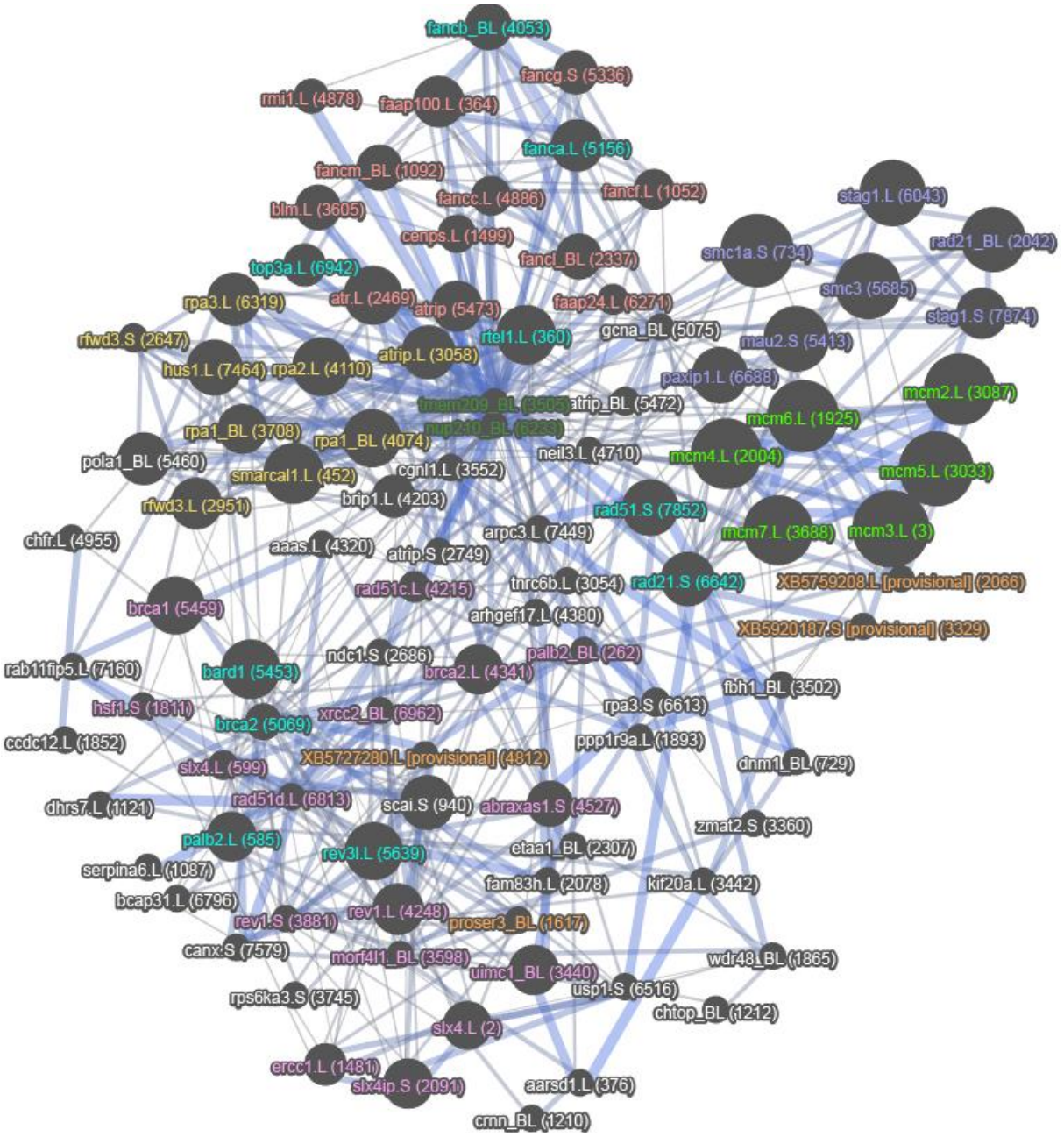


Figure 30 | DNA Repair Network. The network shows the TOP100 scoring hits from the unsupervised random walk with restart for the proteins: FANCA, FANCB, RTEL1, BARD1, TOP3A, RAD21, BRCA2, PALB2, REV3L and RAD51. Each unique protein is represented by a node in the network showing its gene symbol and MQ_id. Each pairwise correlation is represented by an edge. The edge width and blue saturation indicate low or high correlation. The node size is relative to the enrichment score for interstrand crosslinks. Several nodes have been highlighted by color coding (MCM complex: green, FA associated: red, cohesion: blue, DNA damage recognition: yellow, BRCA/HR: magenta, uncharacterized: orange, transmembrane hubs: dark green, input seed: cyan). The subnetwork has been created with the module "*Cluster Search"* in the DRA with a 25% reset probability.

Remarkably, although we used only individual subunits of known DNA repair complexes as seeds, the algorithm reliably identified most of the missing complex components. For example, even though only FANCA and FANCB were present in our starting selection, the algorithm reliably identified nine additional subunits of the thirteen subunit E3 ubiquitin ligase complex (FANCB, C, F, G, L, M, FAAP24, 100 and CENPS, Figure 30, shown in orange). Strikingly, the entire cohesion complex (shown in dark blue) was readily identified, although only one of its subunits (RAD21) was present as an initial seed. This approach also identifies higher-order repair complexes that might assemble only transiently during the repair event. For example, using the set parameters and starting seed proteins, correlation-based clustering readily identified the trimeric single strand binding complex RPA, as well as the associated checkpoint kinase ATR/ATRIP together with its activator ETAA1. These results are well in agreement with the priming function of the ATR kinase, which is known to phosphorylate the FANCI subunit of the Fanconi ID2 complex, which is required for subsequent ubiquitylation by Fanconi E3 ligase [263].

While no direct interaction of the ATR kinase with any of the starting proteins have been reported, RPA was found as a subunit of a nuclear multiprotein complex composed of the Fanconi E3 ubiquitin ligase with the BTR complex and BRCA1 [264]. Functionally it is well known that ATR phosphorylation of the substrate complex FANCI/D2 is required for its ubiquitylation by the Fanconi E3 ubiquitin ligase FANCL. In close proximity to the RPA module we find also several additional proteins known to bind RPA coated ssDNA, including RAD51 and SMARCAL1 (yellow color) in close proximity to BRCA2 (reviewed in [265]).

Next to the identified modules, several highly connected hubs can be seen in the subnetwork, such as the catalytic subunit REV3L (22 connected edges) of the DNA polymerase zeta [266] that correlates with multiple DNA repair factors of homologous recombination or the double-strand break repair protein RAD21 (21 connected edges). Strikingly, despite being only four times significantly enriched for all DNA lesions, the uncharacterized transmembrane protein TMEM209 (43 edges) and the nuclear pore complex protein NUP210 (45 edges) show the highest connectivity in the subnetwork (Figure 30, dark green). While not being directly involved in DNA repair, nucleopore proteins were shown to appear frequently in similar approaches [267] and play an

assisting role in the maintenance of genomic integrity and repair of DNA damage by anchoring damaged DNA strands (reviewed in [268]).

Lastly, the subnetwork shows several uncharacterized proteins, such as the yet unnamed XB5759208 and XB5727280. They associate with the MCM cluster and cohesion or homologous recombination, respectively. For the unnamed protein "XELAEV_18036466mg" (UniProt ID: A0A1L8FPH9), we inferred the name PROSER3_BL by a BLAST search of the human proteome that has shown a 28.9% overlap with "Proline and serine-rich protein 3" of *Homo sapiens*. This network only shows a fraction of the entire dataset, which showcases a global characterization across all investigated DNA lesions.

This characterization can be iteratively expanded to further explore certain modules. Each result of the cluster search returns a list of the top 100 scoring proteins in proximity to the input list next to the plot. Clicking on a member of the list automatically adds it to the field of input proteins for a consecutive search. This allows an iterative refinement of the result, as a longer list of starting proteins improves the fidelity of the algorithm. A short list of input proteins is vulnerable to be influenced by the presence of directly connected hubs, as for example the proteins TMEM209 and the nuclear pore complex protein NUP210 that correlate strongly with a high number of proteins. Those hubs would be visited frequently and diffuse the score away from meaningful clusters due to the high number of further connected nodes. Therefore, to improve the result, the clustering can be easily performed in an iterative manner. By transferring high-scoring nodes to the input list for a consecutive analysis the search can be expanded by meaningful input proteins. Each successive clustering can be used to further refine the input list with proteins that are part of modules of interest, to modulate the final cluster and increase the fidelity of cluster identification.

The resulting cluster can be further refined by overlaying the enrichment scores as relative node size to show selective or unselective enrichment of proteins for specific DNA lesions. In this example the enrichment scores of interstrand crosslinks were chosen. The modules can be further characterized by plotting their DNA lesion specific enrichment as radar plots, to investigate whether the correlation results from all of the four major pathways or only a subset of them (Figure 31).

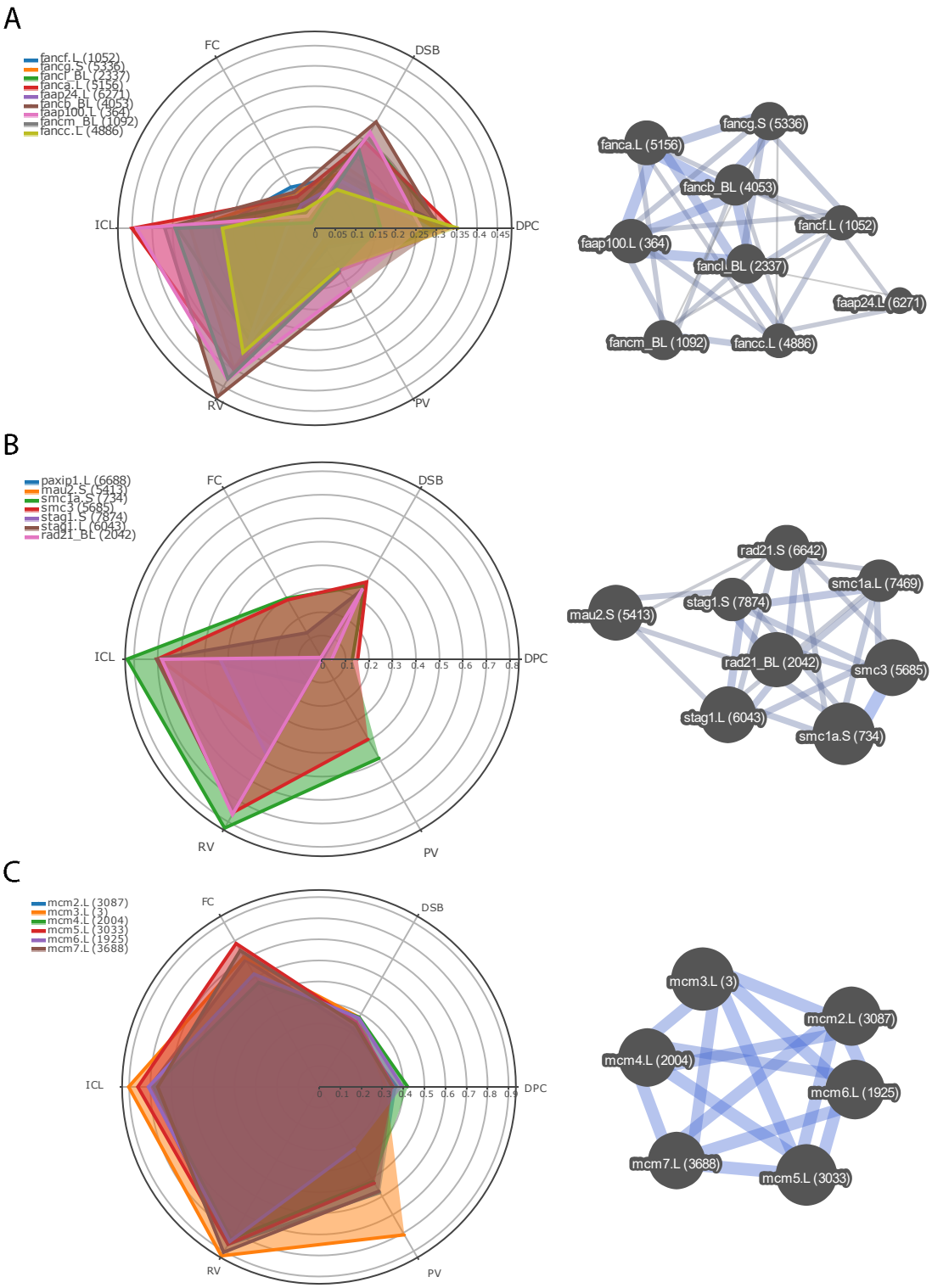


Figure 31 | Enrichment scores of repair modules. The figure shows enrichment scores of three identified modules as radar plots and clusters. (A) Fanconi Core Complex, (B) Cohesion, (C) MCM complex. The nodes represent proteins identified the members of the complex that were identified in the previous analysis. Size of the node shows the enrichment score of the highest overall enrichment per DNA lesion (fork collapse: FC, double-strand break: DSB, interstrand crosslink: ICL, DNA-protein crosslink: DPC, replication – various treatments: RV, plasmid – various treatment: PV). Enrichment scores have been plotted with the modules "*Protein Search"* and *"Cluster Search"*. Labels show gene symbol and MQ_id in parenthesis.

The Fanconi Anemia core proteins shows a distinctly high enrichment pattern in the radar plot for the major DNA lesion ICL (Figure 31) and a moderate enrichment for DPC and DSB. The role of this complex for the ICL is well-described by several different publications [269-271] and can be found in the corresponding combined volcano plots (Figure 19 & 20). In recent studies, the highest enriched member of the core complex for DSB in Figure 31, FANCA, showed a catalytic activity for single-strand annealing and exchange for double-strand break repair, which can explain the observed enrichment [269]. The identified members of the cohesin complex show high enrichment in ICL repair, in agreement with the findings that human developmental disorders related to mutated cohesion factor genes [272]. Lastly, the MCM complex overall shows highest enrichment in ICL repair, but also other DNA lesions, which agrees with the findings that the complex first has to be uncoupled at stalled replication forks to facilitate DNA repair [140]. Its presence on other lesions is surprising, considering that there is little known about the relevant functions of the helicase complex.

In summary, the cluster analysis approach shows a high reliability for the characterization of different DNA repair mechanisms with the DNA Repair Atlas. This is shown for DNA lesion specific recruitment through the creation of combined volcano plots and calculation of enrichment scores for recruited proteins. The analysis also could characterize known, representative factors on a global scale, by creating a representative subnetwork of correlating DNA repair proteins that highlights several repair mechanisms and interaction partners of the set of investigated proteins, as well as several potentially novel interactions.

Novel DNA Repair factor identification with the DRA

The DRA can also facilitate a bottom-up analysis starting from a single protein of interest. The subnetwork in Figure 30 includes several still uncharacterized factors that were identified in our dataset, such as ”XELAEV_18036466mg”, which we named PROSER3_BL due to its 28.9 % identity overlap by a BLAST search to *Homo sapiens* with an E-value of $2.3e^{-31}$, and the proteins XB5727280 and XB5759208. Supplementary table 4 shows the annotation database filtered for all proteins named with ‘provisional’ in name and at least four significant enrichments to highlight the presence of multiple potentially interesting factors in the database. This list includes both the .S and .L sub-genomic variant of XB5727280 with 4 and 7 significant enrichments over all DNA lesions and one variant of PROSER3_BL with 10 significant enrichments. To showcase a bottom-up investigation of PROSER3 in order to predict its function or find potential interaction partners, we used the random walk with restart with this protein as a single input. The restart probability was increased to 35 % to increase the overall number of restarts and diffuse the score in a closer proximity to the input protein. The resulting cluster of the 30 highest scoring hits is shown in Supplementary Figure 7. To draw more robust conclusions about the list of identified proteins in the proximity, GSEA with EnrichR [74] was performed for the top 100 scoring hits (Figure 32).

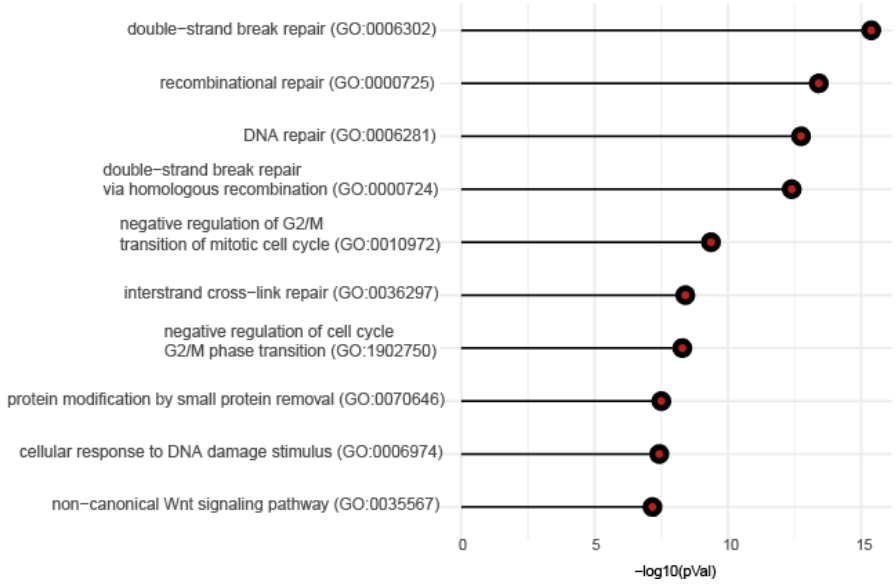


Figure 32 | GSEA of top 100 scoring hits for PROSER3. The p-values from the gene set enrichment performed with EnrichR were transformed to $-\log_{10}$ scale, ranked and plotted as bar plots. The pathways are derived from Gene Ontology Biological Process (2021).

The gene set enrichment has shown multiple DNA repair related mechanisms as most significantly enriched pathways. To validate that this finding is not due to a bias introduced by a dataset consisting of proteins that are significantly enriched in DNA repair processes, the analysis was repeated for XB5885669, the uncharacterized 'provisionally' named gene with the highest number of significant enrichments (Supplement Table 4). This has yielded no DNA repair related pathways (Supplemental Figure 8). To find a broader proximity of PROSER3 in the network, we used the RWR in a second analysis of the proteins included in the term *double-strand break repair* from the gene set enrichment (Figure 33).

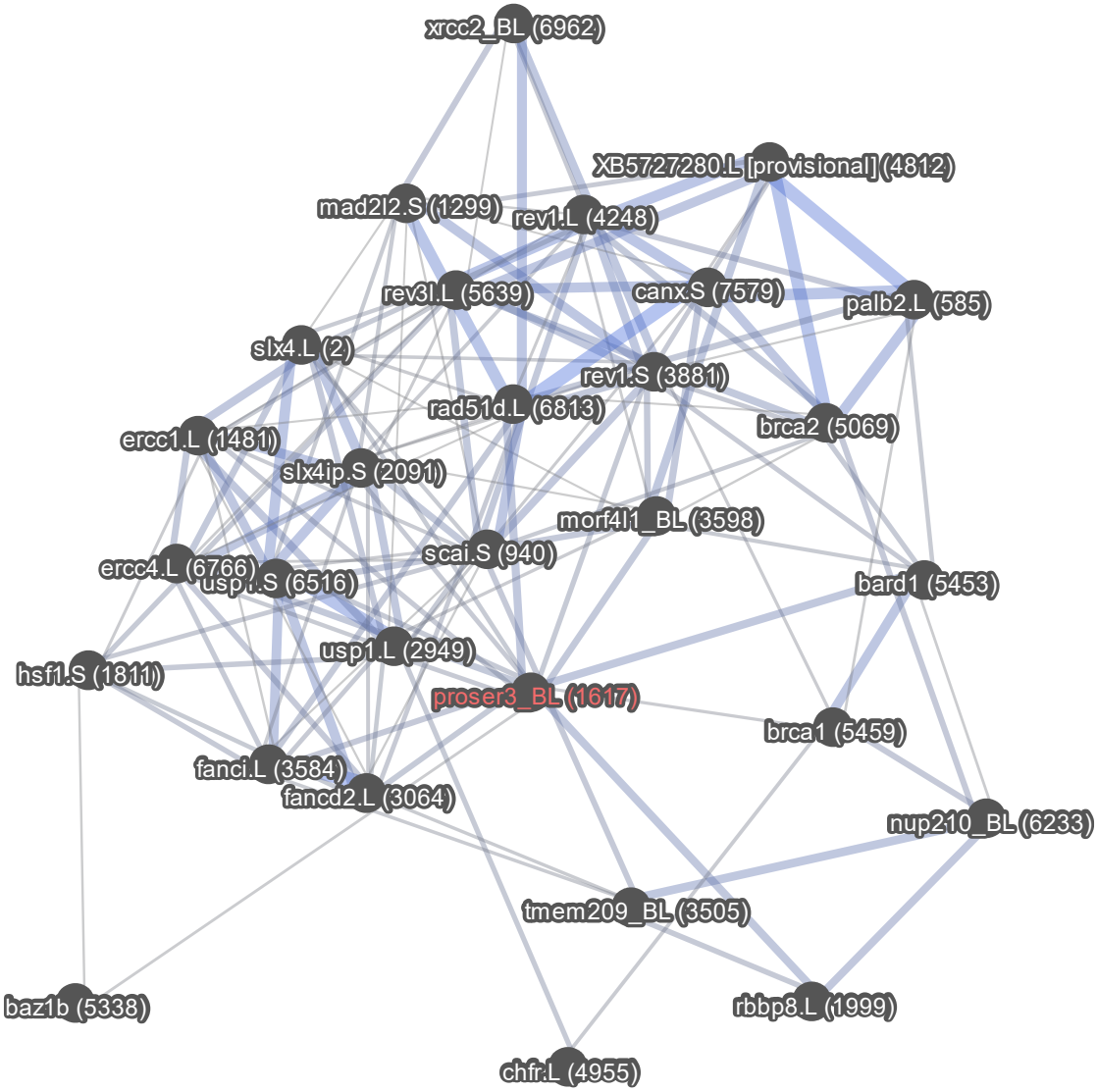


Figure 33 | DSB repair cluster derived from GSEA. The cluster shows the 30 highest scoring hits for the proteins BARD1, FBH1, XRCC2, BRCA2, BRCA1, BAZ1B, PALB2, MAD2L2, MORF4L1, RAD51D, SLX4, ERCC4 ERCC1, REV3L and RBB8, as identified for the GO term *"double-strand break repair"* in the previous GSEA. The protein PROSER3 is highlighted in red.

This reverse analysis has identified PROSER3 again in the direct proximity of the DSB related input list of proteins, together with further factors that play an essential role in DNA damage signaling and repair. Its highest correlating direct neighbors were RBBP9, BARD1, MORF4L1, RAD51D and ERCC4. Many of the shown interactors, among them BRCA1 and BRCA2, BARD1, RAD51, MORF4L1 (MRG15) and PALB2 were all shown to have a functional relationship in the repair of double-strand breaks [273,274]. This finding indicates that the investigated correlation between PROSER3 and the found interactors is not an artifact that occurs due to only a single, highly connected edge.

We investigated the closest correlating factors together with PROSER3 in more detail by plotting the individual t-test, in which it shows the highest p-value as volcano plot (Figure 34) and the underlying z-score and LFQ raw data. All of the 10 significant enrichments of PROSER3 have been identified in pDPC repair experiments. The shown experiment investigated the recruitment of proteins to a plasmid containing a site-specific DNA-protein crosslink (pDPC^{Lead}) against a geminin treated control.

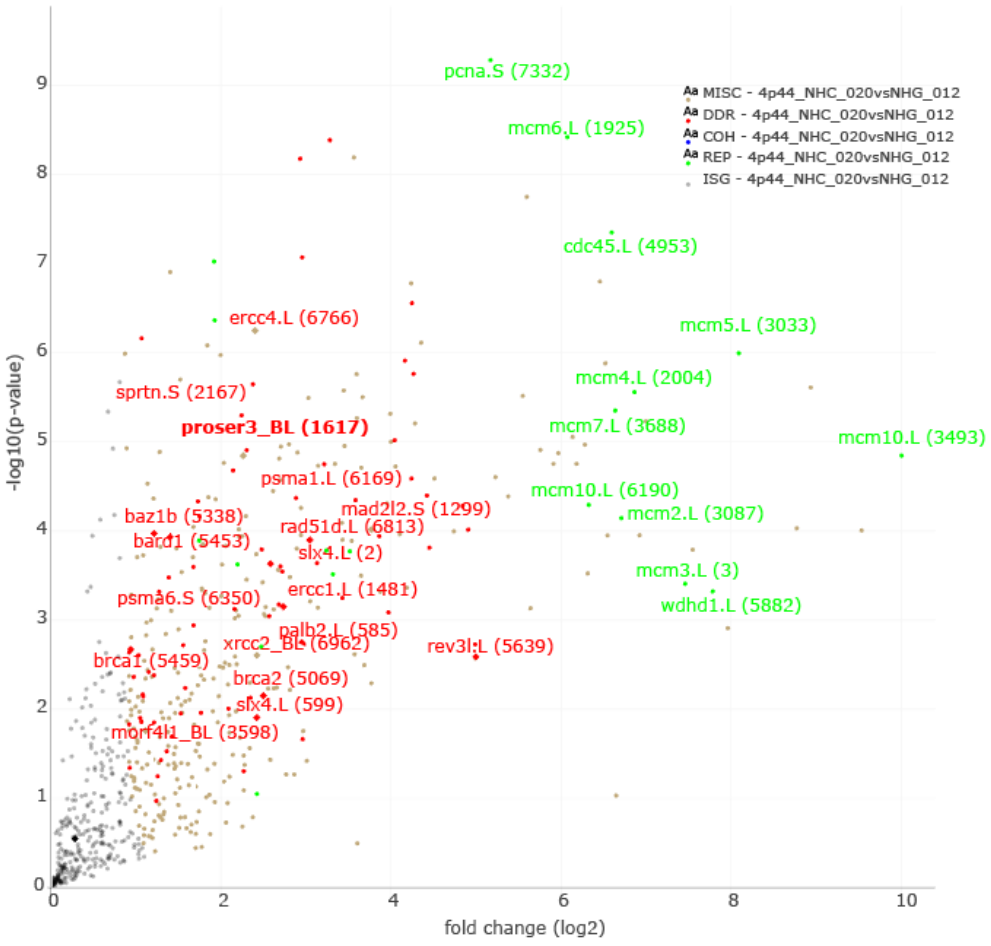


Figure 34 | Volcano plot showing PROSER3 and potential interactors. The volcano plot shows the recruitment of proteins to a plasmid containing a site-specific DNA-protein crosslink (pDPC^{Lead}) at 20 min against a geminin treated control. PROSER3 as well as potential interaction partners identified in the previous analysis are highlighted by red labels. Highest enriched replisome proteins are marked green.

The volcano plot shows the recruitment of repair factors to site-specific DNA-protein crosslinks after 20 minutes against a geminin treated control, where none of the replisome component are recruited to the plasmid. After 20 minutes we expect that the replisome already encountered the DPC on the leading strand template. This leads to an arrest of the replisome, as the CMG helicase is stalled by the adduct [166]. In agreement with this, the replisome components marked in green in Figure 34 are present. To bypass the DPC the MCM helicase is uncoupled to move past the adduct, leaving a gapped molecule behind [247]. This induces degradation of the DPC by SPRTN (Figure 34, marked in red) or the proteasome, which is likely recruited later [163]. The proteins PSMA1 and PSMA6 as well as the TLS polymerases REV3L and MAD2L2 are already present and marked in red. The presence of the HR proteins in the plot could also indicate a repair by template-switching and coupled recombinational repair. The exact mechanism of this repair pathway (reviewed in [275]) needs to be explored in further studies including the role of PROSER3 in this context, which has not been described yet. The related LFQ data and thereby measured intensity is comparably low (20.5 - 21) in both related treatments, indicating a lower overall abundancy and recruitment to the damaged plasmid (Supplemental Figure 9, 10). In the combined processing of all measurements in MaxQuant it was identified by 14 peptides, 13 of them unique, leading to the conclusion that it is not a false positive. Overall, this points out PROSER3 as potentially playing a role in the nuclease dependent repair of DNA-protein crosslinks, possibly by interacting with homologous recombination proteins.

In summary, the DNA Repair Atlas' intended function is to mine and visualize DNA repair data. This can be achieved by investigation of global or DNA lesion specific enrichment of significantly recruited DNA repair factors either in individual, or combined volcano plots or by creating subnetworks with a random-walk with restart from a dataset of significantly enriched repair factors. Alternatively, the DRA facilitates a bottom-up analysis starting from one or a list of proteins of interest by associating them to related proteins that play a significant role in the repair of various DNA lesions. This was demonstrated by a characterization of the dataset by combining series of

individual t-tests for several DNA lesions and further characterizing an exemplary subnetwork based on multiple described repair factors. The bottom-up analysis predicted the function of the still uncharacterized protein ”XELAEV_18036466mg”, or proser3_BL, that was shown as significantly enriched for the repair of pDPCs multiple times and found by independent cluster analysis in a module with multiple proteins involved in homologous recombination.

3.2 Systems approaches identify the consequences of monosomy in somatic human cells

The results presented in this chapter are a part of the Nature Communications publication "*Systems approaches identify the consequences of monosomy in somatic human cells*" (2021) [231].

Narendra Kumar Chunduri[1], Paul Menges[1], Vincent Leon Gotsmann[2], Xiaoxiao Zhang[3], Balca R. Mardin[4], Christopher Buccitelli[4], Jan O. Korbel[4], Felix Willmund[2], Maik Kschischo[3], Markus Raeschle[1], Zuzana Storchova[1]

[1] Dept. of Molecular Genetics, TU Kaiserslautern, Paul-Ehrlich-Strasse 24, 67663 Kaiserslautern, Germany.

[2] Group Genetics of Eukaryotes, TU Kaiserslautern, Paul-Ehrlich-Strasse 23, 67663 Kaiserslautern, Germany.

[3] University of Applied Sciences Koblenz, Joseph-Rovan-Allee 2, Remagen, Germany.

[4] European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstr. 1, 69117, Heidelberg, Germany.

doi: 10.1038/s41467-021-25288-x

3.2.1 Generation and analysis of monosomic human cell lines

Generation of monosomic cell lines

To create a system that allows the study of the consequences of monosomy in somatic human cells, cell lines that were derived from a TERT-immortalized human **R**etinal **P**igment **E**pithelium (RPE1) were generated in collaboration with Jan Korbel (EMBL, Heidelberg) and Rene Medema (NKI, Netherlands). Two approaches were used that involved the knock-out of p53 using the CRISPR-Cas9 or TALENs, as depicted in Figure 35 A, top, or knock-down depletion by stable integration of shRNA against TP53, as depicted in Figure 35 A, bottom. This depletion established several viable clones that were subjected to whole genomic sequencing. This revealed that several clones lost a chromosome, which means that they became monosomic.

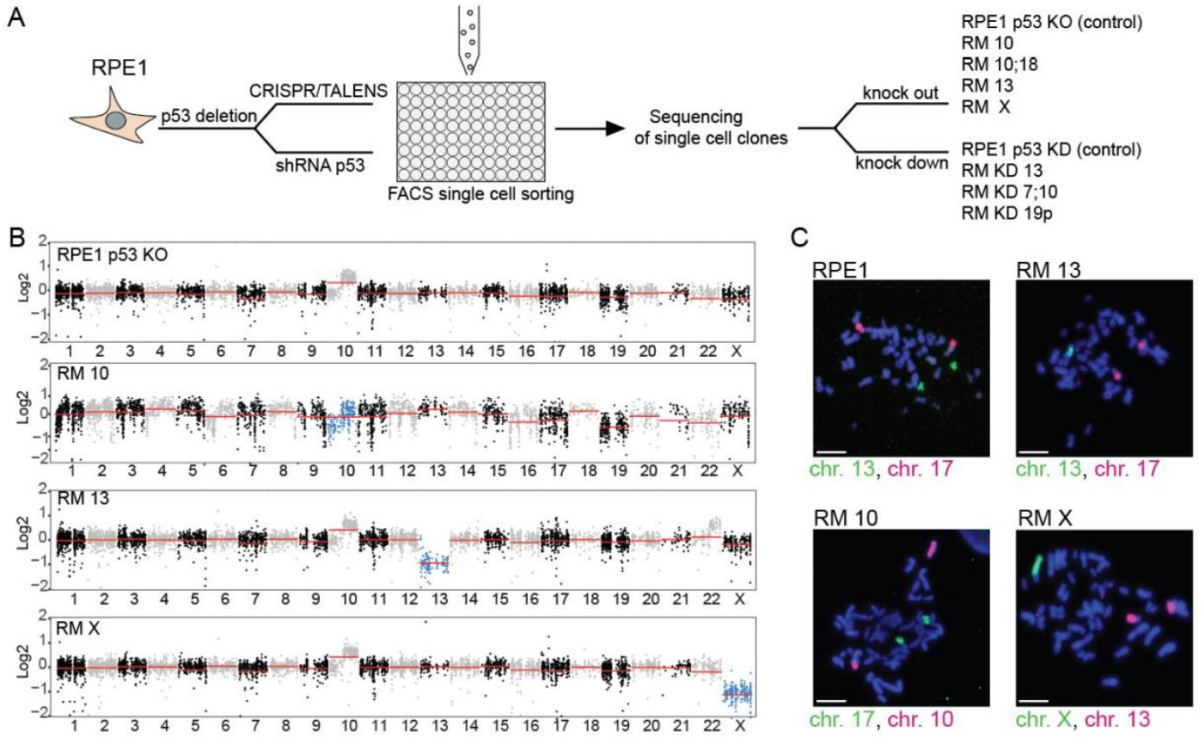


Figure 35 | Creation and validation of monosomic cells. (A) Schematic depiction of the construction of monosomic cells. p53 was mutated or depleted in RPE1-hTERT cell line via CRISPR/Cas9, TALENS, or shRNA expression. Clones arising from single cells were subjected to whole genome sequencing and monosomic clones were selected to be used in further experiments. (B) Read depth plots of all chromosomes in control and RM samples. Chromosome losses are marked in blue. Red lines indicate the copy number of each individual chromosome. Note that the parental RPE1 contains an extra copy of 10q that is preserved in all monosomic derivatives. (C) Chromosomal paints of monosomic cell lines. The painted chromosomes are indicated with respective colors. Scale bar – 10 µm (by Narendra Chunduri).

Whole genome sequencing (WGS) was performed as described in the Methods. The values were converted into log_2 data and the median coverage per read was calculated for all known genes with at least one mapped coverage. To shift the values around zero, the median coverage of each cell line was subtracted for all values, resulting in a normalized population centered on 0, as shown in Figure 35 B.

The shown “Location plots” were created by plotting the mapped non-zero coverage values of all gene positions per chromosome. A characteristic gain of the q-arm of chromosome 10, which is typical for the cell line RPE1, is also observed in these analyses, as well as in the previously published WGS analysis of the KD clones in [276]. Due to the increased genomic instability as a direct consequence to aneuploidy or the loss of p53, some additional partial aneuploidies can be observed, such as a partial loss of chromosome 19 for RM 10 and a gain of the q-arm of chromosome 22 for RM 13. The arising monosomic p53 knock-out (KO) cell lines were named according to their parental cell line **R**PE1-derived **M**onosomy (RM) and respective monosomic chromosome RM 10, RM 10;18, RM 13, RM X, as well as the knock-down (KD) cell lines RM KD 13, RM KD, 7;19 and RM KD 19p.

Transcriptome analysis reveals buffering of gene expression of monosomy-encoded genes towards diploid levels

Transcriptome analysis of the same monosomic and parental cell lines by NGS was conducted at the NGS-Integrative Genomics Core Unit (NIG), Institute of Human Genetics, University Medical Center Göttingen (UMG), as described in the Methods. The resulting dataset contained around 14,000 transcripts. The data was transferred to log_2 scale and the Fold Change (FC) to the parental changes was calculated.

In previous characterizations of gain of chromosome cell lines, it was shown that the protein expression of approximately 20 - 27% of proteins encoded on extra chromosome was adjusted towards the diploid level. In theory if all genes on tetrasomic chromosome would be expressed exactly according to their DNA copy number, then the median log_2 fold change to the diploid wildtype should be 1, while the diploid median should be 0. However, this level of expression was lower than expected, hence closer to the diploid level. This effect was called 'gene dosage adjustment' [79]. If genes encoded on a monosomic chromosome would be expressed according to the DNA copy number at 50%, the median log_2 fold change should be -1.

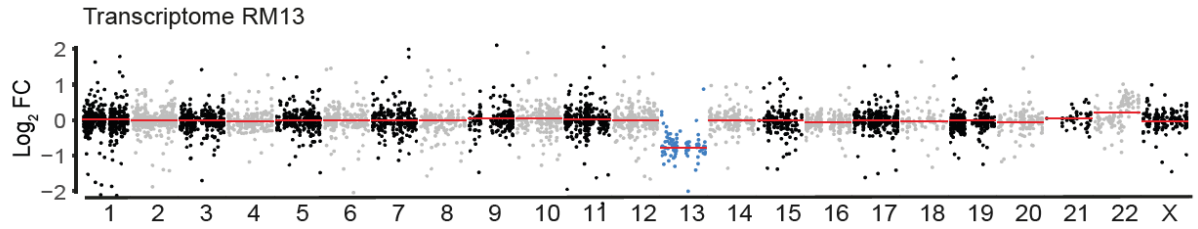


Figure 36 | Expression of mRNA encoded on the monosomic chromosome. The location plot shows the relative abundance of the transcriptome normalized to the diploid parental control for the cell line RM 13. The monosomic chromosome has been highlighted in blue, the median expression for each chromosome is indicated by a red line.

In our transcriptome analysis, we investigated whether the gene expression shows such an adjustment, or whether it is expressed according to the DNA copy number (Figure 36, Supplementary Figure 10). To facilitate this analysis, I calculated log_2 fold changes to the parental diploid cell line, grouped the gene expression by chromosome and plotted a 'location plot' for all cell lines: A grouped scatter plot that shows relative expression of each gene according to their chromosome location, as shown in Figure 36 and supplementary Figure S11.

Next, I calculated the median abundance of genes encoded on the aneuploid chromosome. Strikingly, the expression of mRNA on chromosome 13 did not decrease to the expected level of log_2 fold change -1, rather averaged around -0.747. This can trend be observed for all monosomic cell lines with the exception of RM X (Supplementary Figure S11).

p53 pathway activation does not alter the global response to monosomy

To address the question if p53 has an effect on the cellular response to monosomy, p53 expression was restored in RM 13 and RM 10;18 by using a doxycycline inducible expression system. The new cell lines were labeled with the suffix “ip53”. Transcriptome changes in monosomic and an RPE1 ip53 cell lines without and with restored ip53 after 48h of doxycycline treatment were measured by RNA sequencing.

The resulting transcriptome data was normalized to the corresponding parental control cell line to identify changes induced by p53 in monosomies. The data was filtered for log_2 fold changes outside the range of -2 to 2, hierarchically clustered by Euclidean distance and visualized as a heat map. Further, calculation of the median expression of all genes encoded on the monosomic chromosome of RM13 ip53 showed a median log_2Fc of -0.73 without and -0.8 with doxycycline, comparable to RM 13 shown above.

In conclusion, the similarity of the cell lines with restored p53, that is observable in Figure 37, suggests that functional p53 has only a mild effect on the global response to monosomy. This analysis was performed together with Narendra Chunduri.

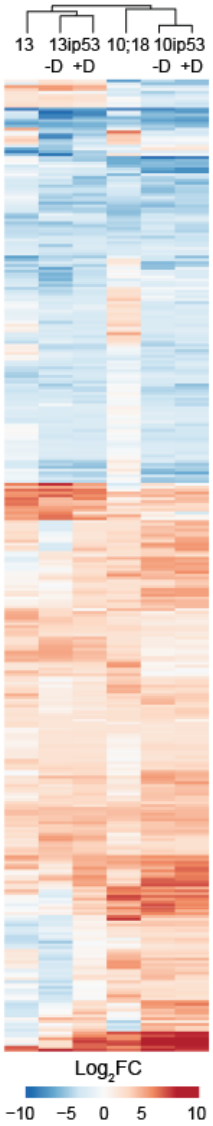


Figure 37 | Global effect of p53 expression in monosomies. The heat map shows differentially regulated mRNA expression compared to the diploid control. +/- D indicates doxycycline. The data was clustered by euclidian distance and shows FC bigger or smaller than +/- 2 (with Narendra Chunduri).

Proteomics analysis reveals further buffering of gene expression of monosomy-encoded genes towards diploid level

To investigate the effect of the loss of different chromosomes on the proteome of the cell, both quantitative label-free and multiplexed Tandem-Mass-Tag (TMT) mass spectrometry analysis was performed for RM X, RM 10;18, RM 13, RM 19p and respective parental controls, as described in the Methods.

In brief, all proteomics data was processed with MaxQuant, version 1.6.3.3. All data was searched against the *Homo sapiens* reference proteome database (UniProt: UP000005640) with a peptide and protein FDR of less than 1%. The identified protein groups were filtered to remove contaminants, reverse hits and proteins identified by site only. Next, protein groups which were identified more than two times in at least one group of replicates (N=4) were kept for further processing resulting in a set of 5887 Protein groups in total for the TMT labeled measurement. For LFQ, Protein groups which were identified more than three times in at least one group of replicates (N=4) were kept for further processing, resulting in a set of 5727 Protein groups in total.

TMT reporter intensities were cleaned for batch effects using the R package LIMMA [58] and further normalized using variance stabilization [59]. Figure 38 A shows an observable batch effect in principal component 2 on the left as well as its removal via normalization on the right side.

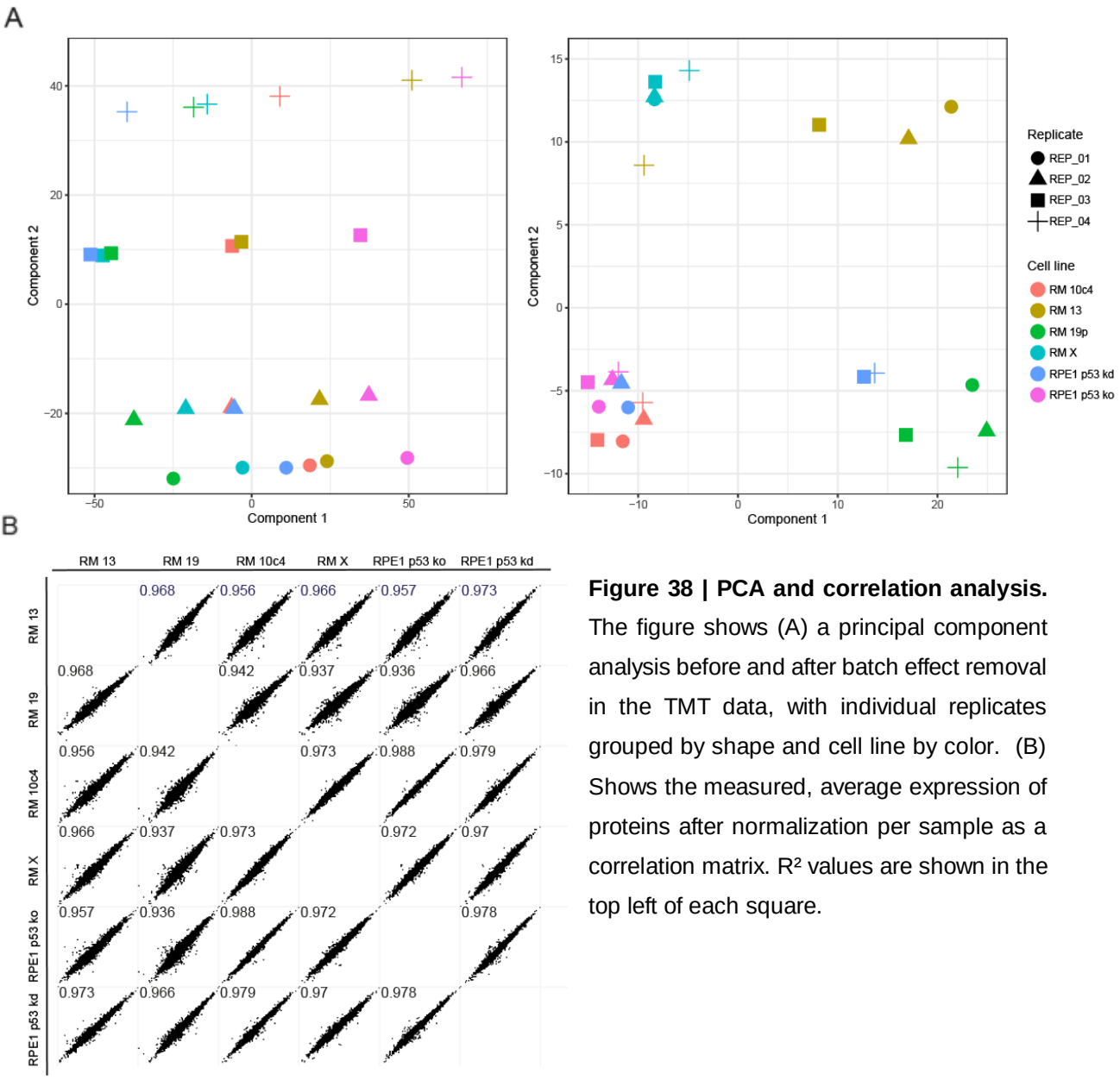


Figure 38 | PCA and correlation analysis. The figure shows (A) a principal component analysis before and after batch effect removal in the TMT data, with individual replicates grouped by shape and cell line by color. (B) Shows the measured, average expression of proteins after normalization per sample as a correlation matrix. R^2 values are shown in the top left of each square.

For all monosomic cell lines, the $\log_2$ median intensity of the replicates was calculated. This resulted in a set showing overall a low variance between the individual samples, indicated by R^2 values between 0.936 and 0.988 as shown in Figure 38 B. The $\log_2$ median intensity of the replicates of the wild type parental cell line was subtracted to calculate comparable fold changes.

The calculated fold changes show an overall strong similarity between the LFQ and TMT measurement (Figure 38), indicated by a Spearman Rank correlation: 0.634 for all expressed proteins, 0.67 for only chromosome 10 and 18 (Figure 39).

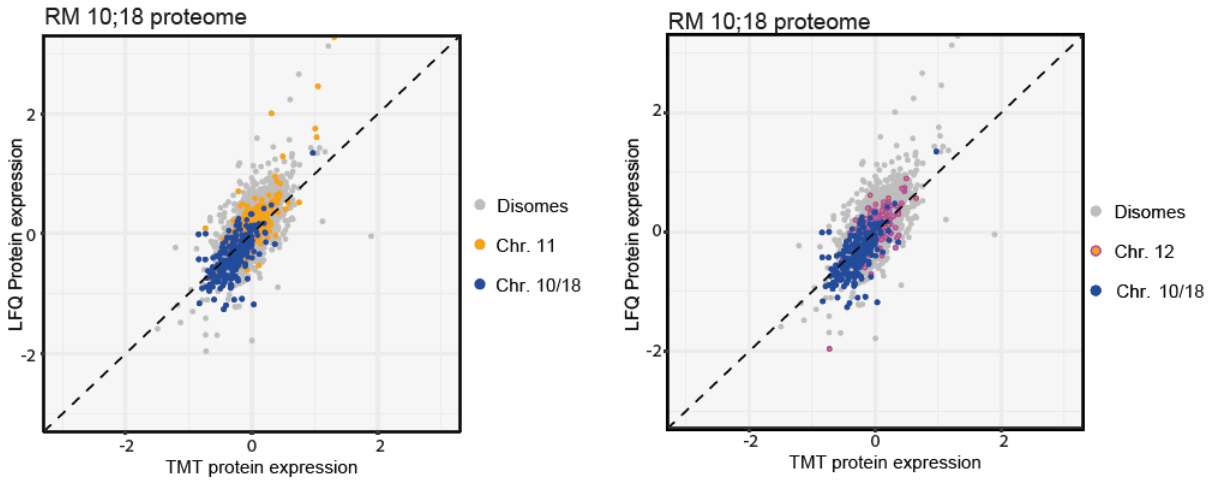


Figure 39 | Comparison of LFQ and TMT. The scatter plots show the log_2 fold changes of RM 10;18, measured by LFQ and TMT, towards a diploid parental cell. The monosomic chromosomes are highlighted in blue, two disomic chromosomes are colored in yellow (Chr.11) and pink (Chr.12).

In conclusion the proteome measurement and normalization results in a highly comparable dataset, that facilitates the following integrative, systematic analysis of monosomic human cells. To investigate if the effect of the loss of chromosome on gene expression is comparable to the previously shown transcriptome levels I calculated the median protein expression similarly as for the transcriptomics data and plotted them as 'location plot' (Figure 40, Supplementary Figure 11).

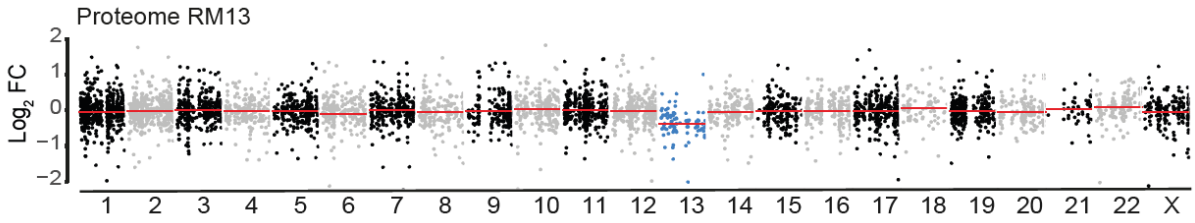


Figure 40 | Expression of proteins encoded on the monosomic chromosome. The location plot shows the relative abundance of the proteome normalized to the diploid parental control for the cell line RM 13. TMT data was used for this analysis. The monosomic chromosome has been highlighted in blue, the median expression for each chromosome is indicated by a red line.

Strikingly, the calculated median protein expression of genes encoded on the lacking chromosome of RM 13 shows an even further increase towards diploid levels than transcriptome with a median of -0.37 instead of -1 according to the DNA copy number. This can be observed for all cell lines (Supplementary Figure S11).

3.2.2 Expression of genes encoded on the monosomic chromosomes is adjusted by transcriptional and post-transcriptional mechanisms

To investigate both protein and mRNA expression on the monosomic chromosome together in more detail, I created density plots for all monosomy-encoded proteins and mRNA together with the calculated median expression (Figure 41, top). As a control, the respective disomic chromosome expression is plotted similarly (Figure 41, bottom).

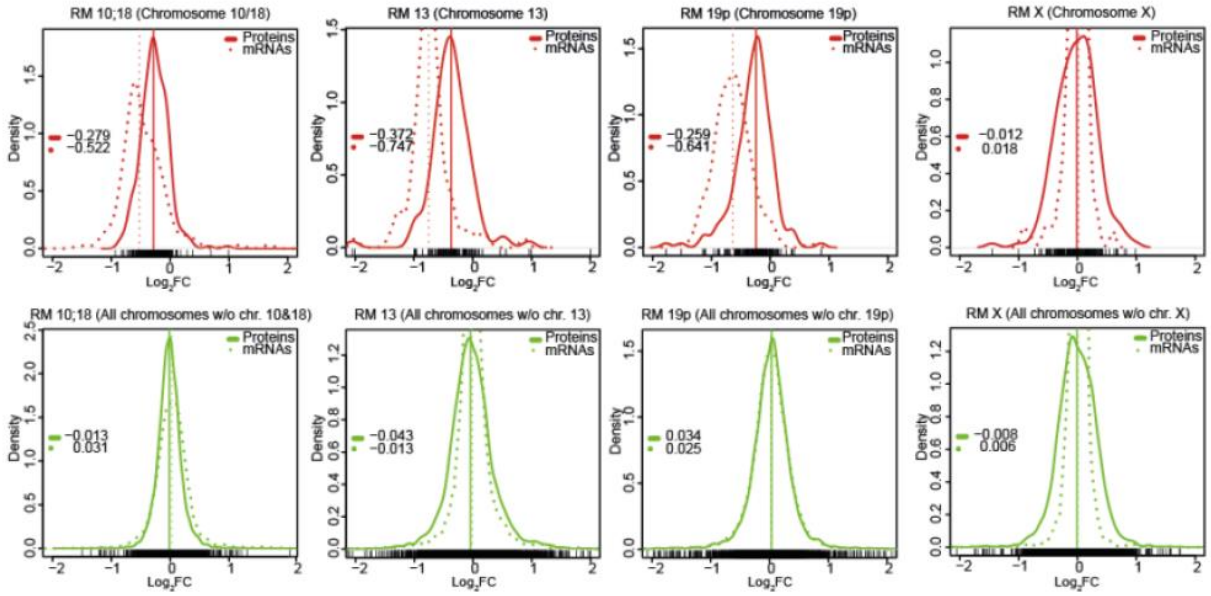


Figure 41 | Density plots of protein and mRNA abundance of monosomic cell lines. The figure shows the monosomy-encoded protein and mRNA expression (top, red) of all cell lines with the disomic-encoded ones as control (bottom, green). Dotted lines show mRNA, solid lines relative protein expression as density plot.

Indeed, the mRNA level of monosomically encoded genes ranges from -0.52 to -0.75, higher than expected from the DNA copy number. The proteome levels range between -0.26 and -0.37, even closer adjusted to the diploid level and higher than expected from DNA copy number. The exception is RM X, which remains close to zero, as expected, as in XX cells one copy of RM X is already transcriptionally inactivated by XIST mediated silencing with the exception of a set of escaping genes [277].

In a consecutive analysis, we combined all expression values encoded on a monosome in all cell lines, with the exception of chromosome X, which is shown separately. The measured monosomically encoded ‘total’ median of protein abundance in this analysis was -0.25 and the median of all corresponding transcripts -0.59. Both values are significantly higher than expected from the DNA copy number, as shown in Figure 42 A.

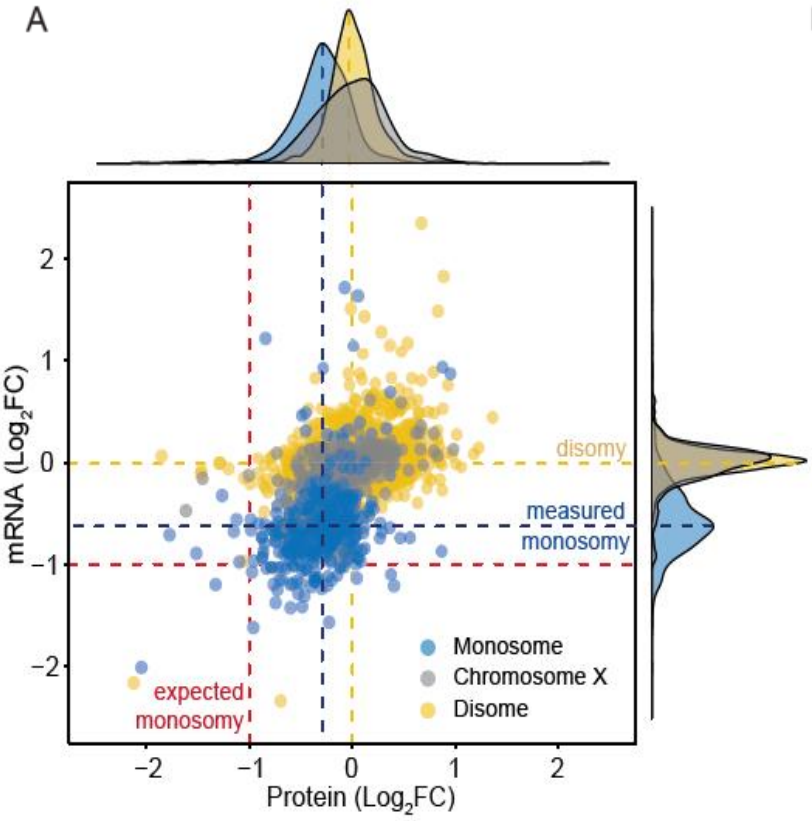


Figure 42 | Combined expression of genes encoded on monosomic cell lines. The left side (A) shows a scatter plot of the combined $\log_2$ fold change mRNA and protein expression of genes encoded on monosomic chromosomes (blue), disomes (yellow) and chromosome X (grey). The density plots located on the axis of the plot show the distributions of mRNA and protein values.(B) shows a bar plot indicating the percentage of genes that could be assigned to groups based on a $\log_2$ fold change cutoff: -0.5; up;up (both mRNA and protein more than -0.5), down;up (RNA<-0.5, protein>-0.5), down;down (both mRNA and protein log2 fold changes less than -0.5), up;down (mRNA>-0.5, protein< -0.5). This analysis was performed with Markus Raeschle and Narendra Chunduri.

The monosomically encoded genes were assigned into four groups based on a log2 cutoff, as shown and described in Figure 42 B. This showed that the expression of approximately 30% of genes encoded on monosomies is adjusted towards the diploid level transcriptionally and 45% post-transcriptionally. Furthermore, only 20% of monosomically encoded genes were expressed at a relative protein and mRNA abundance with a $\log_2$ fold change lower than -0.5. This suggests that the gene

expression of monosomically encoded genes is adjusted at both transcriptional and post-transcriptional level.

Previous analyses [79] could associate this adjustment closely to proteins that are part of subunits of multimolecular complexes, as annotated by the CORUM database [72]. We analyzed our proteome data similarly, by creating two subsets of the monosomically encoded proteins that are part of any CORUM annotation and part of none (Figure 43).

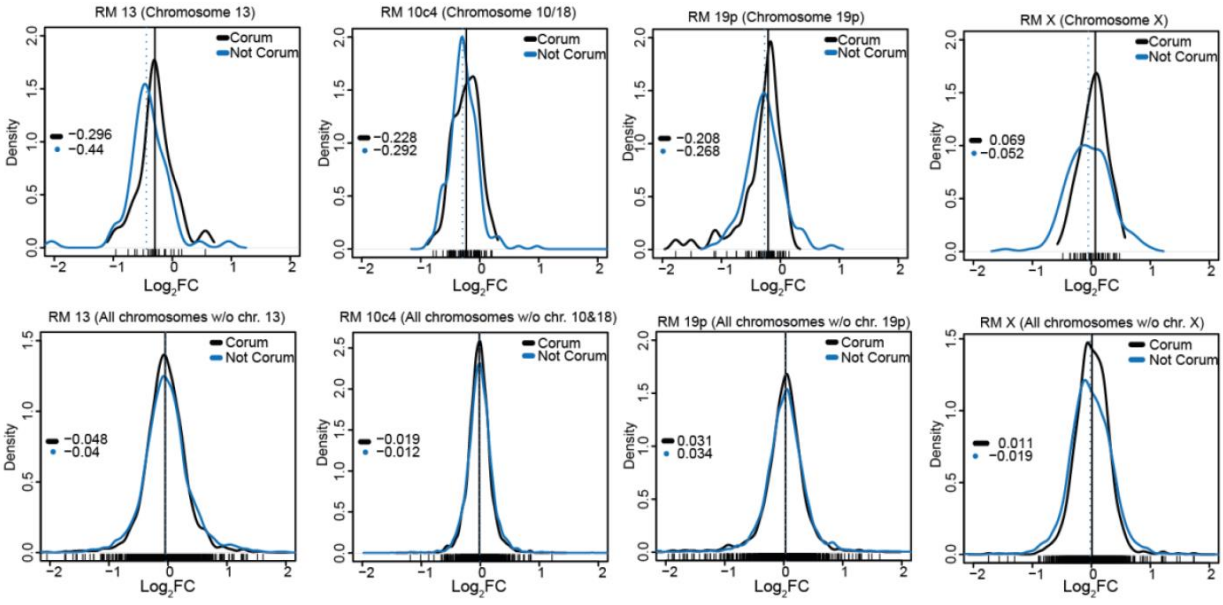


Figure 43 | Compensation for proteins that are part of multimolecular complexes. Shown are density plots of all monosomically encoded proteins of the related cell lines that are part of a CORUM annotation (top), (black line) or not (blue line). As control all disomically encoded proteins are plotted (bottom).

This analysis revealed only a subtle shift that indicates that this mechanism does not translate from trisomic to monosomic and doesn't play an important role in compensation of monosomically encoded genes. We concluded that the mechanisms responsible for the "dosage compensation" differ for chromosome loss compared to chromosome gain. To identify monosomy-specific trends I further investigated the annotations for all cytosolic and membrane proteins encoded on monosomic chromosomes (Supplementary Figure 12). This analysis also shows a subtle shift in median, that similarly to CORUM does not link to a monosomy specific dosage compensation mechanism.

To better analyze the dosage compensated proteins on all monosomies and adjust for chromosome specific expression changes, a subset of genes was defined as significantly compensated on post-transcriptional level if (i) the encoding gene is located on a monosomy, (ii) the log_2 fold change of the protein is below 0, (iii) the difference in log_2 fold change expression between protein and transcript is at least 0.34, a value defined as difference between the global median of transcript (-0.59) and protein expression (-0.25) of monosomically encoded genes. This yielded a set of 194 genes derived from in total 600 monosomically encoded genes with both detected transcript and protein expression value. The top 50 genes based on this difference is shown in Figure 44, together with the relative proteome (TMT) and transcriptome expression.

A gene set enrichment analysis of all 194 dosage compensated genes was performed with the program EnrichR [74,278] for Gene Ontology biological process and cellular compartment (2021). The results of this analysis are shown in Figure 45. *RRNA metabolic process*, *processing* and the *ribosome biogenesis* were identified as the most significantly enriched, compensated biological processes with

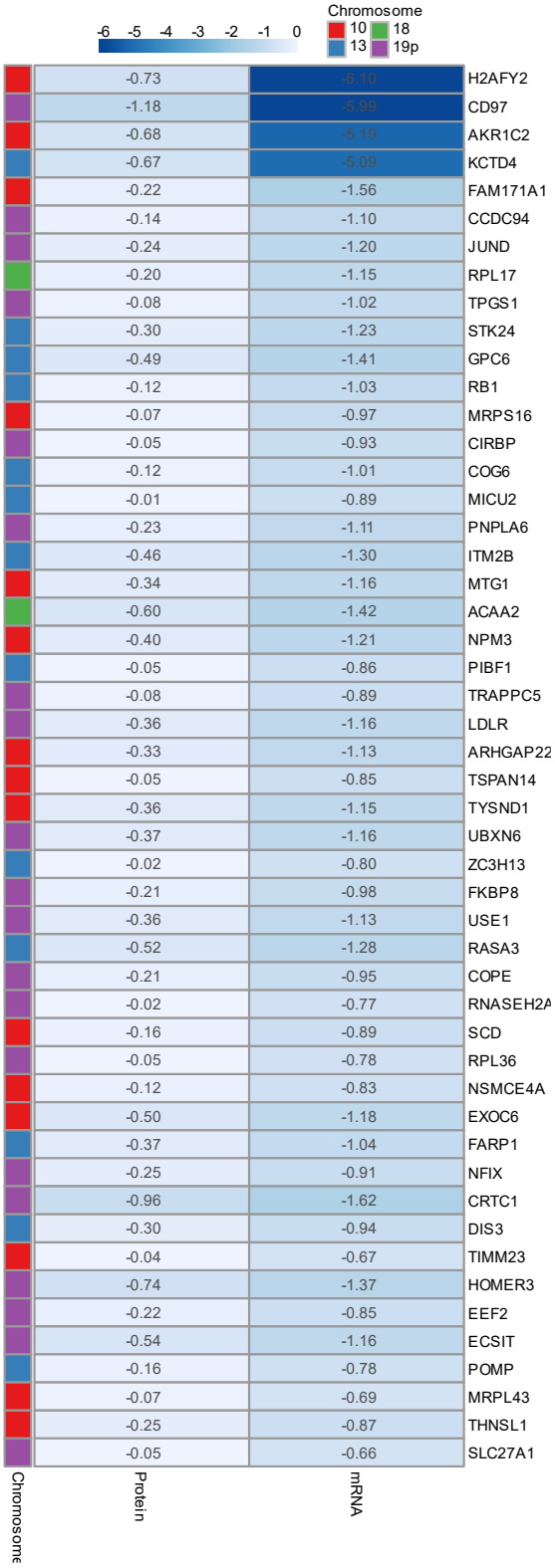


Figure 44 | Top 50 post-translationally compensated genes. The heatmap shows the compensated 50 genes with the highest difference in log2 fold change of protein - mRNA expression, sorted by difference. Chromosome location has been color-coded (red:10, green:18, purple :19p, blue:13).

RPS15 (19p), CUL4A (13), PDCD11 (10), DIS3 (13), BMS1 (10), WDR18 (19p), RPL36 (19p), RPL17 (18), RPS24 (10), EXOSC1 (10) as the genes responsible for this enrichment. *Mitochondrial* and *organelle inner membrane* as well as *mitochondrial membrane* and the *uniplex complex* were the most enriched, compensated cellular compartments.

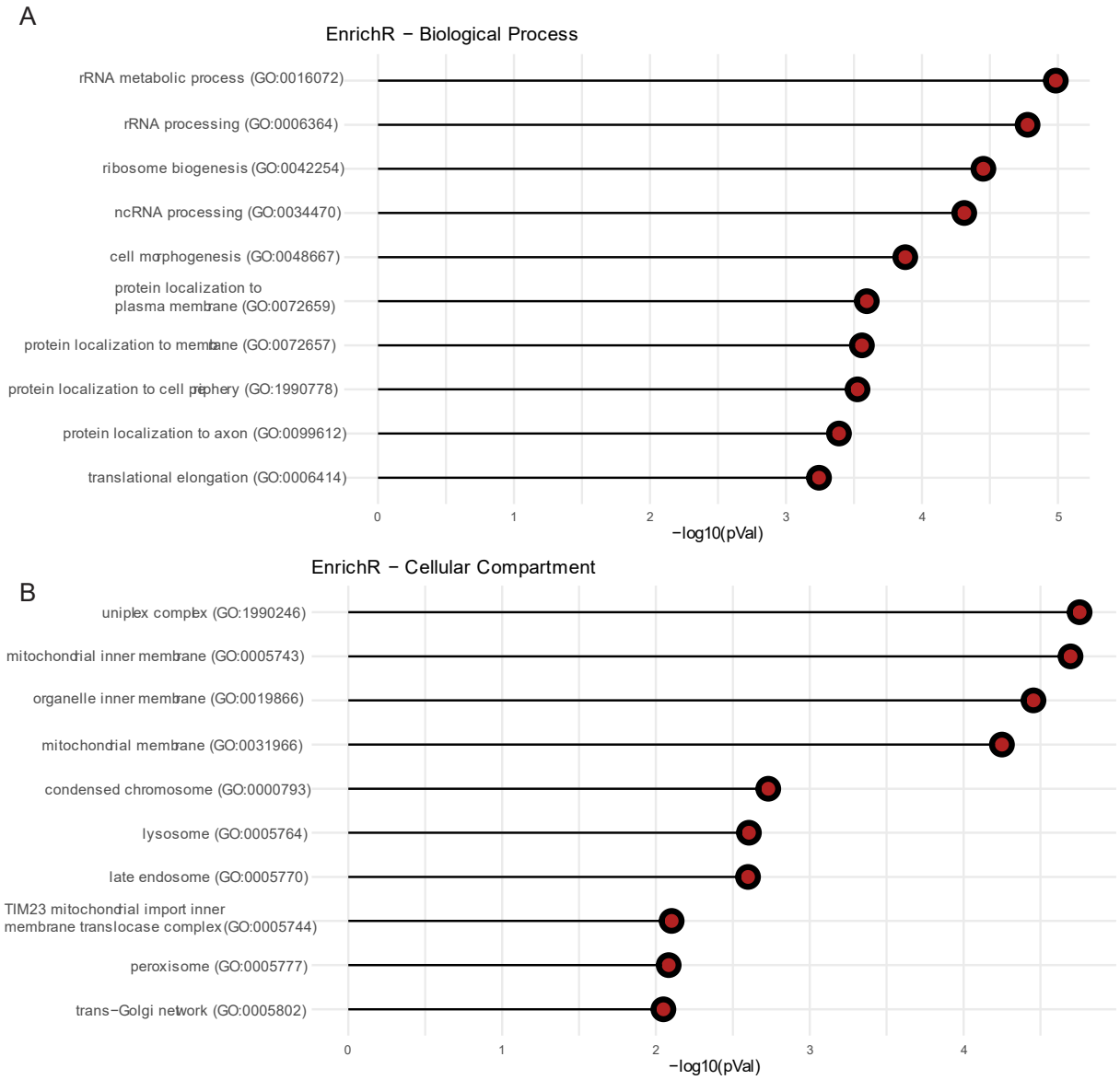


Figure 45 | Gene Set enrichment analysis of dosage compensated genes. The two bar plots show the result of a Gene Set enrichment performed with EnrichR, for a list of 194 compensated genes. Top shows enriched biological processes, bottom cellular compartments, as defined in Gene Ontology (2021). Significance has been calculated based on a Fisher Exact Test. The resulting p-value has been transferred to -$\log_{10}$ scale.

Genes responsible for the enrichment of these cellular compartments again overlap and are SLC27A1 (19p), MRPS16 (10), TIMM13 (19p), TIMM23 (10), ECSIT (19p), MRPL43 (10), MRPL4 (19p), MTG1 (10), YME1L1(10), MICU2 (13), MICU1 (10),

CHCHD1 (10), MCU (10). It is at this point unclear, whether the mechanism of dosage compensation specifically adjusts genes involved in ribosomal biogenesis, to compensate for the loss of chromosome and ribosomal haploinsufficiency or whether this enrichment is based on individual deregulation of genes on protein and transcript level as consequence of chromosome loss and the resulting stresses.

Furthermore, the identified cellular compartments are based on genes that localize, with the exception of MICU2 (13), to chromosome 10 and 19p. It is unclear, whether a mechanism specifically re-adjusts the protein content of genes that are localized on those chromosomes, or whether the adjustment specifically targets a part of the mitochondrial membrane or electron transport chain.

Intriguingly, if the filter is reversed to investigate significant transcriptionally dosage compensated genes, only 7 genes pass the (iii) filter criterion as shown in Figure 46. The subset of transcriptionally regulated genes in total consists of 84 genes if filtered for any compensation (protein > mRNA expression) for mRNA expression of $\log_2$ fold change below 0. The strong drop-off in difference confirms a broader regulation on a post-transcriptional than transcriptional level for monosomic cell lines. The resulting set of genes shows two members associated to the *mitochondrial respiratory chain complex I,* yet no significant other enrichment was found due to a small sample size.

In conclusion, there are mechanisms to alleviate the consequences of the effect of a chromosome loss, but the exact mechanisms and scope needs to be further investigated.

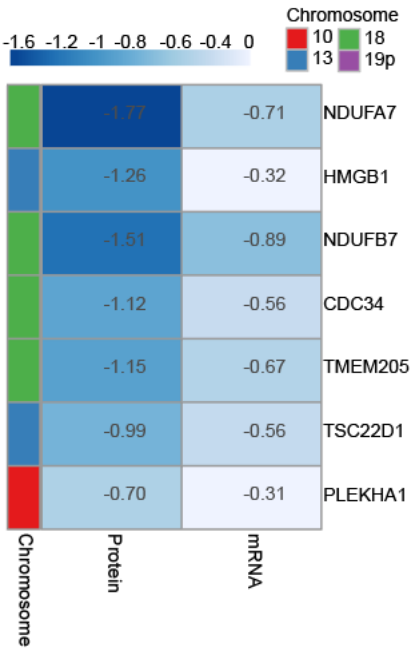


Figure 46 | 7 transcriptionally compensated genes. The heatmap shows 7 compensated genes with the differences in log2 fold change of mRNA - protein expression above 0.34, sorted by difference. Chromosome location has been color-coded (red: 10, green: 18, purple 19p, blue: 13).

Comparison of dosage adjustment in monosomic and trisomic cells

In a subsequent analysis, we compared all mRNA and protein abundancies localized on monosomy against those localized on the trisomic chromosomes 5 and 12 of RPE 5/3 12/3 and 21 of RPE 21/3 to determine the scope and scaling of dosage compensation effects.

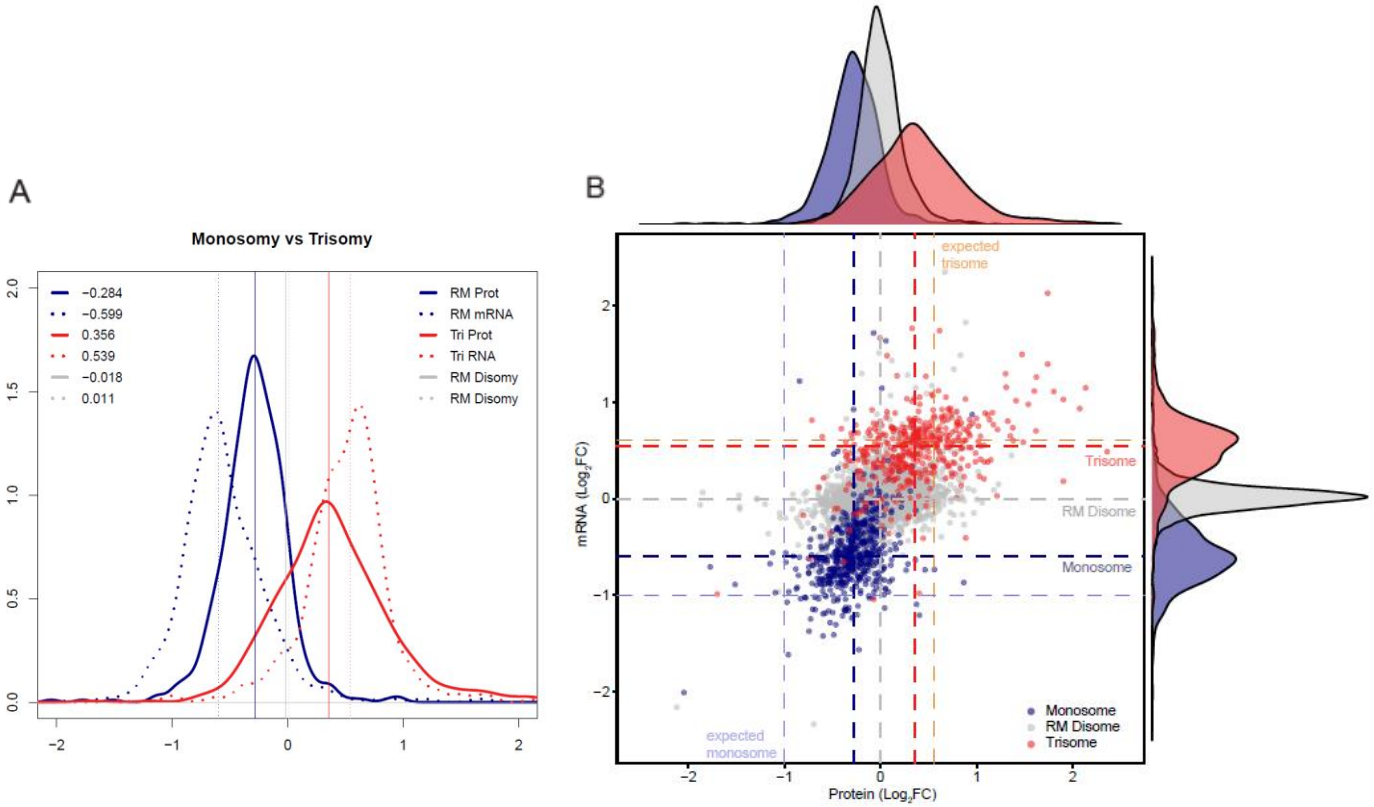


Figure 47 | Scaling of dosage compensation in mono- and trisomic cells. A comparison between mRNA and proteome expression of all monosomically and trisomically encoded genes was depicted in a density plot in (A) and scatter plot in (B). The density plot shows mRNA and protein encoded on trisomic chromosomes in red, on monosomies in blue and disomies derived from the TMT measurement in grey (dotted: mRNA, solid: protein). (B) Shows the scatter plot with the expected median and measured abundancy.

The trisomically encoded proteins and mRNA were combined similarly to the monosomically encoded values. In Figure 47 A & B the median expression of matched monosomically and trisomically encoded genes is shown, indicating a stronger dosage compensation effect on mRNA and protein level for a loss of chromosome compared to a gain of chromosome, indicated by the median relative expression for monosomically encoded protein of $\log_2$ fold change of -0.28 (relatively compensated by 72% towards diploid levels) and mRNA of -0.59 (compensated by 41%) with an

expected expression of -1 and trisomically encoded protein of 0.36 (compensated by 38%) and mRNA 0.53 (compensated by 9%) with an expected expression of approximately 0.58.

This indicates a compensation mechanism affecting mainly the protein expression levels in trisomies, as further indicated by the compensated proteins being part of multimolecular complexes [79], and a more expansive, mixed dosage compensation mechanism for monosomies. They not only show a generally stronger gene dosage compensation effect, but it also affects both protein and mRNA expression, hence acting on a transcriptional and post-transcriptional level.

Of note, the comparison is based on p53 proficient trisomy against p53 deficient monosomy cell lines. As shown in Figure 37, neither the global cellular response to monosomy, nor the median expression of genes encoded on the monosomic chromosome did show significant changes in response to p53 activation. Therefore, a connection of p53 expression and dosage compensation is unlikely.

Technical bias does not affect dosage compensation

As the proteome data was acquired by multiplexed TMT labeling it suffers from ratio compression, an effect that leads to a systematic underestimation of peptide/protein ratios [41]. The investigation of gene dosage compensation relies on high accuracy measurement, because only mild differences between mRNA and protein abundancy can be expected. Hence, to exclude any impact of technical factors, we repeated the analysis for proteome data acquired for the label free measurement of the cell lines RM 10;18, RM 13 and RM 19p. As shown in Figure 48 A & B the difference between TMT and LFQ measured monosomically encoded, relative protein expression is below 0.3, hence dosage compensation is observable in both datasets. This is further shown in the scatter plot in Figure 48 C.

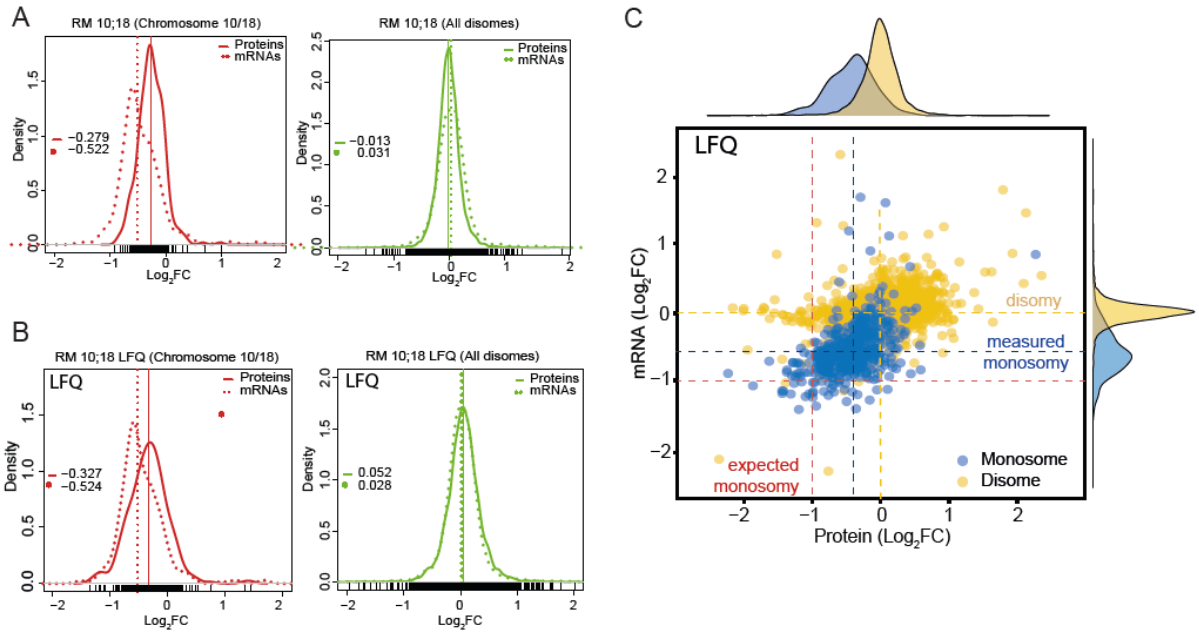


Figure 48 | Comparison of dosage compensation in TMT and LFQ. The figure shows an overview of dosage compensation in label free quantified proteome data compared to TMT (A) and (B) show the density plots for monosomically encoded protein (solid line) and mRNA (dotted line) expression values for TMT(A) and LFQ (B) with the disomically encoded genes as control (green). (C) shows log2 fold changes of monosomic LFQ and mRNA data in a scatter plot, showing the measured and expected median expression (blue/red dotted line) and the disomically encoded proteins (yellow dotted line).

A loss of precision was shown in a subsequent analysis of the raw measured intensities, which were not label-free quantified.

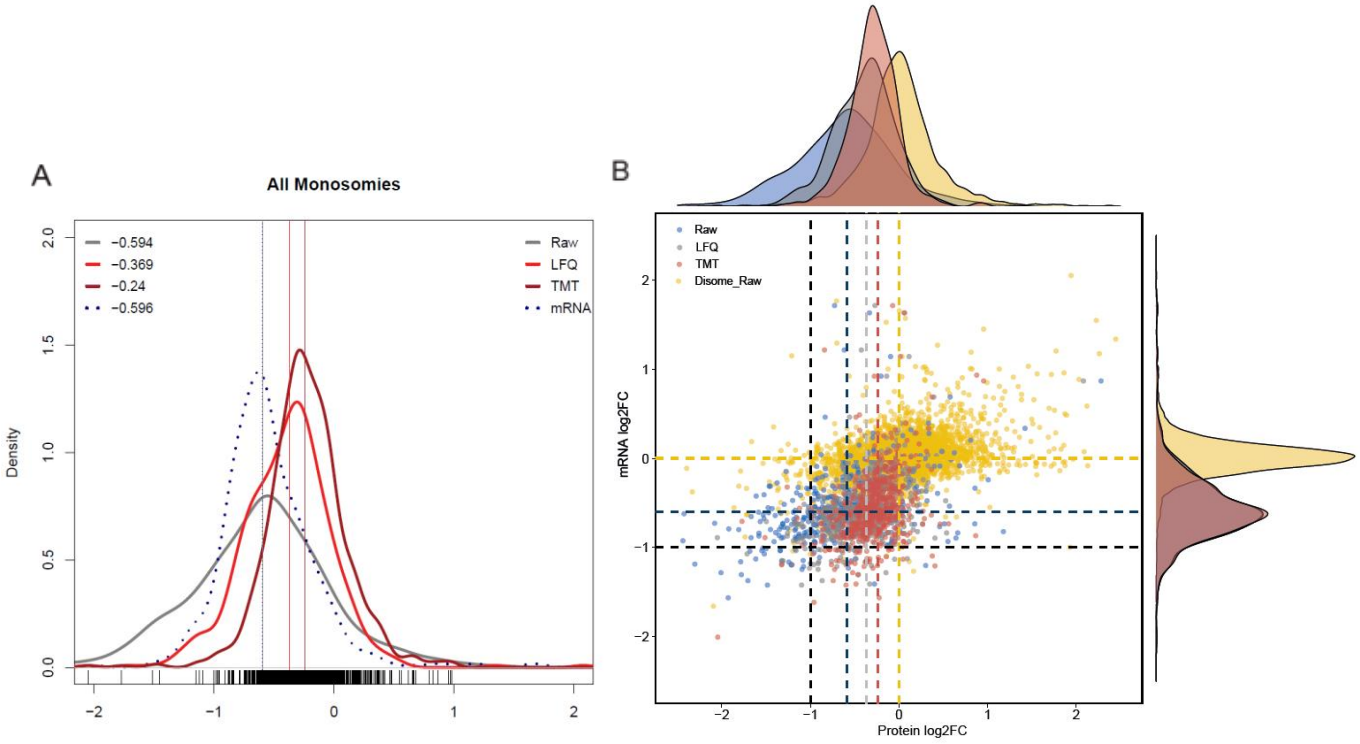


Figure 49 | Comparison of data acquisition methods for monosomically encoded proteins. The figure shows a density plot in (A), which shows TMT (dark red), LFQ (red) and raw intensity (grey) based relative protein expression of monosomically encoded proteins. (B) shows log2 fold changes of monosomic TMT, LFQ, raw intensity as well as mRNA data in a scatter plot, showing the measured and expected median expression (blue/red dotted line) and the disomically encoded proteins (yellow).

I performed this analysis to exclude any potentially introduced normalization bias that might affect dosage compensation effects. As shown in Figure 49 A & B, the observed difference between the median mRNA expression of all monosomically encoded transcripts to the related protein differs in TMT (0.356) from LFQ (0.227), and cannot be observed when raw intensities are analyzed (0.002).

The overall increasing width together with the lower peak of the distribution of each data acquisition and normalization method indicates a higher variance in protein group intensity, which overall has a negative effect on the functional analysis. Concluding, a systematic bias both by ratio compression or label-free normalization, that specifically affects dosage compensation could not be shown.

In summary, a gene dosage effect which adjusts both the monosomically encoded protein as well as mRNA expression towards diploid level was observed in all monosomic cell lines except RM X. This effect likely takes place on both a transcriptional as well as post-transcriptional level and appears to be stronger regulated for loss than gain of chromosome. The exact mechanisms that determine the scope and specific targets remain to be investigated.

3.2.3 Integrative systems analysis reveals pathway changes in proteome and transcriptome in response to loss of chromosome

Pathway deregulation in response to monosomy

A general difficulty in the functional interpretation of integrated omics data from sequencing and spectrometry is is a low direct correlation between the different measurements. Consistent with previous studies, which analyzed mRNA and protein correlation in diverse cancer and healthy cells [279] , the Spearman Rank correlation of the monosomic protein and mRNA was rather low and ranged in the lower positive region between 0.282 and 0.524 (Figure 50 A).

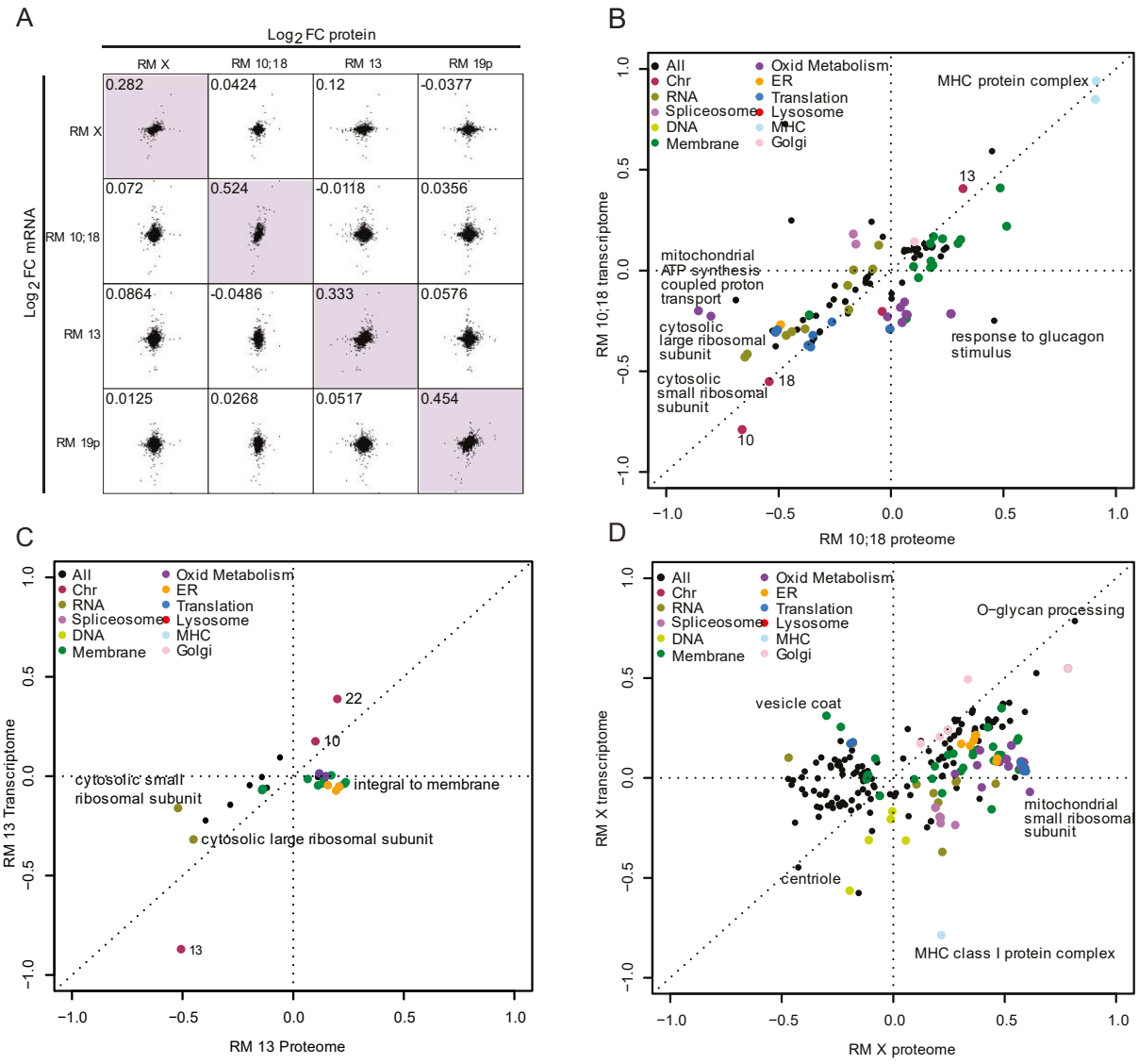


Figure 50 | Correlation and 2D annotation enrichment of proteome and transcriptome data. The figure shows the correlation of proteome and transcriptome data as well as pathway enrichment analysis performed in Perseus. (A) shows the Spearman rank correlation of relative abundances of mRNA and protein values for all monosomies as correlation matrix. (B) to (D) show the 2D annotation enrichment scores of RM 10;18, RM 13 and RM X for Gene Ontology biological process and cellular compartment, as well as the chromosome annotation. The enrichment is controlled by a Benjamini Hochberg FDR threshold of 0.02. A positive score indicates up-, and a negative downregulation. For easier comparison, annotations have been manually assigned to a color-coded legend, as indicated in the top left of each plot.

To gain information about the global consequences of the loss of a chromosome despite this, we performed two-dimensional pathway enrichment analysis for the transcriptome and proteome data. This gene set enrichment analysis calculates enrichment scores for each annotated gene set, derived from the ranked relative expression of all members of the annotated set [78]. This has the advantage of yielding easily comparable results as the datasets are not directly compared with each other, but rather via independent, rank-based scores. This allows the comparison of transcriptome and proteome of the monosomic cell lines, as shown in Figure 50.

Overall, the pathway analysis shows a positive correlation of enriched pathways for proteome and transcriptome of all monosomic cell lines, indicated by the proximity of enrichments towards the dotted, diagonal line (Figure 50 B to D). Strikingly, loss of chromosome shows both a heterogeneous pathway deregulation based on which chromosome is lost, as well as a common trend in the overall deregulation between proteome and mRNA. The two ribosomal subunits appear to be commonly impaired in response to any chromosome loss in both protein and mRNA level, as observed for RM 10;18 proteome and transcriptome (Figure 50 B), RM 13 proteome and transcriptome (Figure 50 C) and RM X proteome (indicated here by the enrichment scores of 0.12 for transcriptome and -0.46 for proteome for the term *cytosolic small ribosomal subunit*) (Figure 50 D). The pathway *O-glycan processing* is specifically enriched in proteome and transcriptome in response to loss of chromosome X and *MHC protein complex* for loss of chromosome 10 and 18.

Additionally, for the cell lines RM 10;18 and RM 13 we observed a trend that a part of the enriched pathways are oriented slightly above the diagonal correlation line, indicating a set of genes that show a slightly less deregulated protein than mRNA, which is not observed in RM X. This trend is in line with the findings described in the

previous section. Overall, the results indicate variable responses to individual chromosome loss. The only shared pathway deregulation in all monosomic cells indicated impaired translation and ribosome.

The proteome of the monosomic cell lines has been comparatively analyzed in the same way as shown in Figure 50 and Supplementary Figure 13. This further confirmed that the deregulation of pathways appears to differ depending on individual chromosome loss, indicated by the amount of annotated biological processes and cellular compartments located in the anticorrelating regions (top left/bottom right) in the 2D plot.

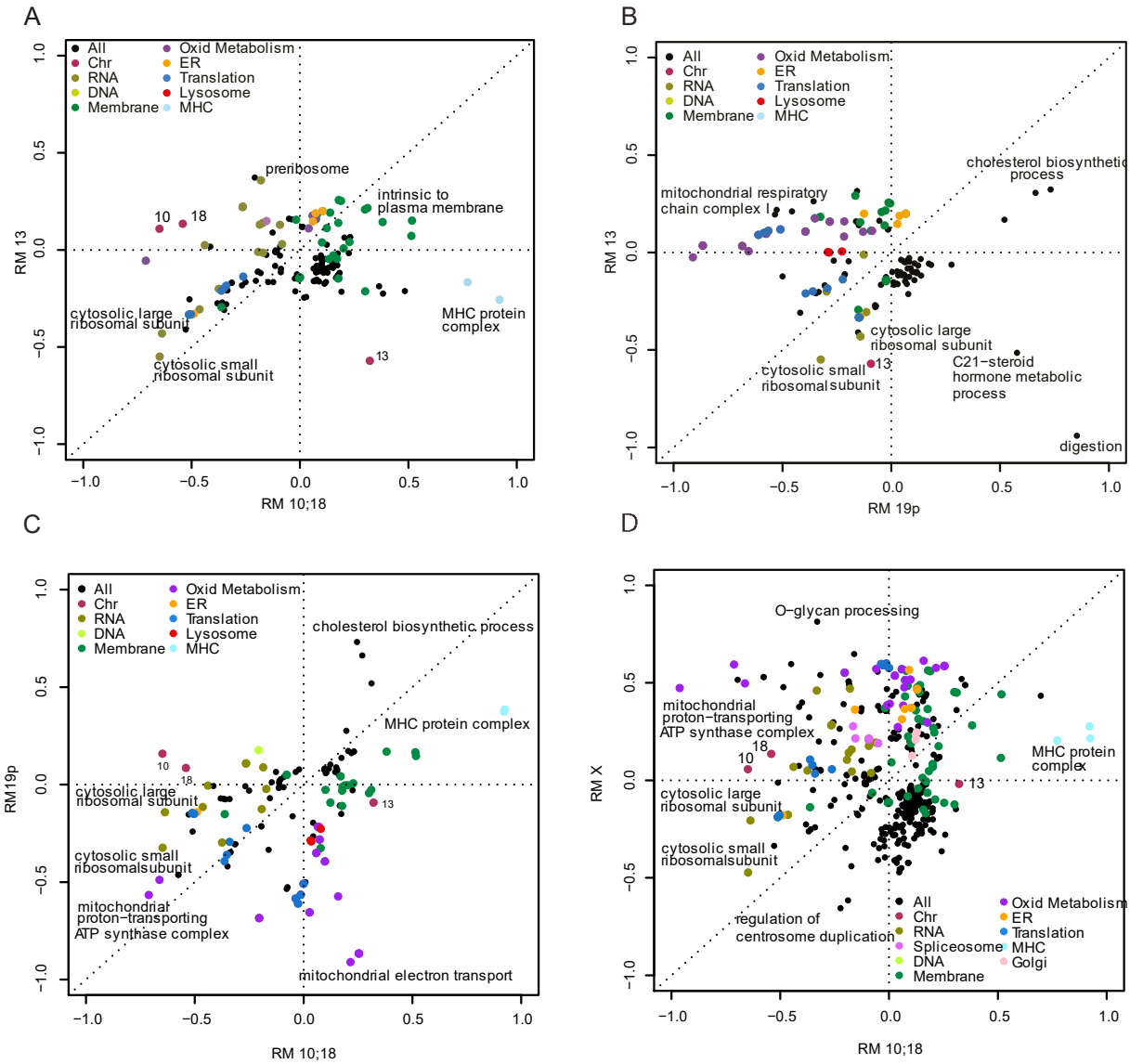


Figure 51 | 2D annotation enrichment of the proteome of monosomic cells. The figure shows different comparisons of 2D annotation enrichments of the proteome of monosomic cells for gene ontology biological process and cellular compartment annotations, as well as the chromosome annotation. (A) RM 10;18 vs RM 13. (B) RM 19p vs RM13 (C) RM 19p vs RM 10;18 (D) RM X vs RM 10;18. For the remaining combinations see Supplement Figure S13. Pathways are controlled by a Benjamini Hochberg FDR threshold of 0.02, grouping and color coding has been performed similar as in Figure 16.

Similar to the previous analysis, the proteome of ribosomal subunits and translational related pathways were commonly downregulated in response to chromosome loss in all investigated cell lines (Figure 51 A to D). Yet, the analysis confirmed that individual monosomic cell lines deregulate unique pathways. Immune related pathways, as *MHC protein complex* and *MHC class I protein complex,* were upregulated in RM 10;18 but downregulated in RM 13 (Figure 51 A). Oxidative metabolism associated terms, such as *ATP synthase complex* or *mitochondrial respiratory chain complex I* were downregulated in RM 19p and RM 10:18, yet upregulated in RM X (Figure 51 B, C, D), while intriguingly the *mitochondrial electron transport*, which mainly consists of NADH dehydrogenase proteins appears to be specifically downregulated in RM 19p (Figure 51 C). Lastly, *O-glycan processing* was upregulated for RM X, yet downregulated for RM 10;18 (Figure 51 D). These insights lead to our hypothesis that monosomy in human cell induces a specific, heterogeneous response depending on which chromosome was lost, with a ribosome biogenesis and translation downregulation presenting the only obvious shared feature.

Comparison of cellular response to monosomy and trisomy

To investigate whether the observed differential regulation is comparable to cell lines that gained a chromosome, we analyzed the gene expression changes together with changes observed in the trisomic cell lines RPE 5/3 12/3 and RPE 21/3 [79]. In these aneuploid cell lines, previous analysis revealed a distinct “aneuploidy specific response”, which consists of an upregulation of genes linked to endoplasmic reticulum, Golgi apparatus and lysosomes, paired with a downregulation of DNA replication, transcription as well as ribosomal subunits [202]. We integrated the proteome data to our set and performed clustering as well as the pathway enrichment analysis, as shown in Figure 52.

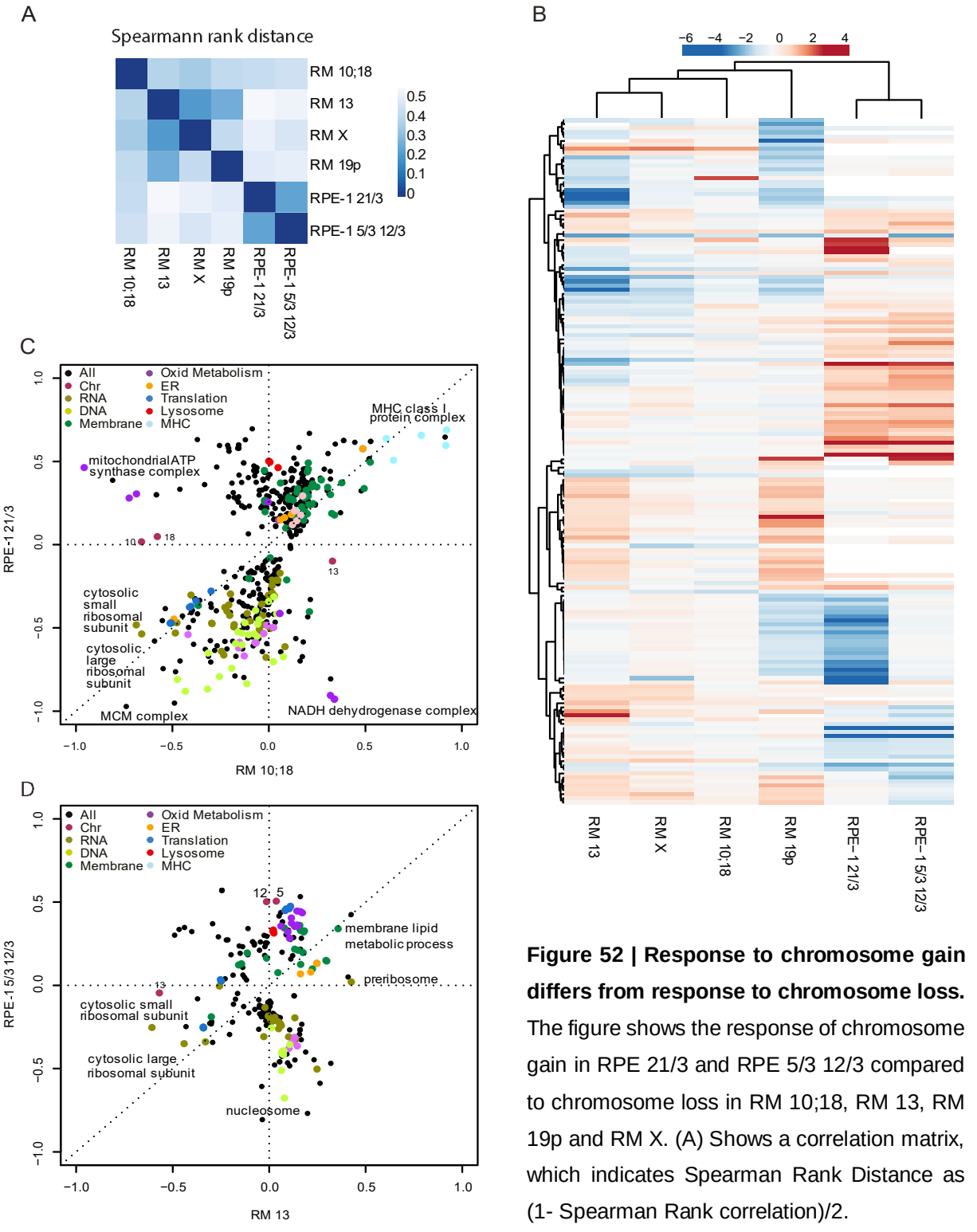


Figure 52 | Response to chromosome gain differs from response to chromosome loss. The figure shows the response of chromosome gain in RPE 21/3 and RPE 5/3 12/3 compared to chromosome loss in RM 10;18, RM 13, RM 19p and RM X. (A) Shows a correlation matrix, which indicates Spearman Rank Distance as (1- Spearman Rank correlation)/2. (B) shows a Spearman rank based clustering analysis and heatmap of $\log_2$ fold changes of the tri- and monosomies to the diploid wild type. (C) Shows the 2D annotation enrichment, performed similarly as in the previous analysis steps, for RPE 21/3 vs RM 10;18 (D) for RPE 5/3 12/3 vs RM 13.

The initial correlation analysis of the relative expression of the proteome in aneuploid cell highlights a strong heterogeneity in gene expression between gain and loss of chromosome, as indicated in the correlation matrix in Figure 52 A, which shows the Spearman Rank distance between all cell lines. Trisomic cell lines generally show a stronger correlation than monosomic cell lines with each and within each other. This indicates that the “aneuploidy specific response” previously described reflects only the gene expression changes in trisomies. Additionally, the correlation between individual monosomies was lower than between the trisomic cell lines, further confirming our hypothesis that responses to chromosome loss are more heterogeneous. Although, the limitation should be noted that only two different p53 proficient trisomic cell lines were analyzed, while the monosomic cell lines are p53 deficient.

The clustering analysis was performed for proteins with at least two valid values with higher $\log_2$ fold changes than 1 to parental, diploid wildtypes in at least two cell lines, to investigate differentially expressed genes. The Spearman correlation also confirmed that the trisomic cell lines cluster both together and away from the monosomic cell lines (Figure 52 B). Comparison of the proteome pathway deregulation for RM 10;18 vs RPE 21/3 shows also a stronger anticorrelation of enriched pathways (Figure 52 C). This effect can be observed even more drastically for RPE 5/3 12/3 vs RM 13 (Figure 52 D), due to the stronger response to the presence of two extra chromosomes, indicated by the accumulation of clusters in the top left/bottom right section of the plot, inverse to the correlating region.

This analysis confirms the hypothesis that the general response to chromosome gain differs from the response to the loss of chromosome, and that monosomic cell lines show more heterogeneous expression.

Expression genes of individual factors

One of the limitations of pathway enrichment is that the interpretation of the result down-weights deregulation of individual gene products by scoring for entire gene sets. To continue the analysis, we therefore investigated differentially expressed genes as general outliers and members of deregulated pathways. First, the dataset was filtered for significantly deregulated proteins, defined as proteins expressed with at least two $\log_2$ fold changes to the diploid wildtype, and outside the range of -1.5 to 1.5 in all monosomies. This only identified 5 up (Figure 53 A) and 13 down-regulated proteins (Figure 53 B), which suggest that in general the gene expression has little overall similarity in cell lines that lost different chromosomes. Further, we identified ribosomal and translation associated gene sets as commonly downregulated in the proteome of all aneuploid cell lines (Figure 50 to 52).

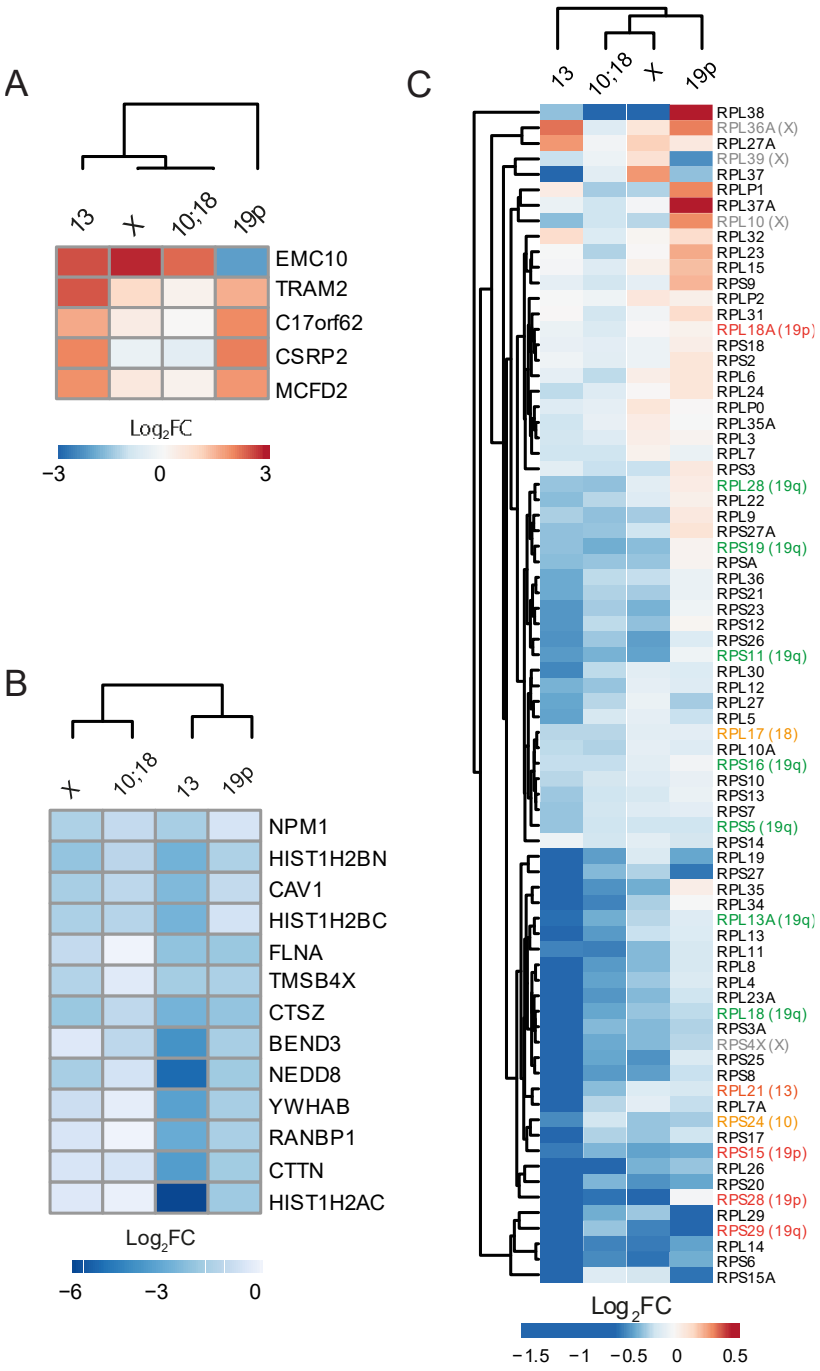


Figure 53 | Differential gene expression in monosomic cell lines. The figure shows significantly up and downregulated proteins in (A), defined as having at least two values outside -1.5 and 1.5. (B) shows the $\log_2$ fold change all members of the GO cellular compartment terms cytosolic large ribosomal subunit and cytosolic small ribosomal subunit. Chromosome location has been added to the gene names and monosomies indicated by different colors (red: 19p, green: 19q, yellow: 10;18, grey: X, orange: 13).

To investigate whether this downregulation is affected by a few highly deregulated proteins or an overall lower expression, we filtered the dataset for the gene ontology cellular compartment terms *cytosolic large ribosomal subunit* and *cytosolic small ribosomal subunit* and plotted the log_2 fold changes as heatmap. This reveals an overall highly uniform signal for ribosomal proteins, independent of which chromosome was lost or the gene localization on monosomic or disomic chromosomes, which indicated by the color coding of gene names on monosomies in (Figure 53 C).

Chromosome specific consequences

The previously described pathway analysis highlighted heterogeneous responses to the loss of different chromosomes in both proteome and transcriptome, as well as in comparison to gain-of-chromosome RPE cell lines. For example, the GO terms related to oxidative metabolism, such as *mitochondrial electron transport*, showed a strong downregulation in RM 19p, but upregulation for RM X (Figure 51 B, C and D). To further investigate the chromosome specific response to monosomy, cluster analysis of deregulated pathways was performed.

Table 2 | 1D Annotation enrichment for RM 19p. The table shows the Top 10 most significantly deregulated pathways for GO biological process and cellular compartment in RM 19p, sorted by P-value. Further included are the enrichment score, p-value, corrected Benjamini Hochberg FDR and the median log_2 fold change expression of the complex.

Gene Ontology	#Members	Enrichment Score	P-value	Benj. Hoch. FDR	Median expression
mitochondrial inner membrane	296	-0.349705	3.20E-24	3.68E-21	-0.173622
electron transport chain	88	-0.616724	2.67E-23	2.24E-19	-0.593283
respiratory electron transport chain	88	-0.616724	2.67E-23	1.12E-19	-0.593283
organelle inner membrane	317	-0.325154	1.82E-22	1.05E-19	-0.141033
mitochondrial electron transport, NADH to ubiquinone	35	-0.911578	1.25E-20	3.51E-17	-0.950473
mitochondrial respiratory chain complex I	38	-0.872784	1.58E-20	6.05E-18	-0.95068
respiratory chain complex I	38	-0.872784	1.58E-20	4.54E-18	-0.95068
NADH dehydrogenase complex	38	-0.872784	1.58E-20	3.63E-18	-0.95068
mitochondrial translational initiation	78	-0.605041	3.83E-20	8.04E-17	-0.304295
mitochondrial membrane	386	-0.278675	4.97E-20	9.53E-18	-0.111792

In Table 2, the top 10 of all significantly deregulated pathways in RM 19p are shown, identified by one-dimensional pathway enrichment. Log_2 fold changes of all proteins

associated with the term *mitochondrial inner membrane* are shown in the hierarchically clustered heatmap in Figure 54 A. As gene ontology terms function similar to a hierarchical tree structure [69] and annotations of the enriched biological pathways and cellular compartments in our analysis intentionally overlap, this term was chosen as it contains proteins of multiple of the smaller, more specific enriched pathways that have shown anticorrelating deregulation in the previous 2D annotation enrichments (Figure 50-52), including *NADH dehydrogenase complex*, *ATP synthase complex* and its carrier proteins and the *mitochondrial respiratory chain* (Figure 54 A).

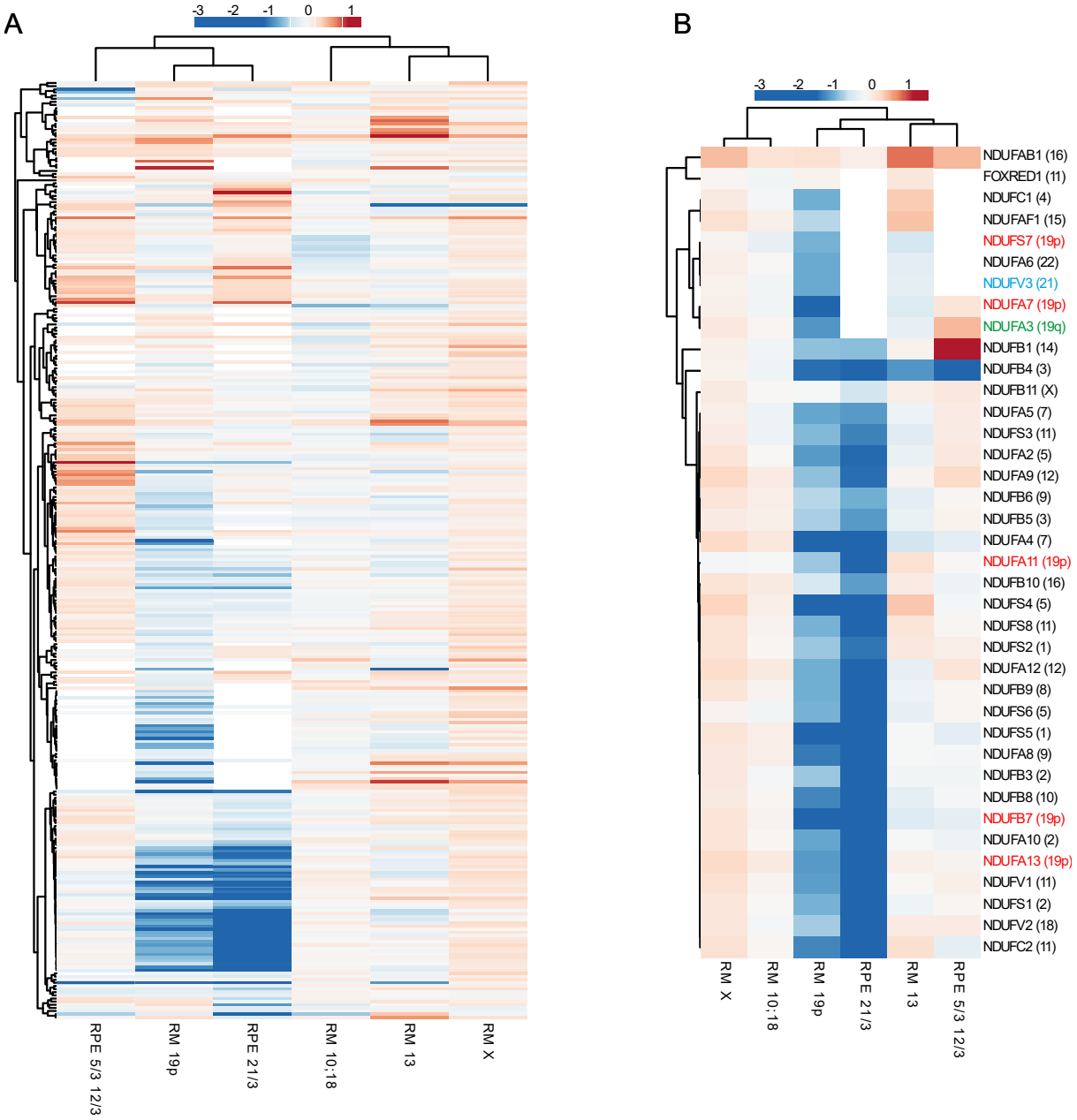


Figure 54 | Oxidative metabolism related pathways in aneuploidies. The figure shows heatmaps of the two GO cellular compartment terms mitochondrial inner membrane (A) and NADH dehydrogenase complex (B). Shown are all the $\log_2$ fold changes of proteins associated to the terms. Localization of the encoding gene to chromosome 21 or 19 have been color coded (red: 19p, green: 19q, blue: 21).

Strikingly, the clustering of mitochondrial membrane proteins in aneuploid cell lines shows that mitochondrial proteins of RM 19p appear closer to the two trisomic cell lines than the monosomic ones, despite the partial loss of a chromosome. RM 10;18, RM 13 and RM X cluster separately away and together. Additionally, a distinct set of proteins can be identified in the lower section of the heatmap in Figure 54 A, which

shows a clear downregulated signal. This set of genes can be associated to the NADH dehydrogenase complex (GOBP), or mitochondrial electron chain (GOCC) (Table 2), which consist largely of the same set of proteins shown in Figure 54 B. Intriguingly, this deregulation was drastic (median fold change of -1.9 for RPE 21/3 and -0.95 for RM 19p), yet highly specific for those two cell lines with 5 encoding genes localized to the p-arm of chromosome 19 and one to chromosome 21. The underlying causes for this specific expression signature remain to be investigated.

Another distinct deregulation was observed for the immune response related pathway *MHC protein complex*, which was upregulated on proteome and transcriptome level in RM 10;18 (Figure 50 B), but downregulated on proteome level in RM 13 (Figure 51 A). A cluster analysis of the *MHC protein complex* has confirmed a similar effect of the level of individual deregulated genes as for the pathways, indicated by clustering of the cell lines RM 10;18 and RPE 21/3 that shows a correlating upregulation of the MHC proteins in Figure 52 C. Similarly, a slight downregulation for related proteins can be observed for RM 13 for both the MHC complex proteins as well as for the interferon stimulated genes (ISG) in Figure 55 B.

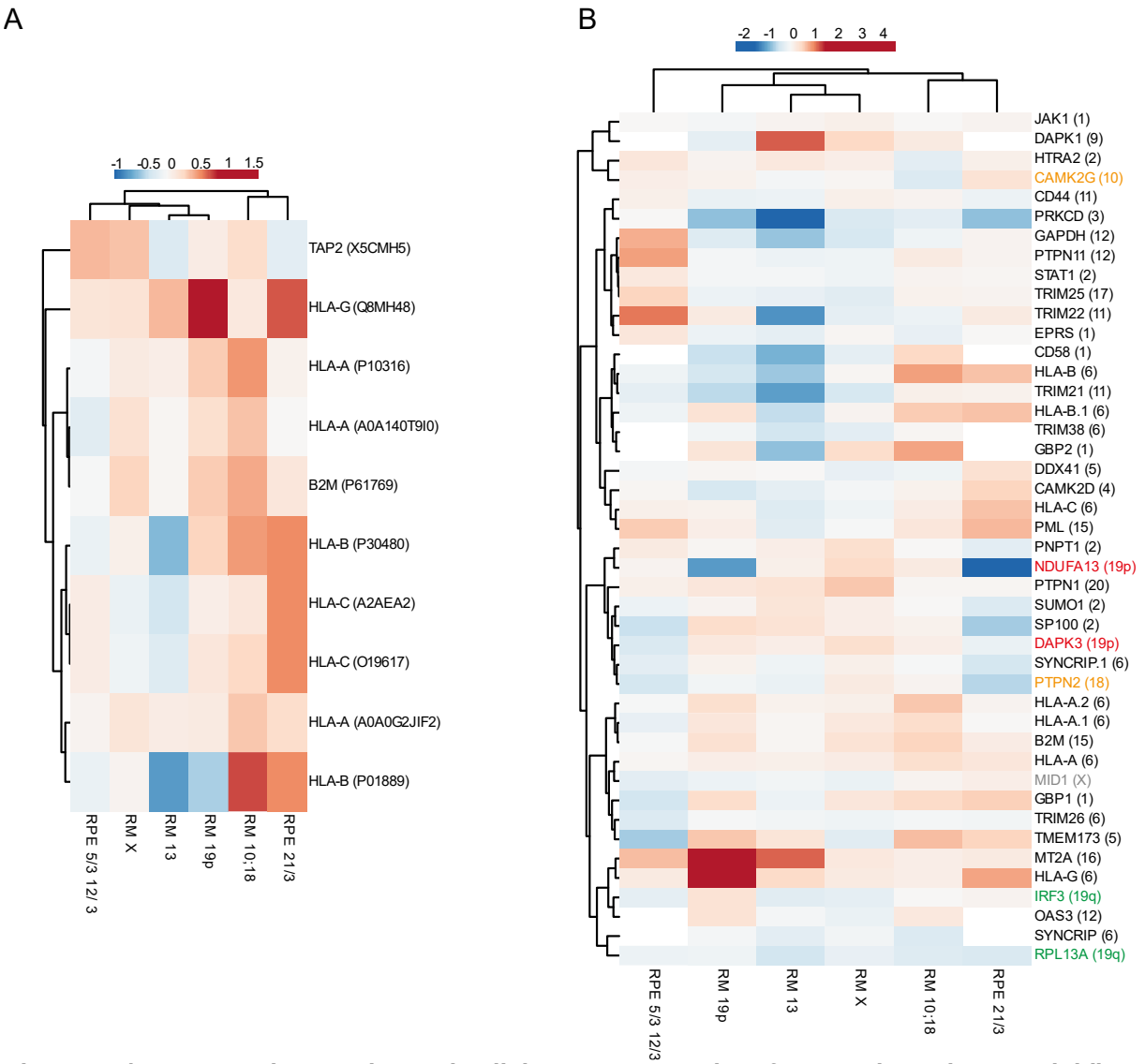


Figure 55 | MHC protein complex and cellular response to interferon pathway in aneuploidies. The figure shows the $\log_2$ fold changes of proteins of the MHC protein complex (A) and cellular response to interferon β and γ (B). Proteins encoding on monosomies have been color coded (red: 19p, green 19q, yellow 10;18, grey: X).

Interferon stimulated genes (ISG) play a role in the defense against various pathogens. The interferon (IFN) mediated immune response is triggered by many different stimuli, such as invading pathogens, and leads to IFN production. IFN molecules bind to receptors at the surface of a cell and initiate the transcriptional regulation of ISGs via the JAK-STAT signaling pathway, to directly inhibit infection and facilitate pathogen resistance [280].

In summary, these findings highlight that the overall correlation of gene expression in monosomic cell lines is low between proteome and transcriptome or compared to the proteome of trisomic cell lines. This indicates a highly heterogenous expression. Pathway analysis revealed that an “aneuploidy specific response”, similarly as described in previous works for gain of chromosome cell lines [79,202], could not be observed for monosomic cell lines. A general downregulated signal was related to ribosomal subunits and translation that could be shown in the enrichment analysis of all monosomic cell lines. Individual comparison of pathway enrichment analyses and identification of differentially expressed genes in monosomies has shown only a limited overlap of pathways and outliers besides ribosome and translation related proteins. The aneuploidy response largely differs between gain and loss of individual chromosomes, indicating that the differential expression of individual genes located on monosomic chromosomes has a stronger impact than the general response to aneuploidy.

3.3 Scaling of cellular gene expression with ploidy

The following chapter presents data that are part of the BioRxiv preprint publication "*Scaling of cellular proteome with ploidy*" (2021).

Galal Yahya [1,2], Paul Menges [1], Devi Anggraini Ngandiri [1], Daniel Schulz [3], Andreas Wallek [4], Nils Kulak [4], Matthias Mann [4], Patrick Cramer [5], Van Savage [6], Markus Raeschle [1], Zuzana Storchova [1*]

[1] Dept. of Molecular Genetics, TU Kaiserslautern, Paul-Ehrlich-Strasse 24, 67663 Kaiserslautern, Germany.

[2] Department of Microbiology and Immunology, School of Pharmacy, Zagazig University, Egypt.

[3] Institute of Molecular Biology, University of Zurich, Switzerland

[4] Max Planck Institute of Biochemistry, 82152 Martinsried, Germany

[5] Max Planck Institue of Biophysical Chemistry, Goettingen, Germany

[6] Department of Biomathematics, University of California at Los Angeles, Los Angeles, CA 90095, United States

doi: 10.1101/2021.05.06.442919

3.3.1 Transcriptome and proteome analysis of yeast cells with different ploidy

Generation of yeast cells with different ploidy

To asses the cellular consequences of polyploidy, series of isogenic haploid (1N) to tetraploid (4N) yeast strains derived from the By4748 background were used for this study. The strains were all modified to have the mating type MATa to eliminate the effects of the pheromone pathway. Ploidy of the strains was confirmed by flow cytometry (Figure 56 A).

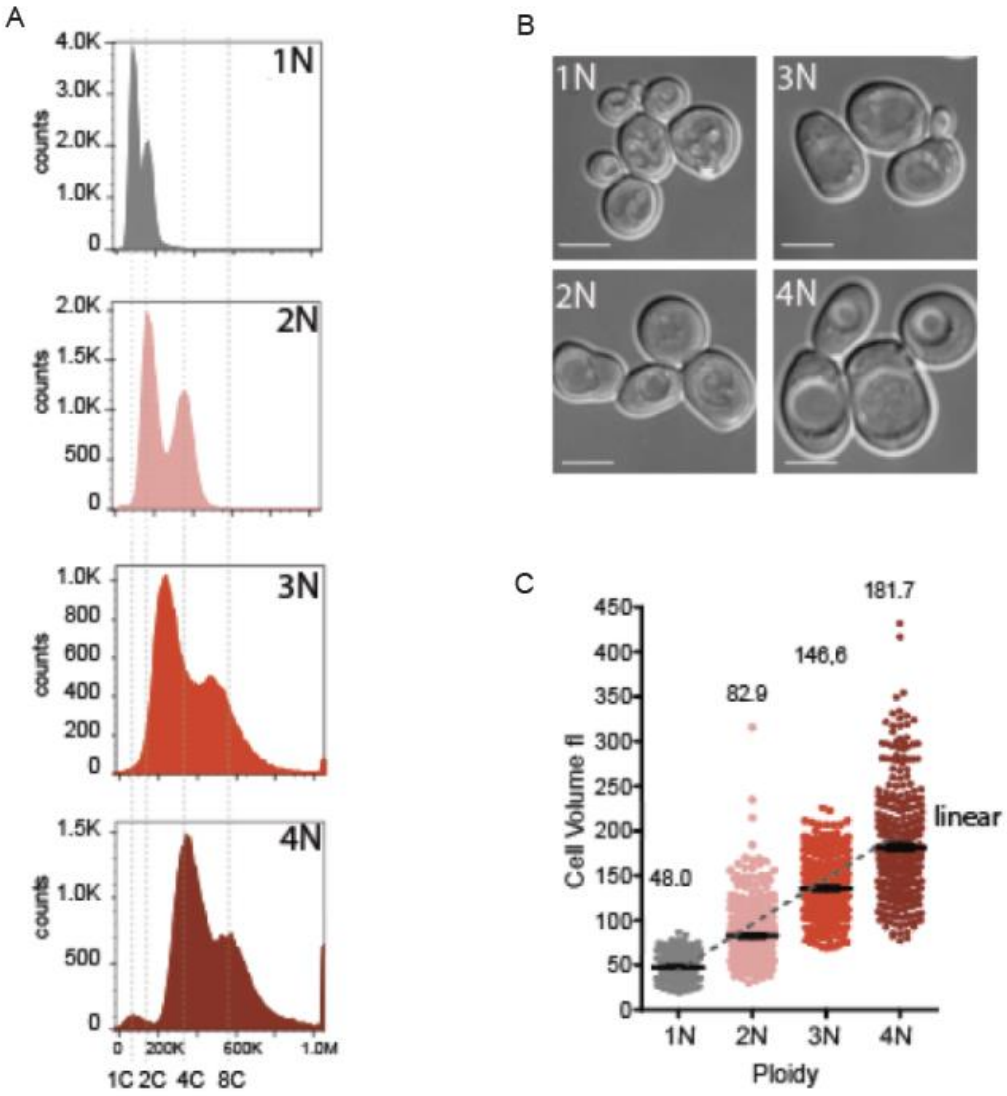


Figure 56 | Cell volume changes in response to increasing ploidy. (A) shows a validation of ploidy by flow cytometry to determine the DNA content of the used strains. (B) Microscopy images of the yeast cells of different ploidy (scale bar: 5 µm). (C) Cell volume measurement for 400 asynchronous cells in 4 independent experiments: median volumes of 48.0 fl for 1N, 82.9 fl for 2N, 146.6 fl for 3N and 181.7 fl for 4N. Dashed line indicates linear scaling (data from Galal Yahya).

Similar as in previous studies [234], a linear increase in cell and nuclear volume was observed (Figure 56 B, C). Additionally, slower cell cycle progression for tetraploid cells was shown, as described in more detail in the related manuscript.

SILAC MS and dynamic transcriptome analysis strategy

To assess the global proteome and transcriptome changes of the yeast strains with different ploidy in order to investigate how transcriptome and proteome levels scale with cell volume, we performed stable isotope labeling of amino acids in cell culture (SILAC) and dynamic transcriptome analysis (DTA) (Figure 57). For both approaches an equal number of cells was analyzed, what allowed us to draw conclusions about scaling of proteome and transcriptome content with increasing ploidy.

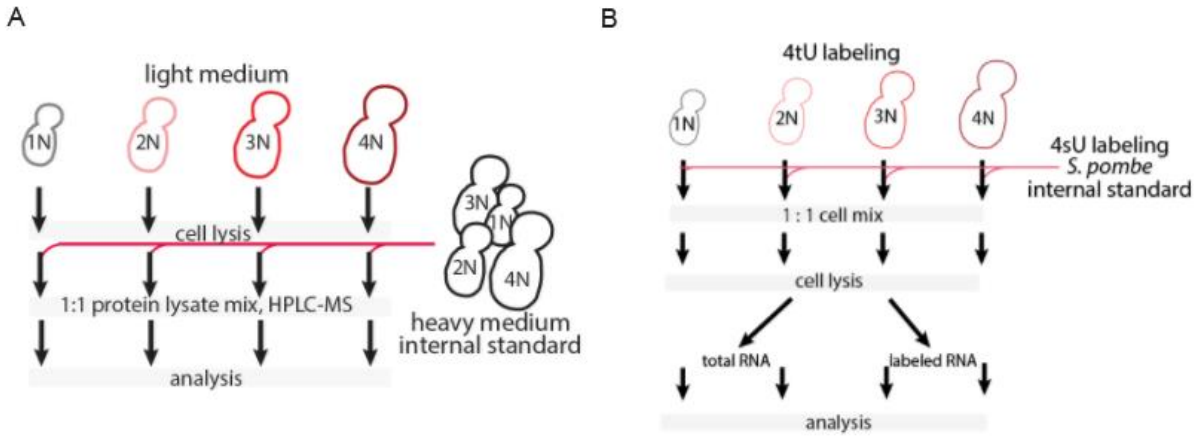


Figure 57 | Labeling and internal standard strategies for SILAC and DTA. The figure shows the labeling and internal standard strategy for SILAC (left) and DTA (right). SILAC: The heavy-labeled standard represented an equal protein mix from cells of different ploidy and was added to equal number of cells of different ploidy. For DTA: mRNA was extracted from *S. cerevisiae* cells of different ploidy (1N, 2N, 3N, 4N) labeled with 4-thiouracil (4tU) and mixed it with mRNA of the distantly related haploid fission yeast *Schizosaccharomyces pombe* labeled with 4sU (4-thiouridin).

In brief, dynamic transcriptome analysis was performed as previously described in [281]. For normalization the distantly related haploid fission yeast *S. pombe* was used as internal standard. This allowed us to obtain data on abundance changes for 5656 mRNAs for all four ploidies. Figure 57 B schematically highlights the used labeling and internal standard strategies. Transcriptomics analysis was performed by Daniel Schulz and Andreas Wallek.

The SILAC analysis provided quantitative information for 70% of all verified open reading frames in each strain. This set was normalized to an internal heavy labeled standard, which represented an equal protein mix from cells of different ploidies (Figure 57 A). This resulted in a set of 3109 protein groups. The haploid cell line 1N shows the highest variance (Figure 58 A), both between its replicates as well as to the other ploidies. This effect appeared slightly amplified by the normalization to the SILAC mix, which consists an equal part of protein from all ploidies. Despite this, the correlation between consecutive ploidies (indicated in blue in Figure 58 C) was higher as the between non-consecutive ones, with 1N to 2N showing the highest correlation. Taken together, this processing yielded two highly reliable datasets, that facilitated the following analysis. SILAC measurement was performed by Nils Kulak.

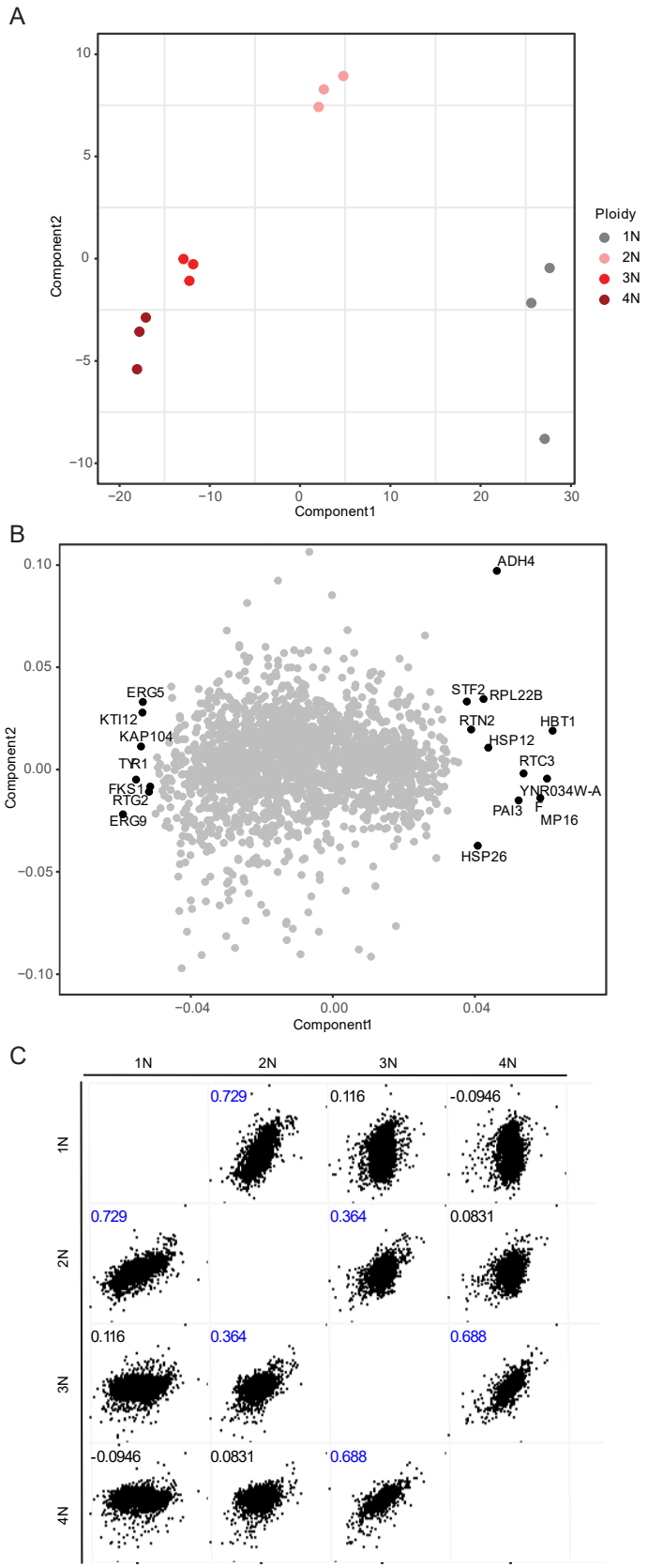


Figure 58 | PCA and correlation analysis of SILAC data. The figure shows a (A) PCA analysis of the three replicates of each ploidy (color coded from grey (1N) to dark red (4N)) as well as the loading of component 1 (B), to highlight several relevant proteins. (C) shows a Spearman Rank correlation matrix of the individual log_2 fold changes of the ploidies against the SILAC standard and in the top left corner the Spearman Rank correlation.

3.3.2 Global transcriptome and proteome changes in response to increasing ploidy

Proteome changes in response to ploidy reveal ploidy-specific protein scaling - "PSS"

Analysis of the 3109 identified protein groups revealed that, surprisingly, the overall protein abundancy does not scale linearly with increasing ploidy as the cell volume (Figure 56 C), but rather shows a distinct allometric ploidy-specific protein scaling (PSS). This can be seen in Figure 59, which shows the median $\log_2$ FC protein expression of each ploidy.

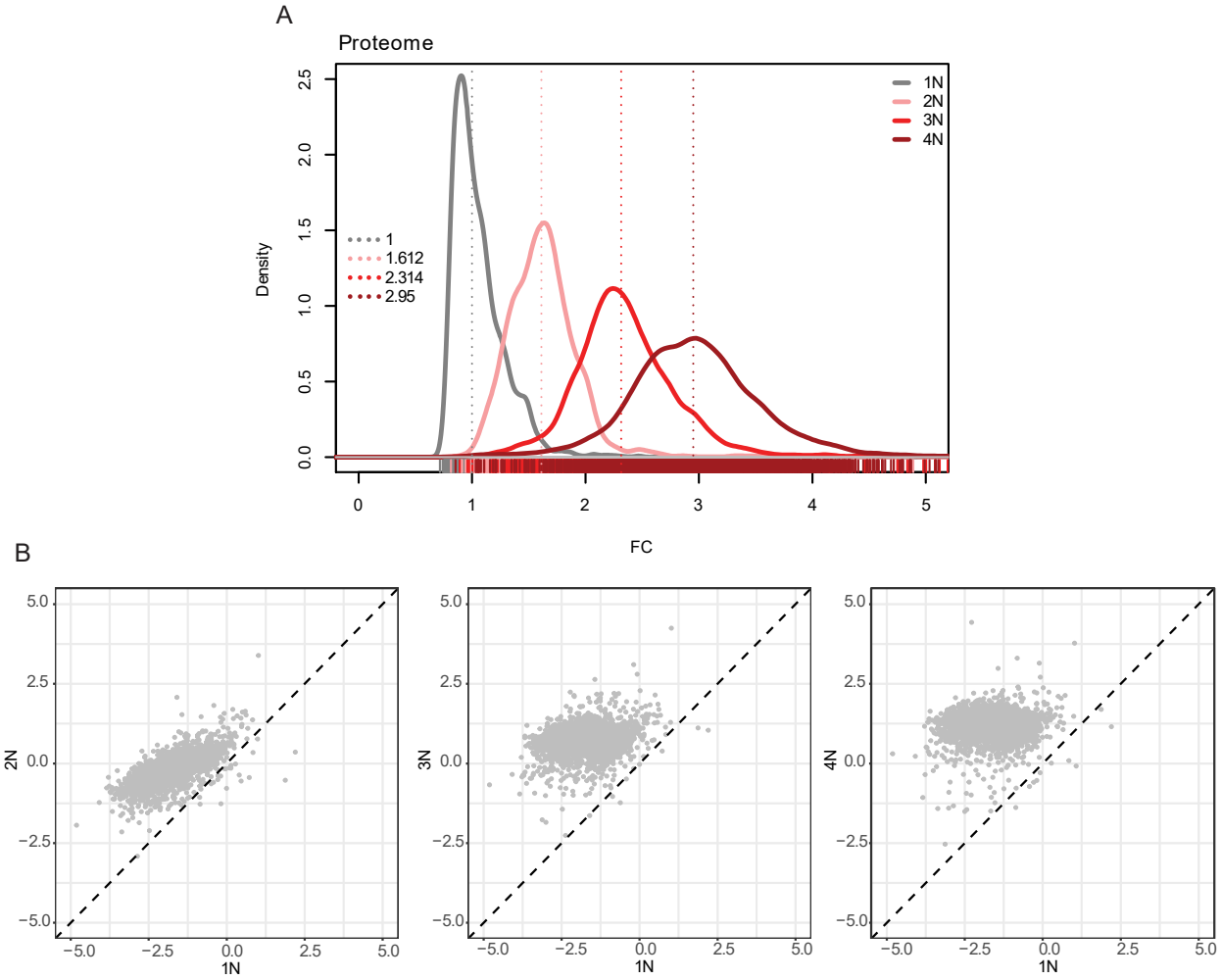


Figure 59 | Scaling of cellular proteome with ploidy. Quantification of 3109 protein groups revealed non-linear scaling of proteome with increasing ploidy, as highlighted in the density plot in (A). The curves show the $\log_2$ fold change of the ploidies 1N to 4N, shifted to a haploid median of 1 for easier comparison. The scatter plots in (B) show the individual $\log_2$ FC protein expressions of 2N to 4N vs 1N, with the dotted line indicating linear scaling.

We validated this scaling by independent measurements of protein concentration, which confirmed that the relative protein abundance per genome decreases with increasing ploidy. Furthermore, we also found comparable scaling of mRNA abundance to the observed protein levels (Supplementary Figure 13). Next, we investigated a set of selected proteins by comparing immunoblotting to the relative log_2 fold change expression to the SILAC standard after shifting the median expression of each individual ploidy to zero. As highlighted in Figure 60, the results of immunoblotting match the measured proteome.

This confirms that our results are neither an artifact of SILAC labelling, which similar to TMT, described in the previous chapter, is affected by a degree of ratio compression [282], nor any other normalization artifact.

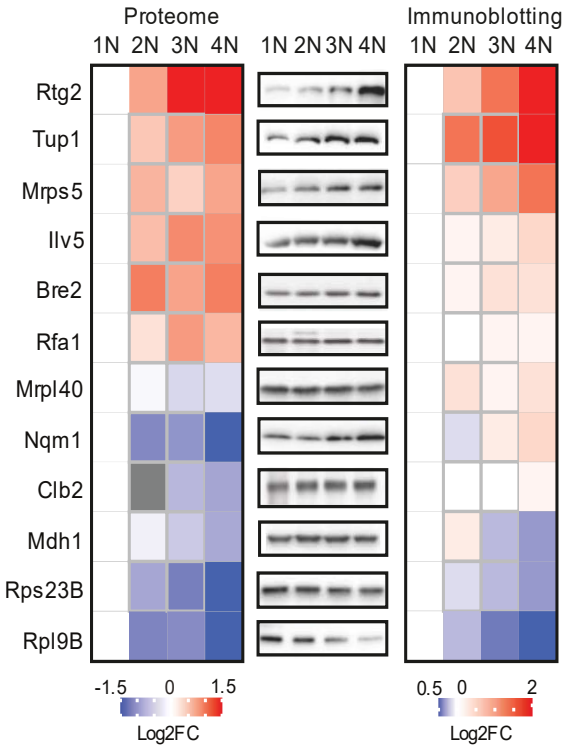


Figure 60 | Validation of abundance changes. A validation of the abundance changes by immunoblotting is shown. The left side shows the relative log_2 fold changes of selected proteins, normalized to 1N. In the middle a representative immunoblot of the candidates is shown. The right side shows the quantification of three replicates of immunoblotting. The WB was performed by Gala Yahya.

Transcriptome changes in response to increasing ploidy.

In a consecutive analysis we investigated if the observed ploidy-specific scaling affects differential gene expression on transcriptome and proteome level. The analysis of the 5656 mRNAs of the four different yeast ploidies has shown overall only marginal changes to differential mRNA expression. This is in agreement with previous analyses of the transcriptome of yeast strains of different ploidies [234,245].

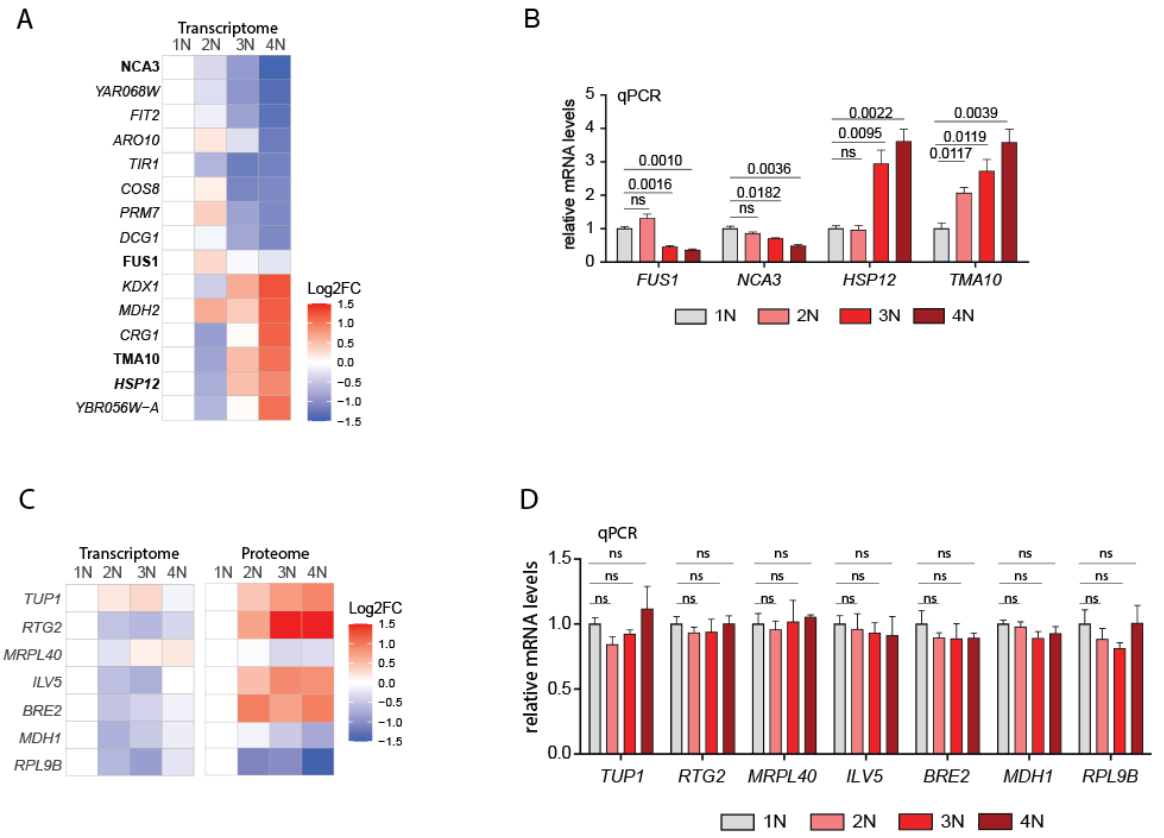


Figure 61 | Transcriptome abundance changes differ with ploidy. The heatmaps show transcriptome changes in response to increasing ploidy. (A) shows the most differentially deregulated transcripts normalized as “ratio of ratio” to 1N. Transcripts marked in bold were validated in (B), that shows qPCR of selected, deregulated transcripts normalized to 1N. (C) highlights candidates that were not deregulated on transcript, but proteome level and validated in (D). For qPCR means and SD of three replicates are shown. QPCR was performed by Galal Yahya and Devi Ngandiri.

Filtering for outliers yielded only 13 mRNAs that changed significantly over ± 2 FC with increasing ploidy (Figure 61 A, B). Those candidates contain transcripts associated with plasma membrane (COS8, FUS1) and cell wall synthesis (TIR1, KDX1), again in agreement with previous findings [234,245].

Comparably, qPCR validated that there are no changes in mRNA abundance of the corresponding transcripts of differentially regulated proteins (Figure 61 C, D). Taken together, this raises the question if ploidy-dependent regulation of protein abundance does occur post-transcriptionally.

Proteome analysis reveals Ploidy-dependent protein regulation - "PDR"

To investigate how the differential expression on proteome level changes with increasing ploidy, we used two consecutive filters. In brief, we calculated the 'ratio of ratio' between the log_2 fold changes of each ploidy normalized to the SILAC standard. (i) This ratio was filtered for outliers based on highest difference between 4N to 1N. (ii) A smoothing filter was applied that further removes 'expression spikes' between ploidies to show deregulation across ploidy. This was achieved by removing proteins with a higher FC difference than 1 between consecutive ploidies. The differentially regulated proteins with the highest differences were visualized as heat maps in Figure 62.

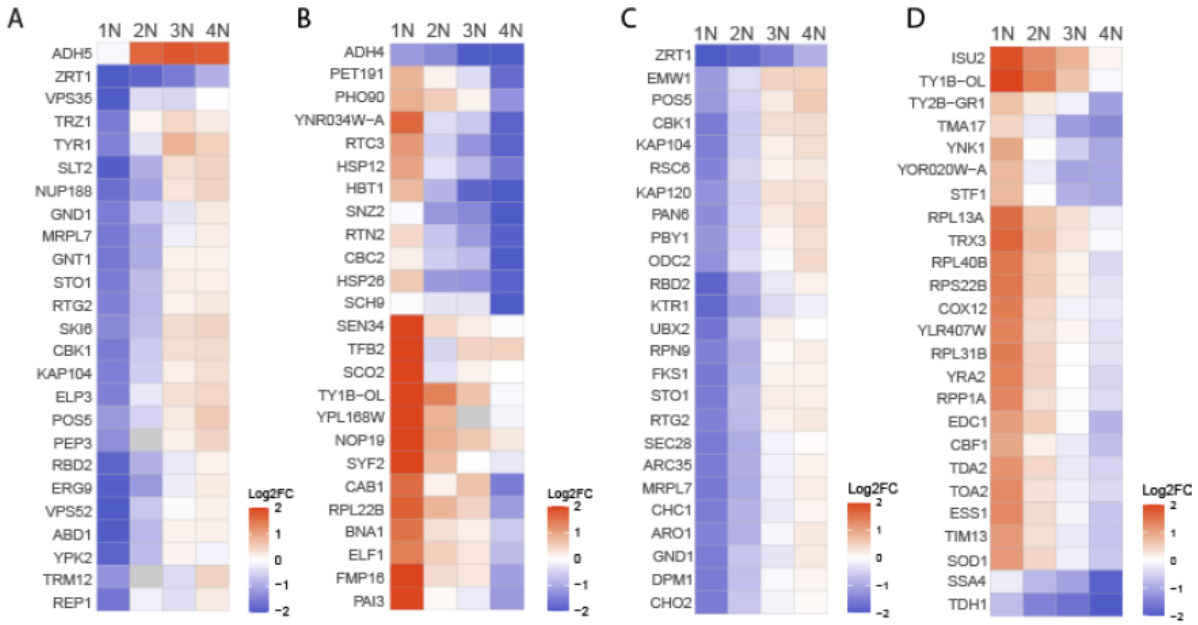


Figure 62 | Proteins differentially regulated with ploidy. The heatmaps show normalized log_2 FC expression changes of the strongest deregulated proteins between ploidies, sorted by 4N/1N difference. (A) Top 25 proteins upregulated in tetraploids (4N) compared to haploids (1N). (B) Top 25 proteins downregulated in 4N compared to 1N. (C) Top 25 proteins upregulated across ploidies without expression spikes. (D) Top 25 proteins downregulated across ploidies without expression spikes.

Intriguingly, while the largest part of the proteome follows the ploidy-specific protein scaling, several proteins are regulated differently by ploidy. We termed this effect ploidy-dependent regulation (PDR). Among those proteins we identified multiple cell wall integrity proteins that were upregulated with increasing ploidy, such as CBK1, PRT1, EMW1, FKS1. Meanwhile, multiple proteins associated with a translational function or ribosomal subunits, such as RPL13A, RPL40B, RPS22B or mitochondrial proteins, such as ISU2, TMA17, TIM13 were overall downregulated.

To validate that the observed changes are not facilitated by increased volume in polyploid cells, haploid mutants with altered cell size have been analyzed: cln3Δ that lacks a G1 cyclin, and a respiration deficient rho0 mutant. The cell volume of those strains is respectively comparable to 2N and 3N strains. Yet, the protein abundance of a set of selected candidates, as quantified through immunoblotting performed by Galal Yahya and Devi Ngandiri, did not resemble the measured protein abundancy in the corresponding yeast ploidies (Figure 63). Overall this analysis confirms, that the changes in ploidy are responsible for the differential expression of genes on protein level.

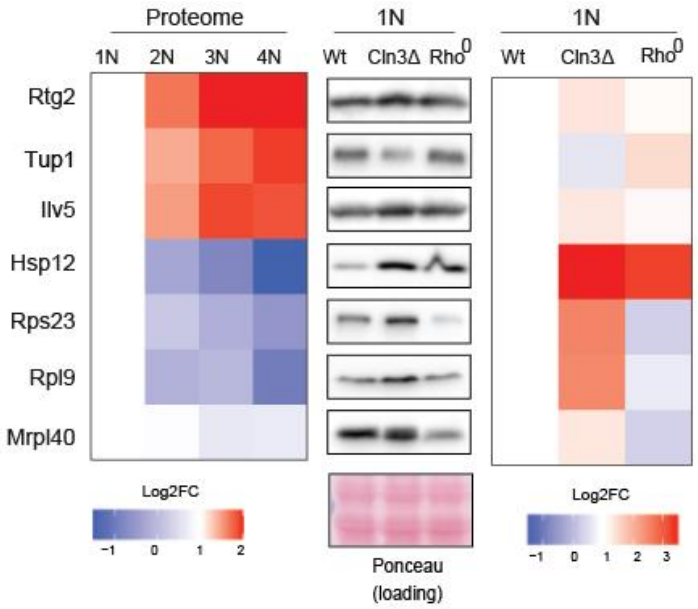


Figure 63 | Protein expression in haploid mutants with altered cell size. The figure shows $\log_2$ FC protein expression normalized to 1N of 1N to 4N yeast strains (left), compared to a representative immunblot (mid) and the quantification of protein levels based on immunoblotting of whole cell lysates.

In summary, proteome and transcriptome analysis revealed that proteome content scales non-linearly with increasing ploidy by “ploidy-specific protein scaling”, which doesn’t correlate with the linear increase in cell volume and with the gene copy number. While the abundance of most proteins is regulated according to the observed allometric scaling, there is also a “ploidy-dependent protein regulation”. This observed deregulation includes ribosomal subunit or biogenesis as well as mitochondria related proteins. Overall weak differential expression on mRNA level, that does show a similar scaling as proteome, could not explain this protein deregulation, indicating post-transcriptional, ploidy dependent regulation of protein abundance.

3.3.3 Pathway changes in response to increasing ploidy

In a consecutive analysis, we investigated how the differential expression on proteome level effects regulation of pathways. Therefore, two-dimensional pathway analysis was performed comparatively for all ploidies. Unsurprisingly, 1N vs 4N (Figure 64 A) shows the highest anticorrelation of deregulated pathways while 1N vs 2N (Figure 64 C) shows still an overall correlation. This further confirms a consecutive, differential expression across ploidies.

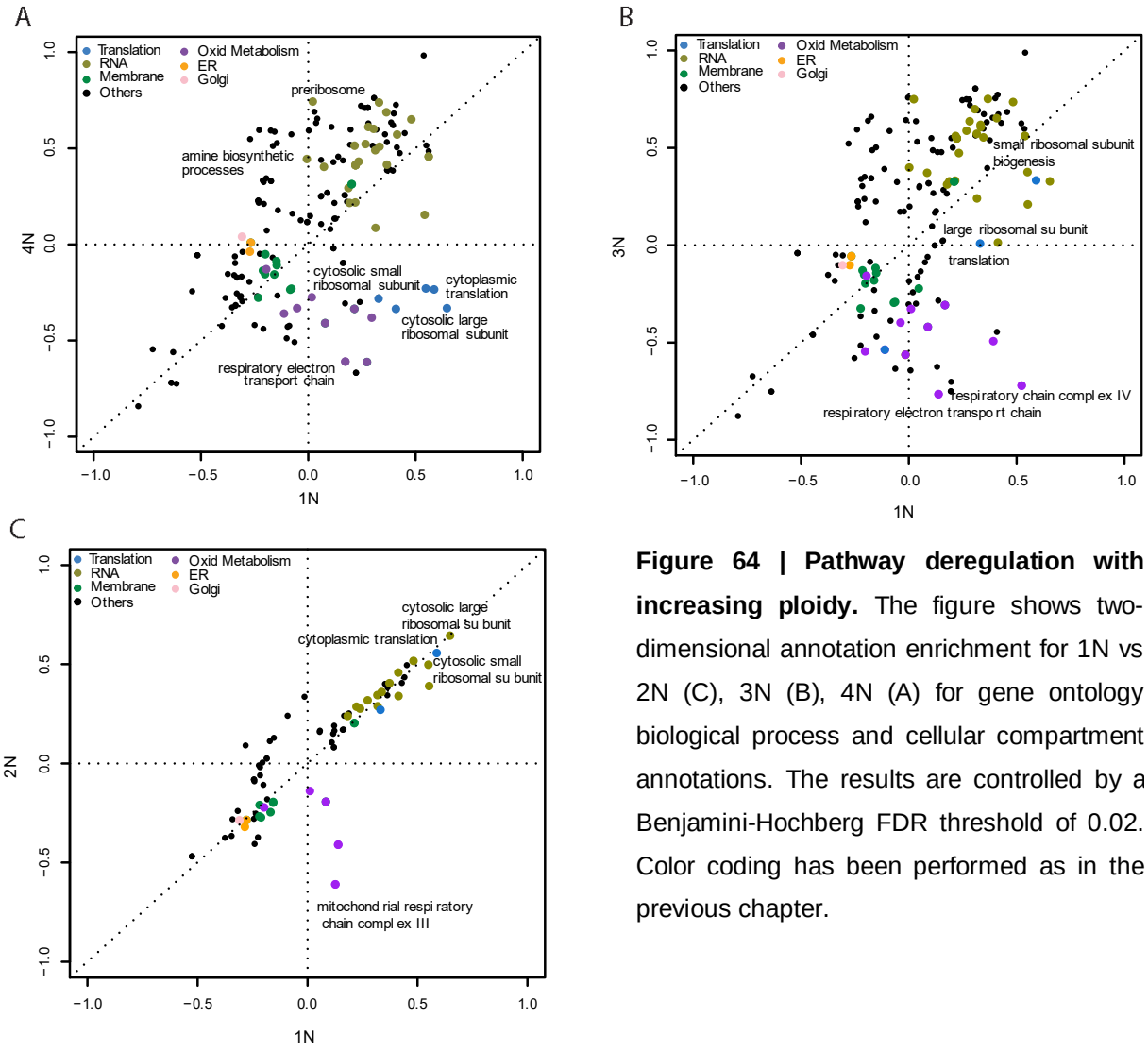


Figure 64 | Pathway deregulation with increasing ploidy. The figure shows two-dimensional annotation enrichment for 1N vs 2N (C), 3N (B), 4N (A) for gene ontology biological process and cellular compartment annotations. The results are controlled by a Benjamini-Hochberg FDR threshold of 0.02. Color coding has been performed as in the previous chapter.

Strikingly, the only consistently downregulated pathways in all ploidies except 1N are associated with the respiratory electron chain complex. The ribosomal subunits and biogenesis, as well as general translation-associated pathways show increasing deregulation with ploidy from 2N to 4N, when compared to 1N (Figure 64 A to C).

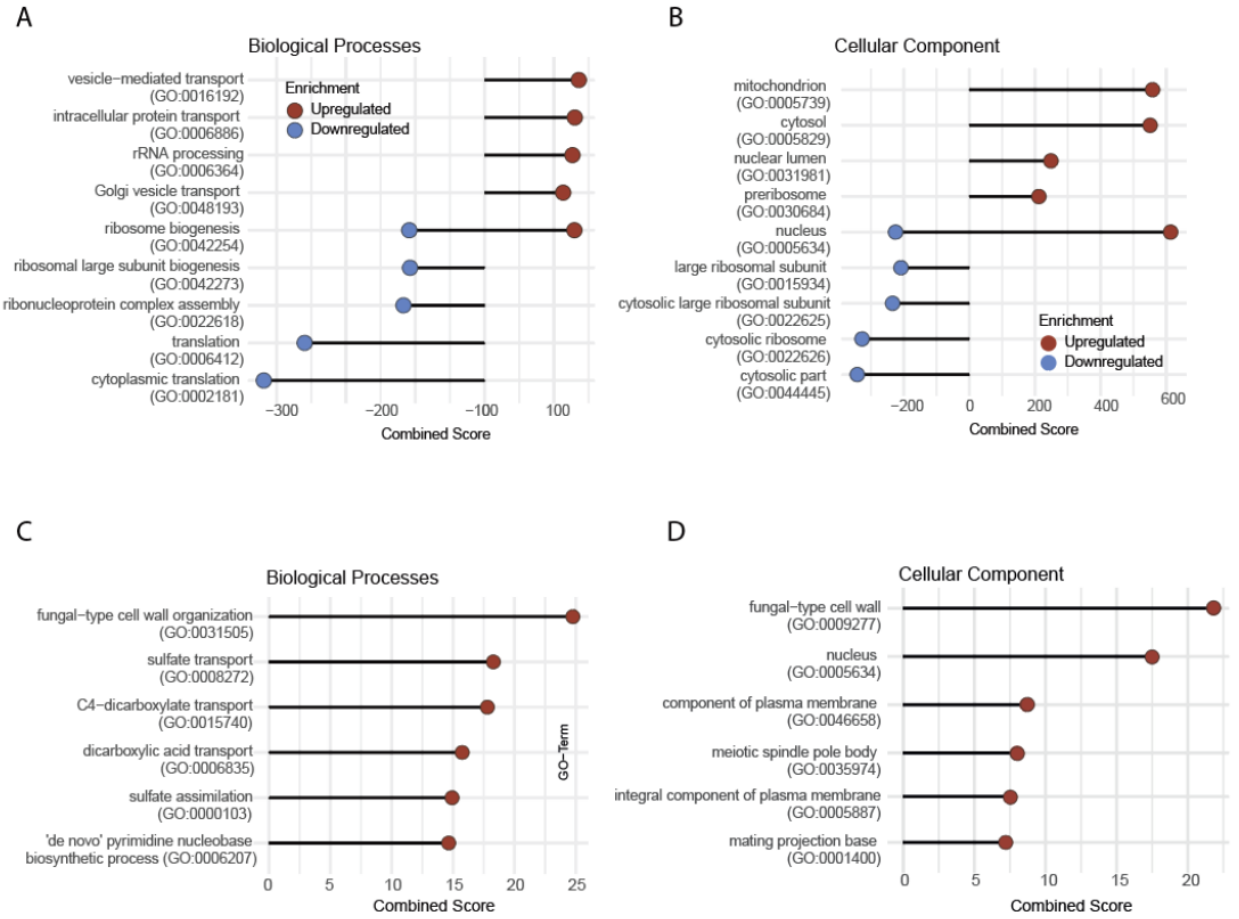


Figure 65 | GSEA of significantly deregulated genes. The figures show a gene set enrichment performed by the yeast sub-module of EnrichR of significantly deregulated genes of all ploidies normalized to 1N. Shown on the axis is the combined score, calculated as a product of the p-value resulting from the Fisher exact test of the enrichment and the z-score of the deviation from the expected rank. (A) Differentially regulated Biological Process and (B) Cellular Components identified by Gene Set Enrichment Analysis of proteome. (C) Differentially regulated Biological Processes and (D) Cellular Components identified by Gene Set Enrichment Analysis of mRNAs. No downregulated pathways were identified by transcriptome.

To further validate our findings, pathway enrichment was repeated for only significantly deregulated genes. This set of genes was determined by performing individual two-sample t-tests of all ploidies against 1N to identify genes that are significantly deregulated with increasing ploidy. The resulting GSEA, performed by EnrichR [74], confirmed the repression of pathways related to cytoplasmic ribosomes and translation with increasing ploidy (Figure 65 top). Vesicle trafficking, cytosolic and intracellular transport were upregulated with increasing ploidy, yet slightly weaker than the signal of the downregulated pathway.

On mRNA level pathway deregulation was much weaker, as observable by the low combined scores (Figure 65 C, D). No significant enrichment was determined for downregulation and only few pathways showed an upregulation, which did not overlap with the identified proteome enrichment. This further confirms both the overall marginal changes on mRNA level in response to polyploidy, as well as the low correlation of transcriptome and proteome changes, and further strengthens the notion that the ploidy-dependent regulation of protein abundance is regulated post-transcriptionally.

The deregulation of both cytoplasmic translation related pathways and ribosome biogenesis could explain the allometric scaling of protein content in response to ploidy increase. As observable in Figure 66 the deregulation of both ribosomal subunits shows a very homogeneous signal, as well as a nearly linear trend in protein decrease.

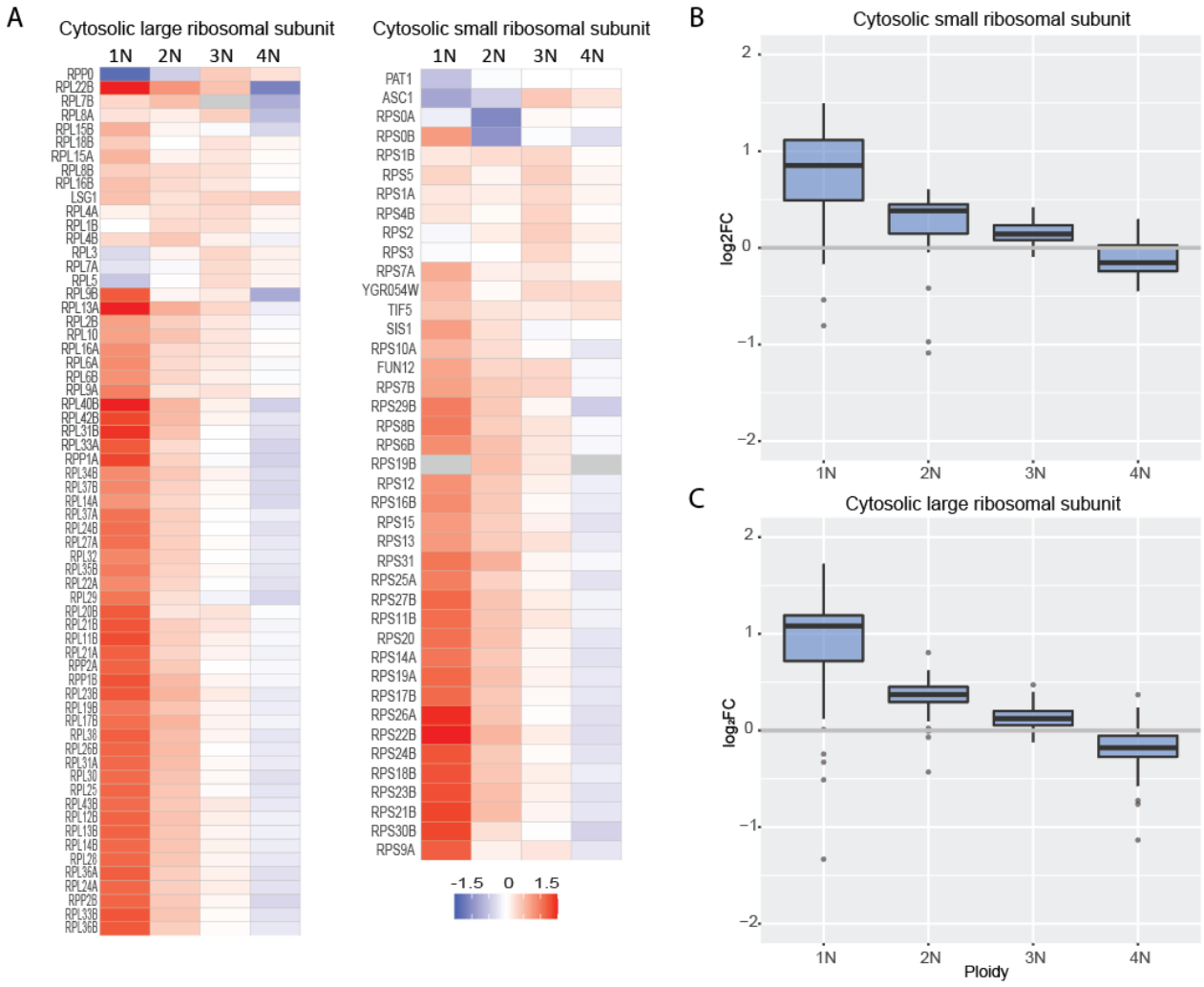


Figure 66 | Translation downregulation in cells with increasing ploidy. The figure shows the normalized log_2 fold change to the SILAC standard of ribosomal proteins of the cytosolic large and small ribosomal subunit, as (A/B) heatmaps and (B/C) box plots, highlighting the downregulation with increasing ploidy.

No decrease in mRNA levels, nor deregulation of pathway in the transcriptome could explain this deregulation in ribosomal protein abundance, therefore we hypothesize that production of rRNA is decreased with ploidy.

Proteome analysis has further shown respiratory electron chain related proteins to be reduced in all ploidies compared to 1N (Figure 67). Consistently, the abundance of RTG2, that plays a role in retrograde signaling of mitochondrial dysfunction to the nucleus was increased with ploidy (Figure 62 & 63). On media with a non-fermentable carbon source or in the presence of the oxidant diamide polyploid cells also proliferated poorly. This indicates a polyploidy specific deregulation of mitochondrial functions, what also could be shown in recent studies in the pathogenic yeast *Candida albicans* [283].

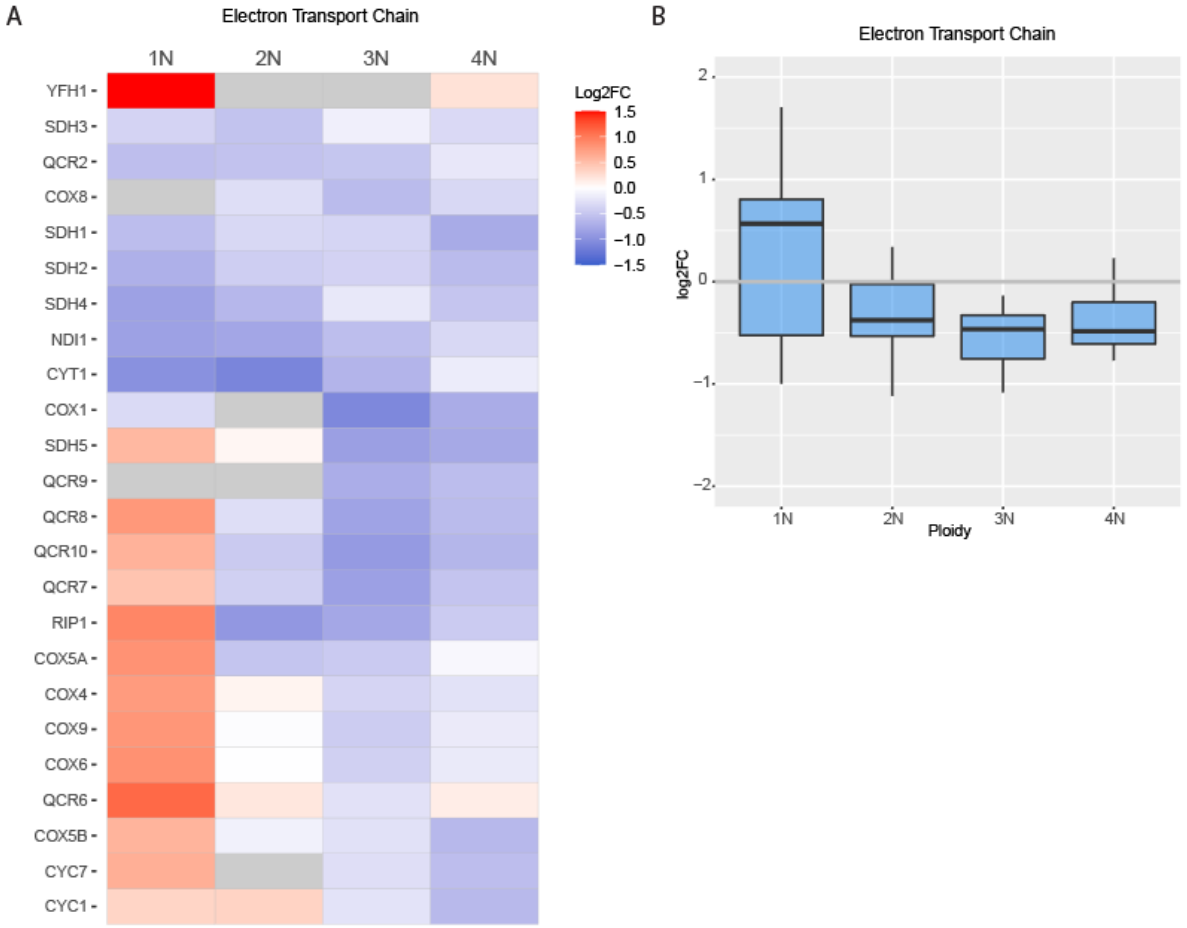


Figure 67 | Electron transport chain deregulation. The figure shows the normalized log_2 fold changes to the SILAC standard of electron transport chain proteins as heatmap (A) and boxplot (B).

In summary, we observed that the ploidy-dependent deregulation significantly downregulates ribosomal subunits and biogenesis. As this deregulation cannot be explained by changes to mRNA expression, we conclude that the production of rRNA is decreased with ploidy. Furthermore, a deregulation of mitochondrial function with changing ploidy could be observed.

4. Discussion

Maintenance of genome integrity is crucial to ensure faithful replication of the genome in each cell of the body. The cell division cycle functions despite thousands of DNA lesions occurring in each cell, each day [112]. Keeping the genome intact despite continuous genotoxic stress requires the concentrated action of cellular metabolism, cell cycle and DNA damage response. Unrepaired DNA lesions lead to impaired genome integrity through errors and a loss of function in any of the maintenance processes [284]. The consequences thereof are highly diverse and lead to both beneficial mutations that drive evolution and various detrimental phenotypes through genomic instability and alterations in gene expression. Genomically unstable cells are often predisposed for malignant growth and tumorigenesis, where genomic instability turns into a driving factor through the selection of more aggressive clones [285]. Aneuploidy [286] and polyploidy [232] are both poorly tolerated in somatic cells, but frequently observed hallmarks of cancer that represent an attractive field of study for potential cancer treatments [287]. The rapid development in omics technology and computational biology in the recent years facilitated a better understanding and hypotheses generation for the exact mechanisms underlying the causes and consequences of impaired genome integrity. During my phD, I focused on the investigation of these diverse causes and consequences by functional analysis of -omics data through a variety of different analysis approaches.

4.1 Causes of impaired genome integrity: DNA damage

Unrepaired DNA lesions have a wide range of detrimental phenotypes. Therefore, cells evolved an intricate toolbox of mechanisms to maintain genomic integrity during all phases of the cell cycle. Defects in DNA repair pathways lead to severe structural and numerical changes and chromosome aberrations (Figure 8) [125]. The identification of DNA repair factors as well as the characterization of DNA repair pathways is crucial for understanding how cells maintain genome integrity. Therefore, I analyzed a dataset consisting of 664 mass spectrometry measurements from DNA repair experiments and created a web-application for the visualization and mining thereof.

4.1.1 Characterization of the combined DNA damage repair dataset

To investigate the recruitment of proteins to different DNA lesions, proteomics measurements resulting from CHROMASS and PP-MS studies were combined in an expansive dataset. The experiments measured the time-resolved recruitment profiles of DNA repair factors for different lesions, including interstrand crosslinks, double-strand breaks, DNA-protein crosslinks and fork collapse (Figure 16). This allowed for highly robust temporal profiling of DNA repair factors under many different experimental conditions and treatments. The significance analysis of this combined dataset (Figure 18) highlighted the presence of a set of universally enriched factors (Table 1) consisting of many known and well-described DNA damage factors, such as SMARCAL1 [288], the regulating single-strand binding factors RPA1, RPA2 and RPA3 [249], as well as FANCD2 and FANCI [251].

These factors can also be found again in the related 'combined volcano plots' (Figure 19 & 20, Supplementary Figure 2 & 3), which were created by grouping the combined dataset by DNA lesion and calculating individual enrichment scores per protein. This shows a high reliability for the identification of lesion-specific DNA damage response factors and agrees with previous studies that contributed to the combined dataset [162,163,247]. The combined plot for ICL shows the RAD18-SLF1-SLF2 complex, that recruit SMC5/6 to stalled replication forks [162] (Figure 11 and 19). The combined plot for pDPC shows the degradation mediator SPRTN and the ubiquitin ligase TRAIP [163], excision factors such as ERCC1 and ERCC4, and the DPC bypass facilitating helicase RTEL1, which unwinds the DNA past the DPC for the CMG complex (CDC45, MCM2-7, GINS) [247] among the highly enriched factors (Figure 12 and 20). Overall the identification of well-described factors, both universally enriched and specific for the repair of different DNA lesions, shows a high fidelity of the dataset for DNA repair factor identification and characterization. As the computationally acquired results stand in conclusion with previously published findings, novel insights drawn from the dataset can be used to predict the function of proteins of interest or investigate their behavior in different DNA repair pathways.

A novelty of the dataset is that we can investigate the recruitment of DNA repair factors while distinguishing between the sub-genomic origins of most genes. In all shown plots it can be observed that these factors are expressed from both sub-genomes of the

allotetraploid model organism *Xenopus laevis,* as shown by the suffix ".L" (large) and ".S" (small) in the gene names. I investigated if there is a selective bias towards the expression of DNA repair factors for one of the two sub-genomes (Figure 21). In agreement with the sequencing data reviewed in [256], which found that additional copies of genes associated with DNA repair are lost at a high rate, we found only limited overlap of repair proteins expressed from both sub-genomes (Figure 21 B). Despite this, the global distribution of DNA repair factor expression shows no conclusive trend towards one of the two sub-genomes (Figure 21 A). We conclude that the expressed proteins from both sub-genomes have to be considered as there is no overall bias towards the expression of either one for the DNA damage response.

Taken together, we present a robust combined proteomics dataset originating from a vast collection of DNA repair experiments, which shows both a high fidelity for the identification of universal DNA damage response proteins as well as factors recruited specifically for different DNA lesions and is therefore suited for the identification of novel repair factors.

4.1.2 The DNA repair atlas as a tool for the visualization and mining of DNA repair data

The creation of a web-application for visualization and mining of the dataset solved many issues of a traditional analysis of the data. By merging and preprocessing the data of all PP-MS and CHROMASS experiments, an easily accessible system-wide analysis can be facilitated for scientists without a background in computational biology. Through the correlation of data and calculation of lesion-specific enrichment scores, the combined dataset is dynamically approachable in different ways. The presented web-application, which consists of several independent modules, allows scientists to access this dataset in real-time starting from a protein of interest and investigate its immediate proximity in the network of all significantly recruited proteins, or in combined volcano plots that show lesion specific enrichment. Therefore, the DRA can serve as an entry point to generate hypotheses for further experimental studies.

A bottleneck in previous *Xenopus laevis* studies was the incomplete annotation of expressed genes. This forced scientists to manually search different databases, which

use individual identifiers, to collect information about their protein of interest. To alleviate this process, I created a combined annotation database by mapping the resources UniProtKB [17], XenBase [260], NCBI [18] and Ensembl [289] and filled the missing gene names by using BLAST [261] to the human UniProtKB FASTA. This complete annotation facilitates a more accessible analysis with the DRA, as every visualized protein is mapped to a single identifier and interactively linked to all related resources that were used for data annotation. The module 'Annotation Data' in the DRA thereby can function as hub to cross-reference multiple commonly used *Xenopus laevis* web-resources for easier data mining.

To facilitate a network analysis without the need for computationally demanding hierarchical approaches, we implemented a clustering algorithm, which starts from proteins of interest and investigates only the direct proximity within the network. This clustering allows users to define modules and identify novel members based on the concept of guilt-by-association [88]. Therefore, the DRA can be used to predict novel DNA repair factors by their association with already known DNA repair pathways.

How the DNA repair atlas can be used for the generation of novel hypotheses was showcased by the creation and analysis of a representative subnetwork from 10 well-described and frequently enriched DNA damage factors associated with inter-strand crosslink repair: FANCA and FANCB [290], RTEL1 [291], BARD1 [292], TOP3A [293], RAD21 [294], BRCA2 [295], PALB2 [296], REV3L [297] and RAD51 [298]. Using bottom-up clustering results in a subnetwork, which contains several modules that correlate with known DNA damage response mechanisms despite only individual subunits being used in the list of input proteins (Figure 30). We identified multiple uncharacterized proteins, which are associated with the modules of which PROSER3_BL was investigated in more detail. Both the function of the protein in *Xenopus laevis* as well as the homologous protein in human, to which it shows a 28.9% overlap by BLAST search, are unknown. The gene encoding the human PROSER3 was described in a recent analysis of the TCGA database as a predicted gene associated with the recurrence of papillary thyroid carcinoma [299]. Investigating its proximity in the network by a modified clustering and performing a gene set enrichment of the result showed double-strand break repair proteins as potential interaction partners (Figure 32). Further investigation of those highlighted the candidate not only in proximity in the network, but also in related volcano plots (Figure 33 & 34) showing the recruitment of

repair factors to site-specific DNA-protein crosslinks. Its presence together with HR proteins could be due to impaired MCM helicase bypass and template switching coupled with recombinational repair [275]. Taken together those indications predict a role of the protein in the repair of DPCs by interacting with homologous recombination proteins. This showcase intends to highlight a potential hypothesis generating workflow with the DNA repair atlas. The precise function of PROSER3 and other potential candidates has to be further determined in experimental studies.

We aim to further improve all aspects of the DRA. For example, the clustering currently uses only empirical scores for the diffusion based on a users input, while its functionality would also allow different scores to be diffused. A promising addition would be the integration of further datasets. For example, the identification of clusters based on calculated mutation frequencies in different cancer sets of the cancer genome atlas or alternatively, to allow users diffuse custom list of scores. Further, the clustering is currently based on an RWR from a user-defined list of input proteins. A meaningful way to curate the input list of proteins could be facilitated by shortest path analysis according to Dijkstra or Bellman-Ford, which would allow the identification of all proteins that connect proteins of interest in the network for further cluster analysis [300]. Currently, the identification of shortest paths in a dense network including a high number of edges is computationally too demanding to facilitate real-time analysis in the web-application. The back-end of the application is written in a way that is easily transferable to multiple server solutions. It first was hosted on the Microsoft cloud-computing platform Azure and later migrated to a local Windows server utilizing IIS web-hosting. For future development, moving the application to a more performant server would facilitate further in-depth analyses in real-time as well as the integration of new datasets, while being less limited by computation times. Despite this room for improvement, the DRA altogether provides a well-rounded resource for the mining and visualization of DNA repair data.

Lastly, by utilizing a modularized structure established by the functional programming language F#, the application can be easily developed further, scaled by the integration of new modules, or expanded by novel datasets while ensuring backwards compatibility. This adaptability is demonstrated by the standalone supplementary application ‘PloidEx’, which uses the DRA framework to visualize ploidy dependent regulation of proteins and mRNA [246].

While there are multiple commonly used web-resources for *Xenopus laevis* genomic, epigenomic or transcriptomic data, mostly revolving around XenBase as central hub [301], and multiple resources containing meta-information about identified proteins [17,18], there are currently no resources that allow the interactive visualization and mining of the time-resolved recruitment of proteins to different DNA lesions. The DNA Repair Atlas therefore aims to serve the scientific community as a novel resource that allows scientist to find out more about DNA repair factors in an expansive proteomics dataset, which also has the novelty of containing information about the sub-genomic origin of proteins. As a current limitation of the DRA is that its functionality is restricted to this dataset, future developments should aim at expanding the application by the integration of further meaningful datasets and modules or by facilitating the analysis of external data.

4.2 Consequences of monosomy

Aneuploidy, defined as a state of the cell with unbalanced chromosome number, is generally described as being detrimental for the development, proliferation and viability in most circumstances. Despite this, aneuploidy is also a striking hallmark of cancer [203]. Most of our current knowledge about effects of aneuploidy on human cells is derived from cell lines with gain of a chromosome. Previous multi-omics studies have shown that gain of a chromosome leads to a scaling of mRNA expression according to the increased DNA copy number, while the protein abundance is compensated closer towards diploid level. This characteristic adjustment was linked to proteins that are subunits of multi-molecular complexes [79,193,199]. Further studies highlighted that aneuploidy triggers a unique stress response pattern and genomic instability [202,302]. However, there have been no systematic studies of consequences of chromosome loss. Therefore, we focused in our study on uncovering the cellular consequences of chromosome loss through an integrative proteome and transcriptome analysis of monosomic cells derived from human RPE1-hTERT cells (Figure 35).

4.2.1 Chromosome loss effects on transcriptome and proteome expression

Very little is known about gene expression changes in response to chromosome loss. A recent transcriptomics analysis of human blastocysts investigated altered transcript abundances for several hundred genes located on monosomic chromosomes and genome wide [303]. While a general deregulation of monosomy-encoded genes was shown, this study could not adjust for genetic diversity of unrelated embryonic cells and generally analyzed only a low number of transcripts. Further, it lacked a diploid parental control to study the extent of deregulation and identify exact scaling with DNA copy number. To gain a better understanding of the consequences of chromosome loss, we performed RNASeq and both TMT labeled and label-free mass spectrometry to quantify both transcriptome and proteome of monosomic and diploid parental controls.

In this analysis we showed that the overall expression of both proteome and transcriptome on the monosomic chromosomes did not scale to the expression level of $\log_2$ fold change -1, which was expected according to its DNA copy number

(Figure 36, 40, 41 Supplementary Figure S11). This systematic analysis reveals that only around 20% of monosomically encoded proteins scale to the expected expression level, while the majority of genes is compensated closer towards the diploid expression. Analysis of gene dosage compensation in previous studies in Drosophila cell lines suggest gene specific mechanisms adjusting the mRNA levels [304]. We observed both transcriptional and post-transcriptional mechanisms compensating towards diploid level (Figure 42).

Strikingly, this is in contrast to previous studies of aneuploidy in gain-of-chromosome cell lines, that found only a marginal compensation of mRNA level, while approximately 25% of proteins were present at diploid level [79]. In a direct comparison of combined monosomic and trisomic cell lines I found both an overall stronger relative compensation for monosomically encoded proteins and mRNA than trisomically encoded ones (Figure 47). Our findings are in agreement with a recent pan-cancer study that investigated a collection of human cancer cell lines and also identified both transcriptional and a stronger post-transcriptional dosage compensation upon chromosome loss [305].

We conclude that the overall compensation effect in monosomies is stronger than in trisomies and happens on both protein and mRNA level, while protein level appears stronger adjusted than mRNA. The exact underlying mechanism thereof remains to be investigated, yet we imagine two likely explanations: Translation of mRNA encoded on monosomic chromosomes is selectively increased or protein degradation of the related proteins reduced. The shown findings indicate that cells utilize multiple routes to alleviate the consequences of gene expression changes in response to chromosome loss.

Previous studies of gain of chromosome cell lines also linked this dosage compensation to proteins belonging to multi-subunit complexes as annotated by the CORUM database [79,306]. Stabilization and decreased degradation of proteins upon incorporation into complexes could contribute to post-translational buffering [306]. I repeated this analysis, but could not detect a comparable effect in individual monosomic cell lines (Figure 43). Similar analyses for membrane or cytosolic proteins also did not reveal a selective trend linking to dosage compensation (Supplementary Figure 12). Therefore, I used another approach to characterize

dosage compensation in monosomic cells. I created subsets of compensated genes of all monosomic chromosomes depending on overall expression and difference between mRNA and protein levels (Figure 44). Enrichment analysis of genes with a higher protein than mRNA expression, that are likely post-transcriptionally compensated towards diploid levels, showed a specific enrichment for pathways associated with ribosomal biogenesis and rRNA processing. Intriguingly, the recent study that performed a comparable analysis of dosage compensation in aneuploid cancers of the CCLE also identified ribosomal biogenesis and rRNA processing enriched for proteins buffered upon chromosome loss [305]. Recent studies suggested that haploinsufficient genes reduce cellular fitness when expressed outside of the range of their usual expression levels [307]. Even subtle shifts in the availability of ribosomes impair cellular growth and proliferation [308]. It is at this point still unclear, whether there is a specific dosage compensation mechanism that adjusts genes involved in ribosomal biogenesis to specifically compensate for this highly sensitive ribosomal haploinsufficiency, or if the found enrichment is based on individual deregulation of genes on protein and transcript level as consequence of chromosome loss and the resulting stresses. Further, the study mentioned above [305] also identified a trend for protein complex subunits as buffered on chromosome loss, which was not detected in our analysis. This study has the limitation of investigating a sensitive effect by an analysis across a dataset of genetically-diverse cell lines from a variety of cancer lineages, including various degrees of aneuploidy and sub-chromosomal alterations. Despite this, they observe a significant difference for buffered and CORUM annotated proteins compared to other proteins and further identify the trend by gene set enrichment. This was found significant with a precision of 36% for buffered proteins annotated by any CORUM annotation (2128/5934). For comparison, in our combined subset of post-transcriptionally compensated genes 31% can be annotated for any CORUM annotation (61/194). The difference in buffering effect potentially can be explained by cancer cell evolution that did not take place in our cell lines. As the CCLE data is derived from patient cells, cancers potentially acquired a stronger dosage compensation mechanism for multi-molecular protein complexes.

There are also limitations to our analysis. Recent advances in mass spectrometry allowed us to investigate protein expression with a high sensitivity, yet the overall number of investigated monosomic cell lines is still low. While I alleviate this by

combining all monosomically encoded genes in a combined analysis (Figure 44 onwards), specifically for an analysis of gene dosage compensation a higher sample size of genes encoded on monosomic chromosomes would be highly beneficial. Additionally, while there is no technical bias that specifically effects monosomic gene expression and the analysis overall benefited greatly from the increased sensitivity of TMT labelling (Figure 48), isobaric labelling suffers from a systematic underestimation of peptide/protein ratios due to precursor ions co-isolating with other precursors, leading to a distortion of the reporter ion patterns [41]. Future proteomics measurements to investigate dosage compensation effects therefore could use alternative labelling approaches, such as EASI tags that provide similar sensitivity and show no such ratio compression effects [42]. It should also be noted that the investigated monosomic cell lines were p53 deficient, while the trisomic cell lines used for comparison were p53 proficient. This deficiency had little influence on the global cellular response to monosomy (Figure 37), hence it is unlikely to influence this analysis. Lastly, the monosomic cell lines were created by several single cell cloning steps. Therefore, the cell lines that adapted to cellular stresses associated with monosomies potentially also were selected for dosage compensation mechanisms.

We can conclude that cells can adjust both mRNA and protein expression levels to compensate for the loss of a chromosome to alleviate the detrimental consequences on a cellular level. The exact mechanism of how and which genes are targeted, as well as what influences the extent of the regulation on transcriptional or post-transcriptional level are to be investigated in further studies.

4.2.2 Chromosome loss effects effects on global gene expression and pathway regulation

Chromosome loss is not only reflected in the expression of genes on the monosomic chromosome, but also effects the global cellular gene expression of the cell. In our multi-omics analysis, we investigated the global effect of the loss of chromosome 10/18, 13, X and 19p on both mRNA and proteome level and compared them to trisomic gain of chromosome cell lines. For gain of chromosome a characteristic “aneuploidy specific stress response” was proposed in previous studies that involves downregulation of ribosomal subunits and biogenesis and cell cycle regulation,

together with an upregulation of lysosomal pathways, membrane metabolism and regulation of autophagy [79,202]. I performed two-dimensional pathway enrichment [78] to investigate if loss of chromosome has a similar response pattern. Strikingly, the pathways enriched on proteome and transcriptome did show only a limited overlap between different monosomic cell lines (Figure 50 & 51). Among the enriched biological processes and cellular compartments, only the terms associated with ribosomal subunits and translation did show a downregulation in all monosomies and overlapped with the aneuploidy stress response in trisomies (Figure 52 C, D). It is further observable that the similarities between different monosomies are low, due to the little overlap of differentially regulated genes (Figure 53 A, B). The comparison with the trisomic cell lines revealed only a low overall correlation between expression changes in cellular response to a gain and loss of chromosome on a global level (Figure 52 A, B). As explained in the previous section, it is unlikely that this is due to the deficiency of p53 in monosomies.

The enrichment analysis overall showed a heterogenous cellular response to the loss of different chromosomes. Immune related pathways, such as *MHC protein complex,* were upregulated in RM 10;18 but downregulated in RM 13 (Figure 51 A). *ATP synthase complex* and *mitochondrial respiratory chain complex I* were downregulated in RM 19p and RM 10:18, yet upregulated in RM X (Figure 51 B, C, D), while the *mitochondrial electron transport*, which mainly consists of NADH dehydrogenase proteins, appears to be specifically downregulated in RM 19p (Figure 51 C). Similarly, other pathways were deregulated in individual monosomic cell lines.

These deregulations can be explained by the specific genes encoded on the lost chromosome. For example, for oxidative metabolism we found several genes encoding for NADH dehydrogenase proteins localized on the p-arm of chromosome 19 (Figure 54 B). Likely the loss of a copy of several participating genes and the resulting lower expression leads to a downregulation of the entire complex. We therefore could not identify a shared cellular response pattern to chromosome loss, as it could be proposed in previous studies for chromosome gain [79,202,302], but rather a heterogeneous global gene expression based on which chromosome is lost.

The only consistently downregulated pathways across different monosomies were cytoplasmic ribosome subunits and translation. In further agreement with the

observation that genes lost on monosomic chromosomes are responsible for the differential regulation of pathways, every human chromosome except chromosome 7 and 21 carries at least one ribosomal gene [309]. This is also true for chromosome X, where the haploinsufficiency of RPS4X, a protein that escapes dosage compensation, contributes to the pathophysiology of Turner syndrome [310]. We hypothesized that the loss of a chromosome likely leads to a reduced abundance of ribosomal subunit genes. As other pathways that affect translation efficiency were not changed in response to chromosome loss, we overall propose impaired ribosome biogenesis caused by haploinsufficiency of ribosomal protein coding genes as a shared consequence of monosomy. Ribosomal biogenesis was identified as enriched dosage compensated pathway in both our analysis (Figure 45) and the study of CCLE data [305]. Therefore, we conclude that this consequence is alleviated by gene dosage compensation. This further suggests that the targeting of ribosomal biogenesis as shared consequence to chromosome loss might prove to be an attractive approach in cancer treatment, which should be investigated in future studies.

The analysis of the differential regulation of pathways in response to chromosome loss revealed several up- and downregulated pathways and complexes that were subsequently experimentally validated. For example, we experimentally confirmed the regulation of interferon stimulated genes by quantitative PCR analysis, which leads to the hypotheses that there is a chromosome 13 specific regulation of the interferon response that is yet to be further explored. The impact of chromosome loss on ribosomal gene expression was investigated by a puromycin-incorporation assay, which confirmed a significantly decreased translation rate. In recent studies of Down syndrome mice models, a similar translation defect was observed, which could be accredited to an “integrated stress response” (ISR) [311]. In monosomic cells immunoblotting of the eukaryotic initiation factor 2 alpha (eIF2α), whose increased phosphorylation is used as a marker of ISR, did not confirm increased phosphorylation in our model. Investigation of mTOR, the regulator eukaryotic protein synthesis and ribosomal subunit gene expression [312], by immunoblotting of the mTOR target p70S6K and its phosphorylation did show no reduction in activity in monosomic cell lines. In conclusion, the decreased translational activity and ribosomal gene expression is not due to changes in mTOR or ISR. Furthermore, polysome profiling revealed higher accumulation of ribosomal subunits and polysomes, composed of one ribosome

residing on an mRNA as well as unassembled small (40S eukaryotes, SSU) and large (60S in eukaryotes, LSU) ribosome subunits, in monosomies than diploid cell lines. Altogether, the hypotheses that impaired ribosome biogenesis is a general consequence to loss of chromosome could be validated. The experimental procedures were performed by Narendra Chunduri and Vincent Leon Gotsmann.

Future approaches could aim at increasing the number of different monosomic cell lines, to get a broader picture of the global response to different chromosome losses. Copy number changes of oncogenes and tumor suppressors are hypothesized to be a driving factor in the evolution of cancers [313]. Hence, it would be promising to analyze more cell lines with chromosome losses that encode for tumor suppressors and oncogenes to investigate how chromosome loss specifically effects the expression of these cancer-relevant genes. Similarly, the functional consequences of individual pathway deregulation for different chromosome losses have to be explored in further experimental studies. As aneuploidy is an important prognostic factor across all cancer types [314,315], a better understanding of how cells react to the loss of different chromosomes may lead to better therapeutic approaches for cancer treatment. Lastly, previous attempts to rescue the translation defect in monosomic cells by restoring levels of RPL21 were not successful. This is likely due to the strong regulation of ribosomal gene expression that makes ribosomal subunits resistant to overexpression [307,316] and requires novel, experimental approaches to ultimately validate our hypothesis that impaired ribosomal biogenesis by haploinsufficiency of ribosomal genes is the general consequence of monosomy.

4.3 Consequences of polyploidy

Polyploidy is commonly found in different eukaryotic organisms. It plays an important role in speciation and can facilitate resistances against environmental stress that would not be tolerated by diploid cells, by increasing the adaptive potential on the cost of potentially disrupting effects [232]. This is facilitated by nuclear and cell enlargement and higher levels of genomic instability [233-235]. Polyploid cell divisions can lead to aneuploidy and whole genome doubling is commonly found in human cancer cells [236]. Here, polyploidy impairs genome integrity, increases adaptability and is a driving factor of tumorigenesis [237,238]. To further investigate the effect of increased ploidy on cellular gene expression, we analyzed the global proteome and transcriptome changes of yeast cells in response to increasing ploidy.

4.3.1 Proteome and transcriptome scale non-linearly in response to increasing ploidy

Previous analyses of model organisms of different ploidies revealed that increased ploidy has a variety of effects on cellular physiology, including lower fitness [244,317], genomic instability [235], reduced proliferation [241,318], altered metabolism [236] and an increase in both cell size and nuclear volume [234]. The mechanisms leading to these effects are only partially understood. To investigate how gene expression changes with increasing ploidy, we investigated a series of isogenic haploid to tetraploid yeast cells (1N to 4N), derived from the BY4748 strain background. As expected, the overall cellular volume did increase linearly with ploidy (Figure 56) [234,241]. I analyzed the proteome, measured by SILAC, and integrated the available transcriptome data, measured by dynamic transcriptome analysis, to assess the gene expression changes in the strains with different ploidy. Due to the data acquisition strategies (Figure 57), namely using equal amounts of cells and global, internal standards it was possible to draw conclusions about scaling of proteome and transcriptome content with gene copy number.

This first study of the proteome of yeast cells with different ploidy revealed that the overall protein abundancy did not scale linearly with increasing ploidy like the cell volume. (Figure. 59, 2N: 1.61, 3N: 2.31, 4N: 2.95). The scaling is rather allometric,

similar as described for the scaling of metabolic rate with body size with an exponent close to ¾ [319,320], or with surface area with an exponent of 2/3 [321]. This is surprising, as linear scaling of both genome content and cell volume accompanied by an allometric scaling of proteome leads to a reduced ratio of DNA-binding proteins to DNA and cell dilution. Stronger increasing volume compared to proteome content leads to an unbalanced availability of proteins required for cellular processes and structures and therefore can manifest in a range of detrimental effects, from impaired gene induction to cell signaling and cell cycle progression [322]. Tetraploid yeast cells often undergo ploidy reduction [237,318] and stable tetraploid cells reduce their cell size again over the course of multiple generations when maintaining the karyotype [323]. This indicates that polyploid cells select for lower ploidy and size, potentially to re-balance protein content.

We termed the observed allometric scaling 'ploidy-specific protein scaling' (PSS) and further investigated, whether protein and mRNA expression strictly follow the PSS trend or are differentially regulated. This analysis showed that the expression of mRNA with increasing ploidy overall followed the PSS (Supplemental Figure S14) and as expected, we found only marginal changes in differential mRNA expression (Figure 61). In agreement with previous publications, only 13 mRNA were found to be differentially regulated with increasing ploidy, containing transcripts associated to plasma membrane and cell wall synthesis. This likely represents an adaptation to lower surface-to-volume ratio in larger polyploid cells [238,241,245]. Overall, the minimal differential regulation of mRNA raised the question if ploidy-dependent regulation occurs post-transcriptionally.

I therefore analyzed differentially regulated proteins with increasing ploidy. We observed that the largest part of the proteome follows the PSS, but several proteins are differentially regulated with increasing ploidy, which we identified by different filtering approaches. This revealed ploidy-dependent upregulation for multiple proteins associated with cell wall integrity, and downregulation for translationally involved, ribosomal and mitochondrial proteins (Figure 62). We termed this 'ploidy dependent regulation' (PDR). Considering the scaling of size and volume with ploidy, cells likely upregulate cell wall integrity proteins in response to this size increase. We further observed downregulation of multiple mitochondrial proteins and consistently, an upregulation of RTG2 with increasing ploidy (Figure 62 A, C), which plays a central

role in the retrograde signaling of mitochondrial dysfunction to the nucleus [324]. This indicates a ploidy-dependent deregulation of mitochondrial function.

There are limitations to this analysis. I identified a higher variance for the replicates of the haploid cell line in the SILAC analysis compared to cell lines with higher ploidy (Figure 58 A). This increased variance manifests two-fold, once in between the replicates of 1N and once between 1N and higher ploidies. This effect is slightly amplified by the normalization against a SILAC standard that consists of an equal part of protein from all ploidies and potentially influences the higher expression range of 1N compared to other ploidies (Supplementary Figure S15). For this analysis identifications of differentially regulated proteins by comparison with 1N therefore may include a slight, technical bias (Figure 62 A, B). To alleviate this effect, I created a second filter that removes expression spikes between consecutive ploidies, which did not exclude any of the discussed differentially regulated proteins in the previous paragraph (Figure 62 C, D).

The findings could be experimentally validated. Protein scaling was validated by immunoblotting of abundance changes to determine independent measurements of protein concentration (Figure 60). Transcriptome abundance changes were validated by qPCR (Figure 61 B, D) and the identified protein abundance changes were validated by immunoblotting in both a haploid G1 cyclin cln3Δ lacking mutant and a respiration deficient rho0 mutant (Figure 63). The experimental validation was performed by Galal Yahya and Devi Ngandiri. This proved that the observed ploidy-dependent scaling is not due to a technical limitation due to the SILAC analysis, nor normalization. Further, as the analysis of haploid mutants with cell volume comparable to diploids did not show protein abundance changes of selected candidates comparable to polyploids, the differential regulation of proteins is induced by increasing ploidy instead of cell volume. This observation is intriguing, as previous studies of yeast transcriptome showed cell size dependent regulation [245], or in a more recent hypothesis size-to-DNA ratio dependent regulation [325], unlike the observed ploidy dependent protein regulation. Overall, the levels of protein expression for differentially expressed candidates were not found on transcriptome level, hence the ploidy-dependent regulation of protein abundance appears to occur largely post-transcriptionally. This is likely due to proteome changes being adapted more rapidly than changes to transcriptome, which has to be investigated in future studies.

4.3.2 Polyploidy effects on differential pathway regulation

Increased ploidy has a variety of effects on cellular physiology. To investigate the pathway regulation with increasing ploidy, I performed two-dimensional pathway analysis and gene set enrichment of significantly deregulated genes. In comparisons of different ploidies, only the respiratory electron transport chain was found as consistently downregulated cellular compartment in all ploidies except 1N (Figure 64 & 67). This is in agreement with deregulation of mitochondrial functions in cells with higher ploidy, which recently was observed in studies in *Candida albicans* [283], as well as with the identification of differential regulation with increasing ploidy of multiple mitochondrial proteins and RTG2, described in the previous paragraph (Figure 62). Furthermore, ribosomal and translational pathways showed a strong downregulation with increasing ploidy. It has to be noted, that the technical limitation, which potentially introduces a slight bias in comparisons with 1N has to be considered for this analysis, similar as in the previous paragraph. To alleviate any influence of technical effects, I performed gene set enrichment analysis of proteins that were significantly deregulated with increasing ploidy, as identified by two-sample t-tests. As the potential technical bias leads to higher variance in replicates of 1N, affected proteins are unlikely to be significant. This analysis confirmed the downregulation of ribosomal subunits and translation with increasing ploidy (Figure 64 & 65). The expression changes of ribosomal proteins highlighted a homogeneous downregulation with increasing ploidy across most ribosomal proteins (Figure 66). Taken together, any technical bias is unlikely to influence the findings.

The identification of ribosomal and translational deregulation is striking, as ploidy dependent regulation thereof could explain the allometric scaling of protein content. The gene set enrichment of transcriptome data showed only mild overall enrichment and no overlap with proteome (Figure 65 C, D), hence mRNA changes could not explain altered ribosomal protein abundance. We therefore hypothesized that rRNA production is reduced with increasing ploidy.

This was confirmed by qRT-PCR that has shown reduced abundance of 25S and 5.8S rRNA in polyploids. Further, we used pulse labelling with puromycin to measure translational efficiency. This revealed that the relative translation rate non-linearly increases with ploidy for equal cell loading, confirming the observed proteome

quantification. Cells of higher ploidy were more sensitive to the translation inhibitors puromycin and rapamycin, an regulator of mTOR that regulates ribosome biogenesis in eukaryotes [326]. Experimental studies by Galal Yahya have linked this to the novel mTORC1-SCH9-TUP1 signaling pathway. With increasing ploidy TOR-SCH9 activity is reduced, which leads to an accumulation of TUP1, homolog of the human mediator TLE1 [327], which negatively regulates rDNA expression.

In previous studies it was shown, that ribosome biosynthesis in response to various stresses, such as heat, osmotic or environmental stress or nutrient deprivation is heavily regulated [328,329]. As cell volume is increased with ploidy the uptake of nutrients could be impaired. Nutrient deprivation is sensed by the cell and leads to reduced mTOR activity (reviewed in [330]). We hypothesize that the increased cell volume therefore leads to the reduced translation and rDNA expression. Additionally, mitochondrial defects, indicated by the observed deregulation of respiratory electron transport chain and upregulation of RTG2 with increasing ploidy (Figure 62 & 64), could contribute to the PDR of ribosome biogenesis by alteration of mTOR activity [331]. In future studies it should be further explored, which mechanisms exactly lead to the reduced expression of mTORC1 and SCH9 in polyploid cells. Hence, an interesting analysis would also be to integrate metabolome data of polyploid cells, to identify how the global cellular metabolism is adjusted in response to the observed proteome remodeling with increased ploidy.

In an analysis of near-tetraploid HPT2 (HCT116, **P**ost **T**etraploid) cells by Galal Yahya, four of six tested ribosomal proteins showed a reduced expression compared to diploid HCT116, which by analysis of previously obtained transcriptome data were also found to be regulated post-transcriptionally [332]. Additionally, the relative rRNA expression was found to be reduced compared to parental controls. Recent studies in Salmonidae revealed that genes encoding for ribosomal subunits and mitochondrial proteins are frequently lost during evolution after whole genome doubling [333]. Taken together these findings suggest a conserved response to increasing ploidy to ensure the survival of the cell. For future approaches our findings may help to explain phenotypes of tetraploid cells, which previously could not be explained by the minor changes on transcriptome level, hence provide new hypotheses for the understanding of ploidy-specific mechanisms.

To summarize, this comprehensive study presented the use of multi-omics data analysis as a versatile tool to investigate the various causes and consequences of impaired genome integrity. An analysis of an expansive collection of CHROMASS and PP-MS proteomics experiments was performed and a novel tool presented, which allows scientists without any computational background to interrogate this dataset. Furthermore, multi-omics data of somatic human cells suffering from chromosome loss were analyzed time for the first time and the cellular consequences as well as adaptions to tolerate the severe effects of monosomy shown. Lastly, this study presented a first analysis of proteome and integrated transcriptome data of yeast cells with different ploidy and showed both the scaling of proteome with ploidy and ploidy-dependent regulation of proteins. Altogether this study intends to show how state-of-the-art multi-omics analysis can uncover cellular responses to impaired genome integrity in a highly diverse field of research.

5.1 The DNA Repair Atlas

Processing of a combined proteomics dataset of DNA repair data

To prepare the data for the combined dataset, .raw files resulting from proteomics experiments were processed with the MaxQuant software version 1.6.17 using the label-free algorithm [24]. 664 mass spectrometry files were processed in a single run on a Hyper-V server cluster using parallel processing on 32 cores and searched against the UniProtKB FASTA of *Xenopus laevis* (UP000186698, Swiss-Prot and TrEMBL). Cysteine carbamidomethylation was set as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. Trypsin/P was set as protease and a maximum of two missed cleavages was accepted. False discovery rate (FDR) was set to 0.01 for peptides (minimum length of 7 amino acids) and proteins and was determined by searching against a generated reverse database. Peptide identification was performed with an allowed initial precursor mass deviation up to 7 ppm and an allowed fragment mass deviation of 20 ppm. Match-between-runs was turned on in each of the experiment groups. MaxLFQ normalization was performed independently for each parameter group. Groups are shown in Table 3.

Table 3 | Parameter groups. The table shows the used parameter groups for processing of the combined dataset.

Parameter group	Experiment groups	Treatments
1	1p85, 1p86, 2p176	Aphidicolin
2	1p90, 1p111, 1p120, 1p170, 2p108, 3p20, 3p34, 3p38, 3p47, 3p112	Psoralen plus CTR plus BRC, UBVS
3	1p111, 1p120, 1p170	Psoralen Mock
4	1p118	Topo Tecan
5	3p42, 3p132, 3p140, 3p158, 3p180, 3p184, 4p8, 4p83	DSB HSS or HSS/NPE
6	3p172, 3p178	DSB LSS
7	3p186	MMS /CPT
8	3p75	UV treated sperm in HSS or HSS/NPE
9	4p70, 4p107	pNICK
10	4p33, 4p44	pDPC
11	4p44	pDPC - Mock
12	4p62	pICL
13	4p76	Termination Set 1 (Dewar)
14	4p184	Termination Set 2 (Sen)

Analysis of MS data

To identify significantly deregulated proteins, the raw data were grouped based on the experimental conditions they were acquired in replicates as well as by treatment for each experiment and their respective control. This list of 139 categorized measurements was used to compute an independent series of Student's t-tests for all proteins identified. The null-hypothesis S0 was set in a data dependent manner using 'SamR, significance analysis of microarray data' [62] after filtering for three valid values in the tested condition. This applies a modified t-test with an optimized false discovery rate (FDR) to adjust for multiple testing. Significance is controlled by a false discovery rate threshold of >5 % and a minimal fold change of at least 1.5. Results were combined and the product of $\log_{10}$-transformed p-values was computed for each protein over all experiments and for each type of DNA lesion individually where it has shown a positive fold change of treated sample to the used control. This yielded a set of a DNA lesion specific 'enrichment scores' similar in concept to the statistics 'combined score' in Perseus [49]. These scores were then scaled in an interval of 0 to 1 to improve comparability between damage conditions. Additionally, the count of comparisons, in which each protein was significantly enriched in the sample was calculated and a filter applied to create a subset with only significant proteins with at

least 4 significant comparisons resulting in a subset of 2418 proteins. The subset was z-scored across similar treatments. Pearson and Spearman Rank correlation datasets were created of the set of significant proteins as an input for the DRA, together with a table of all LFQ data and a similarly z-score normalized set of all data. The tables were uploaded to the supplementary data module in the DRA (password: DNARepairAtlas21).

Network creation and cluster analysis

Upon the application's start-up, an undirect network graph *G* is created from the correlated Pearson and Spearman Rank datasets. This graph is a paired list of nodes and edges $G = (V, E)$, where V is a set of nodes, and $E = \{(u, v, w) | \, u, v \in V\}$ the edges which connect them. Each node represents a protein significantly enriched in the dataset. An edge is defined as a node pair consisting of a start and end note. The weight w of the edge presents the information about the strength of the pairwise correlation and is stored as a third parameter with the edge.

Bottom-up clustering is based on a random walk with restart (RWR) algorithm [257]. On initialization of the cluster analysis, every start node representing a set input protein will be assigned a score of 1, indicating that the assumed function of the protein is already empirically known by the user. Any other node in the network will be assigned an initial score of 0. By default, the random walk with restart will operate with a 25 % reset probability, which determines the chance to return to a random member of the list of start nodes before a further step is performed. After passing the reset probability the random walker will take one step from its current node by choosing from all directly connected neighbors, whereas the chance a node is chosen is proportional to the weight of the edge connecting them. This is calculated by transforming the weight of all connected edges to a percentual representation of the total weight of each connected neighbor and generating a system-time derived random number between 0 and 1 to choose from the considered edges. By default, the walker will perform 100.000 steps to diffuse the initial score to a steady-state. After the set number of steps is performed, the final score is calculated for each node in the network according to the amount of times the walker visited it.

Back to front end communication and real-time data analysis for the DRA

The functional programming language F# was used to implement a web-server relying on a request/response architecture to steer the distribution of the data using the library Suave (https://suave.io/ Version: >2.2.0), which controls the data in form of a asynchronously computed type 'WebPart'. Preprocessed data is imported on application startup by custom parsers as well as the library FSharp.Data (http://fsharp.github.io/FSharp.Data/ Version: >2.2.3). Following real-time processing based on the users input the data it is stored as JavaScript Object Notation (JSON) to establish a straightforward communication with the front-end JavaScript libraries. The NewtonSoft.Json.Fsharp (https://github.com/haf/Newtonsoft.Json.FSharp Version: >10.0.3) library is used for the conversion of results to JSON. This allows real-time communication between front and back end, as well as handling of user requests and searches without the need of manually refreshing the website. Online usage of the DNA Repair atlas does not require additional installations, programming or statistical training. The DNA repair atlas as suave.io-based, standalone application can also be started offline via .NET Framework 4.6.1. The pre-compiled version can be started with the DNARepairAtlas.exe in the Release folder. It will by default open http://localhost:8083// in the systems default browser a function similar to the online version, by re-routing requests and processing to the local machine. Testing has been performed in Windows 10 (v1909) with Mozilla Firefox (ESR 78.5.0) in a 1080 p display setting.

5.2 Consequences of monosomy

LFQ and TMT MS analysis of monosomic cell lines

Sample preparation for LFQ and TMT mass spectrometry and data acquisition is described in [231]. Sample preparation was performed with Naren Chunduri. The following part of the methods is included in the same publication. In brief, the concatenated and TMT-labeled peptide mixtures were analyzed using nanoflow liquid chromatography (LC-MS/MS) on an EASY nano-LC 1200™ system (Thermo Fisher scientific), connected to a Q Exactive HF (Thermo Fisher scientific) through a Nanospray Flex Ion Source (Thermo Fisher Scientific). Three microliters of each

fraction were separated on a 40 cm heated reversed phase HPLC column (75 µm inner diameter with a PicoTip Emitter™, New Objective) in-house packed with 1.9 µm C18 beads (ReproSil-Pur 120 C18-AQ, Dr. Maisch). Peptides were loaded in 5 % buffer A (0.5 % aqueous formic acid) and eluted with a 3 h gradient (5–95 % buffer B (80 % acetonitrile, 0.5 % formic acid) at a constant flowrate of 0.25 µL/mL. Mass spectra were acquired in the data-dependent mode. Briefly, each full scan (mass range 375–1400 m/z, resolution of 60,000 at m/z of 200, maximum injection time 80 ms, ion target of 3E6) was followed by high-energy collision dissociation based fragmentation (HCD) of the 15 most abundant isotope patterns with a charge state between 2 and 7 (normalized collision energy of 32, an isolation window of 0.7 m/z, resolution of 30,000 atm/z of 200, maximum injection time 100 ms, AGC target value of 1E5, fixed first mass of 100 m/z and dynamic exclusion set to 30 s).

LFQ data was measured on the same experimental setup with adapted parameters: Each full scan (mass range 300–1650 m/z, resolution of 60,000 atm/z of 200, maximum injection time 20 ms, ion target of 3E6) was followed by high-energy collision dissociation based fragmentation (HCD) of the 15 most abundant isotope patterns with a charge state between 2 and 7 (normalized collision energy of 28, an isolation window of 1.4 m/z, resolution of 15,000 atm/z of 200, maximum injection time 80 ms, AGC target value of 1.6E3, no fixed first mass and dynamic exclusion set to 20 s).

MS Data processing

The MS data was processed with the MaxQuant software, version 1.6.3.3 [24]. All data was searched against the human reference proteome database (UniProtKB FASTA UP000005640) with a peptide and protein FDR of less than 1 %. Cysteine carbamidomethylation was set as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. Trypsin/P was set as protease and a maximum of two missed cleavages was accepted. False discovery rate (FDR) was set to 0.01 for peptides (minimum length of 7 amino acids) and proteins, and was determined by searching against a generated reverse database. Peptide identification was performed with an allowed initial precursor mass deviation up to 7 ppm and an allowed fragment mass deviation of 20 ppm. For TMT the label mass was corrected by

mass deviations as specified by the manufacturer. All raw files as well as all MaxQuant output tables and parameters have been uploaded to PRIDE.

Analysis of proteome data of monosomic cell lines

Identified protein groups were filtered to remove contaminants, reverse hits and proteins identified by site only. Next, protein groups that were identified more than two times in at least one group of replicates (N=4) were kept for further processing resulting in a set of 5887 protein groups in total. For LFQ, protein groups which were identified more than three times in at least one group of replicates (N=4) were kept for further processing, resulting in a set of 5727 Protein groups in total. Log_2 TMT reporter intensities were cleaned for batch effects using the R package LIMMA [58] and further normalized using variance stabilization [59]. For further analysis, all data were normalized by shifting the replicates to the same median, which was calculated without the monosomic genes to adjust the samples for the loss of a chromosome and the subsequent lower gene expression. For all monosomic cell lines, log_2 median intensity of three replicates of the wild type parental cell line was subtracted to calculate comparable fold changes.

Combined analysis of genomic, transcriptomic, and proteomic datasets

For further analysis comparing genomic, transcriptomic, and proteomic datasets, the DNA and mRNA datasets were matched to the corresponding protein entries and merged into a single table. To compare monosomic and trisomic cell lines, proteome data of trisomic RPE1 cell lines [79] was merged to the dataset. Chromosome/scaffold name, gene start (bp), gene stop (bp) and Ensembl gene stable ID (ENSG) were annotated through BioMart [262]. Perseus was used to add additional annotation (GOBP, GOCC, CORUM) and to carry out 2D annotation enrichment analysis [78]. The figures showing log_2 FC per chromosomes were generated using ggplot2 together with dplyr. Density histograms were generated in R using the library k-density and EQL. The log2 ratios of the mRNA and proteome subsets were plotted as density histograms, including the median of both populations.

5.3 Consequences of polyploidy

SILAC MS Data processing

Sample preparation for SILAC mass spectrometry and data acquisition, as well as the transcriptome analysis is described in [246]. Sample preparation and mass spectrometry was performed by Andreas Wallek and Nils Kulak. The following part of the methods is included in the same publication. In brief, Raw files were analyzed by MaxQuant software version 1.6.3.3 [24] and searched against the *S. cerevisiae* UniProt FASTA database (UP000002311). Lysine-0 (light) and Lysine-8 (heavy) were used as SILAC labels. Cysteine carbamidomethylation was set as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. LysC/P was set as protease and a maximum of two missed cleavages was accepted. False discovery rate (FDR) was set to 0.01 for peptides (minimum length of 7 amino acids) and proteins and was determined by searching against a generated reverse database. Peptide identification was performed with an allowed initial precursor mass deviation up to 7 ppm and an allowed fragment mass deviation of 20 ppm.

Analysis of SILAC proteome data

Identified protein groups were filtered to remove contaminants, reverse hits and proteins identified by site only. SILAC light/heavy ratios were calculated and transformed to log2 scale. Next, protein groups which were identified more than two times in at least one group of replicates were kept for further processing, resulting in a set of 3109 protein groups in total. To determine significance two-sample T-tests of 2N, 3N and 4N to 1N were performed (S0 = 0, permutation based Benjamini-Hochberg FDR threshold = 0.05). A “combined score” was calculated as the product of the q-values of all two sample tests. The median intensity of the replicates was calculated. Additional annotation (GOBP, GOCC) was added and 2D annotation enrichment analysis was performed to identify significantly deregulated pathways [78]. YeastEnrichR was used to perform gene set enrichment analysis of statistically significantly different values of all ploidies to haploid. The data is sorted based on the enrichR - combined score, which is a product of the p-value resulting from the Fisher exact test and the z-score of the deviation from the expected rank [74]. Outliers were calculated twice based

on the (SILAC L/H) ratio of ratios of each individual ploidy to 1N. First, the ratios of 4N to 1N were filtered for the outliers that show the overall strongest up or downregulation. Second, smoothed filtering was performed, which additionally excluded values that showed differences between consecutive ploidies with a FC>1 to remove values which spike between individual ploidies and keep only those that show a consistent trend across ploidies. Shown plots are generated using ggplot2 together with dplyr in R.

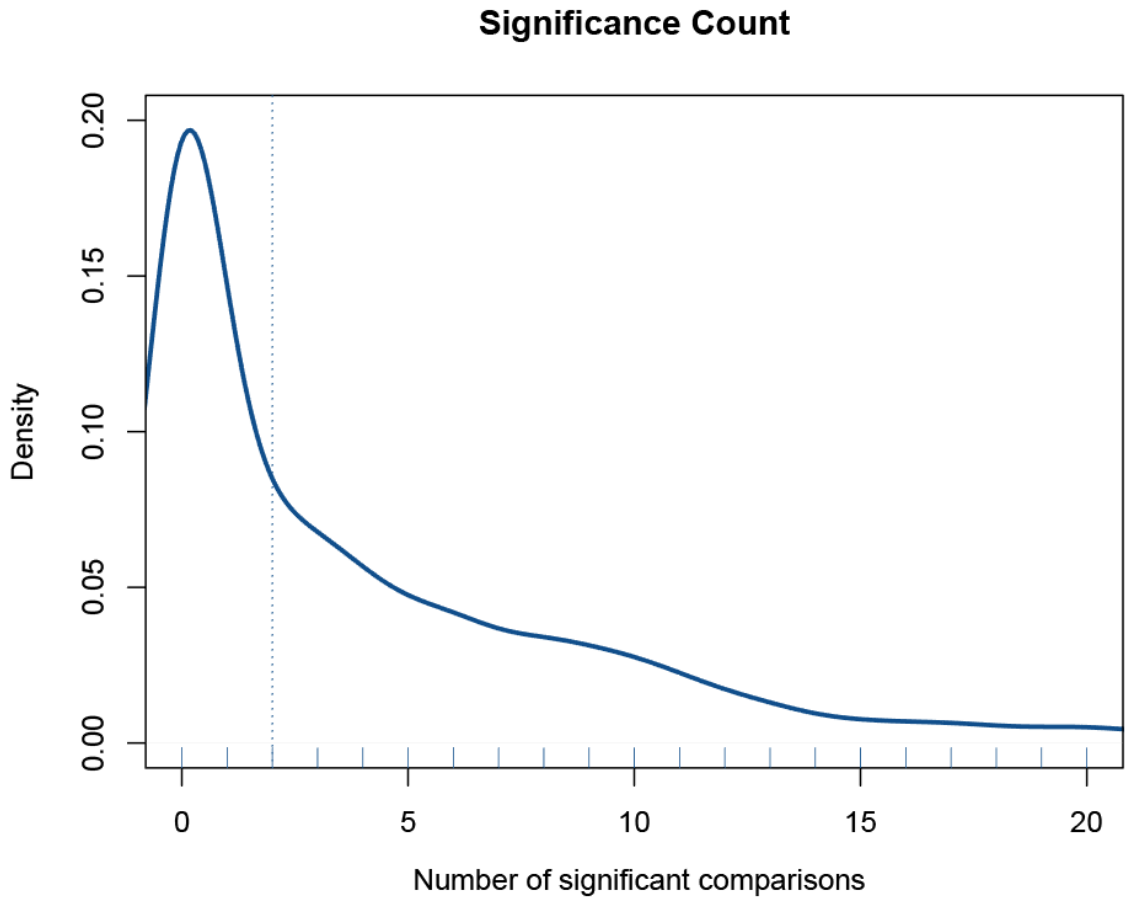


Figure S1 | Density of significance count. The figure shows the density histogram for the distribution of significant enrichment of all proteins in the combined dataset. X axis has been cropped at 20.

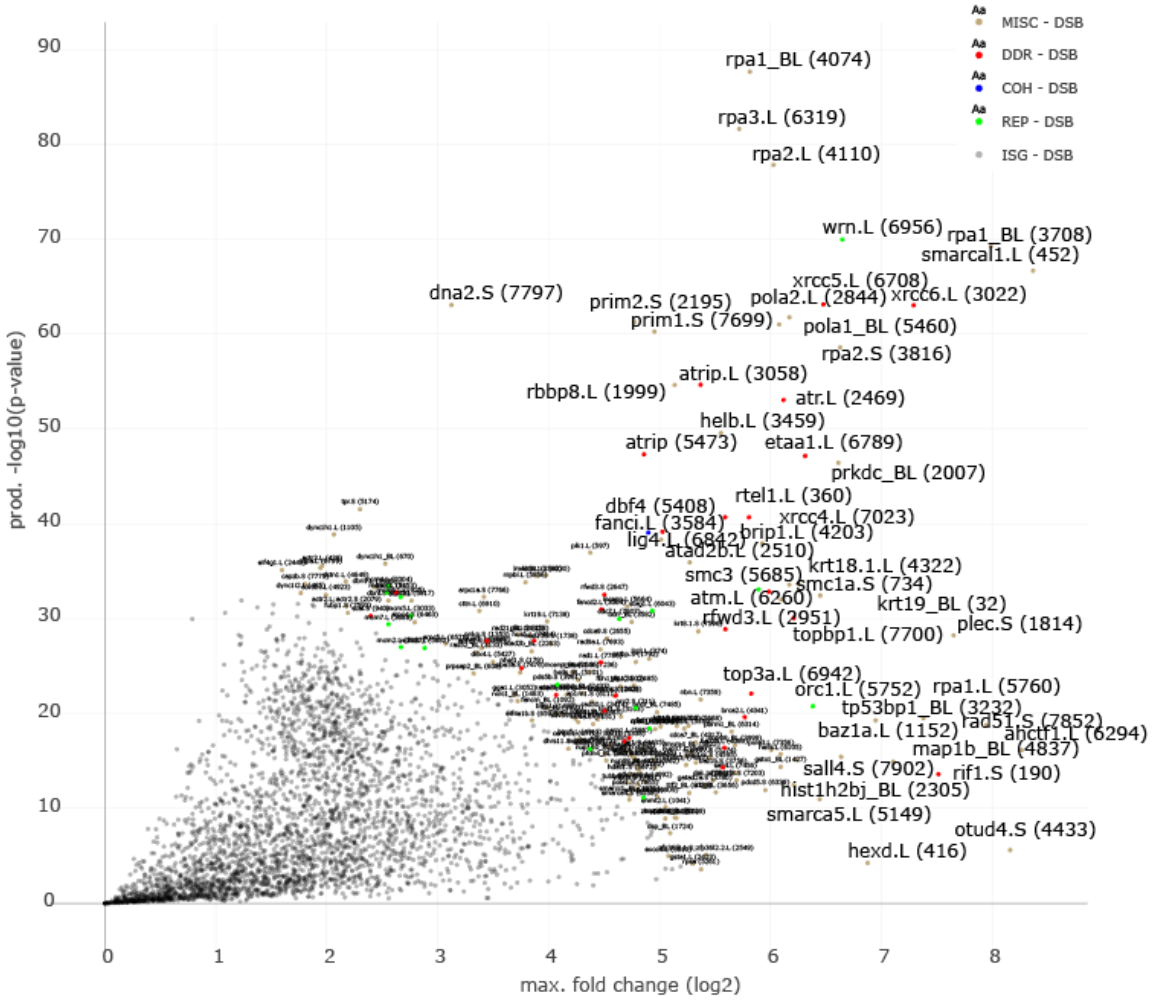


Figure S2 | Double Strand Breaks. The figure highlights the highest enriched factors in all investigated DSB repair experiments. The axis show the maximum fold change per proteins in all experiments and the combined score, calculated as product of the p-values on a -$\log_{10}$ scale. The color coding of the points shows a manually curated grouping of proteins based on their known function in replication (REP, green), cohesion (COH, blue), DNA damage response (DDR, red) or unassigned/miscellaneous (MISC, brown). Labels show gene symbol and MaxQuant id in brackets for unique identification. Several highly enriched proteins are color coded manually. The plot can be interactively visualized in the module *"Volcano Plots"* at http://dnarepairatlas.bio.uni-kl.de/.

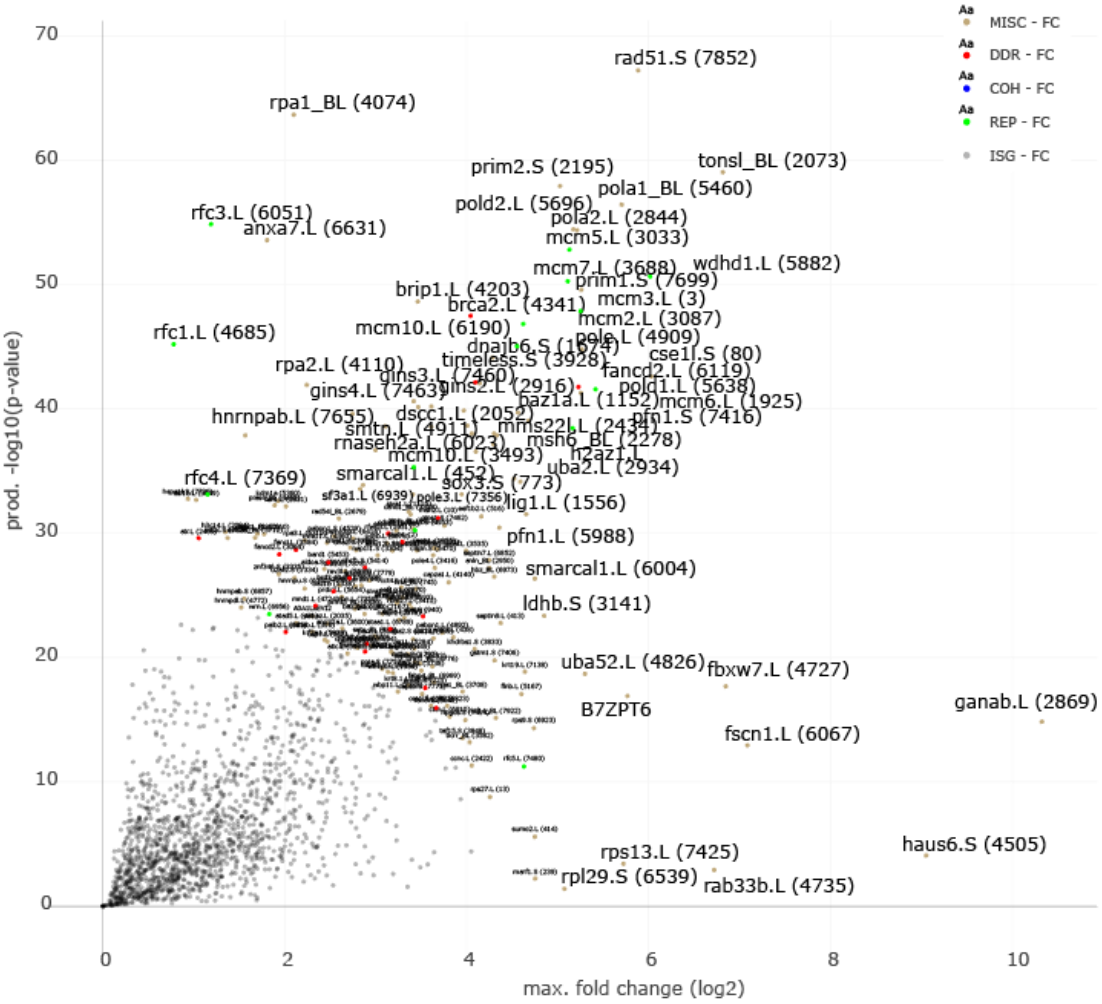


Figure S3 | Fork Collapse. The figure highlights the highest enriched factors in all investigated FC repair experiments. The axis show the maximum fold change per proteins in all experiments and the combined score, calculated as product of the p-values on a -$\log_{10}$ scale. The color coding of the points shows a manually curated grouping of proteins based on their known function in replication (REP, green), cohesion (COH, blue), DNA damage response (DDR, red) or unassigned/miscellaneous (MISC, brown). Labels show gene symbol and MaxQuant id in brackets for unique identification. Several highly enriched proteins are color coded manually. The plot can be interactively visualized in the module *"Volcano Plots"* at http://dnarepairatlas.bio.uni-kl.de/.

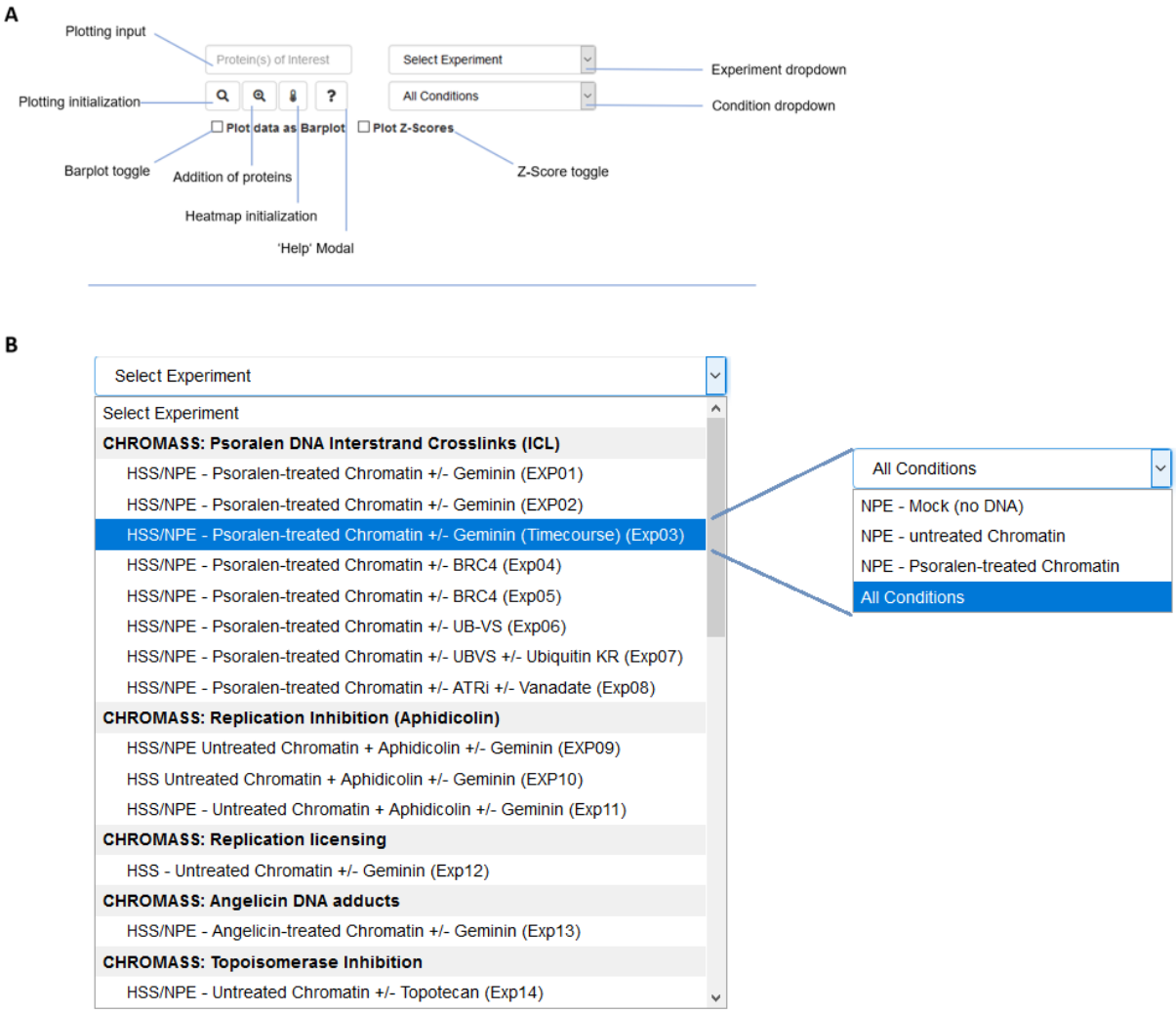


Figure S4 | Control Panel for "*Data Plotting*". The figure shows the control panel of the module "*Data Plotting"* in the DRA. The upper panel shows the input and initialization fields, that differentiate plot type and recognize the user input. The lower panel shows the dynamically filled drop down menu, that pulls the "conditions" field from the back-end server based on the users' choice of "Experiment".

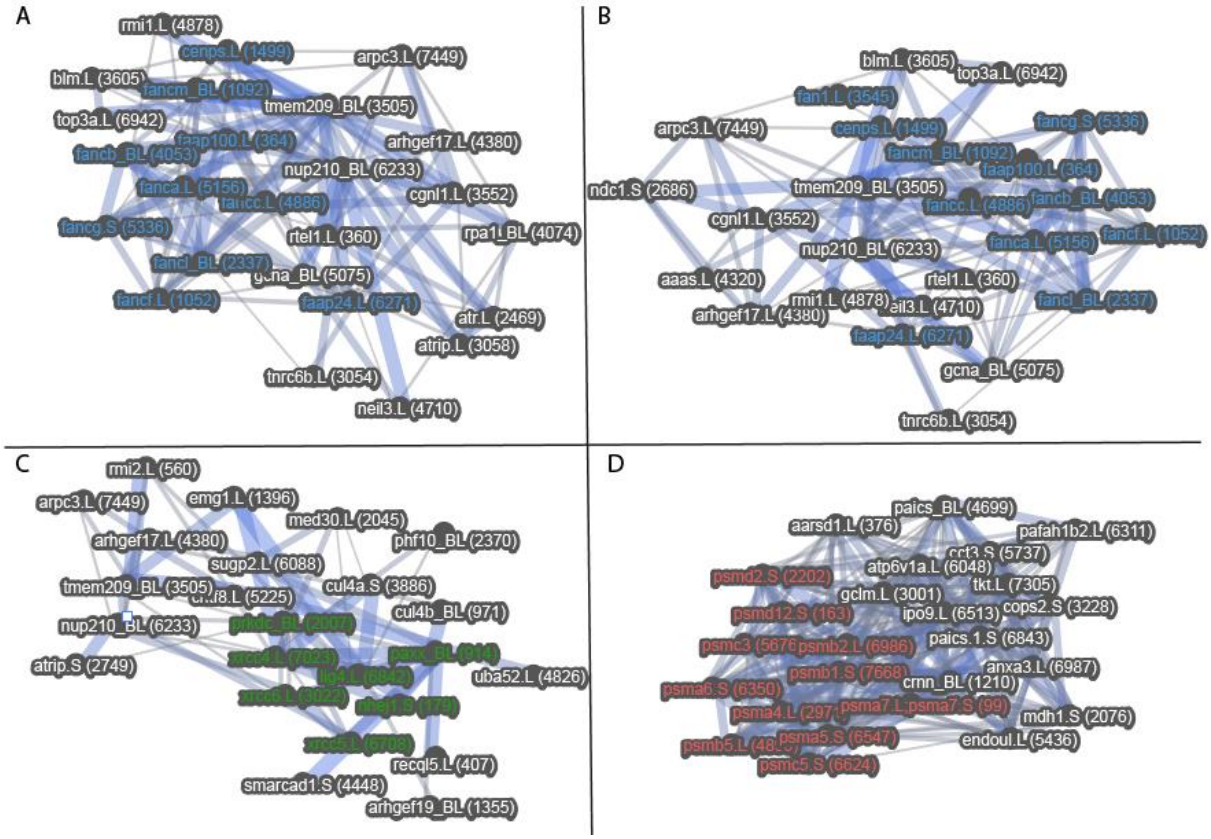


Figure S5 | Cluster identification in the DRA. The figure shows the identification of several clusters. (A) and (B) show the Fanconi Core complex based on the single input proteins FANCA and FANCB respectively. (C) shows the identification of the NHEJ factors (NHEJ1, LIG4, PAXX, XRCC4) and KU proteins (XRCC5/6) and the kinase PRKDC from the input XRCC5. (D) shows the recruitment of several proteasomal subunit proteins starting from the input PSMA5. The figures show the top 25 scoring proteins with a reset probability of 25 % and can be recreated at: http://dnarepairatlas.bio.uni-kl.de/Clustering

Figure S6 | Input proteins for the subnetwork. The figure shows the enrichment scores of all DNA repair factors chosen for the representative subnetwork in Figure 30.

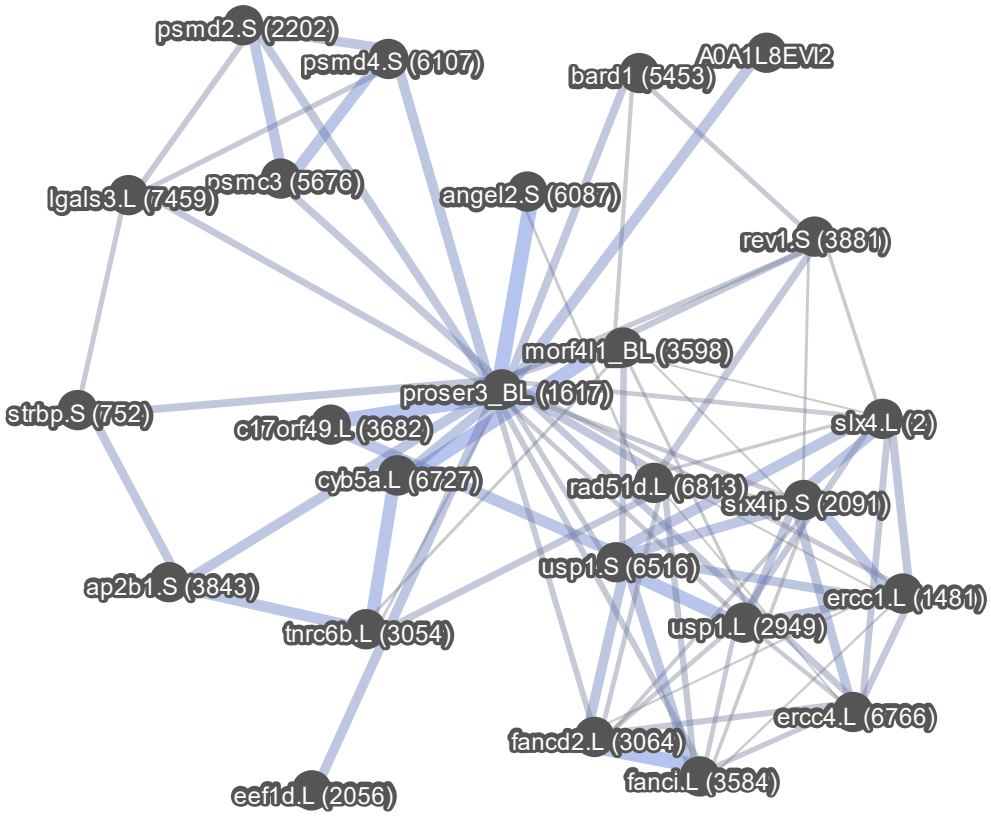


Figure S7 | Top30 Cluster of PROSER3. The figure shows the result of the RWR for PROSER3 with a reset probability of 30%. Plotted are the TOP30 scoring proteins.

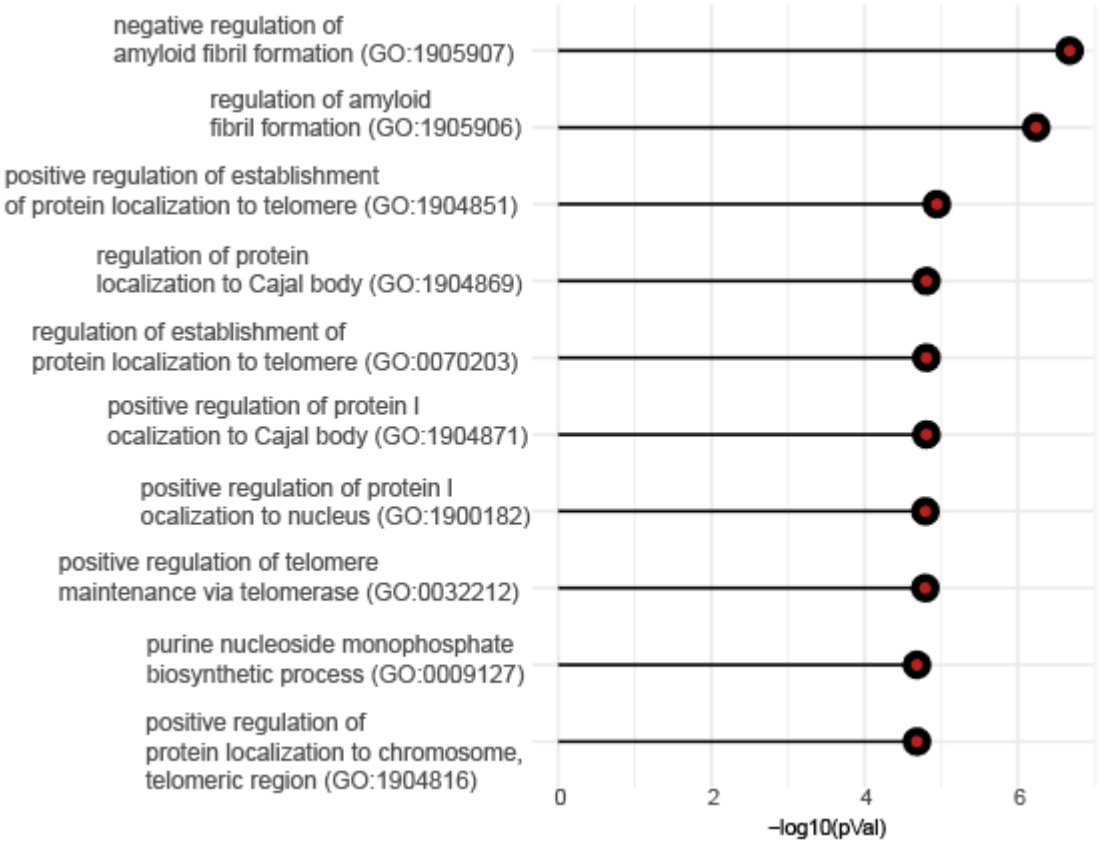
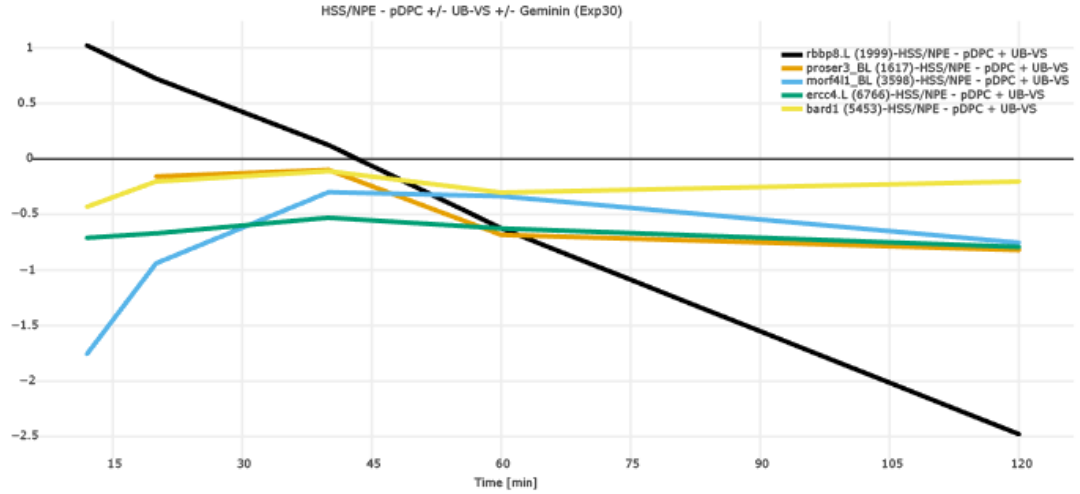


Figure S8 | Enrichment of control. The figure shows the enrichment of the Top 100 scoring proteins for an RWR with a result probability of 30% for XB5885669.

A



B

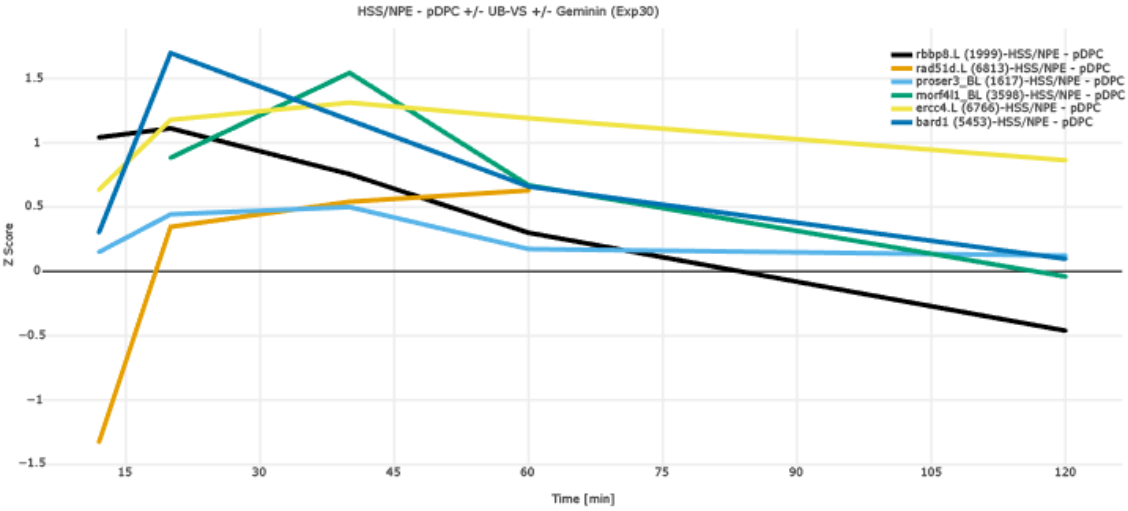


Figure S9 | Z-Scored protein abundancy of PROSER3. The figure shows line plots of z-scored protein abundancy of PROSER3 and highly correlating factors in pDPC +/-UBVS (Exp30).

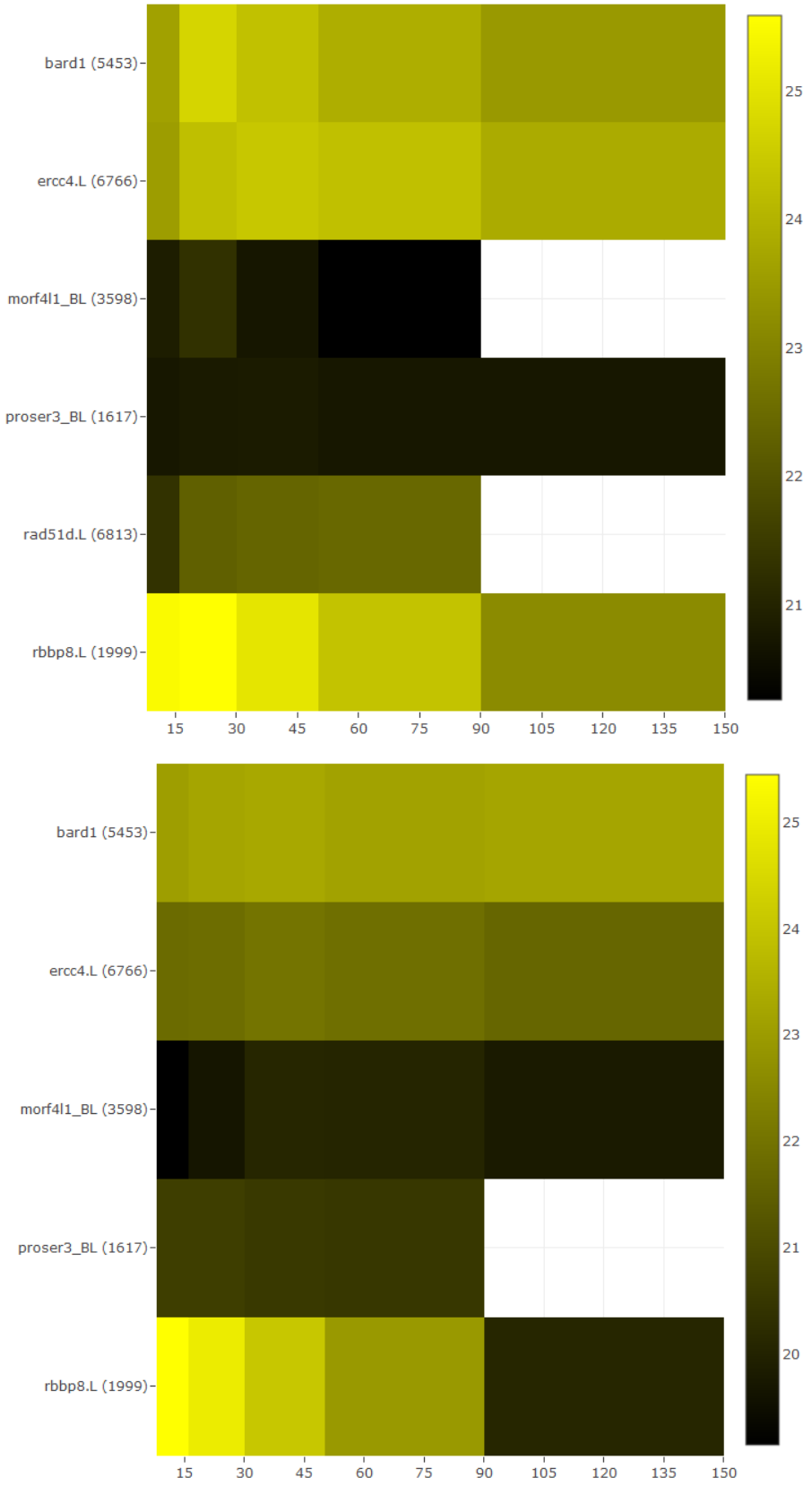


Figure S10 | LFQ protein abundancy of PROSER3 in pDPC + UBVS. The figure shows LFQ abundancy of PROSER3 and closely correlating factors in EXP 30 NPE - pDPC top; NPE- pDPC - UBVS bottom

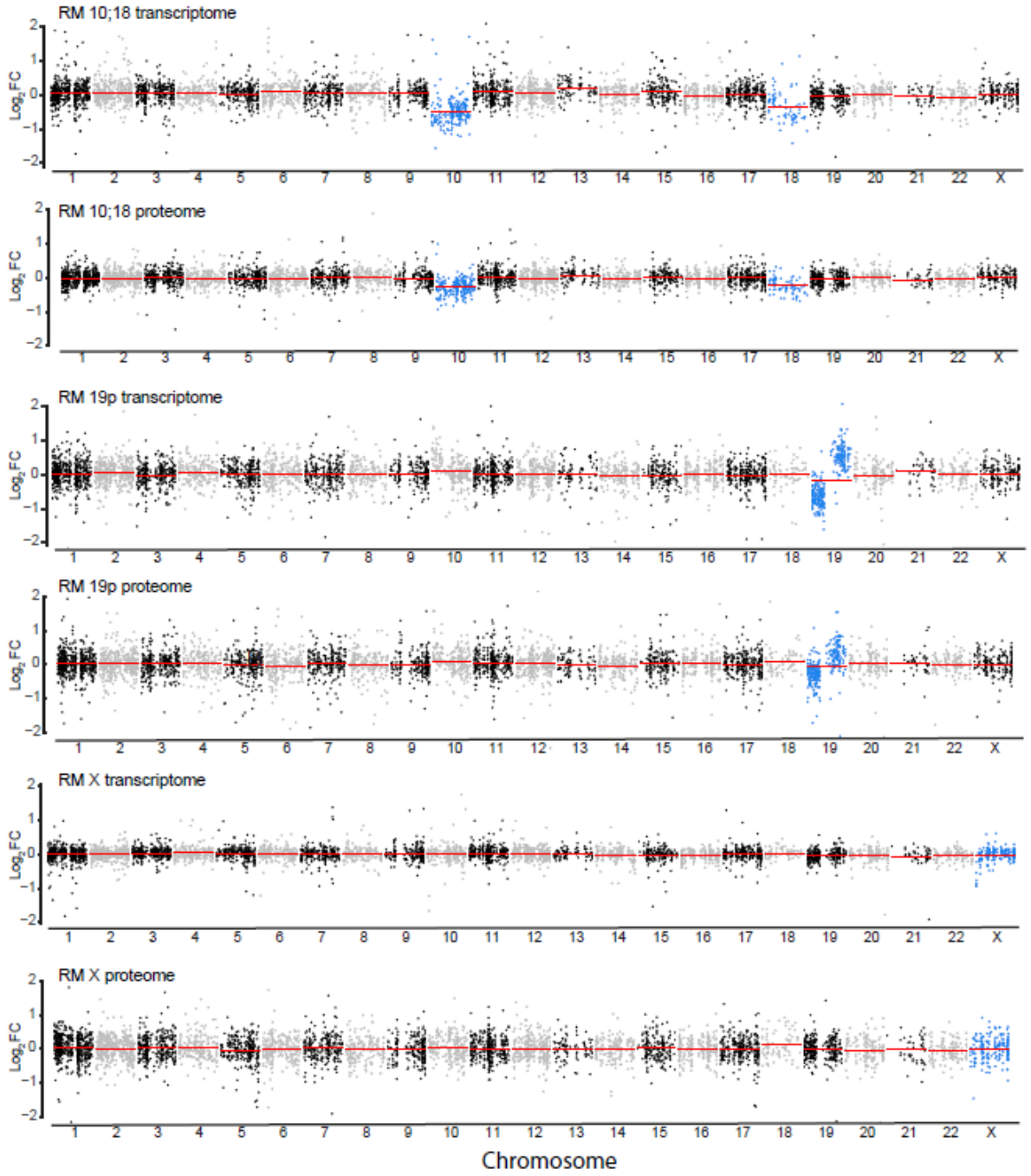


Figure S11 | Location plots of monosomies. The figure shows the location plots of proteome and transcriptome of all monosomies except RM 13. The monosomic chromosome is marked in blue.

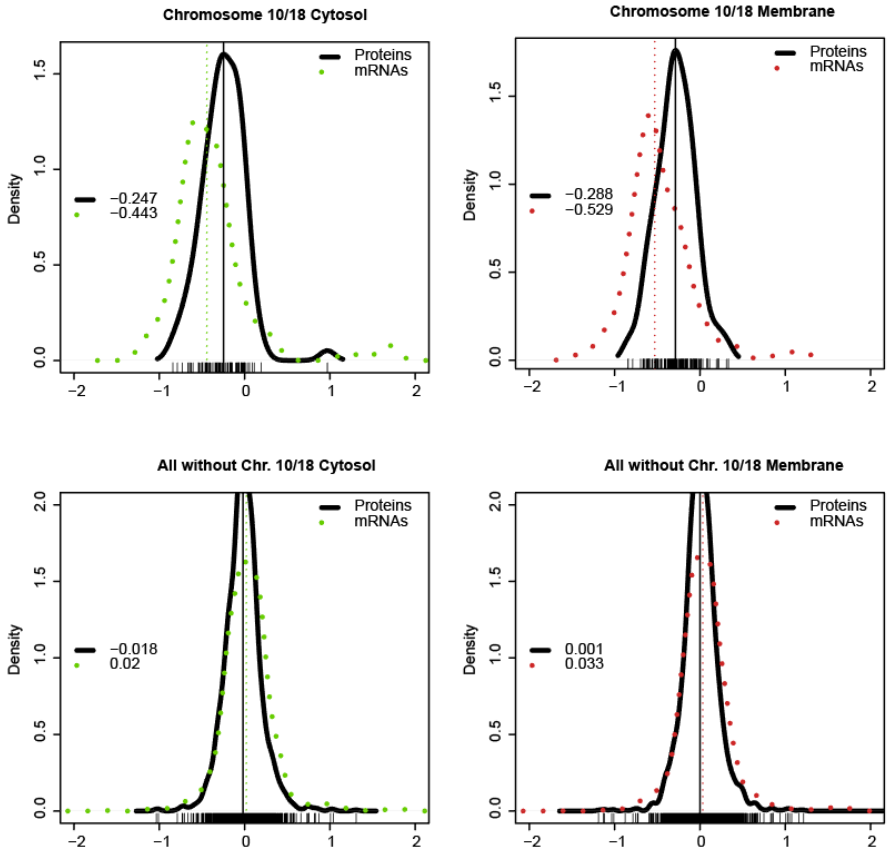


Figure S12 |Dosage Compensation analysis: Membrane, Cytosol. The figure shows the normalized expression of all cytosolic proteins (left) and membrane (right) proteins localized on chromosome 10 or 18 in RM 10/18. As control all cytosolic/membrane proteins not localized on chromosome 10/18 have been plotted at the bottom.

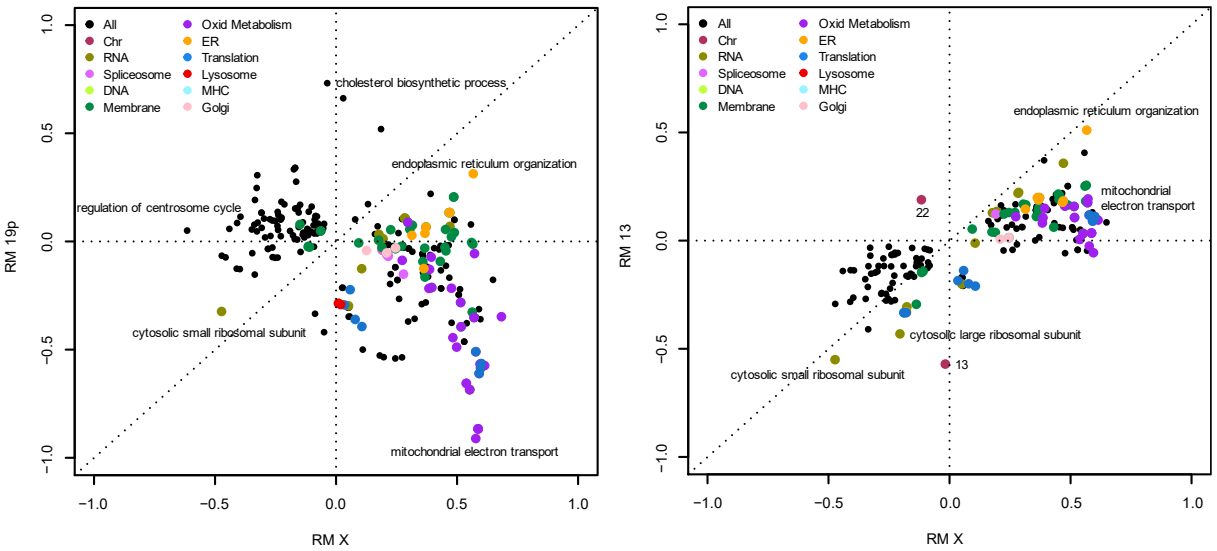


Figure S13 | 2D Enrichment of RM 19p/RM X, RM 13, RM X. The figure shows the 2D enrichment of RM 19p vs RM X (left) and RM 13 vs RM X (right) X for Gene Ontology biological process and cellular compartment, as well as the chromosome annotation. The enrichment is controlled by a Benjamini Hochberg FDR threshold of 0.02. A positive score indicates up-, and a negative downregulation. For easier comparison, annotations have been manually assigned to a color-coded legend, as indicated in the top left of each plot.

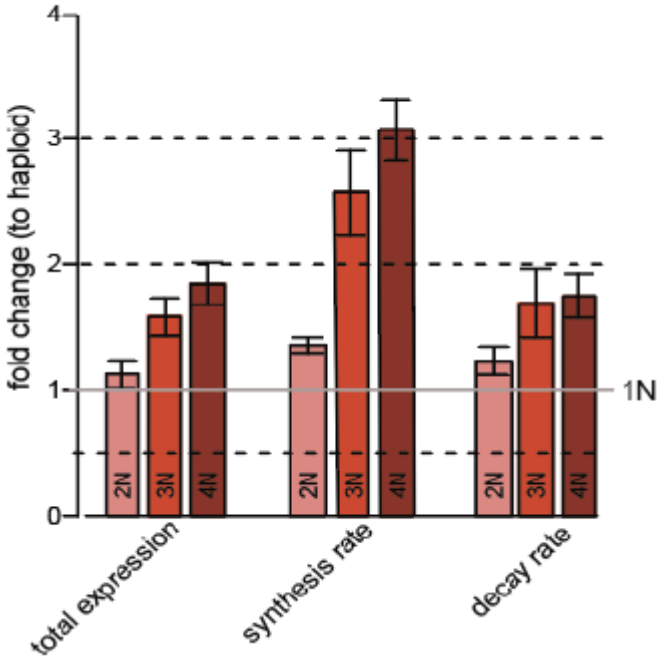


Figure S14 | Scaling of mRNA expression with ploidy. The figure shows the scaling of mRNA total expression, synthesis rate and decay rate with increasing ploidy, normalized to haploidy. The data was calculated based on DTA by Daniel Schulz and Andreas Wallek.

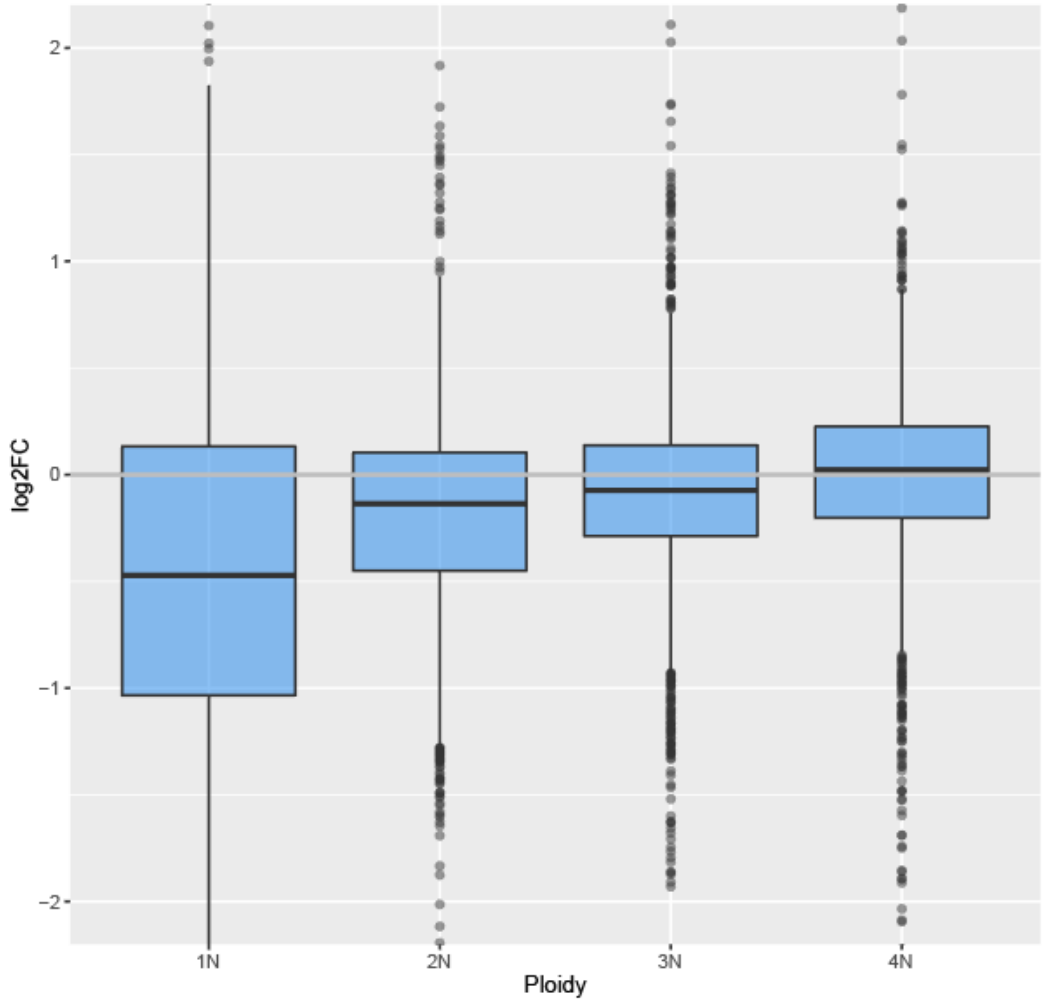


Figure S15 | Protein expression per ploidy. The figure shows the normalized log2 fold changes of all proteins per ploidy against a SILAC standard.

Table S1 | Experiment Series. The table shows a list of all experiment series of the datasets included in the DRA. Of note, EXP20 was excluded from the analysis due to issues with the measurement. An interactive version of this table, including further metadata to the experiments can be found at http://dnarepairatlas.bio.uni-kl.de/Meta

Experiment	*Experiment Series*
EXP01	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) I
EXP02	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) II
EXP03	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) III
EXP04	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) IV
EXP05	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) V
EXP06	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) VI
EXP07	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) VII
EXP08	CHROMASS: Psoralen DNA Interstrand Crosslinks (ICL) VIII
EXP09	CHROMASS: Replication Inhibition (Aphidicolin) I
EXP10	CHROMASS: Replication Inhibition (Aphidicolin) II
EXP11	CHROMASS: Replication Inhibition (Aphidicolin) III
EXP12	CHROMASS: Replication licensing
EXP13	CHROMASS: Angelicin DNA adducts
EXP14	CHROMASS: Topoisomerase Inhibition
EXP15	CHROMASS: MMS and UV
EXP16	CHROMASS: UV 256 nm
EXP17	CHROMASS: MMS and Campthotecin
EXP18	CHROMASS: Double strand Breaks (DSB) I
EXP19	CHROMASS: Double strand Breaks (DSB) II
EXP20	CHROMASS: Double strand Breaks (DSB) III
EXP21	CHROMASS: Double strand Breaks (DSB) IV
EXP22	CHROMASS: Double strand Breaks (DSB) V
EXP23	CHROMASS: Double strand Breaks (DSB) VI
EXP24	CHROMASS: Double strand Breaks (DSB) VII
EXP25	CHROMASS: Double strand Breaks (DSB) VIII
EXP26	CHROMASS: Double strand Breaks (DSB) VIIII
EXP27	CHROMASS: Double strand Breaks (DSB) X
EXP28	CHROMASS: Double Strand Breaks (DSB) XI

EXP29	PP-MS: DNA Protein Crosslinks (DPC) I
EXP30	PP-MS: DNA Protein Crosslinks (DPC) II
EXP31	PP-MS: Fork Collapse I
EXP32	PP-MS: Fork Collapse II
EXP33	CHROMASS: Replication Termination I
EXP34	PP-MS: Replication Termination II
EXP35	PP-MS: DNA Interstrand Crosslinks

Table S2 | Abbreviations and treatments. The table shows a list of all treatments as well as abbreviations of the treatments included in the DRA. Extract systems are LSS/HSS: Low/High-speed supernatant, NPE: Nucleoplasmic extract. An interactive version of this table, including further metadata to the experiments can be found at http://dnarepairatlas.bio.uni-kl.de/Meta

Abbreviation	Treatment
HCA	HSS - Untreated Chromatin + Aphidicolin
HCAG	HSS - Untreated Chromatin + Aphidicolin + Geminin
HCC	HSS - Untreated Chromatin
HCG	HSS - Untreated Chromatin + Geminin
HCO	HSS - Untreated Chromatin + ocadaic acid
HDC	HSS - Untreated Chromatin + PflMI
HDdeltaKU	HSS/NPE - Untreated Chromatin + PflMI + Ku depletion
HDdeltaRPA	HSS - Untreated Chromatin + PflMI + RPA depletion
HDG	HSS - Untreated Chromatin + PflMI + Geminin
HDO	HSS - Untreated Chromatin + PflMI + ocadaic acid
HDPKi	HSS - Untreated Chromatin + PflMI + NU-7441
HUC	HSS - UV-treated Chromatin
LCC	LSS - Untreated Chromatin
LDATRi	LSS - Untreated Chromatin + PflMI + ATRi
LDC	LSS - Untreated Chromatin + PflMI
NAC	HSS/NPE - pICL_AP
NCA	HSS/NPE - Untreated Chromatin + Aphidicolin
NCAG	HSS/NPE - Untreated Chromatin + Aphidicolin + Geminin
NCC	NPE - Untreated Chromatin
NCCPT	HSS/NPE - Campthotecin-treated Chromatin
NCG	HSS/NPE - Untreated Chromatin + Aphidicolin + Geminin
NCMLN	HSS/NPE - Untreated Chromatin + Cul-I

NCMLNG	HSS/NPE - Untreated Chromatin + Cul-I + Geminin
NCMMS	HSS/NPE - MMS-treated Chromatin
NCNMS	HSS/NPE - Untreated Chromatin + p97-I
NCNMSG	HSS/NPE - Untreated Chromatin + p97-I + Geminin
NCT	HSS/NPE - Untreated Chromatin + Topotecan
NCUV	HSS/NPE - UV-treated Chromatin
NDC	NPE - Untreated Chromatin + PflMI
NDRosc	NPE - Untreated Chromatin + PflMI + Roscovitin
NFC	HSS/NPE - Angelicin-treated Chromatin
NFG	HSS/NPE - Angelicin-treated Chromatin + Geminin
NFNMS	HSS/NPE - pICL_FdT + NMS-873
NHC	HSS/NPE - pDPC
NHG	HSS/NPE - pDPC + Geminin
NHU	HSS/NPE - pDPC + UB-VS
NMC	HSS/NPE - Mock (no Chromatin or DNA)
NNB	HSS/NPE - pNICK + BRC4
NNC	HSS/NPE - pNICK
NNG	HSS/NPE - pNICK + Geminin
NPATRi	HSS/NPE - Psoralen-treated Chromatin + ATRi
NPB	HSS/NPE - Psoralen-treated Chromatin + BRC4
NPC	HSS/NPE - Psoralen
NPG	HSS/NPE - Psoralen-treated Chromatin + Geminin
NPUBVS	HSS/NPE - Psoralen-treated Chromatin + UB-VS
NPUBVSplusUBKR	HSS/NPE - Psoralen-treated Chromatin + UBVS + Ubiquitin KR
NPVana	HSS/NPE - Psoralen-treated Chromatin + Vanadate
NQC	HSS - pCTR
NQG	HSS - pCTR + Geminin
NSC	HSS/NPE - pICL_PSO
NUG	HSS/NPE - UV-treated Chromatin + Geminin
pCTR	HSS/NPE - pCTR (no LacI) + LacI
pLacO	HSS/NPE - pLacO
pLacOG	HSS/NPE - pLacO + Geminin
pLacOI	HSS/NPE - pLacO + IPTG
pLagNick	HSS/NPE - pLagNick
pLeadNick	HSS/NPE - pLeadNick + Geminin
pNONICKED	HSS/NPE - pNoNick - LacI

Table S3 | Significance count in percent. The table shows the significance count in percent for all proteins, binned by 1-30 significances and 30+ as well as the number of identified proteins per bin.

Sig_Count	# Proteins	Percentage of total	Sig_Count	# Proteins	Percentage of total
0	2011	34.73	**17**	40	0.69
1	551	9.52	**20**	33	0.57
2	424	7.32	**18**	32	0.55
3	386	6.67	**19**	32	0.55
4	323	5.58	**22**	27	0.47
6	252	4.35	**21**	19	0.33
5	250	4.32	**26**	18	0.31
8	210	3.63	**24**	16	0.28
7	195	3.37	**25**	14	0.24
9	189	3.26	**27**	13	0.22
10	166	2.87	**23**	12	0.21
11	135	2.33	**31**	9	0.16
12	93	1.61	**28**	8	0.14
13	87	1.5	**29**	8	0.14
14	54	0.93	**30**	8	0.14
16	45	0.78	**30+**	95	>=0.1

Table S4 | "Provisionals". The table shows proteins annotated with with provisional names in the DRA, as well as their majority protein ID and significance count. Color coded proteins are found in the subnetwork in Figure 30.

MQ_id	Gene Symbol	Majority protein IDs	Significance Count
526	XB5885669.L [provisional]	A0A1L8EX65	13
3633	XB5867546.L [provisional]	A0A1L8H200	11
3734	XB22164552.S [provisional]	A0A1L8H5M0	11
4377	XB877238.L [provisional:gdpd5]	A0A1L8HJX5	10
5368	XB5871767.L [provisional:gatd3b]	A9UM19	10
6990	XB5909790.S [provisional]	Q6GQ64	9
6484	XB5768883.L [provisional]	Q66IN5	8
1617	**proser3_BL**	**A0A1L8FPH9**	**8**
3053	XB5765667.L [provisional]	A0A1L8GNX4	8
5810	XB5733004.L [provisional]	Q0IH65;Q4V7H4	8
180	XB5731801.S [provisional]	A0A1L8ENU1	8
3329	XB5920187.S [provisional]	B7ZQW3	7
5313	XB5843130.S [provisional]	A4FVF5	7
6083	**XB5727280.S [provisional]**	**Q640B9;Q4V7J3**	**7**
4796	XB5962940.L [provisional:ctdspl2]	A0A1L8HWE0	5
1730	XB5835883.S [provisional]	A0A1L8FS70	4
2066	**XB5759208.L [provisional]**	**A0A1L8G0G0**	**4**
4812	**XB5727280.L [provisional]**	**A0A1L8HWT7**	**4**

7. References

1 Crick, F. H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138-163 (1958).

2 Rodrigo-Brenni, M. C. & Hegde, R. S. Design principles of protein biosynthesis-coupled quality control. *Dev Cell* **23**, 896-907, doi:10.1016/j.devcel.2012.10.012 (2012).

3 Roth, M. J. *et al.* Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. *Mol Cell Proteomics* **4**, 1002-1008, doi:10.1074/mcp.M500064-MCP200 (2005).

4 Harper, J. W. & Bennett, E. J. Proteome complexity and the forces that drive proteome imbalance. *Nature* **537**, 328-338, doi:10.1038/nature19947 (2016).

5 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

6 Edman, P. A method for the determination of amino acid sequence in peptides. *Arch Biochem* **22**, 475 (1949).

7 Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**, 2299-2301, doi:10.1021/ac00171a028 (1988).

8 Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64-71, doi:10.1126/science.2675315 (1989).

9 Chernushevich, I. V., Loboda, A. V. & Thomson, B. A. An introduction to quadrupole-time-of-flight mass spectrometry. *J Mass Spectrom* **36**, 849-865, doi:10.1002/jms.207 (2001).

10 Loboda, A. V., Krutchinsky, A. N., Bromirski, M., Ens, W. & Standing, K. G. A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser desorption/ionization source: design and performance. *Rapid Commun Mass Spectrom* **14**, 1047-1057, doi:10.1002/1097-0231(20000630)14:12<1047::AID-RCM990>3.0.CO;2-E (2000).

11 Williams, J. D., Flanagan, M., Lopez, L., Fischer, S. & Miller, L. A. Using accurate mass electrospray ionization-time-of-flight mass spectrometry with in-source collision-induced dissociation to sequence peptide mixtures. *J Chromatogr A* **1020**, 11-26, doi:10.1016/j.chroma.2003.07.019 (2003).

12 Kaufmann, A. Analytical performance of the various acquisition modes in Orbitrap MS and MS/MS. *J Mass Spectrom* **53**, 725-738, doi:10.1002/jms.4195 (2018).

13 Denisov, E., Damoc, E. & Makarov, A. Exploring frontiers of orbitrap performance for long transients. *International Journal of Mass Spectrometry* **466**, doi:10.1016/j.ijms.2021.116607 (2021).

14 Yang, F., Shen, Y., Camp, D. G., 2nd & Smith, R. D. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Rev Proteomics* **9**, 129-134, doi:10.1586/epr.12.15 (2012).

15 Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355, doi:10.1038/nature19949 (2016).

16 Hebert, A. S. *et al.* Improved Precursor Characterization for Data-Dependent Mass Spectrometry. *Anal Chem* **90**, 2333-2340, doi:10.1021/acs.analchem.7b04808 (2018).

17 UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489, doi:10.1093/nar/gkaa1100 (2021).

18 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65, doi:10.1093/nar/gkl842 (2007).

19 Diament, B. J. & Noble, W. S. Faster SEQUEST searching for peptide identification from tandem mass spectra. *J Proteome Res* **10**, 3871-3879, doi:10.1021/pr101196n (2011).

20 Brosch, M., Yu, L., Hubbard, T. & Choudhary, J. Accurate and sensitive peptide identification with Mascot Percolator. *J Proteome Res* **8**, 3176-3181, doi:10.1021/pr800982s (2009).

21 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805, doi:10.1021/pr101065j (2011).
22 Weisser, H. *et al.* An automated pipeline for high-throughput label-free quantitative proteomics. *J Proteome Res* **12**, 1628-1644, doi:10.1021/pr300992u (2013).
23 Orsburn, B. C. Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **9**, doi:10.3390/proteomes9010015 (2021).
24 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526, doi:10.1074/mcp.M113.031591 (2014).
25 Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat Methods* **16**, 519-525, doi:10.1038/s41592-019-0427-6 (2019).
26 Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **11**, O111 016717, doi:10.1074/mcp.O111.016717 (2012).
27 Martinez-Val, A., Bekker-Jensen, D. B., Hogrebe, A. & Olsen, J. V. Data Processing and Analysis for DIA-Based Phosphoproteomics Using Spectronaut. *Methods Mol Biol* **2361**, 95-107, doi:10.1007/978-1-0716-1641-3_6 (2021).
28 Rost, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* **32**, 219-223, doi:10.1038/nbt.2841 (2014).
29 Tsou, C. C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **12**, 258-264, 257 p following 264, doi:10.1038/nmeth.3255 (2015).
30 Bartels, C. Fast algorithm for peptide sequencing by mass spectroscopy. *Biomed Environ Mass Spectrom* **19**, 363-368, doi:10.1002/bms.1200190607 (1990).
31 Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A* **114**, 8247-8252, doi:10.1073/pnas.1705691114 (2017).
32 Tran, N. H. *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods* **16**, 63-66, doi:10.1038/s41592-018-0260-3 (2019).
33 Tran, N. H. *et al.* Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. *Nature Machine Intelligence* **2**, 764-771, doi:10.1038/s42256-020-00260-4 (2020).
34 Yost, R. A. & Enke, C. G. Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation. *Anal Chem* **51**, 1251-1264, doi:10.1021/ac50048a002 (1979).
35 MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966-968, doi:10.1093/bioinformatics/btq054 (2010).
36 Cardozo, K. H. M. *et al.* Establishing a mass spectrometry-based system for rapid detection of SARS-CoV-2 in large clinical sample cohorts. *Nat Commun* **11**, 6201, doi:10.1038/s41467-020-19925-0 (2020).
37 Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994-999, doi:10.1038/13690 (1999).
38 Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154-1169, doi:10.1074/mcp.M400129-MCP200 (2004).
39 Thompson, A. *et al.* Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**, 1895-1904, doi:10.1021/ac0262560 (2003).

40 Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat Methods* **17**, 399-404, doi:10.1038/s41592-020-0781-4 (2020).
41 Savitski, M. M. *et al.* Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J Proteome Res* **12**, 3586-3598, doi:10.1021/pr400098r (2013).
42 Virreira Winter, S. *et al.* EASI-tag enables accurate multiplexed and interference-free MS2-based proteome quantification. *Nature methods* **15**, 527–530, doi:10.1038/s41592-018-0037-8 (2018).
43 Ong, S. E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376-386, doi:10.1074/mcp.m200025-mcp200 (2002).
44 Zhang, G. & Neubert, T. A. Use of stable isotope labeling by amino acids in cell culture (SILAC) for phosphotyrosine protein identification and quantitation. *Methods Mol Biol* **527**, 79-92, xi, doi:10.1007/978-1-60327-834-8_7 (2009).
45 Lewandowska, D. *et al.* Plant SILAC: stable-isotope labelling with amino acids of arabidopsis seedlings for quantitative proteomics. *PLoS One* **8**, e72207, doi:10.1371/journal.pone.0072207 (2013).
46 Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods* **7**, 383-385, doi:10.1038/nmeth.1446 (2010).
47 Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat Methods* **15**, 440-448, doi:10.1038/s41592-018-0003-5 (2018).
48 Muntel, J. *et al.* Surpassing 10 000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Mol Omics* **15**, 348-360, doi:10.1039/c9mo00082h (2019).
49 Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **13**, 731–740, doi:10.1038/nmeth.3901 (2016).
50 Kuznetsova, I., Lugmayr, A., Rackham, O. & Filipovska, A. OmicsVolcano: software for intuitive visualization and interactive exploration of high-throughput biological data. *STAR Protoc* **2**, 100279, doi:10.1016/j.xpro.2020.100279 (2021).
51 Deracinois, B., Flahaut, C., Duban-Deweer, S. & Karamanos, Y. Comparative and Quantitative Global Proteomics Approaches: An Overview. *Proteomes* **1**, 180-218, doi:10.3390/proteomes1030180 (2013).
52 Morris, J. H. *et al.* Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. *Nat Protoc* **9**, 2539-2554, doi:10.1038/nprot.2014.164 (2014).
53 Tyanova, S. & Cox, J. Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. *Methods Mol Biol* **1711**, 133-148, doi:10.1007/978-1-4939-7493-1_7 (2018).
54 Yang, Y. H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**, e15, doi:10.1093/nar/30.4.e15 (2002).
55 Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193, doi:10.1093/bioinformatics/19.2.185 (2003).
56 Park, T. *et al.* Evaluation of normalization methods for microarray data. *BMC Bioinformatics* **4**, 33, doi:10.1186/1471-2105-4-33 (2003).
57 Callister, S. J. *et al.* Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* **5**, 277-286, doi:10.1021/pr050300l (2006).

58 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).
59 Huber, W., Heydebreck, A. v., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-104, doi:10.1093/bioinformatics/18.suppl_1.s96 (2002).
60 Valikangas, T., Suomi, T. & Elo, L. L. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform* **19**, 1-11, doi:10.1093/bib/bbw095 (2018).
61 Kultima, K. *et al.* Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol Cell Proteomics* **8**, 2285-2295, doi:10.1074/mcp.M800514-MCP200 (2009).
62 Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–5121, doi:10.1073/pnas.091062498 (2001).
63 Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3, doi:10.2202/1544-6115.1027 (2004).
64 Kammers, K., Cole, R. N., Tiengwe, C. & Ruczinski, I. Detecting Significant Changes in Protein Abundance. *EuPA Open Proteom* **7**, 11-19, doi:10.1016/j.euprot.2015.02.002 (2015).
65 van Ooijen, M. P. *et al.* Identification of differentially expressed peptides in high-throughput proteomics data. *Brief Bioinform* **19**, 971-981, doi:10.1093/bib/bbx031 (2018).
66 Webb-Robertson, B. J. *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res* **14**, 1993-2001, doi:10.1021/pr501138h (2015).
67 Wang, J. *et al.* In-depth method assessments of differentially expressed protein detection for shotgun proteomics data with missing values. *Sci Rep* **7**, 3367, doi:10.1038/s41598-017-03650-8 (2017).
68 Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47-52, doi:10.1038/35011540 (1999).
69 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
70 Gene Ontology, C. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325-D334, doi:10.1093/nar/gkaa1113 (2021).
71 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30, doi:10.1093/nar/28.1.27 (2000).
72 Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res* **47**, D559-D563, doi:10.1093/nar/gky973 (2019).
73 Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. & Krawetz, S. A. Global functional profiling of gene expression. *Genomics* **81**, 98-104, doi:10.1016/s0888-7543(02)00021-6 (2003).
74 Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* **14**, 128, doi:10.1186/1471-2105-14-128 (2013).
75 Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-273, doi:10.1038/ng1180 (2003).
76 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
77 Frost, H. R., Li, Z. & Moore, J. H. Spectral gene set enrichment (SGSE). *BMC Bioinformatics* **16**, 70, doi:10.1186/s12859-015-0490-7 (2015).

78 Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC bioinformatics* **13 Suppl 16**, S12, doi:10.1186/1471-2105-13-s16-s12 (2012).
79 Stingele, S. *et al.* Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol Syst Biol* **8**, 608, doi:10.1038/msb.2012.40 (2012).
80 Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212-1226, doi:10.1016/j.cell.2014.10.050 (2014).
81 van Dam, S., Vosa, U., van der Graaf, A., Franke, L. & de Magalhaes, J. P. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief Bioinform* **19**, 575-592, doi:10.1093/bib/bbw139 (2018).
82 Carter, S. L., Brechbuhler, C. M., Griffin, M. & Bond, A. T. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20**, 2242-2250, doi:10.1093/bioinformatics/bth234 (2004).
83 van Someren, E. P. *et al.* Least absolute regression network analysis of the murine osteoblast differentiation network. *Bioinformatics* **22**, 477-484, doi:10.1093/bioinformatics/bti816 (2006).
84 Needham, C. J., Bradford, J. R., Bulpitt, A. J. & Westhead, D. R. A primer on learning in Bayesian networks for computational biology. *PLoS Comput Biol* **3**, e129, doi:10.1371/journal.pcbi.0030129 (2007).
85 Pavlopoulos, G. A. *et al.* Using graph theory to analyze biological networks. *BioData Min* **4**, 10, doi:10.1186/1756-0381-4-10 (2011).
86 Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447-452, doi:10.1093/nar/gku1003 (2015).
87 Schwikowski, B., Uetz, P. & Fields, S. A network of protein-protein interactions in yeast. *Nat Biotechnol* **18**, 1257-1261, doi:10.1038/82360 (2000).
88 Singer, G. A., Lloyd, A. T., Huminiecki, L. B. & Wolfe, K. H. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol Biol Evol* **22**, 767-775, doi:10.1093/molbev/msi062 (2005).
89 Gillis, J. & Pavlidis, P. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Comput Biol* **8**, e1002444, doi:10.1371/journal.pcbi.1002444 (2012).
90 Gascuel, O. & Steel, M. Neighbor-joining revealed. *Mol Biol Evol* **23**, 1997-2000, doi:10.1093/molbev/msl072 (2006).
91 Sharma, A., Lopez, Y. & Tsunoda, T. Divisive hierarchical maximum likelihood clustering. *BMC Bioinformatics* **18**, 546, doi:10.1186/s12859-017-1965-5 (2017).
92 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).
93 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584, doi:10.1093/nar/30.7.1575 (2002).
94 Lu, Y., Lu, S., Fotouhi, F., Deng, Y. & Brown, S. J. Incremental genetic K-means algorithm and its application in gene expression data analysis. *BMC Bioinformatics* **5**, 172, doi:10.1186/1471-2105-5-172 (2004).
95 Wang, Y. & Pan, Y. Semi-supervised consensus clustering for gene expression data analysis. *BioData Min* **7**, 7, doi:10.1186/1756-0381-7-7 (2014).
96 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).
97 Kaur, S., Mirza, A. H., Overgaard, A. J., Pociot, F. & Storling, J. A Dual Systems Genetics Approach Identifies Common Genes, Networks, and Pathways for Type 1 and 2 Diabetes in Human Islets. *Front Genet* **12**, 630109, doi:10.3389/fgene.2021.630109 (2021).

98 Li, B. Q., Huang, T., Liu, L., Cai, Y. D. & Chou, K. C. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. *PLoS One* **7**, e33393, doi:10.1371/journal.pone.0033393 (2012).
99 Navlakha, S., Gitter, A. & Bar-Joseph, Z. A network-based approach for predicting missing pathway interactions. *PLoS Comput Biol* **8**, e1002640, doi:10.1371/journal.pcbi.1002640 (2012).
100 Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet* **18**, 551-562, doi:10.1038/nrg.2017.38 (2017).
101 Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**, 106-114, doi:10.1038/ng.3168 (2015).
102 Reyna, M. A., Leiserson, M. D. M. & Raphael, B. J. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**, i972-i980, doi:10.1093/bioinformatics/bty613 (2018).
103 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995, doi:10.1093/nar/gks1193 (2013).
104 Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res* **48**, D1145-D1152, doi:10.1093/nar/gkz984 (2020).
105 Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).
106 Ledford, H. End of cancer-genome project prompts rethink. *Nature* **517**, 128-129, doi:10.1038/517128a (2015).
107 Liu, C. *et al.* Integrating Multi-omics Data to Dissect Mechanisms of DNA repair Dysregulation in Breast Cancer. *Sci Rep* **6**, 34000, doi:10.1038/srep34000 (2016).
108 Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res* **24**, 1248-1259, doi:10.1158/1078-0432.CCR-17-0853 (2018).
109 Jarnuczak, A. F. *et al.* An integrated landscape of protein expression in human cancer. *Sci Data* **8**, 115, doi:10.1038/s41597-021-00890-2 (2021).
110 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70, doi:10.1016/s0092-8674(00)81683-9 (2000).
111 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).
112 Lindahl, T. & Barnes, D. E. Repair of endogenous DNA damage. *Cold Spring Harb Symp Quant Biol* **65**, 127-133, doi:10.1101/sqb.2000.65.127 (2000).
113 Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen* **58**, 235-263, doi:10.1002/em.22087 (2017).
114 Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168**, 644-656, doi:10.1016/j.cell.2017.01.002 (2017).
115 King, M. C., Marks, J. H., Mandell, J. B. & New York Breast Cancer Study, G. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302**, 643-646, doi:10.1126/science.1088759 (2003).
116 Nielsen, F. C., van Overeem Hansen, T. & Sorensen, C. S. Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat Rev Cancer* **16**, 599-612, doi:10.1038/nrc.2016.72 (2016).
117 Foulkes, W. D., Knoppers, B. M. & Turnbull, C. Population genetic testing for cancer susceptibility: founder mutations to genomes. *Nat Rev Clin Oncol* **13**, 41-54, doi:10.1038/nrclinonc.2015.173 (2016).
118 Lee, E. Y. & Muller, W. J. Oncogenes and tumor suppressor genes. *Cold Spring Harb Perspect Biol* **2**, a003236, doi:10.1101/cshperspect.a003236 (2010).

119 Brooks, C. L. & Gu, W. Ubiquitination, phosphorylation and acetylation: the molecular basis for p53 regulation. *Curr Opin Cell Biol* **15**, 164-171, doi:10.1016/s0955-0674(03)00003-6 (2003).
120 Zhang, X. P., Liu, F. & Wang, W. Two-phase dynamics of p53 in the DNA damage response. *Proc Natl Acad Sci U S A* **108**, 8990-8995, doi:10.1073/pnas.1100600108 (2011).
121 Sablina, A. A. *et al.* The antioxidant function of the p53 tumor suppressor. *Nat Med* **11**, 1306–1313, doi:10.1038/nm1320 (2005).
122 Ohashi, A. *et al.* Aneuploidy generates proteotoxic stress and DNA damage concurrently with p53-mediated post-mitotic apoptosis in SAC-impaired cells. *Nat Commun* **6**, 7668, doi:10.1038/ncomms8668 (2015).
123 Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112-121, doi:10.1038/s41586-019-1913-9 (2020).
124 McGranahan, N., Burrell, R. A., Endesfelder, D., Novelli, M. R. & Swanton, C. Cancer chromosomal instability: therapeutic and diagnostic challenges. *EMBO Rep* **13**, 528-538, doi:10.1038/embor.2012.61 (2012).
125 Keuper, K., Wieland, A., Raschle, M. & Storchova, Z. Processes shaping cancer genomes - From mitotic defects to chromosomal rearrangements. *DNA Repair (Amst)* **107**, 103207, doi:10.1016/j.dnarep.2021.103207 (2021).
126 Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027-1038, doi:10.1016/0092-8674(93)90546-3 (1993).
127 Al-Tassan, N. *et al.* Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet* **30**, 227-232, doi:10.1038/ng828 (2002).
128 Janssen, A., van der Burg, M., Szuhai, K., Kops, G. J. & Medema, R. H. Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. *Science* **333**, 1895–1898, doi:10.1126/science.1210214 (2011).
129 Bakhoum, S. F., Kabeche, L., Murnane, J. P., Zaki, B. I. & Compton, D. A. DNA-damage response during mitosis induces whole-chromosome missegregation. *Cancer discovery* **4**, 1281–1289, doi:10.1158/2159-8290.Cd-14-0403 (2014).
130 Feng, W. & Jasin, M. BRCA2 suppresses replication stress-induced mitotic and G1 abnormalities through homologous recombination. *Nat Commun* **8**, 525, doi:10.1038/s41467-017-00634-0 (2017).
131 Ratnaparkhe, M. *et al.* Defective DNA damage repair leads to frequent catastrophic genomic events in murine and human tumors. *Nat Commun* **9**, 4760, doi:10.1038/s41467-018-06925-4 (2018).
132 Ciccia, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol Cell* **40**, 179-204, doi:10.1016/j.molcel.2010.09.019 (2010).
133 De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**, 169-185, doi:10.1093/mutage/geh025 (2004).
134 Hoeijmakers, J. H. DNA damage, aging, and cancer. *N Engl J Med* **361**, 1475-1485, doi:10.1056/NEJMra0804615 (2009).
135 Wyatt, M. D. & Pittman, D. L. Methylating agents and DNA repair responses: Methylated bases and sources of strand breaks. *Chem Res Toxicol* **19**, 1580-1594, doi:10.1021/tx060164e (2006).
136 Trewick, S. C., Henshaw, T. F., Hausinger, R. P., Lindahl, T. & Sedgwick, B. Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. *Nature* **419**, 174-178, doi:10.1038/nature00908 (2002).
137 Phillips, D. H., Hewer, A., Martin, C. N., Garner, R. C. & King, M. M. Correlation of DNA adduct levels in human lung with cigarette smoking. *Nature* **336**, 790-792, doi:10.1038/336790a0 (1988).
138 Lenaz, L. Mitomycin C in advanced breast cancer. *Cancer Treat Rev* **12**, 235-249, doi:10.1016/0305-7372(85)90007-6 (1985).

139 Vaz, B., Popovic, M. & Ramadan, K. DNA-Protein Crosslink Proteolysis Repair. *Trends Biochem Sci* **42**, 483-495, doi:10.1016/j.tibs.2017.03.005 (2017).
140 Byun, T. S., Pacek, M., Yee, M. C., Walter, J. C. & Cimprich, K. A. Functional uncoupling of MCM helicase and DNA polymerase activities activates the ATR-dependent checkpoint. *Genes Dev* **19**, 1040-1052, doi:10.1101/gad.1301205 (2005).
141 Ward, I. M., Minn, K. & Chen, J. UV-induced ataxia-telangiectasia-mutated and Rad3-related (ATR) activation requires replication stress. *J Biol Chem* **279**, 9677-9680, doi:10.1074/jbc.C300554200 (2004).
142 David, S. S., O'Shea, V. L. & Kundu, S. Base-excision repair of oxidative DNA damage. *Nature* **447**, 941-950, doi:10.1038/nature05978 (2007).
143 Mok, M. C. Y. *et al.* Identification of an XRCC1 DNA binding activity essential for retention at sites of DNA damage. *Sci Rep* **9**, 3095, doi:10.1038/s41598-019-39543-1 (2019).
144 Ronson, G. E. *et al.* PARP1 and PARP2 stabilise replication forks at base excision repair intermediates through Fbh1-dependent Rad51 regulation. *Nat Commun* **9**, 746, doi:10.1038/s41467-018-03159-2 (2018).
145 Gillet, L. C. & Scharer, O. D. Molecular mechanisms of mammalian global genome nucleotide excision repair. *Chem Rev* **106**, 253-276, doi:10.1021/cr040483f (2006).
146 Scharer, O. D. Nucleotide excision repair in eukaryotes. *Cold Spring Harb Perspect Biol* **5**, a012609, doi:10.1101/cshperspect.a012609 (2013).
147 Peltomaki, P. Role of DNA mismatch repair defects in the pathogenesis of human cancer. *J Clin Oncol* **21**, 1174-1179, doi:10.1200/JCO.2003.04.060 (2003).
148 Jiricny, J. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol* **7**, 335-346, doi:10.1038/nrm1907 (2006).
149 Lieber, M. R. The mechanism of human nonhomologous DNA end joining. *J Biol Chem* **283**, 1-5, doi:10.1074/jbc.R700039200 (2008).
150 Lieber, M. R. NHEJ and its backup pathways in chromosomal translocations. *Nat Struct Mol Biol* **17**, 393-395, doi:10.1038/nsmb0410-393 (2010).
151 San Filippo, J., Sung, P. & Klein, H. Mechanism of eukaryotic homologous recombination. *Annu Rev Biochem* **77**, 229-257, doi:10.1146/annurev.biochem.77.061306.125255 (2008).
152 Moynahan, M. E. & Jasin, M. Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nat Rev Mol Cell Biol* **11**, 196-207, doi:10.1038/nrm2851 (2010).
153 Ciccia, A. *et al.* Identification of FAAP24, a Fanconi anemia core complex protein that interacts with FANCM. *Mol Cell* **25**, 331-343, doi:10.1016/j.molcel.2007.01.003 (2007).
154 Räschle, M. *et al.* Mechanism of replication-coupled DNA interstrand crosslink repair. *Cell* **134**, 969–980, doi:10.1016/j.cell.2008.08.030 (2008).
155 Rennie, M. L., Arkinson, C., Chaugule, V. K., Toth, R. & Walden, H. Structural basis of FANCD2 deubiquitination by USP1-UAF1. *Nat Struct Mol Biol* **28**, 356-364, doi:10.1038/s41594-021-00576-8 (2021).
156 Baker, D. J. *et al.* Nucleotide excision repair eliminates unique DNA-protein cross-links from mammalian cells. *J Biol Chem* **282**, 22592-22604, doi:10.1074/jbc.M702856200 (2007).
157 Ide, H., Shoulkamy, M. I., Nakano, T., Miyamoto-Matsubara, M. & Salem, A. M. Repair and biochemical effects of DNA-protein crosslinks. *Mutat Res* **711**, 113-122, doi:10.1016/j.mrfmmm.2010.12.007 (2011).
158 Quinones, J. L. *et al.* Enzyme mechanism-based, oxidative DNA-protein cross-links formed with DNA polymerase beta in vivo. *Proc Natl Acad Sci U S A* **112**, 8602-8607, doi:10.1073/pnas.1501101112 (2015).
159 Warmerdam, D. O. & Kanaar, R. Dealing with DNA damage: relationships between checkpoint and repair pathways. *Mutat Res* **704**, 2-11, doi:10.1016/j.mrrev.2009.12.001 (2010).

160 Patrick, J. L., Christopher, V. & Karlene, A. C. Analyzing the ATR-mediated checkpoint using Xenopus egg extracts. *Methods* **41**, 222-231, doi:https://doi.org/10.1016/j.ymeth.2006.07.024 (2007).
161 Hoogenboom, W. S., Klein Douwel, D. & Knipscheer, P. Xenopus egg extract: A powerful tool to study genome maintenance mechanisms. *Developmental biology* **428**, 300–309, doi:10.1016/j.ydbio.2017.03.033 (2017).
162 Räschle, M. *et al.* DNA repair. Proteomics reveals dynamic assembly of repair complexes during bypass of DNA cross-links. *Science (New York, N.Y.)* **348**, 1253671, doi:10.1126/science.1253671 (2015).
163 Larsen, N. B. *et al.* Replication-Coupled DNA-Protein Crosslink Repair by SPRTN and the Proteasome in Xenopus Egg Extracts. *Molecular cell* **73**, 574-588.e577, doi:10.1016/j.molcel.2018.11.024 (2019).
164 Diaz, M. & Pecinka, A. Scaffolding for Repair: Understanding Molecular Functions of the SMC5/6 Complex. *Genes (Basel)* **9**, doi:10.3390/genes9010036 (2018).
165 Raschle, M. Proteomics reveals a new DNA repair factor involved in DNA damage signaling. *Mol Cell Oncol* **4**, e1263713, doi:10.1080/23723556.2016.1263713 (2017).
166 Duxin, J. P., Dewar, J. M., Yardimci, H. & Walter, J. C. Repair of a DNA-protein crosslink by replication-coupled proteolysis. *Cell* **159**, 346–357, doi:10.1016/j.cell.2014.09.024 (2014).
167 Stingele, J., Schwarz, M. S., Bloemeke, N., Wolf, P. G. & Jentsch, S. A DNA-dependent protease involved in DNA-protein crosslink repair. *Cell* **158**, 327-338, doi:10.1016/j.cell.2014.04.053 (2014).
168 Lessel, D. *et al.* Mutations in SPRTN cause early onset hepatocellular carcinoma, genomic instability and progeroid features. *Nat Genet* **46**, 1239-1244, doi:10.1038/ng.3103 (2014).
169 Stingele, J. *et al.* Mechanism and Regulation of DNA-Protein Crosslink Repair by the DNA-Dependent Metalloprotease SPRTN. *Molecular cell* **64**, 688–703, doi:10.1016/j.molcel.2016.09.031 (2016).
170 Vaz, B. *et al.* Metalloprotease SPRTN/DVC1 Orchestrates Replication-Coupled DNA-Protein Crosslink Repair. *Mol Cell* **64**, 704-719, doi:10.1016/j.molcel.2016.09.032 (2016).
171 Thompson, L. L., Jeusset, L. M., Lepage, C. C. & McManus, K. J. Evolving Therapeutic Strategies to Exploit Chromosome Instability in Cancer. *Cancers (Basel)* **9**, doi:10.3390/cancers9110151 (2017).
172 Sheltzer, J. M. *et al.* Single-chromosome Gains Commonly Function as Tumor Suppressors. *Cancer cell* **31**, 240–255, doi:10.1016/j.ccell.2016.12.004 (2017).
173 Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676-689 e673, doi:10.1016/j.ccell.2018.03.007 (2018).
174 Sheltzer, J. M. & Amon, A. The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends Genet* **27**, 446-453, doi:10.1016/j.tig.2011.07.003 (2011).
175 Thompson, S. L. & Compton, D. A. Examining the link between chromosomal instability and aneuploidy in human cells. *J Cell Biol* **180**, 665-672, doi:10.1083/jcb.200712029 (2008).
176 Thompson, S. L., Bakhoum, S. F. & Compton, D. A. Mechanisms of chromosomal instability. *Curr Biol* **20**, R285-295, doi:10.1016/j.cub.2010.01.034 (2010).
177 Haering, C. H., Farcas, A. M., Arumugam, P., Metson, J. & Nasmyth, K. The cohesin ring concatenates sister DNA molecules. *Nature* **454**, 297-301, doi:10.1038/nature07098 (2008).
178 Ueki, Y. *et al.* A highly conserved pocket on PP2A-B56 is required for hSgo1 binding and cohesion protection during mitosis. *EMBO Rep* **22**, e52295, doi:10.15252/embr.202052295 (2021).
179 Liu, H., Jia, L. & Yu, H. Phospho-H2A and cohesin specify distinct tension-regulated Sgo1 pools at kinetochores and inner centromeres. *Curr Biol* **23**, 1927-1933, doi:10.1016/j.cub.2013.07.078 (2013).

180 Hara, K. *et al.* Structure of cohesin subcomplex pinpoints direct shugoshin-Wapl antagonism in centromeric cohesion. *Nature structural & molecular biology* **21**, 864–870, doi:10.1038/nsmb.2880 (2014).
181 Mirkovic, M., Hutter, L. H., Novak, B. & Oliveira, R. A. Premature Sister Chromatid Separation Is Poorly Detected by the Spindle Assembly Checkpoint as a Result of System-Level Feedback. *Cell Rep* **13**, 469-478, doi:10.1016/j.celrep.2015.09.020 (2015).
182 Cheeseman, I. M. The kinetochore. *Cold Spring Harb Perspect Biol* **6**, a015826, doi:10.1101/cshperspect.a015826 (2014).
183 Gregan, J., Polakova, S., Zhang, L., Tolic-Norrelykke, I. M. & Cimini, D. Merotelic kinetochore attachment: causes and effects. *Trends Cell Biol* **21**, 374–381, doi:10.1016/j.tcb.2011.01.003 (2011).
184 Pihan, G. A., Wallace, J., Zhou, Y. & Doxsey, S. J. Centrosome abnormalities and chromosome instability occur together in pre-invasive carcinomas. *Cancer Res* **63**, 1398-1404 (2003).
185 Fukasawa, K., Choi, T., Kuriyama, R., Rulong, S. & Vande Woude, G. F. Abnormal centrosome amplification in the absence of p53. *Science* **271**, 1744-1747, doi:10.1126/science.271.5256.1744 (1996).
186 Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460**, 278-282, doi:10.1038/nature08136 (2009).
187 Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nat Rev Genet* **13**, 189-203, doi:10.1038/nrg3123 (2012).
188 Izawa, D. & Pines, J. The mitotic checkpoint complex binds a second CDC20 to inhibit active APC/C. *Nature* **517**, 631-634, doi:10.1038/nature13911 (2015).
189 Topham, C. H. & Taylor, S. S. Mitosis and apoptosis: how is the balance set? *Curr Opin Cell Biol* **25**, 780-785, doi:10.1016/j.ceb.2013.07.003 (2013).
190 Rasmussen, S. A., Wong, L. Y., Yang, Q., May, K. M. & Friedman, J. M. Population-based analyses of mortality in trisomy 13 and trisomy 18. *Pediatrics* **111**, 777-784, doi:10.1542/peds.111.4.777 (2003).
191 de Graaf, G., Buckley, F. & Skotko, B. G. Estimates of the live births, natural losses, and elective terminations with Down syndrome in the United States. *Am J Med Genet A* **167A**, 756-767, doi:10.1002/ajmg.a.37001 (2015).
192 Torres, E. M. *et al.* Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science* **317**, 916–924, doi:10.1126/science.1142210 (2007).
193 Upender, M. B. *et al.* Chromosome transfer induced aneuploidy results in complex dysregulation of the cellular transcriptome in immortalized and cancer cells. *Cancer Res* **64**, 6941–6949, doi:10.1158/0008-5472.Can-04-0474 (2004).
194 Williams, B. R. *et al.* Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. *Science* **322**, 703–709, doi:10.1126/science.1160058 (2008).
195 Lockstone, H. E. *et al.* Gene expression profiling in the adult Down syndrome brain. *Genomics* **90**, 647–660, doi:10.1016/j.ygeno.2007.08.005 (2007).
196 Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379**, 131-137, doi:10.1038/379131a0 (1996).
197 Torres, E. M., Springer, M. & Amon, A. No current evidence for widespread dosage compensation in S. cerevisiae. *eLife* **5**, e10996, doi:10.7554/eLife.10996 (2016).
198 Hose, J. *et al.* Dosage compensation can buffer copy-number variation in wild yeast. *eLife* **4**, doi:10.7554/eLife.05462 (2015).
199 Dephoure, N. *et al.* Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. *eLife* **3**, e03023, doi:10.7554/eLife.03023 (2014).
200 Liu, Y. *et al.* Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. *Nat Commun* **8**, 1212, doi:10.1038/s41467-017-01422-6 (2017).

201 Sheltzer, J. M. A transcriptional and metabolic signature of primary aneuploidy is present in chromosomally unstable cancer cells and informs clinical prognosis. *Cancer Res* **73**, 6401-6412, doi:10.1158/0008-5472.CAN-13-0749 (2013).
202 Durrbaum, M. *et al.* Unique features of the transcriptional response to model aneuploidy in human cells. *BMC Genomics* **15**, 139, doi:10.1186/1471-2164-15-139 (2014).
203 Chunduri, N. K. & Storchova, Z. The diverse consequences of aneuploidy. *Nat Cell Biol* **21**, 54–62, doi:10.1038/s41556-018-0243-8 (2019).
204 Deshaies, R. J. Proteotoxic crisis, the ubiquitin-proteasome system, and cancer therapy. *BMC Biol* **12**, 94, doi:10.1186/s12915-014-0094-0 (2014).
205 Anders, K. R. *et al.* A strategy for constructing aneuploid yeast strains by transient nondisjunction of a target chromosome. *BMC Genet* **10**, 36, doi:10.1186/1471-2156-10-36 (2009).
206 Donnelly, N. & Storchova, Z. Aneuploidy and proteotoxic stress in cancer. *Mol Cell Oncol* **2**, e976491, doi:10.4161/23723556.2014.976491 (2015).
207 Aivazidis, S. *et al.* The burden of trisomy 21 disrupts the proteostasis network in Down syndrome. *PLoS One* **12**, e0176307, doi:10.1371/journal.pone.0176307 (2017).
208 Donnelly, N., Passerini, V., Durrbaum, M., Stingele, S. & Storchova, Z. HSF1 deficiency and impaired HSP90-dependent protein folding are hallmarks of aneuploid human cells. *EMBO J* **33**, 2374–2387, doi:10.15252/embj.201488648 (2014).
209 Torres, E. M. *et al.* Identification of aneuploidy-tolerating mutations. *Cell* **143**, 71–83, doi:10.1016/j.cell.2010.08.038 (2010).
210 Hanna, J. *et al.* Deubiquitinating enzyme Ubp6 functions noncatalytically to delay proteasomal degradation. *Cell* **127**, 99–111, doi:10.1016/j.cell.2006.07.038 (2006).
211 Dodgson, S. E., Santaguida, S., Kim, S., Sheltzer, J. & Amon, A. The pleiotropic deubiquitinase Ubp3 confers aneuploidy tolerance. *Genes Dev* **30**, 2259–2271, doi:10.1101/gad.287474.116 (2016).
212 Santaguida, S. & Amon, A. Aneuploidy triggers a TFEB-mediated lysosomal stress response. *Autophagy* **11**, 2383-2384, doi:10.1080/15548627.2015.1110670 (2015).
213 Kelly, T. & Callegari, A. J. Dynamics of DNA replication in a eukaryotic cell. *Proc Natl Acad Sci U S A* **116**, 4973-4982, doi:10.1073/pnas.1818680116 (2019).
214 Santaguida, S. *et al.* Chromosome Mis-segregation Generates Cell-Cycle-Arrested Cells with Complex Karyotypes that Are Eliminated by the Immune System. *Dev Cell* **41**, 638-651 e635, doi:10.1016/j.devcel.2017.05.022 (2017).
215 Passerini, V. *et al.* The presence of extra chromosomes leads to genomic instability. *Nat Commun* **7**, 10754, doi:10.1038/ncomms10754 (2016).
216 Spurgers, K. B. *et al.* Identification of cell cycle regulatory genes as principal targets of p53-mediated transcriptional repression. *J Biol Chem* **281**, 25134-25142, doi:10.1074/jbc.M513901200 (2006).
217 Blank, H. M., Sheltzer, J. M., Meehl, C. M. & Amon, A. Mitotic entry in the presence of DNA damage is a widespread property of aneuploidy in yeast. *Mol Biol Cell* **26**, 1440–1451, doi:10.1091/mbc.E14-10-1442 (2015).
218 Lamm, N. *et al.* Genomic Instability in Human Pluripotent Stem Cells Arises from Replicative Stress and Chromosome Condensation Defects. *Cell Stem Cell* **18**, 253-261, doi:10.1016/j.stem.2015.11.003 (2016).
219 Meena, J. K. *et al.* Telomerase abrogates aneuploidy-induced telomere replication stress, senescence and cell depletion. *EMBO J* **34**, 1371–1384, doi:10.15252/embj.201490070 (2015).
220 Vigano, C. *et al.* Quantitative proteomic and phosphoproteomic comparison of human colon cancer DLD-1 cells differing in ploidy and chromosome stability. *Mol Biol Cell* **29**, 1031–1047, doi:10.1091/mbc.E17-10-0577 (2018).

221 Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. Cyclic GMP-AMP Synthase Is a Cytosolic DNA Sensor That Activates the Type I Interferon Pathway. *Science* **339**, 786, doi:10.1126/science.1232458 (2013).
222 Kesler, S. R. Turner syndrome. *Child Adolesc Psychiatr Clin N Am* **16**, 709–722, doi:10.1016/j.chc.2007.02.004 (2007).
223 Watson, C. T., Marques-Bonet, T., Sharp, A. J. & Mefford, H. C. The Genetics of Microdeletion and Microduplication Syndromes: An Update. *Annual Review of Genomics and Human Genetics* **15**, 215–244, doi:10.1146/annurev-genom-091212-153408 (2014).
224 Butler, M. G., Meaney, F. J. & Palmer, C. G. Clinical and cytogenetic survey of 39 individuals with Prader-Labhart-Willi syndrome. *Am J Med Genet* **23**, 793-809, doi:10.1002/ajmg.1320230307 (1986).
225 Chen, K. S. *et al.* Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat Genet* **17**, 154-163, doi:10.1038/ng1097-154 (1997).
226 Schwartz, M. *et al.* How chromosomal deletions can unmask recessive mutations? Deletions in 10q11.2 associated with CHAT or SLC18A3 mutations lead to congenital myasthenic syndrome. *American Journal of Medical Genetics Part A* **176**, 151–155, doi:10.1002/ajmg.a.38515 (2018).
227 Egloff, M. *et al.* Whole-exome sequence analysis highlights the role of unmasked recessive mutations in copy number variants with incomplete penetrance. *Eur J Hum Genet* **26**, 912–918, doi:10.1038/s41431-018-0124-4 (2018).
228 Poot, M. & Haaf, T. Mechanisms of Origin, Phenotypic Effects and Diagnostic Implications of Complex Chromosome Rearrangements. *Mol Syndromol* **6**, 110–134, doi:10.1159/000438812 (2015).
229 Ebert, B. L. *et al.* Identification of RPS14 as a 5q- syndrome gene by RNA interference screen. *Nature* **451**, 335–339, doi:10.1038/nature06494 (2008).
230 Liu, Y. *et al.* Deletions linked to TP53 loss drive cancer through p53-independent mechanisms. *Nature* **531**, 471–475, doi:10.1038/nature17157 (2016).
231 Chunduri, N. K. *et al.* Systems approaches identify the consequences of monosomy in somatic human cells. *Nat Commun* **12**, 5576, doi:10.1038/s41467-021-25288-x (2021).
232 Storchova, Z. & Pellman, D. From polyploidy to aneuploidy, genome instability and cancer. *Nat Rev Mol Cell Biol* **5**, 45-54, doi:10.1038/nrm1276 (2004).
233 Comai, L. The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6**, 836-846, doi:10.1038/nrg1711 (2005).
234 Storchova, Z. *et al.* Genome-wide genetic analysis of polyploidy in yeast. *Nature* **443**, 541-547, doi:10.1038/nature05178 (2006).
235 Mayer, V. W. & Aguilera, A. High levels of chromosome instability in polyploids of Saccharomyces cerevisiae. *Mutat Res* **231**, 177-186, doi:10.1016/0027-5107(90)90024-x (1990).
236 Schoenfelder, K. P. & Fox, D. T. The expanding implications of polyploidy. *J Cell Biol* **209**, 485-491, doi:10.1083/jcb.201502016 (2015).
237 Selmecki, A. M. *et al.* Polyploidy can drive rapid adaptation in yeast. *Nature* **519**, 349-352, doi:10.1038/nature14187 (2015).
238 Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131**, 452-462, doi:10.1016/j.cell.2007.10.022 (2007).
239 Quinton, R. J. *et al.* Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* **590**, 492-497, doi:10.1038/s41586-020-03133-3 (2021).
240 Cohen-Sharir, Y. *et al.* Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition. *Nature* **590**, 486-491, doi:10.1038/s41586-020-03114-6 (2021).

241 Galitski, T., Saldanha, A. J., Styles, C. A., Lander, E. S. & Fink, G. R. Ploidy regulation of gene expression. *Science (New York, N.Y.)* **285**, 251–254, doi:10.1126/science.285.5425.251 (1999).
242 Di Talia, S., Skotheim, J. M., Bean, J. M., Siggia, E. D. & Cross, F. R. The effects of molecular noise and size control on variability in the budding yeast cell cycle. *Nature* **448**, 947-951, doi:10.1038/nature06072 (2007).
243 Andalis, A. A. *et al.* Defects arising from whole-genome duplications in Saccharomyces cerevisiae. *Genetics* **167**, 1109-1121, doi:10.1534/genetics.104.029256 (2004).
244 Mable, B. K. Ploidy evolution in the yeast Saccharomyces cerevisiae: a test of the nutrient limitation hypothesis. *J Evol Biol* **14**, 157-170, doi:10.1046/j.1420-9101.2001.00245.x (2001).
245 Wu, C. Y., Rolfe, P. A., Gifford, D. K. & Fink, G. R. Control of transcription by cell size. *PLoS Biol* **8**, e1000523, doi:10.1371/journal.pbio.1000523 (2010).
246 Yahya, G. *et al.* Scaling of cellular proteome with ploidy. *bioRxiv*, 2021.2005.2006.442919, doi:10.1101/2021.05.06.442919 (2021).
247 Sparks, J. L. *et al.* The CMG Helicase Bypasses DNA-Protein Cross-Links to Facilitate Their Repair. *Cell* **176**, 167-181.e121, doi:10.1016/j.cell.2018.10.053 (2019).
248 van der Maaten, L. J. P. & Hinton, G. E. Visualizing High-Dimensional Data Using t-SNE. (2008).
249 Awate, S. & Brosh, R. M., Jr. Interactive Roles of DNA Helicases and Translocases with the Single-Stranded DNA Binding Protein RPA in Nucleic Acid Metabolism. *Int J Mol Sci* **18**, doi:10.3390/ijms18061233 (2017).
250 Bailis, J. M. & Forsburg, S. L. MCM proteins: DNA damage, mutagenesis and repair. *Curr Opin Genet Dev* **14**, 17-21, doi:10.1016/j.gde.2003.11.002 (2004).
251 Sareen, A., Chaudhury, I., Adams, N. & Sobeck, A. Fanconi anemia proteins FANCD2 and FANCI exhibit different DNA damage responses during S-phase. *Nucleic Acids Res* **40**, 8425-8439, doi:10.1093/nar/gks638 (2012).
252 Kobel, H. R. & Du Pasquier, L. Genetics of polyploid Xenopus. *Trends in Genetics* **2**, 310-315, doi:https://doi.org/10.1016/0168-9525(86)90286-6 (1986).
253 Kratz, K. *et al.* Deficiency of FANCD2-associated nuclease KIAA1018/FAN1 sensitizes cells to interstrand crosslinking agents. *Cell* **142**, 77-88, doi:10.1016/j.cell.2010.06.022 (2010).
254 Li, L., Tan, W. & Deans, A. J. Structural insight into FANCI-FANCD2 monoubiquitination. *Essays in biochemistry* **64**, 807-817, doi:10.1042/ebc20200001 (2020).
255 Richardson, C. D. *et al.* CRISPR-Cas9 genome editing in human cells occurs via the Fanconi anemia pathway. *Nature genetics* **50**, 1132–1139, doi:10.1038/s41588-018-0174-0 (2018).
256 Session, A. M. *et al.* Genome evolution in the allotetraploid frog Xenopus laevis. *Nature* **538**, 336-343, doi:10.1038/nature19840 (2016).
257 Haveliwala, T. H. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. *IEEE Trans. Knowl. Data Eng.* **15**, 784-796 (2003).
258 Suhasini, A. N. & Brosh, R. M., Jr. Fanconi anemia and Bloom's syndrome crosstalk through FANCJ-BLM helicase interaction. *Trends Genet* **28**, 7-13, doi:10.1016/j.tig.2011.09.003 (2012).
259 Shigechi, T. *et al.* ATR-ATRIP kinase complex triggers activation of the Fanconi anemia DNA repair pathway. *Cancer Res* **72**, 1149-1156, doi:10.1158/0008-5472.CAN-11-2904 (2012).
260 Karimi, K. *et al.* Xenbase: a genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Res* **46**, D861-D868, doi:10.1093/nar/gkx936 (2018).
261 Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res* **36**, W5-9, doi:10.1093/nar/gkn201 (2008).
262 Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-1191, doi:10.1038/nprot.2009.97 (2009).

263 Liang, C.-C. *et al.* The FANCD2–FANCI complex is recruited to DNA interstrand crosslinks before monoubiquitination of FANCD2. *Nature Communications* **7**, 12124, doi:10.1038/ncomms12124 (2016).
264 Meetei, A. R. *et al.* A multiprotein nuclear complex connects Fanconi anemia and Bloom syndrome. *Mol Cell Biol* **23**, 3417-3426, doi:10.1128/MCB.23.10.3417-3426.2003 (2003).
265 Bhat, K. P. & Cortez, D. RPA and RAD51: fork reversal, fork protection, and genome stability. *Nat Struct Mol Biol* **25**, 446-453, doi:10.1038/s41594-018-0075-z (2018).
266 Tomida, J. *et al.* REV7 is essential for DNA damage tolerance via two REV3L binding sites in mammalian DNA polymerase zeta. *Nucleic Acids Res* **43**, 1000-1011, doi:10.1093/nar/gku1385 (2015).
267 Khoudoli, G. A. *et al.* Temporal profiling of the chromatin proteome reveals system-wide responses to replication inhibition. *Current biology : CB* **18**, 838–843, doi:10.1016/j.cub.2008.04.075 (2008).
268 Bukata, L., Parker, S. L. & D'Angelo, M. A. Nuclear pore complexes in the maintenance of genome integrity. *Curr Opin Cell Biol* **25**, 378-386, doi:10.1016/j.ceb.2013.03.002 (2013).
269 Benitez, A. *et al.* FANCA Promotes DNA Double-Strand Break Repair by Catalyzing Single-Strand Annealing and Strand Exchange. *Mol Cell* **71**, 621-628 e624, doi:10.1016/j.molcel.2018.06.030 (2018).
270 Klages-Mundt, N. L. & Li, L. Formation and repair of DNA-protein crosslink damage. *Sci China Life Sci* **60**, 1065-1076, doi:10.1007/s11427-017-9183-4 (2017).
271 Knipscheer, P. *et al.* The Fanconi anemia pathway promotes replication-dependent DNA interstrand cross-link repair. *Science* **326**, 1698-1701, doi:10.1126/science.1182372 (2009).
272 Revenkova, E. *et al.* Cornelia de Lange syndrome mutations in SMC1A or SMC3 affect binding to DNA. *Hum Mol Genet* **18**, 418-427, doi:10.1093/hmg/ddn369 (2009).
273 Martrat, G. *et al.* Exploring the link between MORF4L1 and risk of breast cancer. *Breast cancer research* **13**, 1-14 (2011).
274 Hayakawa, T. *et al.* MRG15 binds directly to PALB2 and stimulates homology-directed repair of chromosomal breaks. *J Cell Sci* **123**, 1124-1130, doi:10.1242/jcs.060178 (2010).
275 Kühbacher, U. & Duxin, J. P. How to fix DNA-protein crosslinks. *DNA Repair* **94**, 102924, doi:https://doi.org/10.1016/j.dnarep.2020.102924 (2020).
276 Soto, M. *et al.* p53 Prohibits Propagation of Chromosome Segregation Errors that Produce Structural Aneuploidies. *Cell Rep* **19**, 2423–2431, doi:10.1016/j.celrep.2017.05.055 (2017).
277 Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400–404, doi:10.1038/nature03479 (2005).
278 Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97, doi:10.1093/nar/gkw377 (2016).
279 Kosti, I., Jain, N., Aran, D., Butte, A. J. & Sirota, M. Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Sci Rep* **6**, 24799, doi:10.1038/srep24799 (2016).
280 Schneider, W. M., Chevillotte, M. D. & Rice, C. M. Interferon-stimulated genes: a complex web of host defenses. *Annu Rev Immunol* **32**, 513–545, doi:10.1146/annurev-immunol-032713-120231 (2014).
281 Sun, M. *et al.* Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res* **22**, 1350-1359, doi:10.1101/gr.130161.111 (2012).
282 Lau, H. T., Suh, H. W., Golkowski, M. & Ong, S. E. Comparing SILAC- and stable isotope dimethyl-labeling approaches for quantitative proteomics. *J Proteome Res* **13**, 4164-4174, doi:10.1021/pr500630a (2014).
283 Thomson, G. J. *et al.* Metabolism-induced oxidative stress and DNA damage selectively trigger genome instability in polyploid fungal cells. *EMBO J* **38**, e101597, doi:10.15252/embj.2019101597 (2019).

284 Jeggo, P. A., Pearl, L. H. & Carr, A. M. DNA repair, genome stability and cancer: a historical perspective. *Nat Rev Cancer* **16**, 35-42, doi:10.1038/nrc.2015.4 (2016).
285 Yao, Y. & Dai, W. Genomic Instability and Cancer. *J Carcinog Mutagen* **5**, doi:10.4172/2157-2518.1000165 (2014).
286 Potapova, T. A., Zhu, J. & Li, R. Aneuploidy and chromosomal instability: a vicious cycle driving cellular evolution and cancer genome chaos. *Cancer Metastasis Rev* **32**, 377-389, doi:10.1007/s10555-013-9436-6 (2013).
287 Was, H. *et al.* Polyploidy formation in cancer cells: How a Trojan horse is born. *Semin Cancer Biol*, doi:10.1016/j.semcancer.2021.03.003 (2021).
288 Postow, L., Woo, E. M., Chait, B. T. & Funabiki, H. Identification of SMARCAL1 as a component of the DNA damage response. *J Biol Chem* **284**, 35951-35961, doi:10.1074/jbc.M109.048330 (2009).
289 Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res* **49**, D884-D891, doi:10.1093/nar/gkaa942 (2021).
290 Moldovan, G.-L. & D'Andrea, A. D. How the Fanconi Anemia Pathway Guards the Genome. *Annual Review of Genetics* **43**, 223-249, doi:10.1146/annurev-genet-102108-134222 (2009).
291 Uringa, E. J. *et al.* RTEL1 contributes to DNA replication and repair and telomere maintenance. *Mol Biol Cell* **23**, 2782-2792, doi:10.1091/mbc.E12-03-0179 (2012).
292 Tarsounas, M. & Sung, P. The antitumorigenic roles of BRCA1-BARD1 in DNA repair and replication. *Nat Rev Mol Cell Biol* **21**, 284-299, doi:10.1038/s41580-020-0218-z (2020).
293 Martin, C. A. *et al.* Mutations in TOP3A Cause a Bloom Syndrome-like Disorder. *Am J Hum Genet* **103**, 456, doi:10.1016/j.ajhg.2018.08.012 (2018).
294 Watrin, E. & Peters, J. M. Cohesin and DNA damage repair. *Exp Cell Res* **312**, 2687-2693, doi:10.1016/j.yexcr.2006.06.024 (2006).
295 Yoshida, K. & Miki, Y. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Sci* **95**, 866-871, doi:10.1111/j.1349-7006.2004.tb02195.x (2004).
296 Zhang, F. *et al.* PALB2 links BRCA1 and BRCA2 in the DNA-damage response. *Curr Biol* **19**, 524-529, doi:10.1016/j.cub.2009.02.018 (2009).
297 Knobel, P. A., Kotov, I. N., Felley-Bosco, E., Stahel, R. A. & Marti, T. M. Inhibition of REV3 expression induces persistent DNA damage and growth arrest in cancer cells. *Neoplasia* **13**, 961-970, doi:10.1593/neo.11828 (2011).
298 Benson, F. E., Baumann, P. & West, S. C. Synergistic actions of Rad51 and Rad52 in recombination and DNA repair. *Nature* **391**, 401-404, doi:10.1038/34937 (1998).
299 He, J. *et al.* A novel RNA sequencing-based risk score model to predict papillary thyroid carcinoma recurrence. *Clin Exp Metastasis* **37**, 257-267, doi:10.1007/s10585-019-10011-4 (2020).
300 Selim, H. & Zhan, J. Towards shortest path identification on large networks. *Journal of Big Data* **3**, 10, doi:10.1186/s40537-016-0042-7 (2016).
301 Vize, P. D. & Zorn, A. M. Xenopus genomic data and browser resources. *Dev Biol* **426**, 194-199, doi:10.1016/j.ydbio.2016.03.030 (2017).
302 Sheltzer, J. M., Torres, E. M., Dunham, M. J. & Amon, A. Transcriptional consequences of aneuploidy. *Proc Natl Acad Sci U S A* **109**, 12644–12649, doi:10.1073/pnas.1209227109 (2012).
303 Licciardi, F. *et al.* Human blastocysts of normal and abnormal karyotypes display distinct transcriptome profiles. *Sci Rep* **8**, 14906, doi:10.1038/s41598-018-33279-0 (2018).
304 Malone, J. H. *et al.* Mediation of Drosophila autosomal dosage effects and compensation by network interactions. *Genome Biol* **13**, r28, doi:10.1186/gb-2012-13-4-r28 (2012).
305 Schukken, K. M. & Sheltzer, J. M. Extensive protein dosage compensation in aneuploid human cancers. *bioRxiv*, 2021.2006.2018.449005, doi:10.1101/2021.06.18.449005 (2021).

306 McShane, E. *et al.* Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167**, 803-815 e821, doi:10.1016/j.cell.2016.09.015 (2016).
307 Morrill, S. A. & Amon, A. Why haploinsufficiency persists. *Proc Natl Acad Sci U S A* **116**, 11866–11871, doi:10.1073/pnas.1900437116 (2019).
308 Dutt, S. *et al.* Haploinsufficiency for ribosomal protein genes causes selective activation of p53 in human erythroid progenitor cells. *Blood* **117**, 2567–2576, doi:10.1182/blood-2010-07-295238 (2011).
309 Kenmochi, N. *et al.* A map of 75 human ribosomal protein genes. *Genome Res* **8**, 509–523, doi:10.1101/gr.8.5.509 (1998).
310 Fisher, E. M. *et al.* Homologous ribosomal protein genes on the human X and Y chromosomes: escape from X inactivation and possible implications for Turner syndrome. *Cell* **63**, 1205–1218, doi:10.1016/0092-8674(90)90416-c (1990).
311 Zhu, P. J. *et al.* Activation of the ISR mediates the behavioral and neurophysiological abnormalities in Down syndrome. *Science* **366**, 843-849, doi:10.1126/science.aaw5185 (2019).
312 Iadevaia, V., Liu, R. & Proud, C. G. mTORC1 signaling controls multiple steps in ribosome biogenesis. *Semin Cell Dev Biol* **36**, 113–120, doi:10.1016/j.semcdb.2014.08.004 (2014).
313 Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962, doi:10.1016/j.cell.2013.10.011 (2013).
314 Smith, J. C. & Sheltzer, J. M. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. *Elife* **7**, doi:10.7554/eLife.39217 (2018).
315 Shukla, A. *et al.* Chromosome arm aneuploidies shape tumour evolution and drug response. *Nat Commun* **11**, 449, doi:10.1038/s41467-020-14286-0 (2020).
316 Sung, M. K., Reitsma, J. M., Sweredoski, M. J., Hess, S. & Deshaies, R. J. Ribosomal proteins produced in excess are degraded by the ubiquitin-proteasome system. *Mol Biol Cell* **27**, 2642–2652, doi:10.1091/mbc.E16-05-0290 (2016).
317 Weiss, R. L., Kukora, J. R. & Adams, J. The relationship between enzyme activity, cell geometry, and fitness in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences* **72**, 794-798 (1975).
318 Gerstein, A. C., Chun, H. J., Grant, A. & Otto, S. P. Genomic convergence toward diploidy in Saccharomyces cerevisiae. *PLoS Genet* **2**, e145, doi:10.1371/journal.pgen.0020145 (2006).
319 West, G. B., Woodruff, W. H. & Brown, J. H. Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. *Proc Natl Acad Sci U S A* **99 Suppl 1**, 2473-2478, doi:10.1073/pnas.012579799 (2002).
320 Savage, V. M., Deeds, E. J. & Fontana, W. Sizing up allometric scaling theory. *PLoS Comput Biol* **4**, e1000171, doi:10.1371/journal.pcbi.1000171 (2008).
321 Hou, C. *et al.* Energy uptake and allocation during ontogeny. *Science* **322**, 736-739, doi:10.1126/science.1162302 (2008).
322 Neurohr, G. E. *et al.* Excessive Cell Growth Causes Cytoplasm Dilution And Contributes to Senescence. *Cell* **176**, 1083-1097 e1018, doi:10.1016/j.cell.2019.01.018 (2019).
323 Lu, Y. J., Swamy, K. B. & Leu, J. Y. Experimental Evolution Reveals Interplay between Sch9 and Polyploid Stability in Yeast. *PLoS Genet* **12**, e1006409, doi:10.1371/journal.pgen.1006409 (2016).
324 Torelli, N. Q., Ferreira-Junior, J. R., Kowaltowski, A. J. & da Cunha, F. M. RTG1- and RTG2-dependent retrograde signaling controls mitochondrial activity and stress resistance in Saccharomyces cerevisiae. *Free Radic Biol Med* **81**, 30-37, doi:10.1016/j.freeradbiomed.2014.12.025 (2015).
325 Chen, Y., Zhao, G., Zahumensky, J., Honey, S. & Futcher, B. Differential Scaling of Gene Expression with Cell Size May Explain Size Control in Budding Yeast. *Molecular cell* **78**, 359-370.e356, doi:10.1016/j.molcel.2020.03.012 (2020).

326 Mayer, C. & Grummt, I. Ribosome biogenesis and cell growth: mTOR coordinates transcription by all three classes of nuclear RNA polymerases. *Oncogene* **25**, 6384-6391, doi:10.1038/sj.onc.1209883 (2006).
327 Ali, S. A. *et al.* Transcriptional corepressor TLE1 functions with Runx2 in epigenetic repression of ribosomal RNA genes. *Proc Natl Acad Sci U S A* **107**, 4165-4169, doi:10.1073/pnas.1000620107 (2010).
328 Laferte, A. *et al.* The transcriptional activity of RNA polymerase I is a key determinant for the level of all ribosome components. *Genes Dev* **20**, 2030-2040, doi:10.1101/gad.386106 (2006).
329 Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**, 4241–4257, doi:10.1091/mbc.11.12.4241 (2000).
330 González, A. & Hall, M. N. Nutrient sensing and TOR signaling in yeast and mammals. *The EMBO journal* **36**, 397-408, doi:10.15252/embj.201696010 (2017).
331 Kawai, S. *et al.* Mitochondrial genomic dysfunction causes dephosphorylation of Sch9 in the yeast Saccharomyces cerevisiae. *Eukaryot Cell* **10**, 1367-1369, doi:10.1128/EC.05157-11 (2011).
332 Kuznetsova, A. Y. *et al.* Chromosomal instability, tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells. *Cell Cycle* **14**, 2810–2820, doi:10.1080/15384101.2015.1068482 (2015).
333 Gillard, G. B. *et al.* Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biol* **22**, 103, doi:10.1186/s13059-021-02323-0 (2021).

I. List of abbreviations

4sU	4-thiouridin
4tU	4-thiouracil
APC/C	Anaphase-promoting complex/cyclosome
AP-MS	Affinity purification-mass spectrometry
ATM	Ataxia-telangiectasia mutated kinase
ATR	Ataxia telangiectasia and RAD3-related protein
BER	Base excision repair
CCLE	Cancer cell line encyclopedia
CHROMASS	Chromatin mass spectrometry
CID	Collision-induced dissociation
CIN	Chromosomal instability
CNA	Copy number alteration
CNV	Copy number variant
Da	Dalton
DDA	Data-dependent acquisition
DDR	DNA damage response
DIA	Data independent acquisition
DL	Deep-learning
DNA	Deoxyribonucleic acid
dNTP	deoxyribonucleoside 5´-triphosphate
DPC	DNA-protein crosslink
DSB	Double strand break
DTA	Dynamic transcriptome analysis
EASI	Easily abstractable sulfoxide-based isobaric tag
ESI	Electrospray ionization
FA	Fanconi Anemia
FC	Fold change
FDR	False discovery rate
FISH	Fluorescence in situ hybridization
FWHM	Full-width-half-maximum
GG	Global genome
GOC	Gene ontology consortium
GSEA	Gene set enrichment analysis
HBGF	Hybrid bipartite graph formulation
HCC	Hepatocellular carcinoma
HCD	High energy collision chamber
HCT	Human colon tumor
HPLC	High-performance liquid chromatography
HPT	HCT post tetraploid
hPSC	Human pluripotent stem cells
HR	Homologous recombination
HSF	Heat shock factor
HSP	Heat shock protein
HSS	High-speed supernatant
ICAT	Isotope coded affinity tag
ICL	Interstrand crosslink
IFN	Interferon
IR	Ionizing radiation

ISG	Interferon stimulated genes
KD	Knock-down
KNN	K-nearest neighbors
KEGG	Kyoto encyclopedia of genes and genomes
KO	Knock-out
LC-MS	Liquid chromatography - mass spectrometry
LFQ	Label-free quantification
LIMMA	Linear models for microarray data
LOD	Limit of detection
LOESS	Locally estimated scatterplot smoothing
LSA	Least-squares adaptive
LSS	Local least square
MALDI	Matrix assisted laser desorption and ionization
MANOVA	Multivariate analysis of variance
MCC	Mitotic checkpoint complex
MCL	Markov clustering algorithm
MMS	Methyl methanesulfonate
MNNG	N-methyl-N′-nitro-N-nitrosoguanidine
MNU	Methylnitrosourea
MRM	Multiple reaction monitoring
mRMR	Maximum relevance minimum redundancy
MRN	MRE11-RAD50-NBS1 complex
mRNA	(messenger) Ribonucleic acid
MSI	Microsatellite instability
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining
NPE	Nucleoplasmic extract
ORA	Overrepresentation analysis
ORC	Origin recognition complex
ORI	Origin of replication
PP-MS	Plasmid-pulldown mass spectrometry
PRM	Parallel reaction monitoring
PSM	Peptide-spectrum matches
PTM	Post translational modification
QQQ	Triple quadrupole mass spectrometer
RJALS	Ruijs-Aalfs syndrome
RM	RPE-derived monosomy
RNN	Recurrent neural networks
RNS	Reactive nitrogen species
ROS	Reactive oxygen species
RPE	Retinal pigment epithelium
RWR	Random walk with restart
SAM	Significance Analysis of Microarrays
SCNA	Somatic copy number alteration
SGSE	‘Spectral’ gene set enrichment
SILAC	Stable isotope labeling by amino acids in cell culture
SMC	Structural maintenance of chromosome
SSB	Single strand break
SSCC	Semi-supervised spectral clustering
ssDNA	Single stranded DNA

STM	Single reaction monitoring
SWATH-MS	Sequential window acquisition of all theoretical mass spectra
TC	Transcription-coupled
TCGA	The cancer genome atlas
TLS	Translesion synthesis
TMT	Tandem mass tag
TOF	Time-of-flight
t-SND	T-distributed stochastic neighbor embedding
UB-VS	Ubiquitin Vinyl Sulfone
UV	Ultraviolet
VSN	Variance stabilization normalization
WCGNA	Weighted gene correlation network analysis
WGD	Whole genome doubling
WGS	Whole genome sequencing

II. List of figures

III. List of tables

IV. Acknowledgement

The first person I want to thank here is my supervisor Prof. Dr. Zuzana Storchova, for the opportunity to work in her group as well as the steady support of my thesis. I am grateful for the guidance, very patient explanations, discussions and the great working atmosphere in your lab!

A special thanks also to my thesis committee members Prof. Dr. Michael Schroda and Prof. Dr. Johannes Herrmann for evaluating my thesis.

Further, I want to thank every collaborator that contributed directly and indirectly to this work. Prof. Dr. Maik Kschischo, Vanessa Schmitt and Pascal Rihm for providing support with the DNA Repair Atlas and contributing to the hopefully soon-to-be-published manuscript. Dr. Timo Mühlhaus, not only for letting me work on the Hyper-V cluster, but also for teaching me the fundamentals of bioinformatics and steering me on this course back during my bachelor studies. Special thanks to Dr. Narendra Chunduri, for his work on monosomies, being able to make sense of my analysis and generally for being an awesome person. Angela Wieland for helping with the revisions, and for being a great office mate. Dr. Galal Yahya, for being the expert on just about everything yeast related and his work on the yeast ploidy project and Devi Ngandiri for her significant contribution to the project. I owe special thanks to Dr. Markus Räschle for starting the work on the DNA Repair Atlas and teaching me everything mass spectrometry related.

I want to thank all other members of the Storchova lab. Sara, for our combined effort to keep each other sane while both writing our thesis. Sushweta, for always making the office atmosphere more lighthearted. Prince, Kristina, Jan, Karen, Isabel, Ingeborg, Robin and all the students for meaningful discussions and generally a great working atmosphere. Just let me know if there's a computer to fix again, will you? I also owe thanks to my family and my very important group of friends at home, simply for being there and believing in me. After nearly ten years at the TU Kaiserslautern the list of people I need to thank is far longer than what I could possibly write here. I owe thanks to all my friends that now work in other labs and the people I met along the way, the members of the Fachschaftsrat Biologie I worked with for so long and all fellow students that made my studies at the TU the awesome journey that it was. While I can't mention everyone by name, I am certainly more than grateful for all of you.

V. Curriculum Vitae

Name: Paul Robert Menges

Nationality: German

Education:

2018-2021 **phD candidate**, in the Department of Molecular Genetics, Technical University Kaiserslautern.

Dissertation topic: "*Multi-omics analysis as a tool to investigate causes and consequences of impaired genome integrity*"

2015-2018 **Master of Science**: Microbial and Plant Biotechnology, Technical University Kaiserslautern

Master Thesis, "*Identification and visualization of DNA repair modules through network analysis of proteomic data*."; Molecular Genetics, with Prof. Dr Zuzana Storchova.

2012-2016 **Bachelor of Science**: Life Sciences, Technical University Kaiserslautern

Bachelor Thesis, "*A functional graph-based approach for the de novo identification of protein mass spectrometry data*"; Computation Systems Biology, with Dr. Timo Mühlhaus.

2011-2012 Business Administration, *Hochschule Heilbronn*

Publications:

"*Systems approaches identify the consequences of monosomy in somatic human cells*". Nature Communications (2021) Narendra Kumar Chunduri, Paul Menges, Vincent Leon Gotsmann, Xiaoxiao Zhang, Balca R. Mardin, Christopher Buccitelli, Jan O. Korbel, Felix Willmund, Maik Kschischo, Markus Raeschle, Zuzana Storchova

"*Scaling of cellular proteome with ploidy*". Biorxvis (2021). Galal Yahya, Paul Menges, Devi Anggraini Ngandiri, Daniel Schulz, Andreas Wallek, Nils Kulak, Matthias Mann, Patrick Cramer, Van Savage, Markus Raeschle, Zuzana Storchova