TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Masterarbeit

Titel

The DNA Repair Atlas: A user-friendly web resource for the identification of functional DNA repair modules

Titel (Deutsch)

Der DNA Repair Atlas: Eine nutzerfreundliche Internetressource zur Identifikation funktionaler DNA Reparatur-Module

Angestrebter Abschluss

Master of Science (M.Sc.)

Name, Geburtsort

Pascal Rihm, Frankenthal (Pfalz)

Studiengang

Biology - Microbial and Plant Biotechnology

Abteilung

Molecular Genetics

Betreuer

Dr. Markus Räschle

Erstkorrektur

Prof. Dr. Zuzana Storchová

Zweitkorrektur

Jun.-Prof. Dr. Timo Mühlhaus

Datum der Einreichung

15.05.2020

TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Eidesstattliche Versicherung

Rihm Pascal
Name, Vorname

Ich versichere hiermit an Eides statt, dass ich die vorliegende Abschlussarbeit mit dem Titel

The DNA Repair Atlas: A user-friendly web resource for the identification of functional DNA repair modules

selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Kaiserslautern, den 15.05.2020
Ort, Datum, Unterschrift

Zusammenfassung

Die folgende Thesis präsentiert die Optimierung und Validierung einer webbasierten Software zur Analyse und Visualisierung massenspektrometriebasierter Daten von DNS Reparaturmechanismen. Die Rolle von DNS Replikation sowie spezifischer DNS Schäden und deren Reparatur wird beschrieben und die Beziehungen zwischen Proteinen, welche an DNS Reparaturmechanismen beteiligt sind, werden durch Einsatz von Nachbarschaftsähnlichkeits-Netzwerken visualisiert. Zusätzlich wird gezeigt, wie diese Netzwerke zur Analyse individueller Datensätze in Bezug auf die Beteiligung einzelner Proteine an DNS Replikation und Reparatur genutzt werden können. Etablierte Gruppierungsalgorithmen wurden implementiert bzw. optimiert und die Möglichkeit der Identifikation von funktionalen Modulen wird am Beispiel bekannter Reparaturkomplexe und deren Interaktoren erläutert. Um die Leistungsfähigkeit dieser Algorithmen zu beschreiben werden Referenznetzwerke basierend auf Nachbarschaftsähnlichkeiten für jeden einzelnen Reparaturmechanismus sowie Subnetzwerke, welche mittels der genannten Gruppierungsmethoden erstellt wurden, gezeigt. Ein Protein-Anreicherungswert wurde für jede Art von DNS Schaden mittels fortgeschrittener statistischer Methoden berechnet, welcher zur Stilisierung der genannten Netzwerke eingesetzt weden kann. Dies erlaubt die Darstellung und den übergeordneten Vergleich der Beteiligungen verschiedener Proteine an den Reparaturmechanismen einzelner DNS Schäden. Abschließend wird die Relevanz des Projektes präsentiert, zusammengefasst und diskutiert und Anwendungs- sowie Verbesserungsmöglichkeiten beschrieben.

Abstract

The following thesis presents the optimization and validation of a web-based application for the analysis and visualization of mass spectrometry-based data of DNA repair mechanisms. The role of DNA replication and specific lesions as well as their individual repair pathways is described and the relationship between proteins involved in these processes are visualized using neighborhood similarity networks. Additionally it is shown how these networks can be used to mine individual data sets for the role of each included protein in DNA replication and the repair of DNA lesions. Already established cluster analysis tools were implemented or optimized and their ability to identify functional modules is elaborated on the example of known repair complexes and their interactors. To elucidate how these algorithms perform, reference networks based on computed neighborhood similarities for each mechanism of DNA lesion repair are shown as well as subnetworks created using the mentioned clustering methods. An enrichment score for each protein dependent on the DNA lesion present on template DNA was calculated using advanced statistical methods that can be used to style the reference networks. It visualizes the involvement of individual proteins in each repair pathway to allow smooth comparison of factors between different repair networks. In conclusion the relevance is presented, summarised and discussed and an outlook on possible applications and improvements is given.

Contents

List of Figures

1 Introduction

1.1 DNA replication, damage and repair: An overview

All cells experience a large number of DNA modification events per day that have to be processed and repaired in order for the cell to replicate its genome with as few errors as possible to guarantee proper function of daughter cells. Though mutations are not favoured on a large scale, lower levels of mutations in the genome of differentiated cells can have useful effects on daughter cells such as gains of protein functions to adapt better to the environment. Additionally, cell specialization relies in its core on retaining preferable mutations in a cell population or a tissue.
Many of those modifications occur spontaneously but some DNA lesions are caused by endogenous or environmental factors such as reactive oxygen species, ultraviolet light, ionizing radiation as well as chemical agents and drugs (Loeb, Monnat, 2008). Reactive oxygen species in particular are introduced to the cell by metabolic processes such as respiration or from exogenous sources such as smoking and the uptake of reactive oxygen species with nutrients.

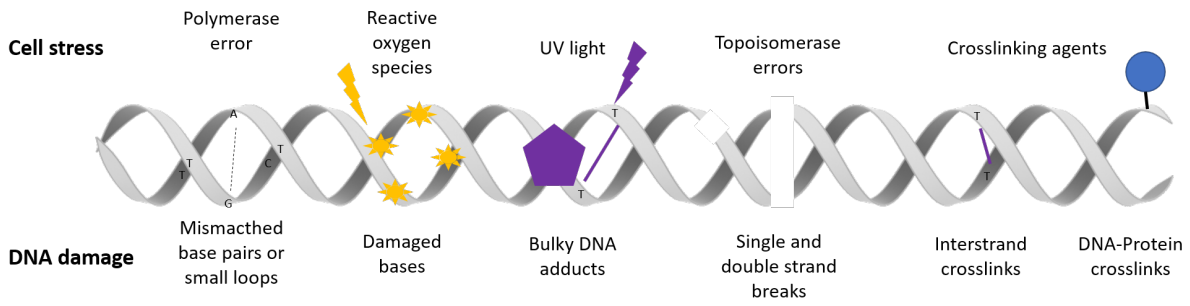

Figure 1: Overview of DNA damage effectors and their main lesion. Errors caused by DNA polymerases are limited to mismatched bases or small insertions or deletions whereas reactive oxygen species cause damaged bases or single strand breaks. UV light often leads to DNA adducts of larger sizes. Crosslinking agents such as Psoralen can introduce links between the DNA strands (Interstrand Crosslink) or between a DNA strand and a bound Protein (DNA-Protein Crosslinks). Modified from Massey, Jones (2018).

Some chemical DNA damage inducers such as Mitomycin C or malphalan are used as cancer therapeutics due to their alkylating properties. Additionally, crosslinking agents such as Psoralen, introduce covalent crosslinks between DNA strands (interstrand crosslinks, ICL) (Ciccia, Elledge, 2010) wherease other crosslinkers such as formaldehyde can form DNA-Protein-Crosslinks (DPC). Formaldehyde can be a byproduct of

DNA methylation and therefore, DPCs are relatively common DNA lesions. Inhibition of topoisomerases can also cause enzymatically derived DPCs by covalently linking the topoisomerases to the DNA.
The inability to repair those alterations leads to pathological phenotypes such as senescence, cancer and cell death. Additionally, accumulation of damaged DNA destabilizes genomic integrity. To prevent an accumulation of lesions they are individually detected and repaired via a specific pathway using a multitude of proteins with varying degrees of specificity. Though many proteins are involved in the repair of different lesion types, each lesion type itself shows a specific footprint over the course of its detection and repair. Some DNA damages are detected during replication and immediately repaired where as others are detected via replication independent factors that induce on-the-fly repair mechanisms (Klages-Mundt, Li, 2017).

As mentioned, changes in the blueprint of life are a key feature of evolution that facilitate diversity but it is crucial to keep those changes at equilibrium with the perpetuation of genetic information. In a multicellular organisms it is possible to transmit genomic instability introduced by mutations to a new generation of cells (gametic cells) and structural components of the organism (somatic cells). It has been shown in humans that genomic instability in somatic cells is closely related to cancer. The importance of DNA repair mechanisms can be demonstrated in a number of human syndromes that are characterized as being caused by defective DNA repair processes. One of those syndromes is xeroderma pigmentosum (XP) which is described as a defect in nucleotide excision repair (see 1.3.1; NER). NER normally removes lesions that are caused by UV exposure which in XP patients leads to a high frequency of tumors in exposed areas of the skin as well as - in more severe cases - developmental problems, neurodegeneration and premature aging. The more severe symptoms are also associated with Cockayne syndrome that is also characterized as a defect in NER (Menck, Munford, 2014). Ataxia telangiectasia or Louis-Bar-Syndrome shows similar symptoms with the addition of motor impairments and an immunological imbalance but is characterized by a mutation in the *ATM*-gene which codes for a protein kinase that is involved in double strand break (DSB) repair. ATM is expressed ubiquitously but seems to be especially involved in the maintenance of Purkinje cells in the cerebellum and other structures of the human brain. Other human syndromes are caused by more complex mutation patterns, one of them being Fanconi anemia (FA). FA is caused by mutations in at least one of 15 genes involved in the repair of interstrand crosslinks (ICL) that lead to developmental imbalances such as malformed appendages as well

as skin discolorations and - in serious cases - bone marrow failure.

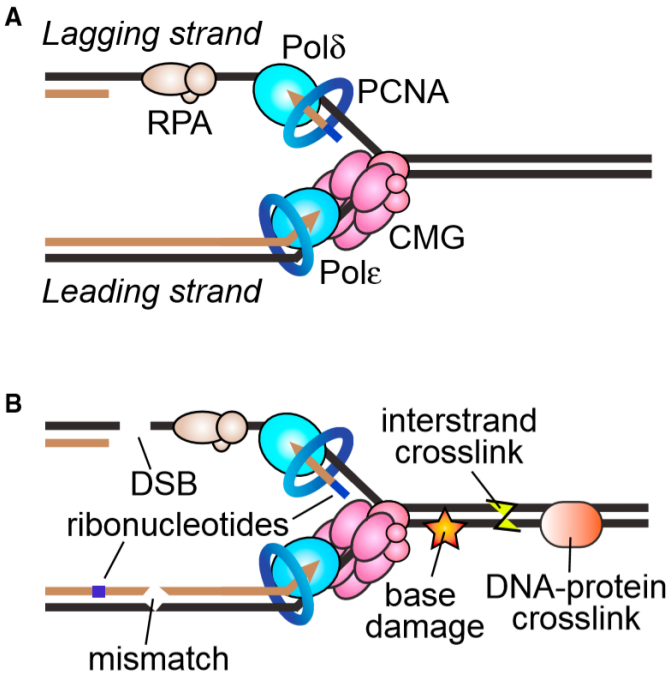

Figure 2: Resolution of DNA Lesions via Replication-Coupled Pathways.
A) Simplified schematic of a eukaryotic replication fork. B) Lesions generated or encountered by the replisome (Cortez, 2019).

All of the syndromes briefly introduced above share an elevated risk for cancer though based on the inability to repair one or more types of DNA lesions which leads to genomic instability and/or uncontrolled growth of affected cells. Understanding the mechanisms of the repair of DNA lesions as well as the pathways regulating them is crucial for preventing cancer and treating syndromes such as the ones mentioned above. Though numerous studies have been published over the last decades that investigated and resolved the mechanisms behind the repair of individual DNA lesions there is no comprehensive resource available to researchers that focuses on DNA repair on a systemic level. Many proteins involved in the repair of DSBs are also associated with the repair of ICLs, even though lesion detection as well as regulation and recruitment of lesion specific proteins differs greatly between those mechanisms. Due to this we wanted to look at the interactions between each repair mechanism to resolve DNA repair as a system instead of a collection of individual pathways. To do so we applied graph theory to a collection of chromatin proteomic data investigating the recruitment of proteins involved in DNA repair and replication in *Xenopus* egg extracts (see 1.4) to specifically damaged DNA templates. Each of the sets used in this thesis has been - be it already published or still in progress - used to resolve or improve the understanding of a single repair mechanism but has never been comprehensively compared to experiments investigating other repair mechanisms. As will be explained in more detail in section 1.4 the *Xenopus* egg extract systems have in the past only

been used to investigate mechanisms that are dependent on replication of the DNA template used. Due to this most of this thesis will focus on replication coupled DNA repair mechanisms (see Figure 2).

1.2 DNA replication and the cell cycle

To understand how DNA lesions are repaired in a replication dependent manner, one has to understand DNA replication itself. Eukaryotic cells undergo different stages during their life cycle to ensure an efficient and error-free cell division that leads to the formation of two identical and functioning daughter cells. A key mechanism during this cycle is the replication of the genome and therefore the error-free synthesis of new DNA molecules based on a template. Before we go into detail about how replication is initiated and carried out, we want to give a short introduction to the cell cycle of eukaryotic cells and the differences between model organisms and humans.

1.2.1 The cell cycle: A brief overview

The circular process of cell life and division is differently regulated for each type of differentiated cell, where some continue to divide indefinitely and others differentiate to the point of being fixed in one phase until cell death is induced. Figure 3 shows a schematic of the stages of the cell cycle. It consists of two major phases - M-Phase and Interphase - where the latter can be split into three main stages that make up the bulk of the cells life (see figure 3). Simplified, Interphase starts with a stage called G_1 where all DNA is organised in a condensed form called heterochromatin that has to be unpacked to be replicated during the following S-Phase. During S-Phase DNA is replicated and the replicated molecules are attached to each other on their centromers to form so called sister-chromatids that are stored in the nucleus during the following G_2 phase. Differentiated cells that are not supposed to undergo mitosis anymore then enter G_0 until signalled otherwise or until cell death (not pictured in Figure 3). Cells that are supposed to continue in cell division enter M-Phase which consists of Mitosis and Cytokinesis. During Mitosis, the nucleus of a cell is dissolved, the sister-chromatids are separated and transported to different poles of the cell. From there, two new daughter nuclei form to complete Mitosis.

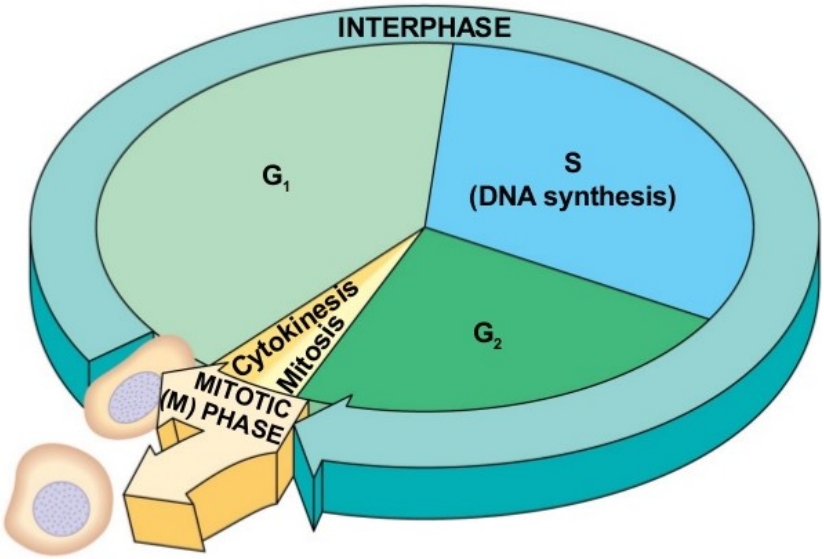

Figure 3: Schematic view of the cell cycle. Cells experience five distinct phases during their life cycle. In G_1 the cell prepares to replicate its DNA in S-Phase and remains in G_2 until it has to divide its nucleus during mitosis. The cell cycle is completed after one maternal cell divides into two daughter cells during cytokinesis (Campbell, Reece, 2015).

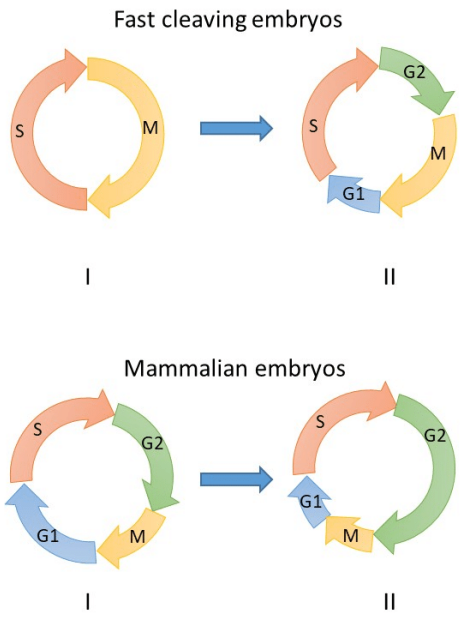

Figure 4: Different course of fast cleaving VS. mammalian embryonic cell cycles.(Kermi et al., 2019)

Cytokinesis completes the cycle by dividing the cytoplasm of the mother cell into two newly generated daughter cells that enter Interphase. These Mechanisms are of course heavily regulated with molecular checkpoints that have to be cleared before entering a new phase. Additionally, regulatory proteins are deployed during S-Phase that inhibit DNA replication to prevent the cell from replicating its genome more than once. Such regulatory proteins can be used to induce stress in a cell or another system used to study replication and DNA repair (see section 1.4).

Embryonic cell cycles show differences in the presence of phases and proteins in comparison to the cell cycle in adults as well as differences between classes of eukaryotes. Fast cleaving embryos such as *Xenopus laevis* don't undergo G_1 and G_2 during Interphase in their early stages but show relatively short G-phases and a longer M-Phase than mammalian embryos that start with

all four phases and only change the duration of each phase during embryonic development (see Figure 4).
All data used in this thesis was acquired using *Xenopus* cell free egg extracts that can be used to study replication dependent mechanisms of DNA repair. Different extract types are used to study different mechanisms but the details of said extracts will be discussed later.

1.2.2 Initiation of DNA replication

The replication of DNA based on the template genome of a cell during S-Phase starts with the recruitment of the Origin Recognition Complex (ORC) consisting of six subunits to an Origin of Replication. Per DNA molecule, many origins are loaded with ORC but not all of them are activated. The mechanisms of choosing which origin stays dormant and which will get activated as well as the function of dormant origins is extremely complex and beyond the scope of this thesis. It has been shown that the dormant origins are involved in the maintenance of genomic stability (Alver et al., 2014).

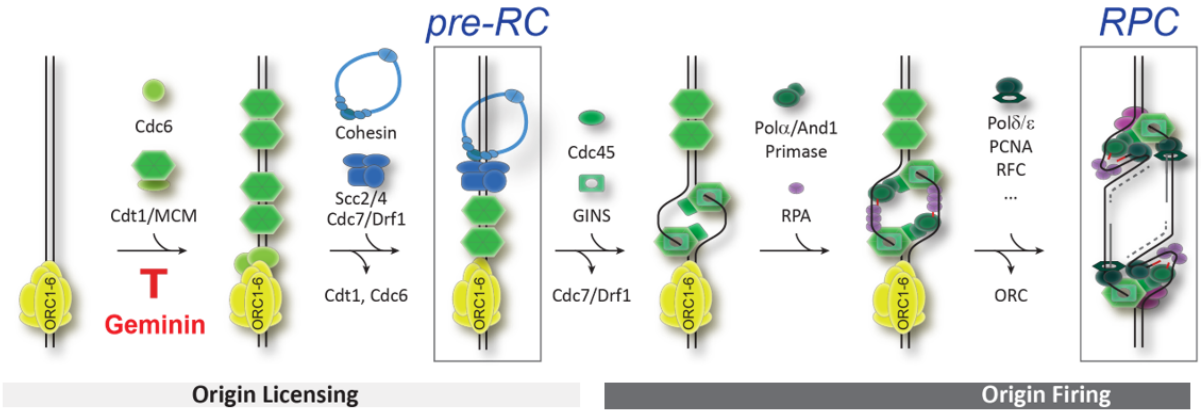

Figure 5: Mechanisms of Replication Initiation. The origin recognition complex (ORC) binds to an origin of replication and recruits Cdc6 and the MCM-helicase that is associated with Cdt1. Cdt1 and Cdc6 dissociate from the origin to give way for the recruitment of the Scc2/4 Cdc7/Drf1 tetramer and Cohesins to complete the pre-Replication Complex (pre-RC). The Cdc7/Drf1 dimer is released and Cdc45 in conjunction with the GINS complex is recruited. With this, the origin is prepared for DNA replication to be initiated. RPA, Primase and Polαare then recruited and after release of the ORC other polymerases as well as PCNA are recruited to form the replication complex (RPC) and start replication (Räschle et al., 2015).

Activated origins are then loaded with Cdc6 as well as the Cdt1/MCM complex. After dissociation of Cdt1 and Cdc6 the MCM helicase complex promotes recruitment of Cohesins and the Cohesin-recruiter proteins Cdc7 and Drf1 to form the pre-Replication

Complex (pre-RC). The Cdc7/Drf1 dimer is replaced by Cdc45 and the GINS complex that bind to and activate the MCM helicase to open up the replication fork. DNA polymerases and the primimg protein Primase as well as the ssDNA stabilizing protein RPA are recruited to the now opened replication fork. To finalize the initiation of DNA replication PCNA is loaded and ORC dissociates from the origin. From there on each DNA molecule is replicated with an extremely low error rate until a potential DNA lesion is reached. each type of DNA lesion is detected by different factors and repair using specific molecular mechanisms that will be explained in the section below.

1.3 An Introduction to the DNA Damage Response (DDR)

Quick and accurate replication of DNA in each cell during every cell cycle is crucial for the survival of the cell. Even though the replisome is a very accurate machinery, it does not function error-free.

DNA replication is enabled by the use of three main polymerases: Polα, which incorporates priming nucleotides, Polδ, which polymerizes the leading strand, and Polε, which polymerizes the lagging strand. These three polymerases are listed in order of descending error-proneness where Polε shows the highest fidelity with only 1 false base incorporations out of 10^6 bases.

1.3.1 Resolution of polymerase blockages

Base lesions such as abasic sites, base oxidation and base methylation often pause polymerases in their path. Spontaneous base loss and glycosylase removal of uracil cause about 10,000 - 20,000 abasic sites each day in each human cell whereas all other base lesions form about 20,000 times each day on average. Environmental factors such as UV light can increase the rate of such base lesions dramatically and even though most of them are repaired by replication independent mechanisms such as Base Excision Repair (BER) or Nucleotide Excision Repair (NER), these systems can not always repair them prior to replication fork arrival.
The response to a base lesion heavily depends on the template strand the lesion is found in. A lagging strand lesion may stall Polα-primase or Polδ but can not stall the replication fork in total because new primers on the lagging strand lead to a natural skippage of the lesion. The gaps created during those skipping maneuvers are repaired after replication by translesion-bypass polymerases. This lesion-skipping model has not yet been verified in vertebrates because of the difficulty of introducing lagging-strand directed lesions (Goodman, Woodgate, 2013). Lesions on the leading strand however are block Polα more effectively because of the lack of new primers on the strand which leads to a functional uncoupling of the CMG helicase and the DNA synthesis. This uncoupling by itself may not be as problematic as one might expect. It has been shown in *E. coli* that there is little coordination between leading and lagging strand synthesis with or without DNA damage. Additionally, helicase activity is reduced by 80% when DNA synthesis stops. This effectively forms a ”dead man’s switch” that prevents the helicase from unwinding too much DNA ahead of synthesis (Goodman, Woodgate, 2013; Marians, 2018). In case of the helicase running too far

ahead, repriming of the leading strand occurs within several minutes and does not require a specific primase in *E. coli* which allows replication to resume. The bacterial replisome is capable of skipping leading strand lesions relatively easily whereas the eukaryotic replisome has the ability to skip them to some extend but it seems to prefer usage of a specialized primase called *PrimPol* (Rechkoblit et al., 2016). An overview of the responses to leading- and lagging strand base damages can be seen in Fig. 6. Fork reversal has been described as an alternative to lesion skipping. This mechanism is catalyzed by enzymes including SMARCAL1 (also called HARP), ZRANB3 and HLTF (Bétous et al., 2012; Kile et al., 2015). Fork reversal happens by migrating the three-way fork junction backwards to anneal nascent DNA strands while forming a so called chicken foot structure (Fig. 6B and Fig. 7).

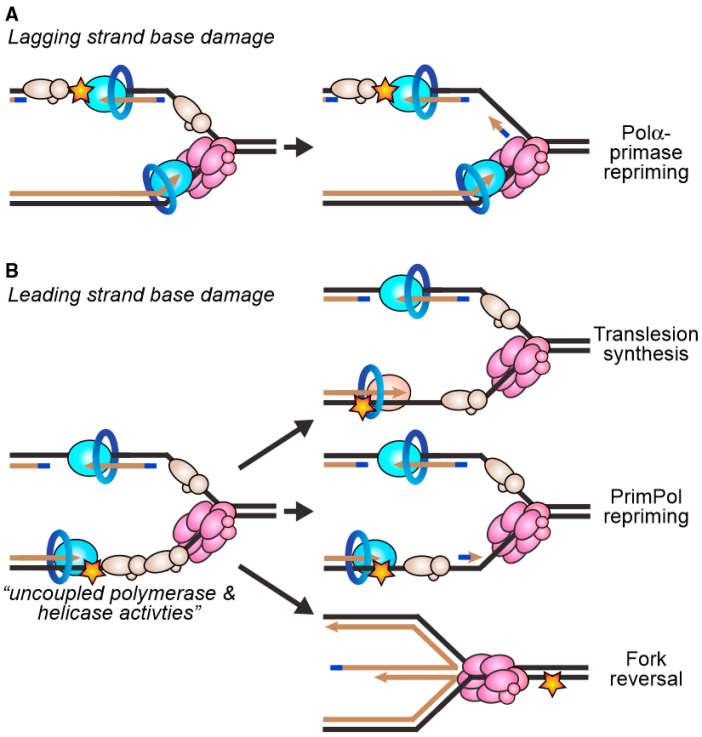

Figure 6: Schematic of the Response to Leading- and Lagging Strand Base Damage.
(A) Lagging strand base damages often create ssDNA gaps that can be repaired by Polα-primase repriming. This does not block DNA elongation.
(B) Leading strand base damages are repaired in one of three ways: Translesion synthesis,*PrimPol* repriming or via Fork Reversal. Those three pathways help the resolution of DNA lesions that could otherwise lead to replisome uncoupling (Cortez, 2019).

Around 25% of all fork defects caused by nucleotide depletion, oxidative base damage, DNA crosslinks and UV photoproducts are resolved by this mechanism (Zellweger et al., 2015). Inactivating one of the mentioned key enzymes changes the cells replication stress sensitivity and alters fork progression while genomic stability is reduced (Ciccia, Elledge, 2010; Yuan et al., 2009). Due to this, fork reversal is now seen as an essential part of effective DNA replication and repair even if the replisome encounters lesion that should be easy to skip. One potential benefit of fork reversal is the placement of template DNA lesions into the context of duplex DNA to enable excision repair mechanisms to function properly. Although there

is little data that suggests a coupling of excision repair to fork reversal, one possibility shown by Weston et al. (2012) is the involvement of ZRANB3 in the removal of DNA lesions due to its endonuclease domain in addition to its fork-remodeling functions. Additionally, fork-reversal is an essential part of repair mechanisms involved in the convergence of two forks on an interstrand crosslink as well as an intermediate in a recombination pathway of fork restart (Amunugama et al., 2018). Recombination proteins such as RAD51, BRCA1 and BRCA2 play crucial roles in fork-reversal pathways. Neelsen, Lopes (2015) proposed that RAD51 binds ssDNA on the template strand to start a coordinated annealing reaction (see Fig. 7). Bhat, Cortez (2018) on the other hand published some evidence on RAD51 capturing nascent ssDNA as its formed by recombination proteins to shift the equilibrium towards fork-reversal.
Reversed forks aren't static constructs but can rather be processed by Double-Strand-Break repair nucleases like MRE11 or DNA2. Processing by those nucleases may remove end-binding proteins and therefore promote fork restart (Teixeira-Silva et al., 2017). It has been shown that the MRE11-RAD50-NBS1 (MRN) nuclease binds to replication forks even in absence of exogenous DNA damage to interact with RPA but end-processing is mostly restricted by RAD51 to prevent nascent strand degradation (Mijic et al., 2017; Dungrawala et al., 2015). The actual actions of RAD51 at forks are heavily regulated by proteins including BRCA1, BRCA2 and the FANC proteins. The mentioned proteins promote RAD51-dependent fork protection but are not essential for fork-reversal. Negative regulation of RAD51 is done through RADX, FBH1 and BLM, indicating that repair mechanisms must not only be activated to act on specific lesion but they also need to be negatively regulated to prevent some effects of over-repair that can actually be more deleterious than the initial lesion (Bhat, Cortez, 2018).
The exact fate of the replication machinery during fork-reversal is not clearly known but there is some evidence for the dissiciation of the replisome (Dungrawala et al., 2015) whereas the CMG helicase most likely remains on the fork. This can be assumed due to the competence of most forks to resume replication quickly after prolonged blockage.

Fork reversal and lesion skipping may be independent mechanisms that could operate individually. How mammalian cells choose between the two pathways is not well studied yet. Post-translational modifications like ubiquitylation and sumoylation, especially to PCNA, are essential in pathway choice. ATR signalling also seems to play a critical role because reversal enzymes like SMARCAL1 are ATR substrates. Phosphorylation of SMARCAL1 by ATR for example has been shown to reduce its fork-remodelling activity (Couch et al., 2013). Mutreja et al. (2018) report an ATR-

promoted stalling of forks that have not encountered any lesions by signalling from stalled forks.How exactly the reversal of undamaged forks benefits the cell even though it delays the completion of DNA replication is unknown.

1.3.2 Resolution of helicase-blocking lesions

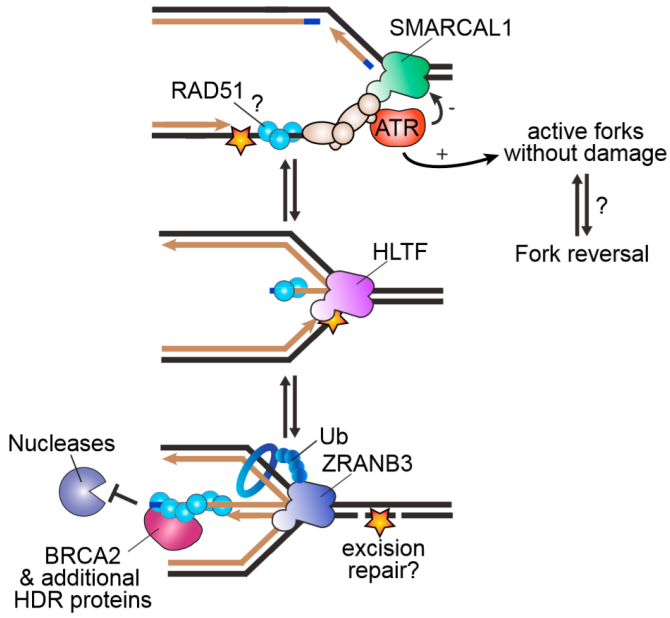

Figure 7: Schematic of Fork-reversal. Through fork reversal the damaged simplex DNA is placed back into a duplex context where the damage can be removed via excision repair. SMARCAL1, HTLF and ZRANB3 are regulated by RPA, 3' DNA ends and poly-ubiquitylated PCNA (see below) respectively (Cortez, 2019).

Interstrand Crosslinks (ICLs) and DNA-Protein Crosslinks (DPCs), especially on the leading strand, pose a potent risk of fork blocking because they interfere with helicase unwinding, although they are much less common than polymerase-blocking lesions.
As already mentioned ICLs are formed in response to DNA-damaging agents that are used in cancer chemotherapy, such as Psoralen and mitimycin C, as well as through ionizing and UV radiation. Many effectors that can cause ICLs also cause DPCs. Additionally, interrupting the catalytic cycle of DNA-binding enzymes can lead to protein-DNA intermediates. Removal of those lesions by replication-coupled repair mechanisms is done through three main pathways. One might expect complete blockage of replication fork movement in response to an ICL or DPC, but that is not the case. The CMG helicase encircles the leading strand and uses MCM10 as an accessory protein that helps to prevent lagging strand DPCs from blocking the helicase (Langston et al., 2017). DPCs on the leading strand, that are to big to physically be accommodated by the CMG channel, can also be traversed by the helicase via unwinding of DNA past the lesion by the accessory polymerase RTEL1 to provide a short ssDNA strand. This allows the MCM complex to bypass the DPC which promotes proteolysis-dependent repair of said lesion (Sparks et al., 2019).

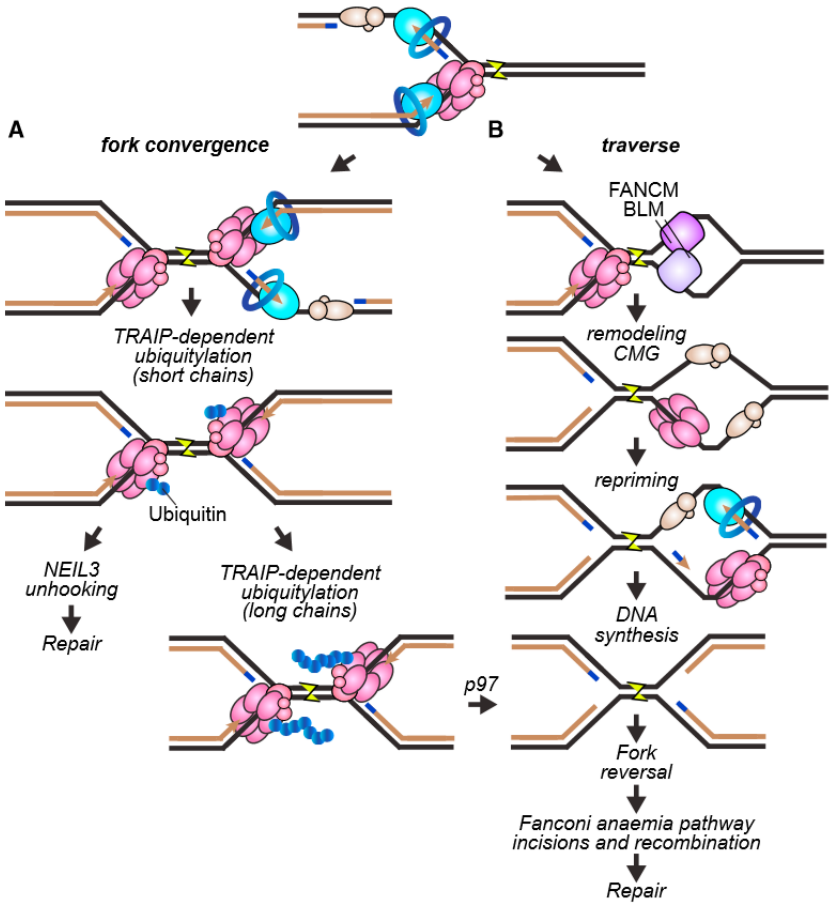

Figure 8: Replication-Coupled ICL Repair Mechanisms.
A) Forks converge on an ICL, TRAIP-dependent short-chain ubiquitylation promotes recruitment of NEIL3 and leads to strand-unhooking. Long-chain ubiquitylation leads to unloading of the replisome followed by fork reversal and DNA incision via the Fanconi anemia pathway.
B) ICL traversion via the use of FANCM and BLM as accessory helicases. Remodeled CMG helicase may allow traversal of the lesion followed by repriming of the MCM complex past the lesion. This leads the ICL to be repaired post-replicatively via the Fanconi anemia pathway (Cortez, 2019).

ICLs also do not block the CMG helicase completely. The two additional helicases FANCM and BLM as well as the ATR kinase and FANCD2 can remodel the CMG helicase (Huang et al., 2013) in a way that allows ”traversal” of interstrand crosslinks, although exact mechanisms for this have not been resolved yet (Huang et al., 2019). The general mechanism of ICL traversal is similar to that of DPC traversal in that the CMG helicase may be blocked from unwinding DNA, but as long as another helicase provides ssDNA the MCM complex can reform past the lesion. Interestingly, ICL traversal can not be observed in *Xenopus* egg extract systems which may be due to the high concentration of replication proteins that could disfavor ICL traversion.

DPC and ICl traversion lead generally to a similar situation to a base lesion which means that polymerases are stalled while the CMG helicase is able to continue unwinding DNA. Therefore repair of said ICL has to occur post-replicatively. A DPC would not be lethal on its own wheras ICLs could interfere with chromosome segregation causing mitosis errors and cell death. Figure 8 shows possible replication-coupled mechanisms for ICL repair.

The two described possibilities for ICL handling during replication both end in X-shaped DNA structures around the crosslinks (see Fig. 8). Unhooking of the crosslink is essential to complete replication and chromosome segregation and takes place through at least two independet pathways.

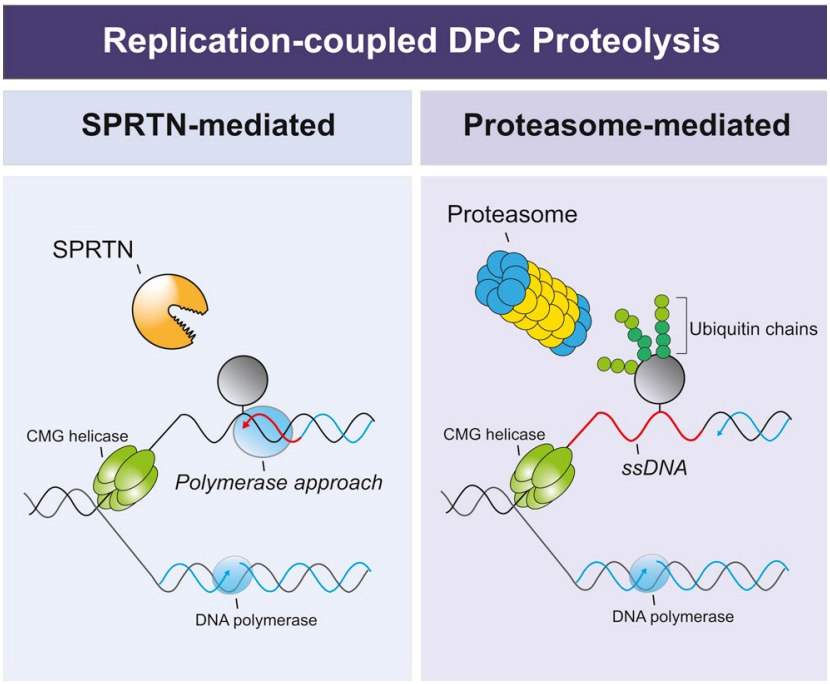

Figure 9: SPRTN- and Proteasome-mediated DPC Proteolysis. SPRTN recognizes polymerase stalling on both the leading and the lagging strand and mediates DPC proteolysis similar to the yeast protease Wss1 (Stingele, Jentsch, 2015). Proteasome-mediated proteolysis requires polyubiquitylation of the DPC (Larsen et al., 2019).

Psoralen- and abasic site induced ICLs are mostly unhooked without DNA incision through NEIL3 (Semlow et al., 2016). Unhooking via the NEIL3 glycosylase leaves an abasic site on one template strand and a mono-adduct or adenosine on the other. Those lesions are then repaired via TLS and some additional steps (Räschle et al., 2015). Lesions introduces via other effectors are not good substrates for NEIL3 and have to be

repaired via an alternative pathway involving a FANC protein-dependent mechanism. The FANC proteins get their name from the disorder that arises when one or more are inactivated called Fanconi anemia, that is characterized by cellular hypersensitivity to crosslinking agents (Ceccaldi et al., 2016). This pathway requires the CMG helicase to be unloaded, the DNA backbone to be incised as well as the involvement of fork reversal (Amunugama et al., 2018). Initially it was thought that BRCA1 was required for CMG helicase unloading but new evidence suggests that TRAIP catalyzes CMG ubiquitylation and therefore promotes p97-dependent unloading Wu et al. (2019).
DPC repair also requires further processing after lesion-traversion but because large DPCs block NER, at least two proteolysis-dependent mechanisms are required to resolve such lesions. Both of those mechanisms rely on SPRTN or the proteaseome (Stingele et al., 2016; Vaz et al., 2016).
Ubiquitylation of DPCs catalyzed by the E3 ligase TRAIP is necessary to trigger proteasome activation. Sumoylation also regulates DPC repair but it is not well understood how nearby DNA-binding proteins are protected from proteasomal removal. Once proteasomal degradation is finished, the remaining small-peptide-crosslink can be excised by NER after the gap in the daughter strand is closed by a combination of Polδ and TLS polymerases. Because of this DPC repair via the proteasome is considered to be mutagenic.
An additional mechanism of DPC proteolysis has been described to be mediated by the metalloprotease SPRTN in eukaryotes (Gao et al., 2018). SPRTN has been shown to act in a similar way to the yeast protease Wss1 (Stingele, Jentsch, 2015) meaning it recognizes polymerase stalling on both the leading and the lagging strand without the need for polyubiquitylation of the DPC the fork is stalling at (Larsen et al., 2019).
There is a possibility of topoisomerase proteins not completing their catalytic cicle which leads to them staying covalently bound to the DNA. Drugs like etoposide are considered topoisomerase poisons and greatly increase the frequency of such covalent bounds occurring. Pommier et al. (2014) showed that those topoisomerase-DNA complexes are removed by tyrosyl-DNA phosphodiesterases (TDP1 or TDP2) that are regulated via ubiquitylation and sumoylation.

1.3.3 Doublestrand break (DSB) repair during replication

Breaks in the DNA backbone are generated via multiple mechanisms during replication where any process that involves strand cutting can introduce DSBs. They can also be caused by structure-specific nucleases such as MUS81 that process persistently stalled forks (Hanada et al., 2007).
In stark contrast to DSBs that form in cells in G_0 or G_1 phase, replication-associated breaks are mostly ”single ended” in nature, meaning that recombination is the preferred mechanisms to resolve them. Breaks in the DNA strands caused during replication can be repaired using a process called Break Induced Replication (BIR) where a DNA resection followed by strand invasion can generate a fork structure that is capable of DNA synthesis using a modified replisome (see Figure 10).

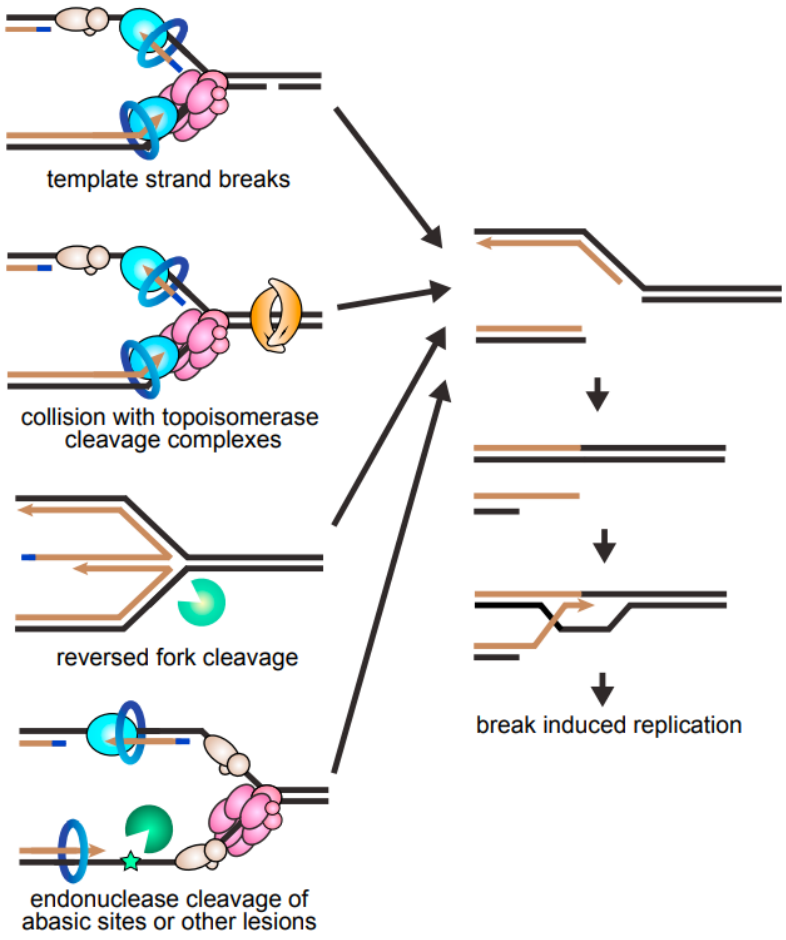

Figure 10: Mechanisms of DSB generation at Replication Forks. Single-ended breaks are repaired by break-induced replication. The individual mechanisms depend heavily on an alternative replisome (Cortez, 2019).

BIR in general is best understood in yeast systems but has been studied in every eukaryotic organisms as well as in model system such as *Xenopus* egg extracts. As a mechanisms involving strand invasion RAD51 plays a major role in BIR as well as an additional subunit of Polδ called POLD3 in humans or Pol32 in yeast. RAD51-dependent BIR mechanisms always involve RAD52 and are engaged at extraordinarily fragile sites. One example of this is when during replication MUS81 cleaves unreplicated DNA during mitosis. This underlines the idea that BIR is a mechanisms to prepare for mitosis and to facilitate chromosome segregation (Bhowmick et al., 2016). Finally, BIR can be useful for telomere extension in cells using ALT where telomere damage induces a BIR replisome to replicate telomere sequences (Dilley et al., 2016).

1.3.4 Mismatch repair

Misincorporation errors are mainly repaired by mismatch repair (MMR). Base mismatches can by definition only occur on the daughter strand. Therefore it is necessary to reliably distinguish the template from the daughter strand. In *Escherichia coli* the template strand has a specific methylation pattern in the palindromic repeat sequence d[GATC] This enables excision and replacement of mismatched nucleotides directed by the detection of a nick on the daughter strand caused by an error in this pattern (Kunkel, Erie, 2005).
Similarly in eukaryotes, specifically in cell free extracts, mismatch repair is directed by a nick located 3' or 5' to the mismatch. The rate of mismatch repair is inverse proportional to the distance between the nick and the error from 125 to 1000 bp although repair was shown at much larger distances (Iyer et al., 2006).
The actual repair of mismatched nucleotides is mediated by the redundant heterodimer MSH6-MSH3 that detects smaller insertion/deletion loops (IDLs) of 1-2 bp in cooperation with their obligate partner MSH2 whereas larger IDLs are detected by MSH2-MSH3. In addition to that MSH6-MSH3 interact with the DNA polymerase processivity factor proliferating cell nuclear antigen (PCNA) to localize MMR heterodimers to replication forks to repair replication errors as they occur. When MSH6-MSH3 encounters a DNA mismatch it undergoes a conformational change to a sliding clamp that diffuses along the DNA to free up mismatched nucleotides for recognition via MSH2-MSH6. This dimer recruits the heterodimer MLH1-PMS2 which in turn regulates the loading of the exonuclease Exonuclease 1 (Exo1) onto the daughter strand to promote excision of error-containing DNA fragments (Gupta, Heinen, 2019).

1.4 *Xenopus* egg extracts as a model system for chromatin proteomics

Studying proteins bound to chromatin is a crucial part of understanding the systemic and molecular mechanisms of DNA Repair and Replication. Even though systems like *Xenopus* egg extracts have been used for decades to study different aspects of DNA processing using more traditional methods such as Western Blots and Immunostaining, the rise of mass spectrometry gave way to the field of high-throughput chromatin proteomics (Blow, Sleeman, 1990; Cupello et al., 2016; Bönisch et al., 2008). One study that should be mentioned in the context of *Xenopus* egg extracts in mass spectrometry is the large collaborative effort of M. Räschle, J. C. Walter, Z. Storchová and M. Mann published in 2015 (Räschle et al., 2015) that explained how DNA crosslinks are detected and repaired using protein-chromatin interaction networks to resolve protein modules involved in crosslink repair.

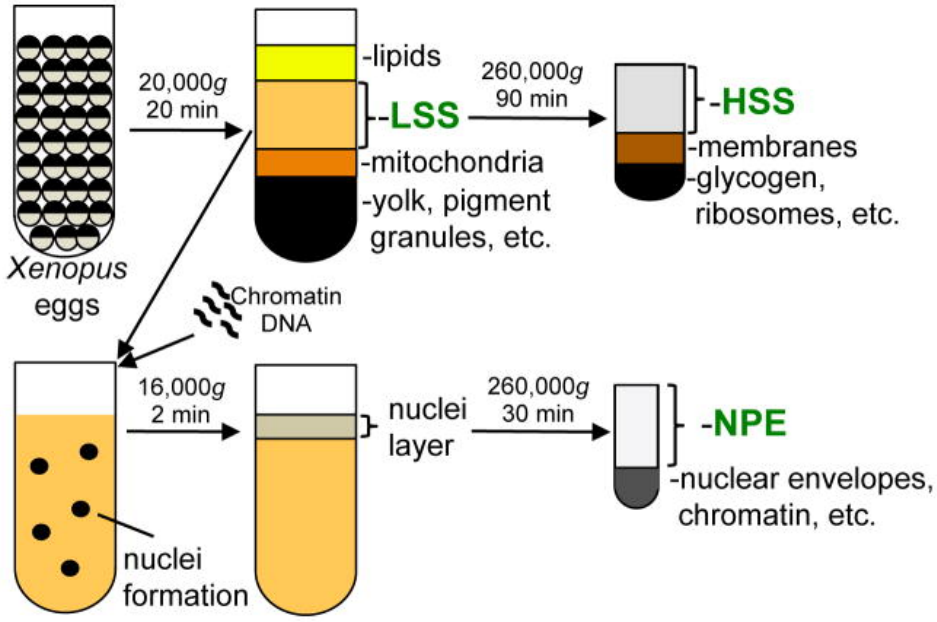

Figure 11: *Xenopus* egg extracts. *Xenopus* eggs are packed and then centrifuged at 20.000g for 20 min to separate lipids, mitochondria and yolk from the Low Speed Supernatant (LSS). Centrifugation of LSS at 260.000g for 90 min isolates the High Speed Supernatant (HSS) from the membranes and ribosomes. The addition of chromatin DNA to LSS leads to the formation of nuclei that can be separated from the rest of the mixture via a short centrifugation at 16.000g. An additional high speed centrifugation of the nuclei layer at 260.000g for 30 min separates the Nucleoplasmic Extract (NPE) from the nuclear envelops and the added chromatin (Cupello et al., 2016).

There are three main ways to prepare cell free extracts from *Xenopus* eggs that are used to study specific aspects of chromatin processing. The easiest extract system to prepare is the so called Low Speed Supernatant (LSS) where eggs are packed and then centrifuged. This separates a yellow-brown mixture of protein, glycogen and membranes, the LSS extract, from the rest of the egg. Centrifuging LSS at high speeds separates the so called High Speed Supernatant (HSS) from membranes, glycogen and ribosomes.

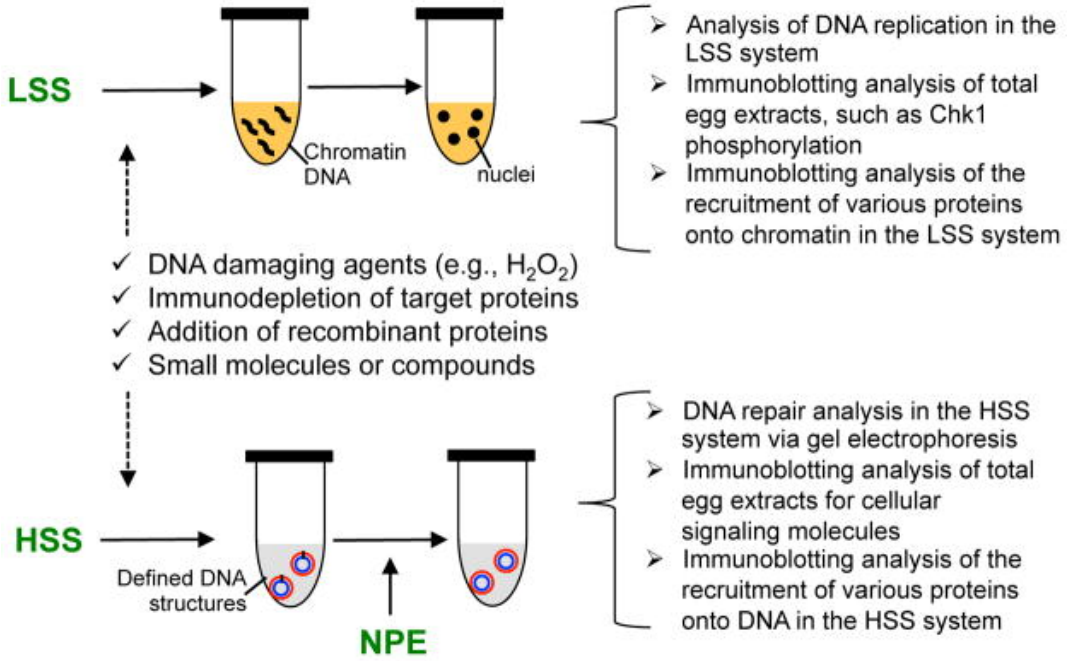

Figure 12: Usecases for *Xenopus* egg extracts. Mixing LSS with chromatin DNA leads to the formation of nuclei that can be used to analyze DNA replication, for immunoblotting of total extracts and for proteomic analysis of protein recruitment to chromatin. Addition of substrate DNA to HSS forms the pre-replication complex (pre-RC) and licenses the added substrate for replication. After adding NPE DNA repair mechanisms can be analysed via gel electrophoresis while it is also possible to investigate protein recruitment and modification via proteomic methods (Cupello et al., 2016).

These extract systems provide reliable tools to scientists studying the mechanisms of DNA repair and replication in a eukaryotic system that is closer to that of humans than fungal systems such as yeast, even though the molecular composition of the *Xenopus* extract systems are not equal to unobstructed cells due to the treatments necessary for preparation. What extract system to choose for the experiment one wants to conduct depends heavily on these molecular differences and will be subject of later parts of this introduction (see Section 1.4.1 and 1.4.2). As an example it should be mentioned that LSS as the simplest extract system is molecularly comparable to the embryonic

Interphase until sperm chromatin is added which initiates replication and therefore moves the extract to S-Phase (see Figure 3.

In studies looking at the replication stress response of egg extracts LSS was treated with the endogenous regulatory protein Geminin which is synthesized during S-Phase to inhibit DNA replication by binding the initiation factor Cdt1 that cooperates with the ORC (origin recognition complex, Cook et al. (2004)) to form the pre-RC complex (see section 1.2.2.

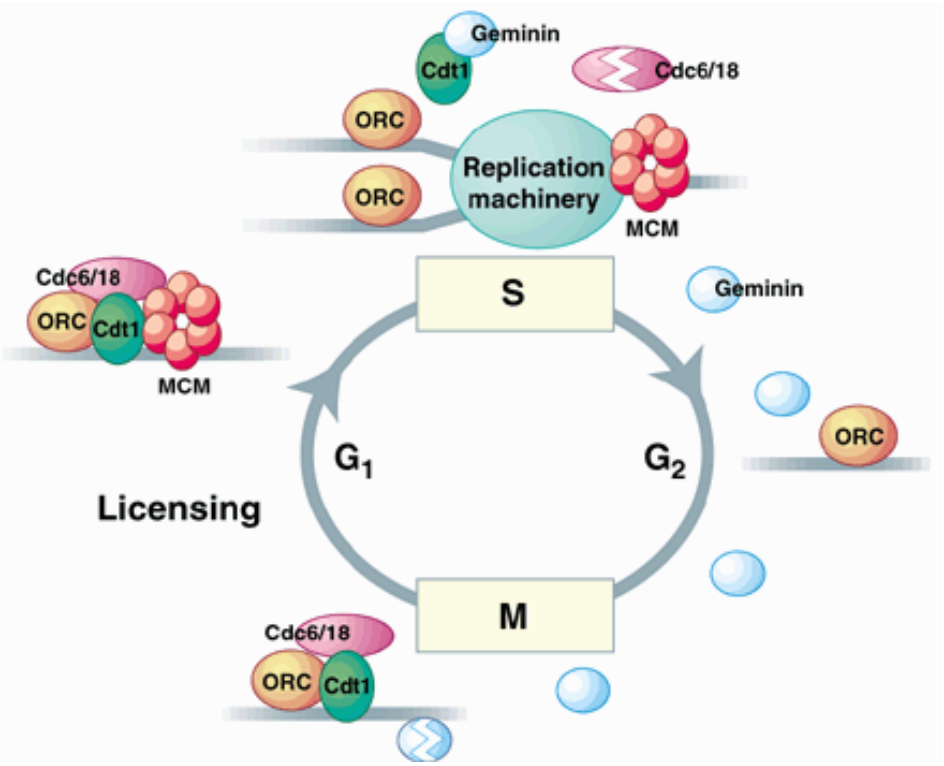

Figure 13: Schematic depiction of a cell cycle showing the action of Geminin. During S-Phase once DNA replication started the regulatory protein Geminin is expressed. Geminin binds to Cdt1 and therefore inhibits the formation of the pre-RC to prevent the replication of already replicating DNA molecules. Geminin expression is kept up throughout G_2 but is decreased in M to allow effective chromatin licensing in G_1 (Lygerou, Nurse, 2000).

Even though HSS and NPE are essentially different extracts, they have to be mixed to study replication-dependent mechanisms due to the fact that HSS mostly contains proteins necessary to form the pre-RC complex and NPE does not. Without the addition of HSS to NPE DNA replication can not start because no origin licensing can occur (Lebofsky et al. (2009) and Figure 5). Therefore they can be considered as one system — referenced henceforth as the NPE/HSS system — even though non-replicating extracts

are rising in popularity due to the possibility to study replication-independent repair mechanisms as well as interaction of different ATPases with DNA (J.C. Walter & M. Räschle).

1.4.1 CHROmatin MASS Spectrometry

The CHROMASS (Chromatin Mass Spectrometry) system established by Räschle et al. (2015) uses *Xenopus* egg extracts described in 1.4 to analyse chromatin bound proteins using mass spectrometry. CHROMASS can therefore be used to analyse the time-dependent recruitment of proteins to chromatin under different conditions to characterize repair- and replication mechanisms.

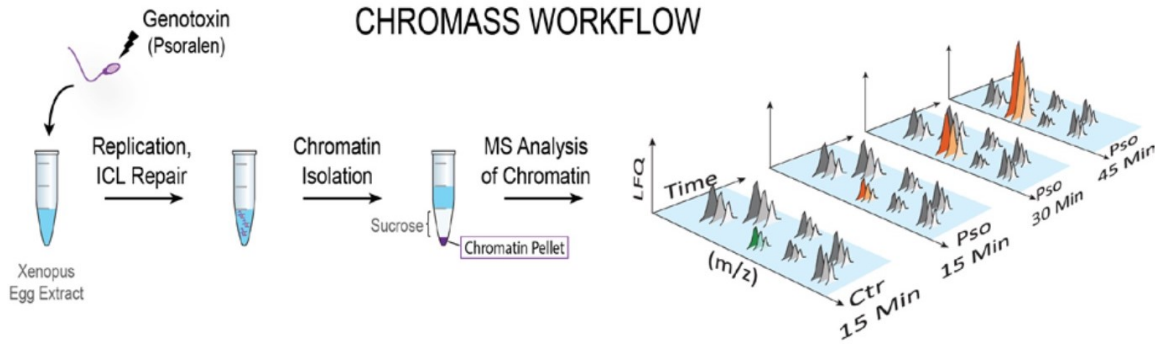

Figure 14: CHROMASS workflow. Genotoxin-treated sperm chromatin is added to *Xenopus* egg extracts to initiate replication and repair. Chromatin is isolated after incubation for 15, 30 and 45 min via centrifugation through a sucrose cushion. The proteins bound to chromatin are isolated, digested, desalted and measured via mass spectrometry (Räschle et al., 2015).

Figure 14 shows the workflow used by Räschle et al. (2015) to characterize the dynamic complex assembly during ICL bypass. Generally, damaged or undamaged sperm chromatin is added to *Xenopus* egg extracts and then kept at room temperature to initiate replication and repair of added DNA. The chromatin is then separated from the mixture after set time points after initial addition by centrifuging through a sucrose cushion. Chromatin-bound proteins are then purified and a tryptic digest followed by a desalting step are performed. The purified peptides are then measured via mass spectrometry. For this kind of study, all of the three extract systems described in section 1.4 can be used to study protein recruitment in a time-dependent manner.

1.4.2 Plasmid-Pulldown Mass Spectrometry

Plasmid-Pulldown Mass Spectrometry (PP-MS) is a system closely related to CHROMASS, although it has been shown that only the NPE/HSS system reliably replicates the plasmid DNA used for this method. Generally, plasmids with defined lesions are added to NPE/HSS and incubated to allow replication and repair. Isolation of this template DNA is achieved via a pull-down system which is enabled by the inclusion of a *lacI*-binding site on the plasmid. The reaction mixture is added to *lacI*-coated beads that bind to the *lac*-operon fragment on the plasmid. The mixture is washed multiple times to remove unbound molecules after which the beads were dried and prepared for mass spectrometry in a similar fashion to the methods described in section 1.4.1.
In addition to the proteomic analysis PP-MS can also be used to process the substrate DNA via gel electrophoresis to visualize replication kinetics (Fig. 15).
This system has recently been used to describe the mechanism of SPRTN- and proteasome-mediated replication-dependent DNA-Protein crosslink repair in *Xenopus* egg extracts (Larsen et al., 2019). The main benefit of PP-MS over CHROMASS is the ability to create specific lesions on the substrate in comparison to mostly unspecific lesion on sperm chromatin treated with genotoxins. Additionally, the repair of the lesion takes place over a shorter time-frame because plasmid DNA has a significantly faster replication kinetic due to its structure and size and its independence of prior chromatin assembly in *Xenopus* egg extracts (Sanchez et al., 1992; Aquiles Sanchez et al., 1995). Using protein binding motifs on the plasmid templates it is also possible to synchronize their replication.

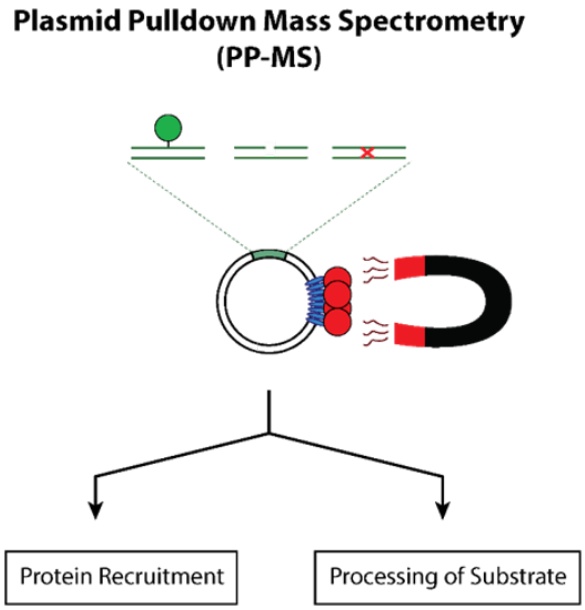

Figure 15: Plasmid-Pulldown Mass Spectrometry workflow. The plasmid substrate contains defined lesions and multiple repeats of a *lacI*-binding sequence. After incubating the reaction mixture for the desired time a defined volume is removed and combined with beads coupled to *lacI* which binds the plasmid substrate and therefore all DNA-bound proteins. The beads are then processed for mass spectrometry or loaded onto gels to visualize substrate processing (modified, Markus Räschle).

1.5 Big Data and Networks in Biological Research

Methods such as the ones described in the sections above create large amounts of data that have to be processed and handled correctly. Big Data, as the aggregate of such large datasets tends to be called, is used to draw conclusions from individual collections of data as well as the entirety of those collections.
Advanced computational methods such as the ones used by NCI in the development of TCGA can help understand biological systems using data gathered by 'omics experiments such as CHROMASS or PP-MS. Identification of key regulatory factors and mechanisms is crucial in resolving those systems and can be directly achieved by for example co-expression analysis in the case of gene expression data. Co-expression analysis algorithms can also be applied to CHROMASS and PP-MS data sets under the assumption that proteins found on chromatin are recruited to it under specific conditions. This enables one to identify functionally similar proteins due to their similar chromatin binding pattern. The similarity of data collected for different proteins in one or more experiments can then be visualized in form of a protein-protein or gene co-expression network.

1.5.1 Network visualization using correlation algorithms

The following section is based on Dytham (2011) unless specified otherwise.

A relatively simple way to tackle co-expression analysis is correlation which is a measure of the relationship between two variables. Pearson's r correlation is the most widely used correlation statistic to measure the degree of relationship between variables that are linearly related. Pearsons correlation coefficient r is calculated using the following equation.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{(\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{1}$$

r_{xy} : Pearson r correlation coefficient between x and y
n : number of observations
x_i : value of x (for ith observation)
y_i : value of y (for ith observation)

Pearson correlation only works based on the assumption that both variables are normally distributed and have a linear relationship. Additionally, the data has to be

homoscedastic, meaning it has to be equally distributed along the regression line. Given these assumptions Pearsons r can be useful to investigate biological mechanisms up to more complex relationships such as protein-chromatin interaction networks.
The calculated correlation matrix for a set of variables can then be visualized in form of a network where the nodes represent variables and the edges their relationship. Additionally it is possible to use grouping algorithms such as Hierarchical Clustering to look at especially highly correlating variables.
To minimize computational effort and to reduce noise in the data set it is necessary to filter the gathered correlation matrix by the coefficient as well as the p-value. Determining the thresholds for filtering is another challenge in itself that can be tackled in different ways.
The simplest method of filtering networks for ease of visualization is setting a fixed correlation threshold. This is known as *hard thresholds* where all edges of the network are filtered in such a way that all edges with $x \geq \tau$ are included. This method produces simple results that come with the danger of losing meaningful information depending on the threshold. If one sets $\tau = 0.7$ it is possible to lose a meaningful edge that has a weight of 0.6985 (Carter et al., 2004).
Another option is filtering for statistical significance of the calculated correlation coefficient based on the assumption that they are normally distributed meaning that the probability density function follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{2}$$

σ : standard deviation
μ : mean

The p-value itself describes the probability of obtaining test results that are at least as extreme as the actually achieved results, meaning that it tells one if an observation could have also been done at random. An observation can be seen as statistically significant if $p < 0.05$ and highly significant if $p < 0.01$, stating that the probability making said observation at random is below 5% or 1% respectively. For Pearson's r this can be done relatively easily because the p-value for this measure of correlation uses the t-distribution. In this case $H_0 : p = 0$ vs. $H_1 : p \neq 0$ where p is the correlation value between two variables.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{3}$$

r : correlation coefficient
n : number of observations

Where the p-value is $2 * P(T > t)$ with T following a t-distribution with $n - 2$ degrees of freedom.
The correlation matrix can then be filtered to include all relationships where $p(x) \geq \rho$ with an established p-value threshold in biological relationships of 0.01.
After filtering the correlation matrix is transformed to an adjacency matrix using a sigmoid function where the correlation or rank interval $[-1, 1]$ is mapped to $[0, 1]$.

$$a_{xy} = signum(s_i, \tau) \equiv \begin{cases} 1 & \text{if } s_{xy} \geq \tau \\ 0 & \text{if } s_{xy} < \tau \end{cases} \tag{4}$$

The two filtering methods work reliably in combination with one another if optimized correctly but can, as already mentioned, lose meaningful information. One way to prevent this is to implement soft thresholds such as the sigmoid function used here.

$$a_{xy} = sigmoid(s_i, \alpha, \tau_0) \equiv \frac{1}{1 + e^{-\alpha(s_{xy} - \tau_0}} \tag{5}$$

1.5.2 Weighted Correlation Network Analysis

Weighted Correlation Network Analysis also known as weighted gene co-expression network analysis (WGCNA) is an established method for the identification of modules and intermodular hubs mostly used in genomic applications (Horvath, 2011).
It was developed by Steve Hovarth and his colleagues at the UCLA Fielding School of Public Health as a means to expand on the previously unweighted methods of network analysis. The benefits of using weighted correlation network analysis over an unweighted analysis start with the use of a sigmoid adjacency function to apply a soft threshold that is data dependent to prevent data loss. Additionally, a topological overlap measure is calculated from a the adjacency matrix that has been shown to more precisely represent gene co-expression. The construction of networks using this method delivers highly robust results if the parameters of the soft threshold are changed meaning that an optimization of the sigmoid adjacency function is not necessary for brief data analysis (Zhang, Horvath, 2005). It can also be used to enhance standard

data-mining methods such as cluster-search since similarity measures can often be transformed to weighted networks (Oldham et al., 2012).

Most recently, WGCNA in conjunction with an optimized adjacency function was used to visualize a co-regulation map of the human proteome with which they could identify the function of novel proteins (Kustatscher et al., 2019). To achieve this goal they combined WGCNA with a tree-based clustering algorithm published by Buttrey, Whitaker (2015) that improves co-regulation analysis by using the Jaccard similarity coefficient as a measure of similarity (Gupta et al., 2018). The Jaccard coefficient is defined as:

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{6}$$

From this coefficient the so called Jaccard metric can be deduced.

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \bigcup B| - |A \bigcap B|}{|A \bigcup B} \tag{7}$$

By applying this combination of methods to a large collection of SILAC experiments they were able to provide an interactive resource for scientist to browse the human protein interactome and to deduce the functions of not well studied proteins by their association using the principle of ”Guilt-by-association” described in Figure 29.

In this study we assumed that the methods applied by Kustatscher et al. (2019) can be used to analyze the regulation of chromatin binding under specific DNA damage conditions to identify novel repair factors or to give mathematical support to experimentally found associations.

1.5.3 Predicting protein functions using networks

The following section is based on Gillis, Pavlidis (2012) unless stated otherwise.

It is mostly thought that gene functions have to be studied in the context of networks. Networks consisting of millions of interactions gathered from RNA coexpression analysis, protein binding assays and other high-throughput methods can be studied using freely available data only have been embedded in a large number of studies pub-

lished by molecular biologists all over the world. Most of these studies combine these networks with codifications of gene function (i.e. Gene Ontology). If the information derived from the network overlaps with the annotation of a gene one might assume that the genes must have a similar function but to understand the actual function of a gene in the whole systemic context one has to look at all interactions of a gene or protein.
Biologists have dealt with this problem by leveraging the ”Guilt-by-association” principle (GBA). GBA states, as mentioned in Figure 29, that it is possible to use a network representation of complex protein relationships to deduce the function of an unknown protein from the ones it correlates with. This can in theory be applied to biological data of any ”level”, meaning it can be used to analyze genomic, transcriptomic, proteomic and metabolomic data (Oliver, 2000).
The performance of GBA in a purely computational applications is commonly assessed using cross-validation where known functions are masked from part of the network followed by measuring the ability to recover the masked information. This is known as the concept of ”Precision-recall”. The area under a Precision-recall curve is useful to evaluate the performance of identification algorithms because it is very sensitive to the effects of a single highly-ranked correct guess. This means that precision-recall rewards methods that provide one good prediction while subsequent errors have much less effect on the evaluation (Gillis, Pavlidis, 2012, S1). In our case we used the ”average precision” (AP) as a metric for GBA performance based on the findings of Gillis, Pavlidis (2012). It is calculated using:

$$AP = \frac{1}{k} \sum_{i=1}^{k} \frac{i}{rank_i} \tag{8}$$

Another option of estimating the performance of grouping algorithms in graph theory is to simply compare computational results with known experimental results. A strong similarity between experimental and computational data given integrity of both data sources suggests a good performance of an algorithm for that specific type of data. For this to be considered a reasonable approach one has to first verify the integrity of the underlying data and check if the sets themselves are comparable.
There are multiple algorithms for the identification of protein modules, two of which will be explained and used in this thesis.

1.6 Clustering algorithms for the identification of functional protein modules

Generally the assumption one has to make before using clustering algorithms on biological networks of any kind is that items in the network that are close to each other are thought to behave similarly. This *closeness* can either be a direct neighborhood or a connection over a low number of nodes between two items. To group individual items of a network together, one has a multitude of options depending on their use case. In the field of chromatin proteomics a network item describes a protein while each edge between two items or nodes describes a measure or co-binding or co-regulation. The simplest method of clustering nodes in biological networks is an algorithm called ”Nearest Neighbor Clustering”. Functionally, it selects all nodes that are directly connected with an input node. While effective, this algorithm has a high false discovery rate due to its reliance on the correct preparation of the network it is used on. When a network contains a lot of edges that are biologically speaking insignificant, the probability of a number of resulting nodes in the cluster being there by chance is high. On the other hand, proteins that are important for the function of another protein but are connected through longer paths to the starting node, they are not included.
An improvement over Nearest Neighbor Clustering are ”Shortest Path” algorithms like Bellman Ford that are used to determine the distance of nodes to an input node - henceforth called ”query”. The resulting path lengths are then ranked to reduce the false positive rate and improve the ability to draw meaningful conclusions from the clustering result (Franke et al., 2006).

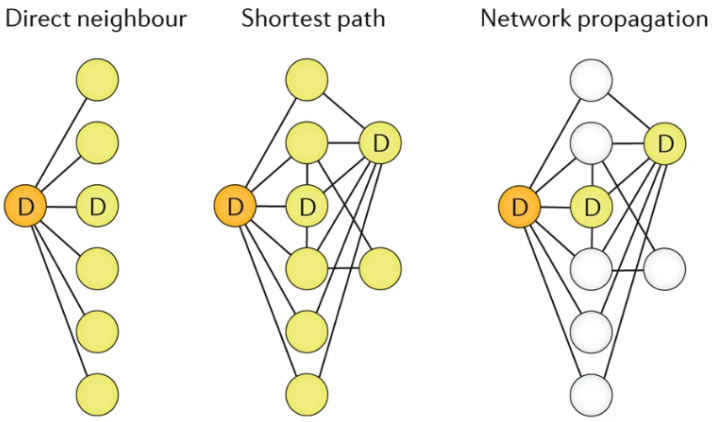

Figure 16: Schematic example of neighborhood clustering, shortest path search and network propagation. A single network node (D, orange) with a known function is used to identify other potential functionally related nodes (D, yellow) (Cowen et al., 2017).

While the two methods described above can yield presentable results for simpler applications they should only be used to quickly analyze more complex data. Predicting relationships and functional interactions in complex biological networks by means of simple mathematical clustering be it scored or not is not feasible due to the high amount of processing that has to be done to the data to decrease the amount of false positives and negatives. Cowen et al. (2017) showed that *Network Propagation* is a better way to mine data using complex biological networks. Functionally, Network Propagation transforms a list of query proteins into a proteome-wide profile of similarly regulated proteins. In our case this method can be applied to find proteins that have a similar chromatin binding pattern under specific damage conditions.

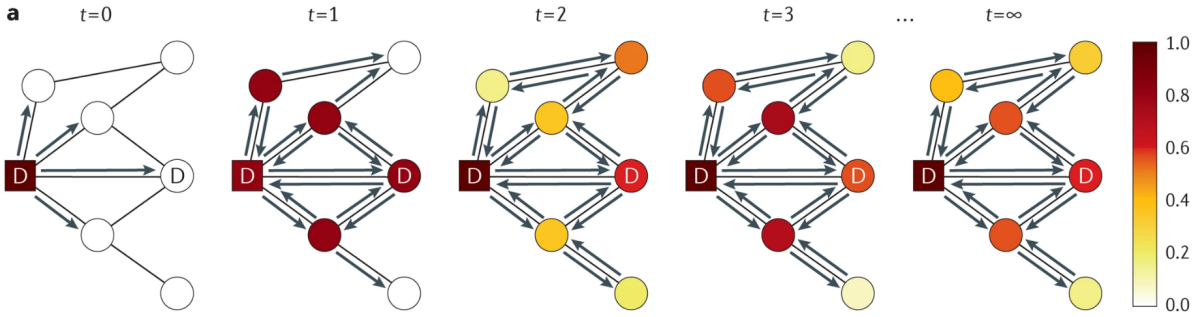

Figure 17: Step-by-step demonstration of network propagation. Propagation process is shown at different time points until a steady state is reached ($t = \infty$) and arrows depict the direction of the walk flow. Nodes are colored according to the amount of flow they receive. D indicates nodes with a known function (square) or with predicted functions (circle) (Cowen et al., 2017).

Network propagation uses a list of proteins with a known function as a query. The have been shown to be capable of identifying previously unknown or predicted disease-associated genes in complex networks ranging from cancer patient data to Alzheimer's research collections (Cowen et al., 2017). One algorithm belonging to this class is the Google PageRank algorithm (Taher Haveliwala et al., 2003). The algorithm gives the query nodes a score of 1.0 and the score of all other nodes to 0.0. It then walks along the edges of the network starting at one of the query nodes and at every step each node diffuses its score to its neighboring nodes where the amount it diffuses is proportional to the weight of the edge in weighted networks. Random walk with restart improves on this by deciding after each step if it teleports back to one of the input nodes. This effectively disperses the score exponentially over the visited nodes. The random walk with restart algorithm has been implemented by Menges (2018) and will be used to mine the DNA Repair Atlas for functional DNA repair modules.

1.7 Dimensionality reduction and its applications in high-throughput data analysis

In addition to analyzing the co-regulation or co-expression of proteins it is often wise to analyze the relationship of the individual experiments that are part of a meta-analysis. For example, a collection of 48 proteomic experiments consisting of about 400 measurements in total with around 5000 proteins identified per measurement can be represented as a matrix with 96,000,000 dimensions. The relationship between the different conditions can teach a data scientist about the integrity of each set and allows, for example, filtering for outliers. In general, dimensionality reduction is used to identify driving components of high-dimensional data sets (i.e. Principle Component Analysis) or to visualize the relationship of individual sets in large collections (i.e. t-Distributed Stochastic Neighbor Embedding) as demonstrated by Kustatscher et al. (2019).

1.7.1 t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction algorithm that has been implemented in different programming languages to be used in multiple data analysis toolkits. It was initially developed as a machine learning algorithm for visualization by L. van der Maaten and G. Hinton (van der Maaten, Hinton, 2008) that is especially useful for visualising relationships between high-dimensional datasets in a low-dimensional space. t-SNE is comprised of two main steps:
A probability distribution over pairs of high-dimensional objects is constructed in such a way that similar objects have a high probability of being selected.
After that similar distributions are constructed for a low-dimensional map and the one with the lowest ”Kullback-Leibler divergence” with respect to the position of map points to the high-dimensional map is selected. Due to simplicity, the mathematical details of the algorithm are not shown here but can be found in the paper linked above.
t-SNE is a useful tool for working with large collections of data because it allows the comparison of individual parts of the collection with one another on a low-dimensional map. This is especially the case if one wants to identify local structures of complex collections i.e. the relationship between different replicates or treatment conditions in one of the included experiment sets while still maintaining information about the global structure of the data collection Kobak, Berens (2019). In this thesis this algorithm was used to see whether the experiment sets included in the atlas can be compared with one another without loosing information about the individual sets themselves.

2 Materials and Methods

All scripts used in this thesis can be found on my Github or under
github.com/BaQBone/DNARepairAtlas.

2.1 Data Collection and Processing

The data used in the creation of this thesis was provided by Dr. Markus Räschle and his collaborators. These data sets have been measured using different mass spectrometers over the last 12 years and has partially been published in the works cited in this thesis. Processing of said data was initially conducted using the analysis tool Perseus developed by the group of Dr. J. Cox of the Max-Planck Institute for Biochemistry in Munich. Further processing and visualization of data as well as the construction of correlation-based networks was conducted using the statistical programming language 'R' under the use of the following packages:

ggplot2	dplyr	data.table
WGCNA	Hmisc	tidyverse
BioCManager	PerseusR	samr

2.2 Computing Enrichment Scores

The raw abundance data grouped based on the experimental conditions they were collected under. A list of comparisons was created that compared treated samples of each experiment with their respective control. This list was used to compute Student's t-Tests for all proteins found in each individual set. The null-hypothesis S_0 was set in a data dependent manner using the R package 'samr'. This applies a modified t-Test with an optimized false discovery rate (FDR) to adjust for multiple testing. 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. These scores were then scaled in an interval of $[0, 1]$ to improve comparability between damage conditions.

2.3 Application of WGCNA to visualize Protein Networks

The R package 'WGCNA' was deployed to construct networks representing behavior similarities of proteins under different damage conditions. Initially, all missing values were replaced by 0 to represent that the protein was not present on chromatin in these

samples. Using the package 'Hmics' the Pearson's correlation coefficient was calculated and a significance test was performed on the correlation coefficients. Relationships with $p > 0.01$ were filtered out and a soft threshold was applied to the resulting correlation matrix (sigmoid adjacency; $\mu = 0.91; \alpha = 37$). From this similarity matrix, the Topological Overlap Measure (TOM) for each relationship was computed using the function *TOMsimilarity* included in 'WGCNA'. The resulting matrix was transformed to a list of relationships using the function *melt* and used to construct protein similarity networks. Network visualization is done using cytoscape.js and the 'cola' layout add-on implementing a physics based force algorithm.

2.4 Development of an interactive Application

An application for visualizing the processed data was developed using the functional programming language F# that uses the library Suave.IO to connect functions written in F# with an HTML and JavaScript based website in an asynchronous and interactive manner. The application was deployed on a publicly accessible virtual machine running Microsoft Windows Server 2018 and the Microsoft Internet Information Services (IIS). This virtual machine has 1 vCPU with access to 4 GB of system memory and a 60 GB virtual storage device. Using IIS, the application is published by routing requests sent to the virtual machine to the port assigned to the Suave.IO interface program. This way, the user interacts with the front-end website and sends requests to receive on-demand results using the functions implemented in the application.
The application can be reached through the following Link using the given login credentials.

http://dnarepairatlas.bio.uni-kl.de
USER: Admin
PASSWORD: TUKLAdmin1118

3 Results

The DNA Repair Atlas (DRA) is an already developed but not yet published web-based resource for the identification of functional DNA repair modules based on quantitative label-free proteomic data collected as part of an ongoing project. It has been in a constant state of development since then (Menges, 2018) and the main goal of this work is to substantially improve its user experience and the feature set. Large parts of this result section are therefore based on optimizations of existing functions as well as improvements to the usability of the administrator. Direct code and runtime comparisons will be omitted to focus on newly implemented features. Most of the functionality implemented has been reused unless stated otherwise. The data used in the DRA used different DNA lesions, inhibitors and extract systems to study the recruitment of DNA repair factors to chromatin under specific conditions in a time-dependent manner by means of the CHROMASS and PP-MS workflows described above. All of the data used in this thesis were already pre-processed and some subsets focusing on individual DNA repair pathways were already published (Räschle et al., 2015; Haahr et al., 2016). While combining such a diverse collection of data to form one comprehensive picture is a challenge in itself, the user-experience has to be considered when developing a resource that is supposed to be used by scientists around the world.

The basic features of the DNA Repair Atlas include the visualization of individual protein raw and normalized abundances over time for each experiment as well as the possibility to check if a protein of interest is significantly enriched in one of the experiments included in the DRA via polar and Volcano plots. The main feature of the DRA are networks that are either heavily pre-filtered and curated to reflect our current understanding of DNA repair or that are mostly unfiltered to allow for the identification of novel interactions of known chromatin-binding proteins under specific DNA lesion conditions using two different clustering algorithms.

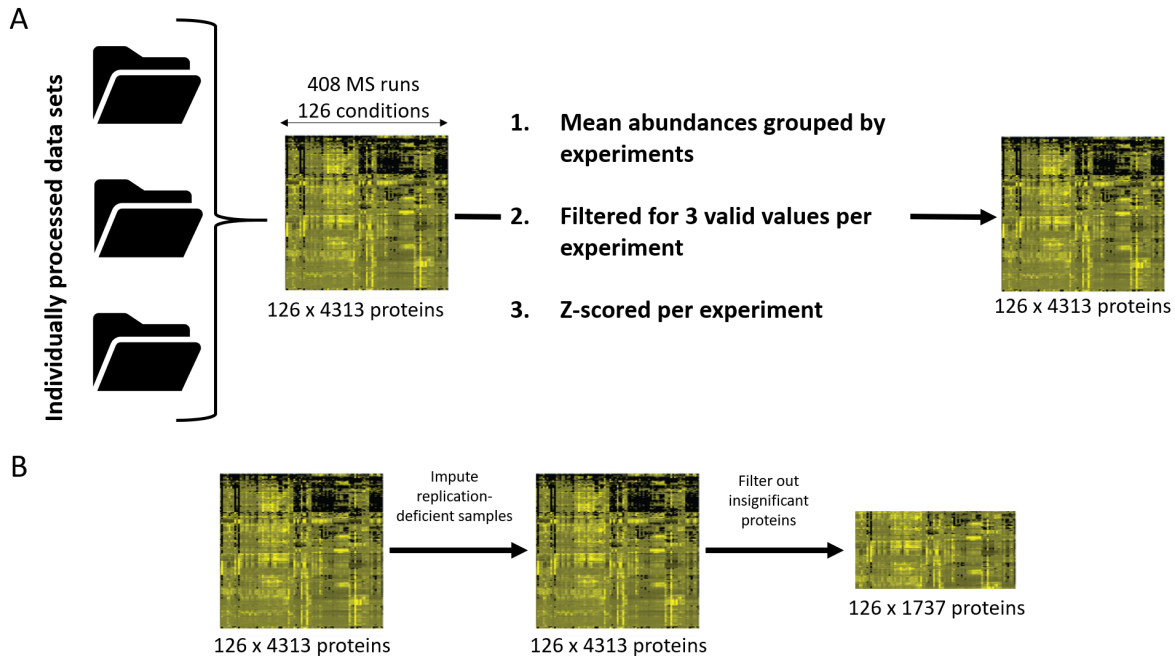

Figure 18: Processing of all datasets for visualization. A) All individually processed data sets were combined into one large matrix containing abundances for 4313 proteins over 127 conditions collected in 408 mass spectrometry runs. The mean abundances per replicate where calculated and used to filter each protein for three valid values. This reduced the matrix to 1272 proteins over all conditions that where then Z-scored per experiment to normalize the abundance variation between them. This matrix was used as the input for all further computational methods. B) To improve performance of correlation and clustering algorithms the normalized matrix was further processed. First missing values in replication-deficient data sets were imputed utilizing a pseudo-random bootstrap method by replacing them with values from a normal distribution calculated from the rest of the abundances measured in the experiment. This imputed matrix was then filtered for proteins found to be significantly enriched under the investigated DNA lesions and already known and annotated DNA repair and replication proteins were added afterwards. The resulting matrix of 127 conditions with 1737 proteins each were used to construct new repair networks using Pearson correlation and Topological Overlap Measures.

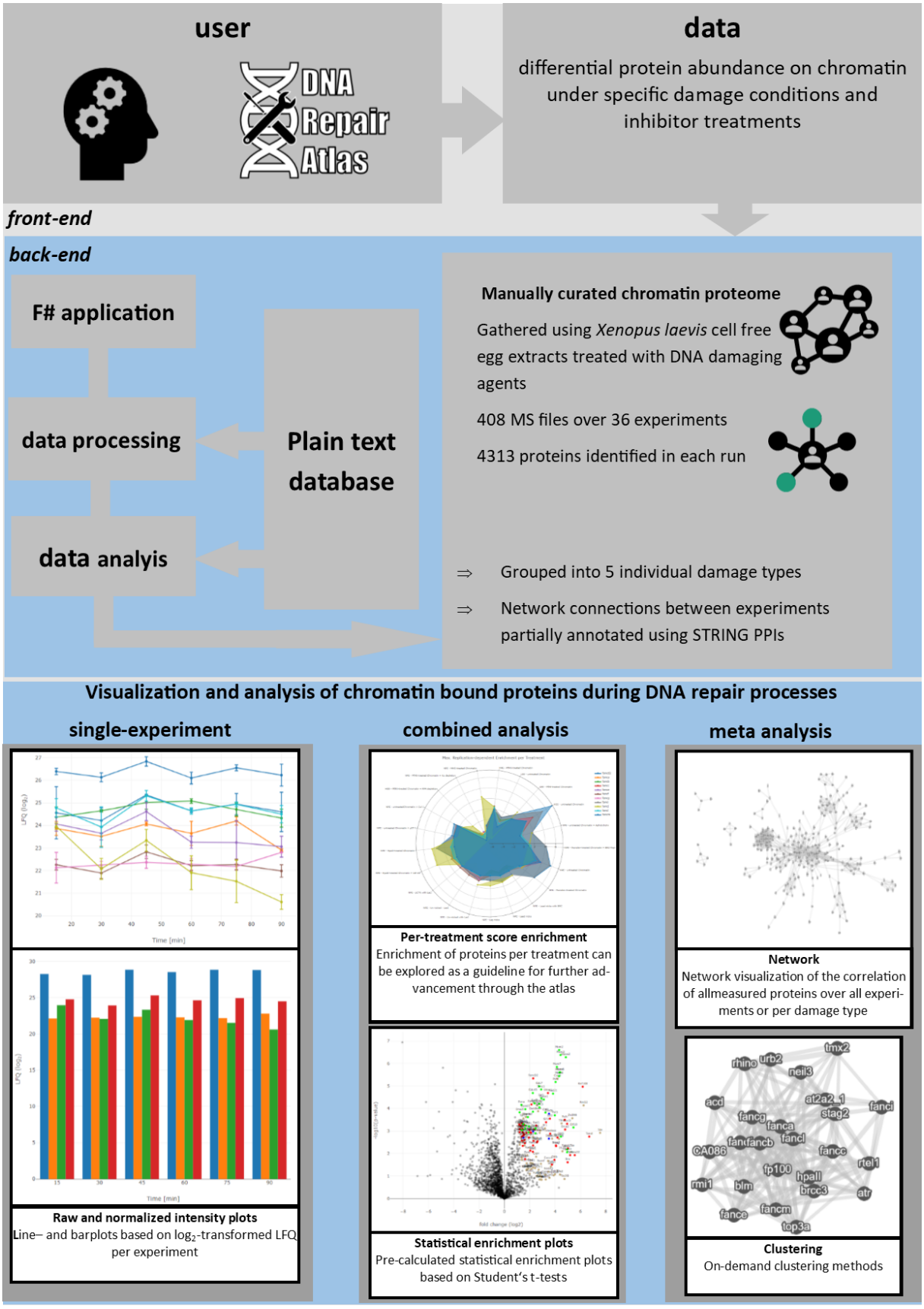
user
DNA Repair Atlas
data
differential protein abundance on chromatin under specific damage conditions and inhibitor treatments
front-end
back-end
F# application
data processing
data analyis
Plain text database
Manually curated chromatin proteome
Gathered using Xenopus laevis cell free egg extracts treated with DNA damaging agents
408 MS files over 36 experiments
4313 proteins identified in each run
⇒ Grouped into 5 individual damage types
⇒ Network connections between experiments partially annotated using STRING PPIs
Visualization and analysis of chromatin bound proteins during DNA repair processes
single-experiment
combined analysis
meta analysis
Raw and normalized intensity plots
Line– and barplots based on log2-transformed LFQ per experiment
Per-treatment score enrichment
Enrichment of proteins per treatment can be explored as a guideline for further advancement through the atlas
Statistical enrichment plots
Pre-calculated statistical enrichment plots based on Student's t-tests
Network
Network visualization of the correlation of allmeasured proteins over all experiments or per damage type
Clustering
On-demand clustering methods

Figure 19: Setup and function of the DNA Repair Atlas. A network representation of protein correlations based on their abundance on chromatin under different damaging conditions represents the central part of the DNA Repair Atlas. The measured *Xenopus* proteins are mapped to the human homologue and annotated with links to protein/gene databases such as UniProt and GeneCard to improve access to the resource for new users. Due to restraints put on the project by the web-deployment method of choice all data needed for the processing and analysis are stored in a plain text database. Data processing and analysis is done on-demand after input from the user using scripts written in the functional programming language F# running the *back-end* of the resource. The user input as well as the information the *back-end* returns are processed using JavaScript scripts at the *front-end*. Several interfaces are implemented into the DNA Repair Atlas that allow easy access to the five main functions of the resource.

The user interacts with an HTML-based front-end deployed on a virtual machine hosted by the RHRK with using Microsoft's Internet Information Service set up as a web-server running our application. This front-end consists of different subpages with each of them focusing on one category of functions mentioned in Figure 19. Connecting this front-end to the back-end are scripts written in JavaScript that listen for pre-specified user-interactions and then send the information provided by the user to the back-end which in itself uses the library **Suave.IO** to listen for a specific set of information. The back-end consists of a collection of modules written in the functional programming language F# built mostly by Paul Menges during his master thesis (Menges, 2018) that are called using the information received by Suave.IO and return an output based on pre-defined permutations based on the input data. Streamlining and optimizing most of those modules and therefore maximize the user experience was one of this works main goals.

3.1 Data (pre-) preprocessing

A large part of this project was to optimize the pre-processing of the 408 input data sets based on proteomic analysis of chromatin bound proteins in *Xenopus laevis* egg extracts. Due to the time span over which the data was collected and because it is advised if one wants to combine different collections of data in one large meta analysis we wanted to see how each group of experiments behaves in respect to all others. To do this the non-linear dimensionality reduction algorithm t-SNE was applied to a matrix of mean protein abundances over all time points and replicates of each experiment where columns represented experiments and rows defined protein abundances.
t-SNE itself uses the parameters *dims*, *theta* and *perplexity* to alter the location of each

data point on the map. *Dims* can either be set to 2 or 3 and defines the dimensions one wants to reduce the high-dimensional data set to. In most biological circumstances two dimensions are sufficient to visualize relationships between data points. *Theta* should always be set to 0.0 for best results as it defines the error between different iterations the algorithm deems to be sufficient. Higher *theta*-values improve computation times with a decrease in accuracy. The parameter *perplexity* (loosely) defines how to balance attention between local and global aspects of the input data and sort of ”guesses” the number of close neighbours of each data point on the map. Due to this it always has to be smaller than the number of data points one wants to look at.

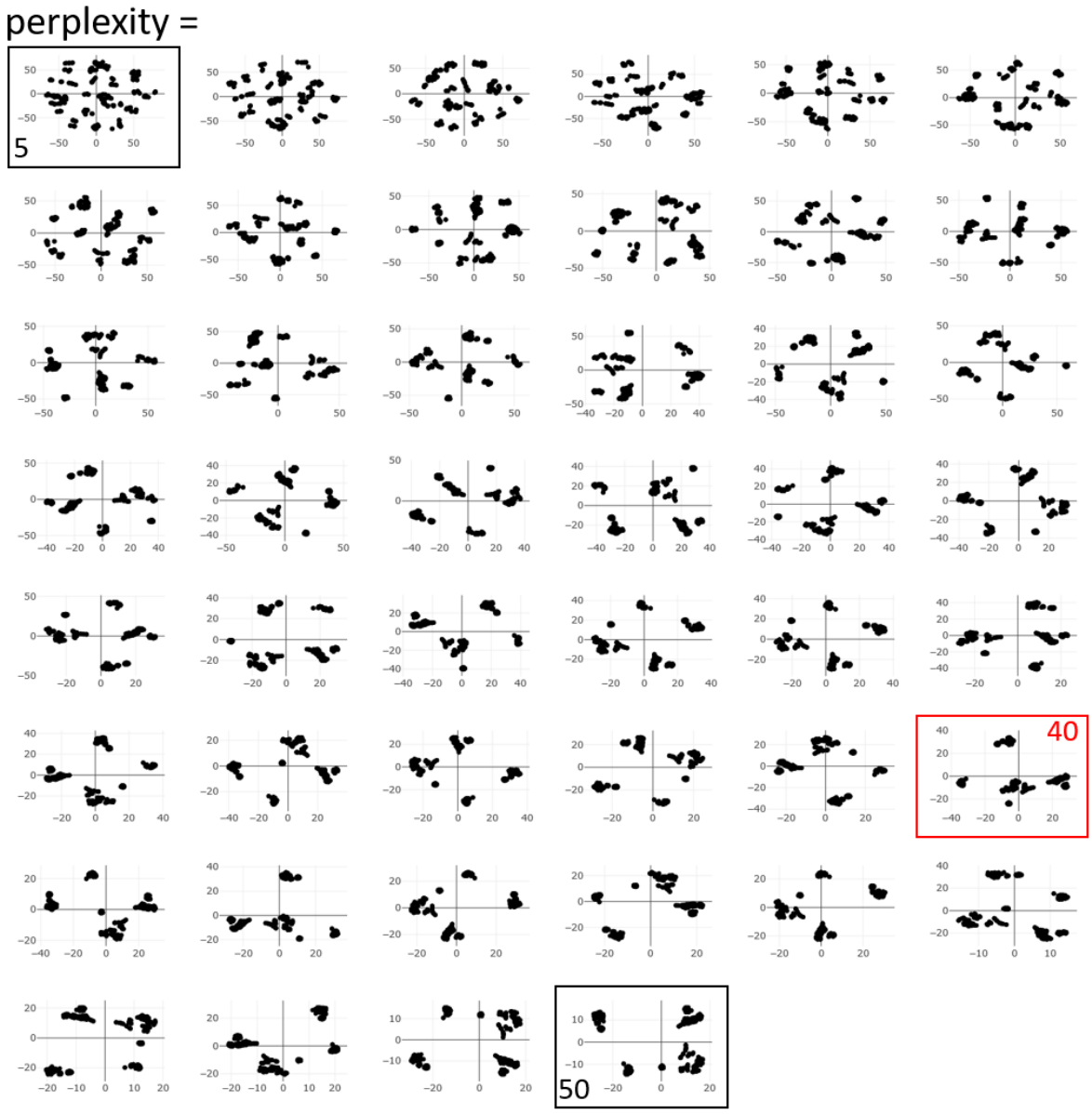

Figure 20: t-SNE Perplexity Optimization. A non-normalized matrix of mean protein abundances per experiment was filtered to include only the 1000 most prominent proteins based on their known ability to bind chromatin to lower computation times. *theta* was set to 0.0 while *perplexity* iterated from 5 to 50 in increments of 1. After comparing all resulting t-SNE maps we decided to continue with $perplexity = 40$ to avoid large indistinguishable groups.

To optimize the *perplexity* for our use case we made the assumption that the result should show our data in two main and seven sub-groups that each represent the two different DNA templates, sperm chromatin or plasmid DNA, and the six different induced damages for each set of experiments. After applying t-SNE with perplexities between 5 and 50 to a pre-filtered data set we decided on $dim = 2$, $perplexity = 40$ and *theta*

$= 0.0$ for the final large scale t-SNE. Those parameters resulted in the final map of all experiments shown below.

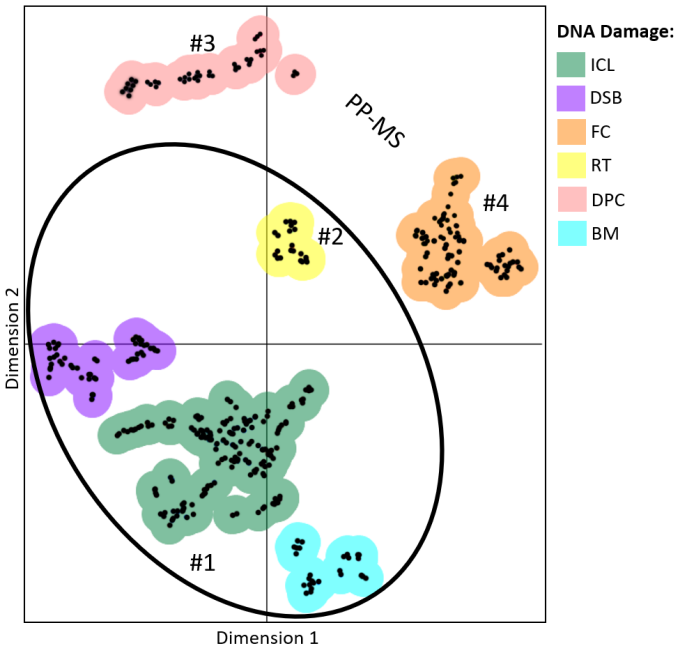

Figure 21: Two-dimensional t-SNE map of all sets included in the Atlas. Non-normalized mean protein abundances for each experiment were used as an input for the t-SNE using the optimized parameters mentioned above. Colored clouds represent DNA lesions with: ICL = Interstrand Crosslink, DSB = Double-Strand Break, FC = Replication Fork Collapse, RT = Replication Termination, DPC = DNA-Protein Crosslink, BM = Base misincorporation. Data points included in the black ellipse represent experiment using sperm chromatin as the DNA while the others use plasmid vectors with defined lesions. Individual points inside the clouds each represent a single experiment and time point. Numbers next to each group indicate the batch the sets were processed in due to their use in different publications.

Figure 21 shows a close relationship between all sets using sperm chromatin as a template (black ellipse) while still retaining closer grouping between each experiment series and the damage type induced respectively (cloud colors).
As already mentioned some of the experiment sets were processed together in one processing session based on the publication they were used in (on the map indicated by #1 through #4). We previously expected this circumstance to have a large impact on the neighborhood on the t-SNE map but this does not seem to be the case. Neighborhood embedding seems to strongly depend on the DNA template and extract type used. We can see this especially for processing group #2 that uses sperm chromatin as a template but does not seem to embed closely to other experiments using the same

substrate. This can be explained due to the experiment setup in general that was focused on investigating the termination of DNA replication specifically. Due to the specific protein fingerprint of replication termination in comparison to all other sets and the synchronous nature of these experiments this group of experiments embeds distant from the other CHROMASS data sets and closer to the PP-MS sets were replication is more synchronized compared to CHROMASS. The PP-MS sets also embed far away from each other as expected due to the synchronized replication and the very strictly defined lesions present on the plasmid template.
From this preliminary data investigation using dimensionality reduction to build a neighborhood map of all experiments included in the atlas we can conclude that DNA template and lesions have a higher impact on the embedding of each experiment series. While still present, the effect of batch processing can be neglected once proper normalization steps are carried out for the whole collection of data sets. This proves that the previous approach to visualizing DNA repair modules using a normalized pre-filtered network on the atlas was valid but could be improved by adding networks for each type of DNA lesion. Implementing such networks can, as will be discussed later in this thesis, noticeably improve the performance of clustering algorithms for the identification of functional DNA repair modules. t-SNE analysis showed that the sets are usable for the purposes of the DNA Repair Atlas but should be reprocessed together once the computational resources are available.
After validating that the data can be combined without batch effects inhibiting the creation of repair networks we moved on to prepare the data sets.

To enable filtering of significantly enriched proteins in later steps we used the combined matrix of protein abundances as an input for an R script utilizing a modified Student's t-Tests with an FDR cut-off to adjust for multiple testing. This script used the package "samr" (https://CRAN.R-project.org/package=samr) to determine the S_0 for each comparison dependent on the data itself. This allowed us to automate the calculation of significance values for each protein regarding its enrichment based on predefined comparisons. The resulting p-values were stored in $\log_{10}$ format and afterwards multiplied over all comparisons to obtain an arbitrary "Enrichment Score" over all experiments as well as for each DNA lesion separately. This score is used in the DRA in network styling as well as the base for the polar enrichment plot of each DNA lesion. As briefly described in Fig 18 the individual data collections were combined into one matrix consisting of all 408 mass spectrometry runs over 126 conditions. Altogether, a total of 4313 proteins could be detected whose mean abundances per replicate

were calculated. For each experiment proteins with less than three valid values were filtered out. This yielded a matrix with 2727 proteins over 126 conditions retaining time point information that was used as an input for the visualization functions of raw and normalized abundances for each experiment. As will later be shown the functions enabling this feature have been optimized to be more modular and user friendly.

To build the aforementioned networks for each type of DNA lesion this matrix was further processed to only include a Z-scored abundance value per protein for each experimental condition. Those conditions then were grouped by the DNA lesion repair they were used to investigate and split into separate matrices while keeping one un-grouped matrix containing all conditions. The ”Enrichment Score” over all DNA lesions was filtered for all proteins with a higher score than the 90^{th} percentile to reduce the number of proteins to 1737. This list of proteins was used to filter the individual DNA lesion matrices before applying WGCNA.

3.2 Improving the user-experience of the DNA Repair Atlas

The DNA Repair Atlas (DRA) is an interactive web resource for the analysis of DNA repair specific proteomic data sets can be used to visualize the time-dependent behavior of specific proteins for each included experiment. If a user wants to find out about the involvements of a single or a group of proteins she or he is interested in, the DRA offers multiple options to mine the included experiments for information about these proteins. In the following section we will explain how the DRA can be used to retain information about specific complexes. As an example we will use the Fanconi core complex, an E3 ubiquitin ligase known to be involved in the mediation of Interstrand Crosslink repair. Throughout the results section we will mention how code has been optimized during this thesis work to improve both user experience and the long-term maintenance of the web resource.

3.2.1 Visualizing protein involvement in DNA Repair

To get a first impression on the representation of a protein of interest and its possible involvement in DNA repair a user can visualize the enrichment of said protein using the newly implemented **Protein Search** subpage. This page plots a radar plot that shows the relative enrichment of a protein of interest over all included experiments

investigating different DNA lesions. Each radial axis represents a particular perturbation corresponding to different DNA repair pathways or replication termination as indicated (see Figure 22). The plotted scores represent the enrichment scores of the input protein within each experiment group (see Section 2.2).

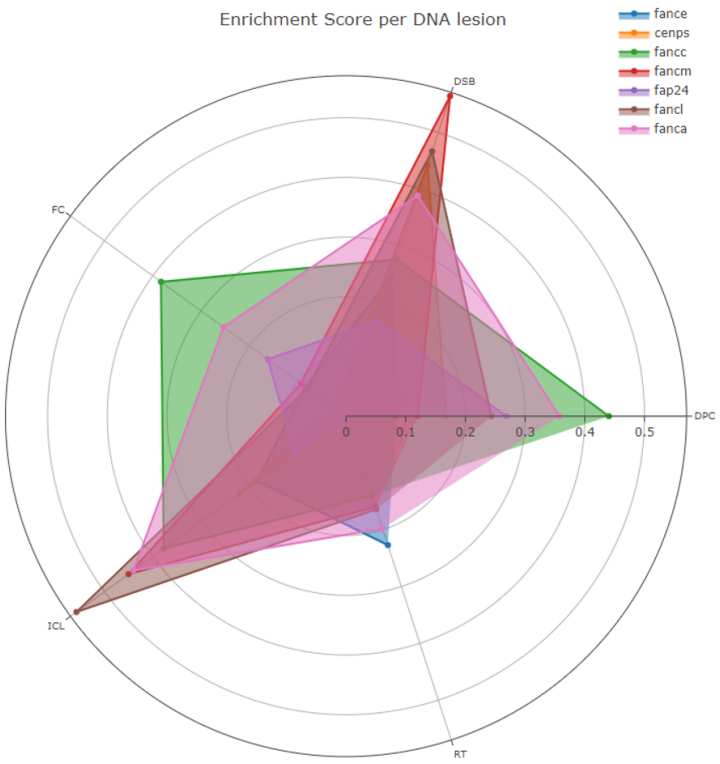

Figure 22: Relative enrichment of Fanconi core Complex subunits. Enrichment scores for each subunit of the Fanconi Core Complex included in the filtered list of enriched factors. Included are only those subunits that are enriched over all data sets but the scores are specific to each DNA lesion. ICL = Interstrand Crosslinks; DSB = Doublestrand Breaks; DPC = DNA-Protein Crosslink, RT = Replication Termination; FC = Fork Collapse
http://dnarepairatlas.bio.uni-kl.de/Summary

On the example of the Fanconi core complex the user can find that most subunits are most prominently enriched under Interstrand Crosslink conditions indicating that they play a role in the repair of this type of DNA lesion. Some subunits (fancm and fancl) also seem to be enriched under Doublestrand Break conditions as well, hinting at functions of these subunits as independent enzymes during DSB repair.
Generally this feature of the DNA Repair Atlas gives the user a good idea of what to expect from the Atlas itself. It provides an overview of which DNA repair pathways can be investigated and where a protein of interest is most prominently enriched. The user can then continue with this information to plot individual time-dependent protein

abundances on chromatin within the experiment group with the highest enrichment score.

```
let rec radialData
    (metaData:(string*float*float*float*float*float*float)list) (poi:string*string) =
        match metaData with
        | [] -> []
        | ((prot,all,dpc,dsb,fc,icl,rt) :: rest) when prot = (fst poi) ->
          (prot,dpc,dsb,fc,icl,rt) :: radialData rest poi
        | (_ :: rest) -> radialData rest poi
```

The code above shows the function the back-end uses to provide the user with the enrichment information starting from a text file including scores for all proteins.

3.2.2 Plotting time-dependent protein abundances

Continuing with the information that the Fanconi core complex is most prominently enriched under Interstrand Crosslink conditions the user can now explore the time-dependent changes in protein abundance within the included experiments investigating ICL repair.

The experiment most fitting for demonstrative purposes is EXP03 (Experiment Series 3) which used Psoralen to introduce ICLs to sperm chromatin and the NPE/HSS system to study the replication dependent mechanisms of ICL repair. Samples were taken in 15 minute intervals starting at 15 minutes and ending at 90 minutes after addition of DNA. The experiment included three major conditions.

- Control (untreated chromatin)
- Psoralen-treated chromatin
- Mock (without added DNA)

Plotting the abundances of the FA core complex subunit **fancl** with the highest enrichment score for all three conditions one can see (Figure 23) that it is enriched on Psoralen-treated chromatin and peaks around 60 minutes after the start of replication and decreases after 75 minutes while its abundance fluctuates heavily around a lower mean value on untreated chromatin.

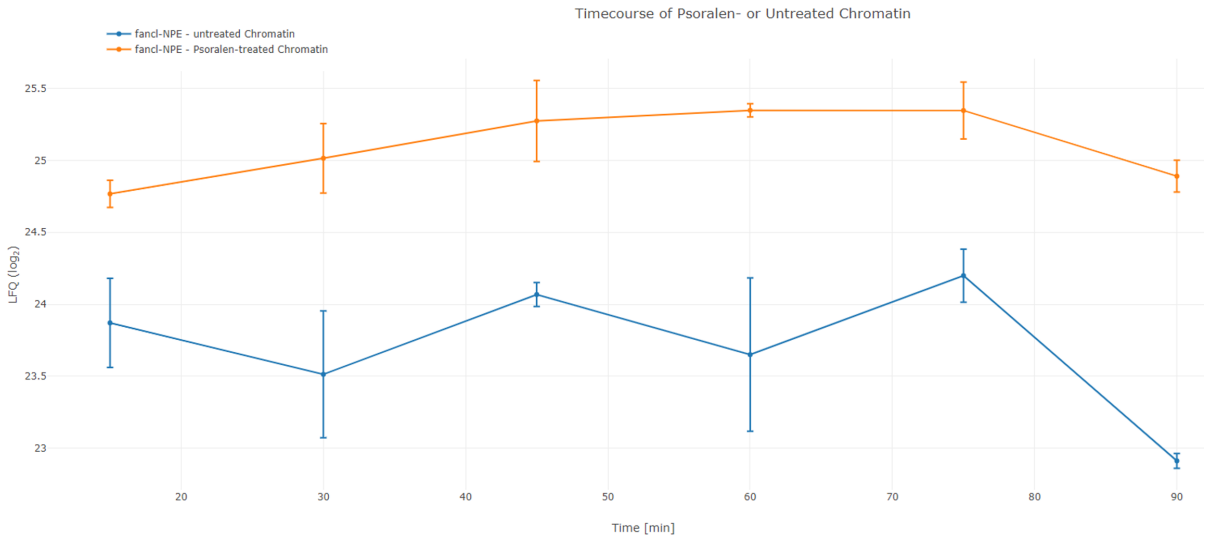

Figure 23: Time-dependent abundance of fancl on Psoralen-treated and untreated Chromatin. Depicted is the difference in fancl abundance between Psoralen-treated and untreated chromatin over time.

From this result the user can deduce that fancl and possibly all other FA core complex subunits have a higher abundance under treatment with Psoralen indicating that they could be involved in ICL repair.

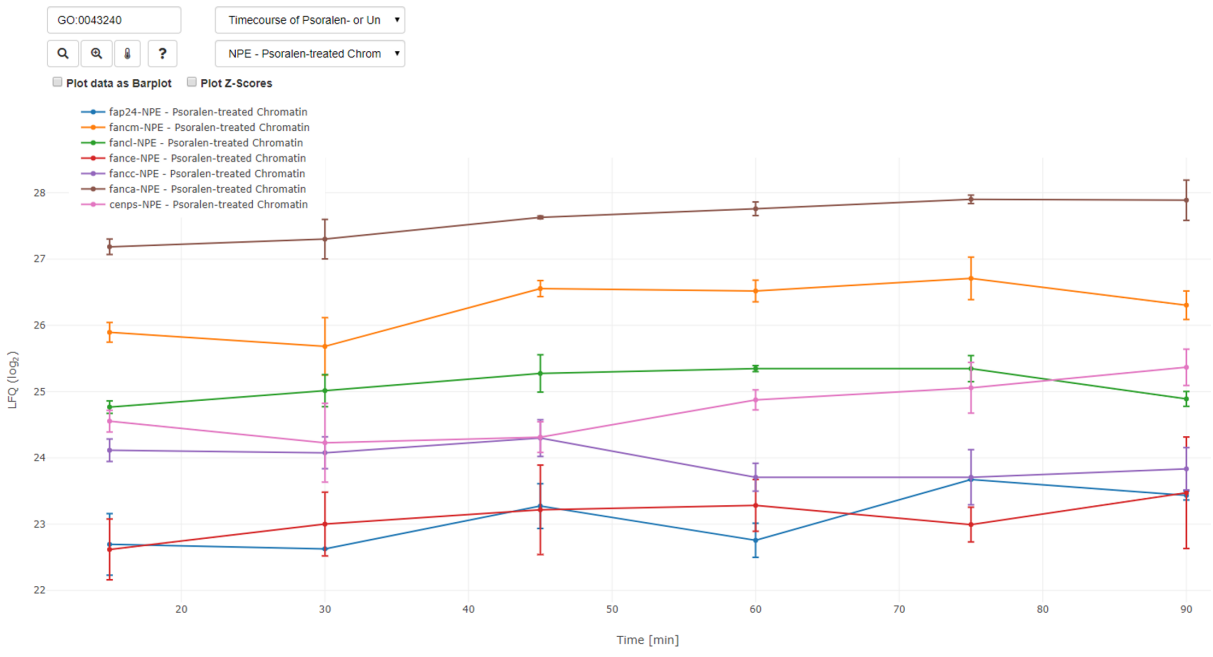

Figure 24: Changes in abundance of FA core complex proteins under Psoralen treatment. The DRA features the option to plot the changes in abundance on chromatin for each included protein, experiment and experimental condition.
http://dnarepairatlas.bio.uni-kl.de/Plotting

To further investigate if this assumption can be consolidated using the DRA we could now plot the changes in abundance under Psoralen treatment for all FA core complex proteins. This plot again shows that all FA core complex subunits behave similarly under ICL conditions further solidifying our assumption of the FA core complexes involvement in the replication-dependent repair of ICLs.
Additionally we've implemented the option for the user to plot z-scores for each protein on demand.

3.2.3 Improving the back-end adaptivity

Previously many features of the DRA were built to only work with the specific plain text database where rows represented individual proteins and each column represented a single proteomic data file. The function written to extract data from this set based on the user input consisting of an experiment is shown below.

```
let chooseExperiment (exp:string) =
    match exp with //matches the input experiments with pre-defined keys in the database
    |"Key" -> let Exp_Time_1 =
                    [for row in textDataBase.Rows -> row.Exp_Time_1]
                    |> Seq.ofList
                    |> Seq.zip miscData
    //and uses FSharp.Data to extract matching intensities per timepoint
              let Exp_Time_2 =
                    [for row in textDataBase.Rows -> row.Exp_Time_2]
                    |> Seq.ofList
                    |> Seq.zip miscData
    //to finally merge to a sequence of two tupled lists containing intensities and time
             seq[Exp_Time_1;Exp_Time_2],["1";"2"]

    |"" -> seq[],[]
```

Put simply, this function matches the experiment chosen by the user with *hard coded* experiments in the database, extracts every row of the column representing this experiment and then merges it with a list of proteins defined elsewhere. In the context of the resource this function is called within another function that gets the experiment and a list of proteins of interest as an input and that outputs data in the format the front-end needs to plot a time course or protein intensities for the selected experiment and treatment. To improve on this function design we looked at the input file and how data is extracted from it. Noticeable was that each time a new data set was added to the collection there were two locations in the code that had to be adapted to fit the new input structure. Not only had the HTML document to be updated but also the matching function in the back-end. We came to the conclusion that it would be easier to store the data in such

a way that new data sets could be added without the need to update most of the code. To achieve this we implemented a function that takes a plain text file, an experiment, a treatment condition and a list of proteins as its input. The list of proteins is fed to the function in string format to simplify the interaction between the back- and front-end.

```
let searchForProteinData (data:CsvFile<CsvRow>) (expIn:string) (protList:string) (treatIn:
    string) =
    let lfqValueSeq =
        let proteins = data.Headers.Value |> Array.tail |> Seq.ofArray
        seq[for i in (addedNames protList) -> proteins |> Seq.filter (fun x -> x.StartsWith(
            fst i))]
        |> Seq.concat
        |> Seq.map (fun p -> proteins
                             |> (fun _ -> data.Rows
                                          |> Seq.map (fun y -> try float (y.GetColumn p) with
                                              | :? Collections.Generic.KeyNotFoundException ->
                                               nan))
                             |> Seq.zip [for row in data.Rows ->
                                           int (row.GetColumn "FILE_ID")],p)
    let filteredMeta =
        let filtered = (metaDataParse.Filter (fun x -> String.Equals((x.GetColumn "Treatment")
            , treatIn) )).Filter ( fun x -> String.Equals((x.GetColumn "Experiment Series"),
            expIn))
        seq[for row in filtered.Rows ->
                int (row.GetColumn "FILE_ID"), row.GetColumn "Treatment", row.GetColumn "
                    Experiment Series", row.GetColumn "Time"]
    let treatment = filteredMeta |> Seq.map (fun (_,treat,_,_) -> treat) |> Seq.head
    let inter =
        filteredMeta
        |> Seq.collect (fun (metaID,_,_,time) -> (lfqValueSeq |> Seq.map (fun (valueSeq,
            protName) -> (valueSeq |> Seq.tryFind ( fun (lfqID,_) -> lfqID = metaID),protName,
            time))))
        |> Seq.map (fun (valOpt,protName,time) -> match valOpt with
                                                  | Some x -> (snd x), protName,time
                                                  | None -> nan,protName,time)
    inter
    |> Seq.filter (fun (lfq,_,_) -> (Double.IsNaN >> not) lfq )
    |> Seq.groupBy (fun (_,y,z) -> y,z)
    |> Seq.map (fun (groupName,myGroup) -> groupName,(myGroup |> Seq.map ( fun (lfq,_,time) ->
         lfq,time)))
    |> Seq.map (fun ((p,t),x) -> (p,t),(List.ofSeq x |> List.unzip))
    |> Seq.map (fun ((p,t),(vL,_)) -> p,vL,t,(List.averageBy float vL),(sDev vL))
    |> Seq.groupBy (fun (p,_,_,_,_) -> p)
    |> Seq.map (fun (groups,data) -> groups,data |> Seq.map ( fun (_,_,t,vm,e) -> (vm,e,t) )
        |> List.ofSeq |> List.unzip3)
    |> Seq.map (fun (n,(y,e,x)) -> (n,treatment,(y,e,x)))
    |> Seq.sortByDescending (fun (prot,_,(_,_,_)) -> prot)
```

Note that in this case the data is split into two files with one containing meta information and the other the LFQ intensities. The two files are matched using unique IDs. This format was chosen to make it easier to maintain the meta data and make changes on the fly without risking to alter the LFQ intensity file. In general applying this function results in the same output as the function mentioned above, but the mechanism by which it works allows easier maintenance.

Additionally this change gave way to a dynamically filled drop-down menu on the front-end. Each time an experiment series is selected (see Figure 25) the drop-down menu

for selecting the experimental condition is updated to show all valid options. To do this the front-end request a JSON string containing a map of experiment series and their valid conditions anytime the experiment selector is changed.

A


```
{
"exP01":"NPG,NPC",
"exP02":"NPC,NPG",
"exP03":"NCC,NMC,NPC",
"exP04":"NPB,NCC,NMC,NPC",
"exP05":"NPB,NCC,NMC,NPC",
"exP06":"NCA,NCAG",
"exP07":"NCA,NCAG",
"exP08":"NCA,NCAG",
"exP10":"HCC,HCG",
"exP11":"LSS,NPE",
"exP12":"Angelicin",
"exP13":"UB-VS",
"exP14":"MMS",
"exP15":"ATRi",
"exP16":"HCC,HDC",
"exP22":"HCC,HDC",
"exP23":"NHEJ-Inhib T1",
"exP24":"NHEJ-Inhib T2",
"exP25":"HDN,HCC,HDC",
"exP26":"LDC,LCC",
"exP27":"LCC,LDC,LDA",
"exP28":"NCC,NDC",
"exP29":"NDC,NDR,NCC",
"exP30":"NCC,NTC,NMMC",
"exP31":"HDK,HDC,HCC",
"exP32":"HDS,HDC,HCC",
"exP33":"NCC,NC-NMS,NC-MLN,NCG,NC-NMSG,NC-MLNG",
"exP34":"NCC,NCG,NPC,NAC,NFC,NFB,NFG",
"exP35":"NCC,NC-NMS,NC-MLN,NCG,NC-NMSG,NC-MLNG",
}
```


B

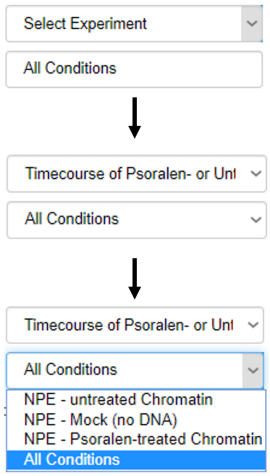

Figure 25: Filling a Drop-Down menu with JSON data. A) Illustrated is the JSON string the front-end request each time the experiment selector is changed. The JSON string represents a map that contains keys representing individual experiments that are matched via a JavaScript function and respective values that are used to fill the second dropdown menu dynamically. B) An overview of the drop-down menus used to select an experiment and the dynamically filled treatment.

The dynamic addition and removal of drop-down options given the described JSON map is handled via a JavaScript function that can be found in the supplemental information on Github. It uses a switch-case statement to match key values of the JSON map requested from the back-end to extract their respective values. It then uses another switch-case statement to match the three-letter treatment codes with human readable treatment descriptions that are added to the treatment selection drop-down menu.

3.2.4 Inclusion of GO term plotting

The last version of the DRA included only the option to use specific predefined protein names as an input for the plotting- and search functions. This has been updated to include the UniProtIDs for each human homologue of the included proteins as well as major Gene Ontology (GO) codes. This gives the user the opportunity to input a certain GO-term and plot all proteins in the database that are annotated with that term. For example if a user is interested in the changes of protein abundance on Psoralen-treated chromatin over time for the Fanconi anemia complex proteins it is not necessary to type each subunit into the input field but the GO term would be sufficient.

Name	Entry	label	GO
orc2	Q13416	Orc2	GO:0000075
orc2	Q13416	Orc2	GO:0000082
orc2	Q13416	Orc2	GO:0000084
orc2	Q13416	Orc2	GO:0000122
orc2	Q13416	Orc2	GO:0000216
orc2	Q13416	Orc2	GO:0000792
orc2	Q13416	Orc2	GO:0000939

Figure 26: Small excerpt from the annotation table used for data mining. Shown is a list of additional names for the Origin Recognition Complex subunit orc2. Note that each row represents a single given combination of all possible inputs that result in orc2 being used as a query protein. Available are a multitude of GO terms (GO) as well as the gene name (Name), an alternative gene name (label), human homologue UniProtID (Entry).

addedNames takes the user input string which can be a single protein of interest, a list of proteins or a GO term and splits it into a sequence of strings. Each element of this sequence is then matched against a protein annotation file containing multiple entries for each protein present in the database based on ever single possible GO term and UniProtID for that protein. If a match is found, the function returns a list of proteins with identification values that match the ones used in the plain text data base of the DRA, effectively translating known alternative names and IDs to usable protein names. As this function can be called inside any other function extracting information for a protein from the data base, we've implemented a way to use multiple identification systems with the DRA that can be easily adapted once new proteins get added or more identifiers are needed. Especially useful is the inclusion of GO terms due to their ease of use and the possibility to use them as an easier input for clustering algorithms using multiple proteins as their input - such as the random walk with restart algorithm implemented in the DRA by Menges (2018) (see Section 1.6).

This feature has been implemented globally so that it functions with all basic input fields on the DRA. Functionally,the back-end stores all annotation data for each protein in a large plain text table with the structure shown in Figure 26 and applies a function called

addedNames to every text input made by the user.

```
let addedNames (inString:string) =
    inString.Split [|',';';'|] |> Array.filter (String.IsNullOrWhiteSpace >> not) |> Array.map
        ( fun x -> x.Trim()) |> Seq.ofArray
    |> Seq.collect ( fun inputProt ->
                     additionalNames
                     |> List.map ( fun x ->
                                   match x with
                                   | (a,_,_,_) when String.Equals (a,inputProt) -> a,inputProt
                                   | (a,b,_,_) when String.Equals (b,inputProt) -> a,inputProt
                                   | (a,_,c,_) when String.Equals (c,inputProt) -> a,inputProt
                                   | (a,_,_,d) when String.Equals (d,inputProt) -> a,inputProt
                                   | _ -> inputProt,inputProt ) )
    |> Seq.distinctBy fst
    |> ( fun x -> if (inString.StartsWith "GO:") || (inString.StartsWith "EC:") then
                    x |> Seq.filter ( fun (a,b) -> not (String.Equals (a,b)) )
                  else
                    x)
    |> Seq.filter ( fun (a,b) -> (String.IsNullOrWhiteSpace >> not) a )
    |> List.ofSeq
```

In addition to the functional changes to the back-end explained above the design of the DRA has been updated to improve the user experience. The starting page now shows a visual representation of all included functionalities as well as descriptive texts for each major sub-page.

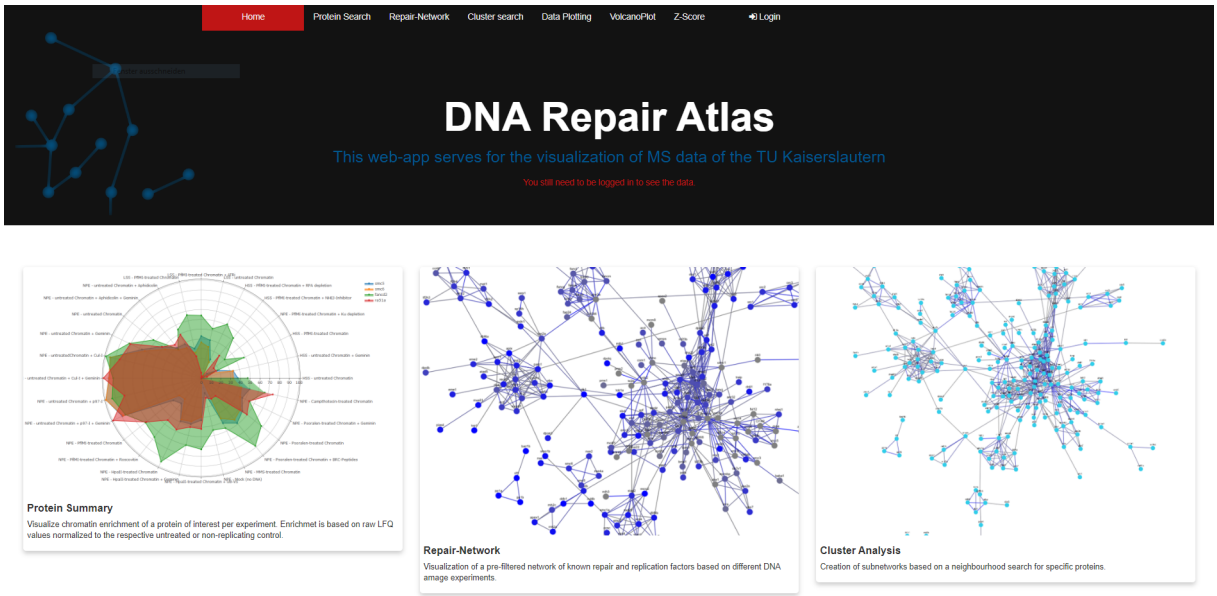

Figure 27: Screenshot of the starting page of the DNA Repair Atlas. Shown is a screenshot of the starting page with the visual represenations of all major functions as well as descriptive texts for each of the sub-pages.
http://dnarepairatlas.bio.uni-kl.de

The remainder of this thesis will focus on the changes made to the **Cluster Search** function. We will briefly describe how networks were optimized to make them more suitable for the identification of functional modules.

3.2.5 Constructing DNA lesion specific networks using modified WGCNA

Previously a reference network was included in the DRA that was constructed using the Pearson's correlation coefficient of all proteins over all included experiments. This network was used to identify functional DNA repair modules using the clustering algorithms mentioned in Section 1.6. Although this yielded reasonable results in the past, we noticed a large amount of proteins in the resulting cluster lists that appeared unlikely to be involved in the respective repair mechanism. We therefore wanted to test, whether the identified modules would more accurately represent known DNA repair complexes, if networks were created for each repair pathway individually.
To further improve the networks, Topological Overlap Measures (TOM). These have been proven to reliably represent co-regulation and co-expression of genes in biological systems (Peter Langfelder, Steve Horvath, 2008) and have since then found their way into proteomic studies due to the similarities between the activation and regulation of proteins and genes. Here we assume that protein recruitment under specific damage inducing conditions behaves similar and therefore TOM can be used to accurately depict a co-binding of proteins to chromatin. To compute these co-binding measures the following steps were conducted using a combination of R packages.
First, the Pearson correlation coefficient for each column was then calculated using the R package ”Hmisc” (https://CRAN.R-project.org/package=Hmisc) that also reports the significance of each correlation value. The correlation matrices were then flattened to obtain lists of the format shown in Table 3.2.5.

Protein A	Protein B	Correlation	p-value
acbp4	abcf1	0.5645	0.03
aldoa	actc	0.9001	4.8e-10

Every correlation coefficient with a p-value below 0.01 was then filtered out and anti-correlations were removed because they were not used in network construction. From the columns ”Protein A”, ”Protein B” and ”Correlation” we reconstructed a correlation matrix and a sigmoid adjacency function included in the R package WGCNA (Peter Langfelder, Steve Horvath, 2008) was used to apply a soft-threshold. The result of this was used to compute the Topological Overlap Measure (TOM) that defines co-expression or in our case co-binding patterns using a geometric measure based on a

similarity matrix. Put simply, TOM transforms similarity values to geometric distances while normalizing for the maximal and minimal similarity, or range, of the input similarity matrix. TOM therefore describes neighborhood similarity of nodes where $TOM = 0$ means that two nodes share no neighbor and $TOM = 1$ means all neighbors are shared. The calculated TOM and Adjacency matrices were flattened and merged together to form lists of adjacencies and TOM values for each protein-protein relationship. Those lists can be considered to be edge lists that are used to create the network on the DRA.

Protein A	Protein B	Adjacency	TOM
anxa5	aimp1	0.7932	0.1133
anxa5	aldoa	0.9723	0.6644

They contain a column that specifies protein A, a column that denotes protein B and one column for Pearson's correlation coefficient and the Topological Overlap Measure for the relationship between A and B. How the F# back-end creates a network from this edge list will be mentioned in Section 3.2. To ensure that our clustering algorithm of choice can be applied to our data we had to check if the newly constructed networks can be considered "scale free". In scale free networks the node degree distribution, meaning the distribution of the number of direct neighbours of each node, follows a power law.

$$P(k) \sim k^{-y} \tag{9}$$

The slope of this model is not really relevant for our observation of scale-freeness but the regression coefficient value shown above each plot indicates how well the networks follow the power law. Networks with an regression coefficient of above 0.6 can be considered scale free in the biological context due to real scale-freeness being very rare in real life data (Broido, Clauset, 2019). Given these thresholds all of our TOM networks can be considered to be scale free (see Figure 28). Networks that do not fit this criterion can of course still be used for clustering algorithms but can yield less reliable results because they do not follow the small world principle of sufficient connectivity between all network nodes.

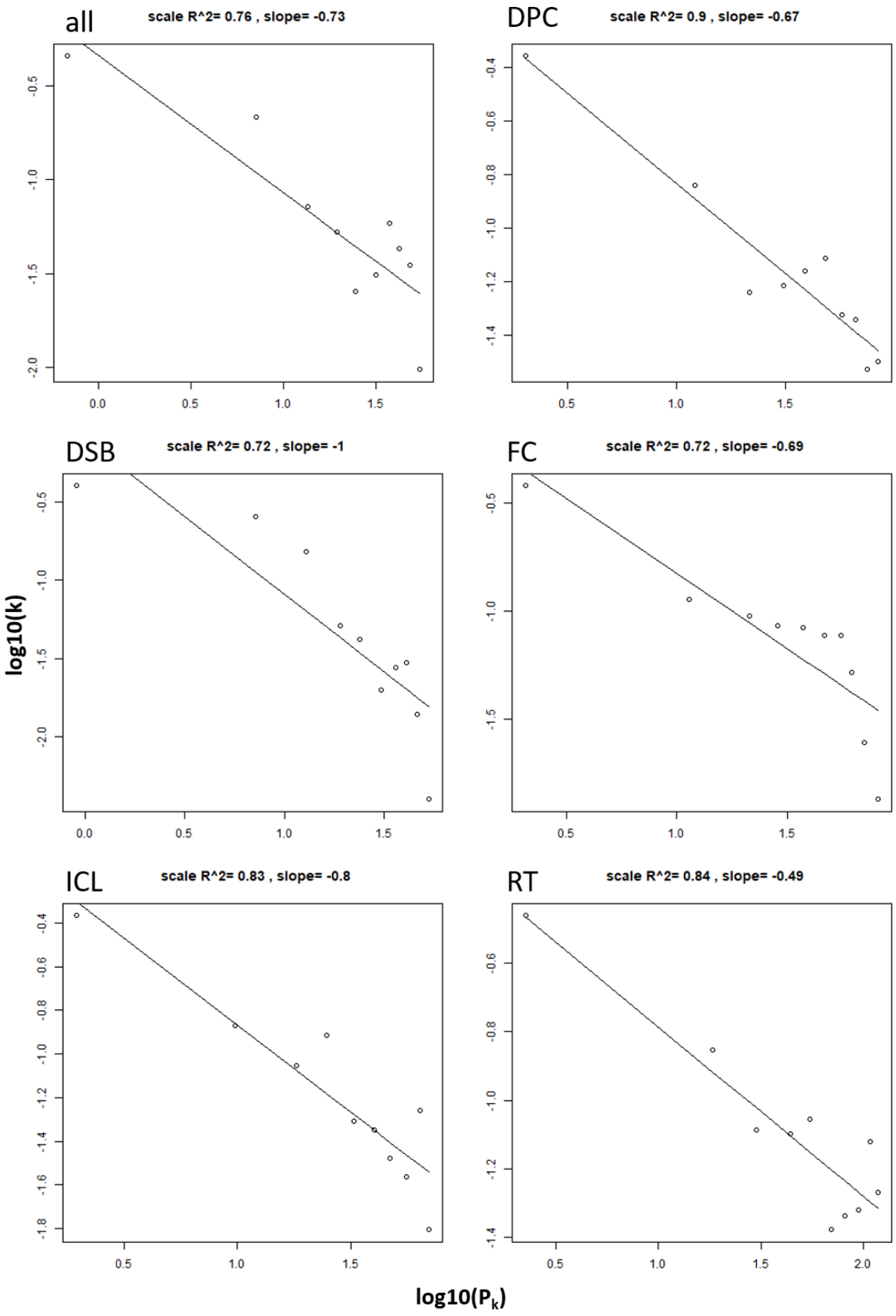

Figure 28: Scale free plots for all newly constructed TOM networks. Plotting of scale free plots was done using the a built-in function of the WGCNA package and native R methods. The top left plot labeled ”all” shows the scale-freeness of the combined correlation reference network.

The constructed networks can either be plotted and browsed using the interactive website or used as a base for nearest neighbor clustering or a random walk with restart algorithm to find functional clusters within the network as shown in Figure 29 on the example of gene co-expression.

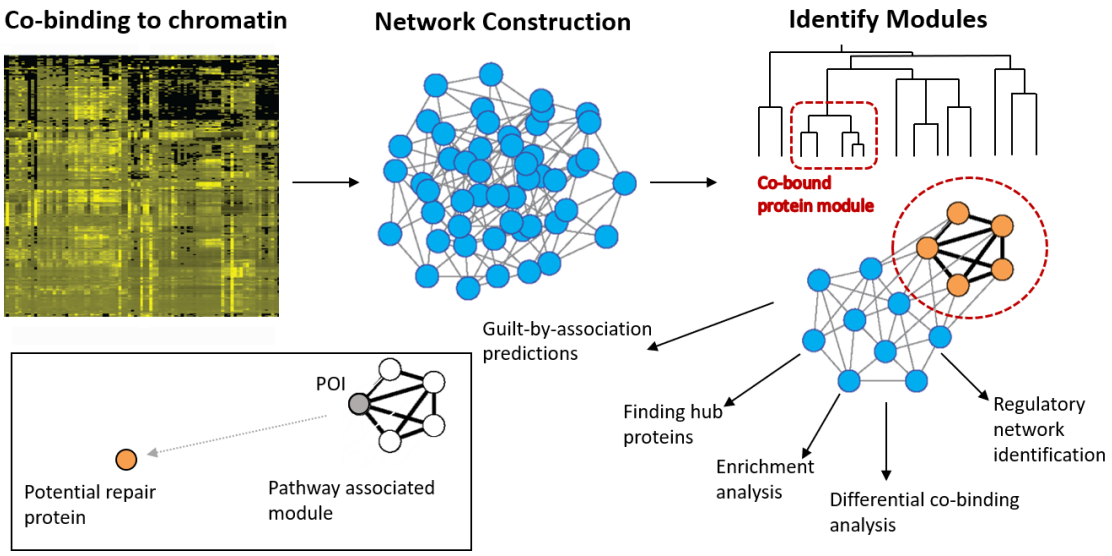

Figure 29: Schematic of co-binding analysis. Schematic of the steps used in co-binding analysis. Pairwise correlation is used to determine the relationship between each possible pair. This is followed by filtering of said relationships in alignment with appropriate thresholds and the visualization the remaining correlation values in form of a network. This network is then used to identify functional modules using grouping methods such as hierarchical clustering. From these modules hubs and regulating factors as well as functional predictions can be drawn. The latter can be based on the ”Guilt-by-association” approach (GBA, highlighted) that identifies the function of unknown variables by the variables they cluster with. Gene co-binding is generally mathematically analogous to the protein co-regulation studied in this thesis.
van Dam et al. (2018), modified

With the created network tables using TOM as a measure of the similarity between chromatin binding behavior of repair and replication proteins it is now possible to apply clustering algorithms such as Random walk based Network Propagation to identify functional DNA repair modules for all sets and for each DNA lesion individually. To ensure that a maximum amount of data is available to the nearest neighbor- and random walk clustering algorithms we calculated TOM maps for all DNA lesion sets including all proteins found in each run ($N_{max} = 4313$). Those edge lists are only available to the clustering algorithms and can not be plotted as networks on the DRA itself. Scale-freeness for these networks was analyzed as explained above and the corresponding plots can be seen in the supplementary collection on Github.

3.3 Module identification using TOM networks

In addition to the previously included reference network consisting of 295 known DNA repair and replication proteins we now included DNA lesion specific repair networks based on topological overlap measures for DNA-Protein Crosslinks (DPC), Doublestrand Breaks (DSB), Interstrand Crosslinks (ICL), Replication Fork Collapse (FC) as well as a network visualizing protein involvement in Replication Termination (RT). These networks can be visualized using the JavaScript package cytoscape.js on demand including only the 1737 most abundant proteins under damage conditions. In addition to the processing steps mentioned in Section 3.2.5 a soft threshold for plotting is applied to the raw edge list before plotting. The back-end is configured in such a way that it only plots edges that have a weight higher than $0.7 * EdgeWeight_{max}$ to improve plotting performance.

When visualizing the networks one has to consider that each network is based on a specific set of experiments using different methods, DNA templates and especially treatments that can greatly influence the appearance and accessibility of a network plot itself. We've already mentioned that not all of the unfiltered networks can follow the scale free principle. Especially networks created from a small list of conditions and that therefore had less information available to the correlation- and TOM algorithms are very interconnected. Even though all our networks fit our criteria for biological scale free-like networks we see large differences in the plotted filtered networks (see Figures 30 through 40) with especially the ICL and DPC networks being relatively inaccessible to the users naked eye. We defined accessibility of a network as the ability of a trained user to distinguish groups of functionally similar proteins in the network plot solely by means of observation.

Before starting to mine the constructed networks for functional protein modules we determined that the algorithm parameters should be set to $iter = 1000000$ and $reset = 0.25$ while the number of proteins to be plotted after random walk varies for each query. These parameters gave us results representing known modules with a great balance between computation time and accuracy.

3.3.1 Analysis of the DNA-Protein Crosslink Network

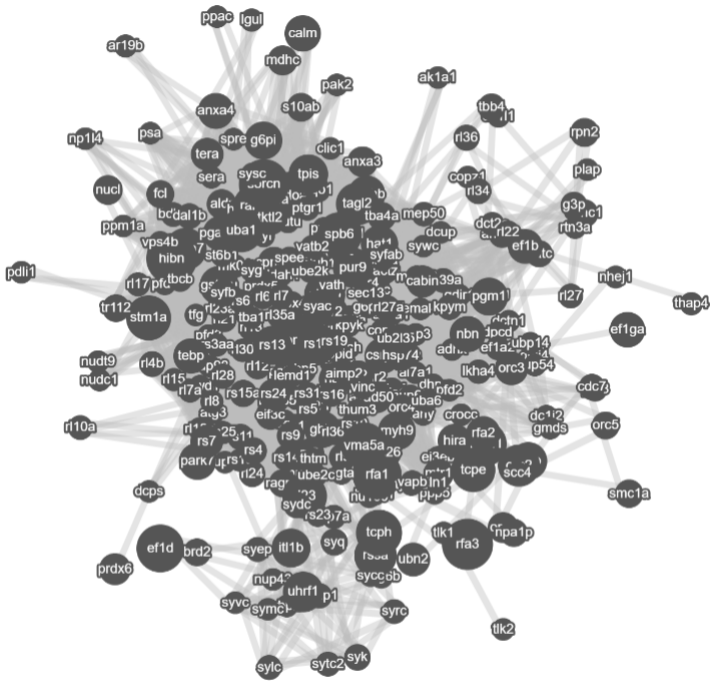

Figure 30: DNA-Protein Crosslink TOM Network. Depicted is the visual representation of proteins loaded onto chromatin under DNA-Protein Crosslink conditions. Edges are filtered using a soft threshold of 70% and their width is proportional to their score. Node size represents the protein enrichment scores.
http://dnarepairatlas.bio.uni-kl.de/Clustering

As mentioned in the introduction of this work DNA-Protein Crosslinks are mostly repaired in a replication dependent manner through mechanisms that are mediated by the proteasome and the metalloprotease SPRTN. Using the network shown in Figure 30 one can observe replication factors on the outer right and bottom parts of the network as well as many subunits of the proteasome in the center that are grouped together. Due to the soft threshold applied for plotting SPRTN is not included in the network plot but is still available for the random walk algorithm. To investigate whether the unfiltered network used for clustering represents published protein interactions and DNA binding mechanisms in DNA-Protein Crosslink repair we used SPRTN, PCNA and the MCM helicase as a query for the random walk algorithm. Plotting the 20 best results yielded us the cluster shown in 31.

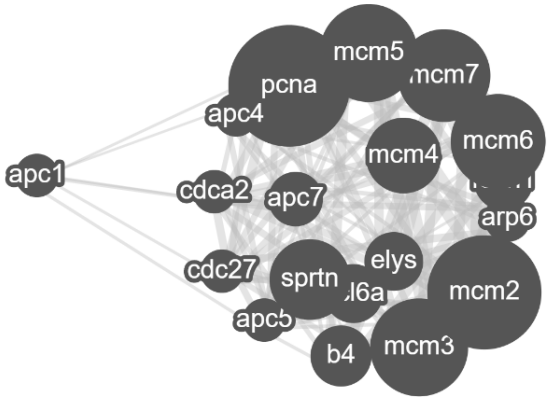

Figure 31: Random walk result for SPRTN and replication associated proteins under DPC conditions. Shown is the resulting cluster graph of the random walk algorithm based on the unfiltered DPC network.
Query: SPRTN, PCNA, MCM2-7; 20 nodes

This random walk output validated that the subunits of the MCM helicase have a high similarity over all conditions of the DPC experiment series. Additionally, they seem to have a highly similar chromatin binding profile to PCNA and other DNA replication factors and checkpoint proteins such as CDC23 and CDCA2. Interestingly there is no direct edge between PCNA and SPRTN despite Vaz et al. (2016) showing that SPRTN binds directly to the replication fork where PCNA is always loaded during replication.

3.3.2 Analysis of the Doublestrand Break Network

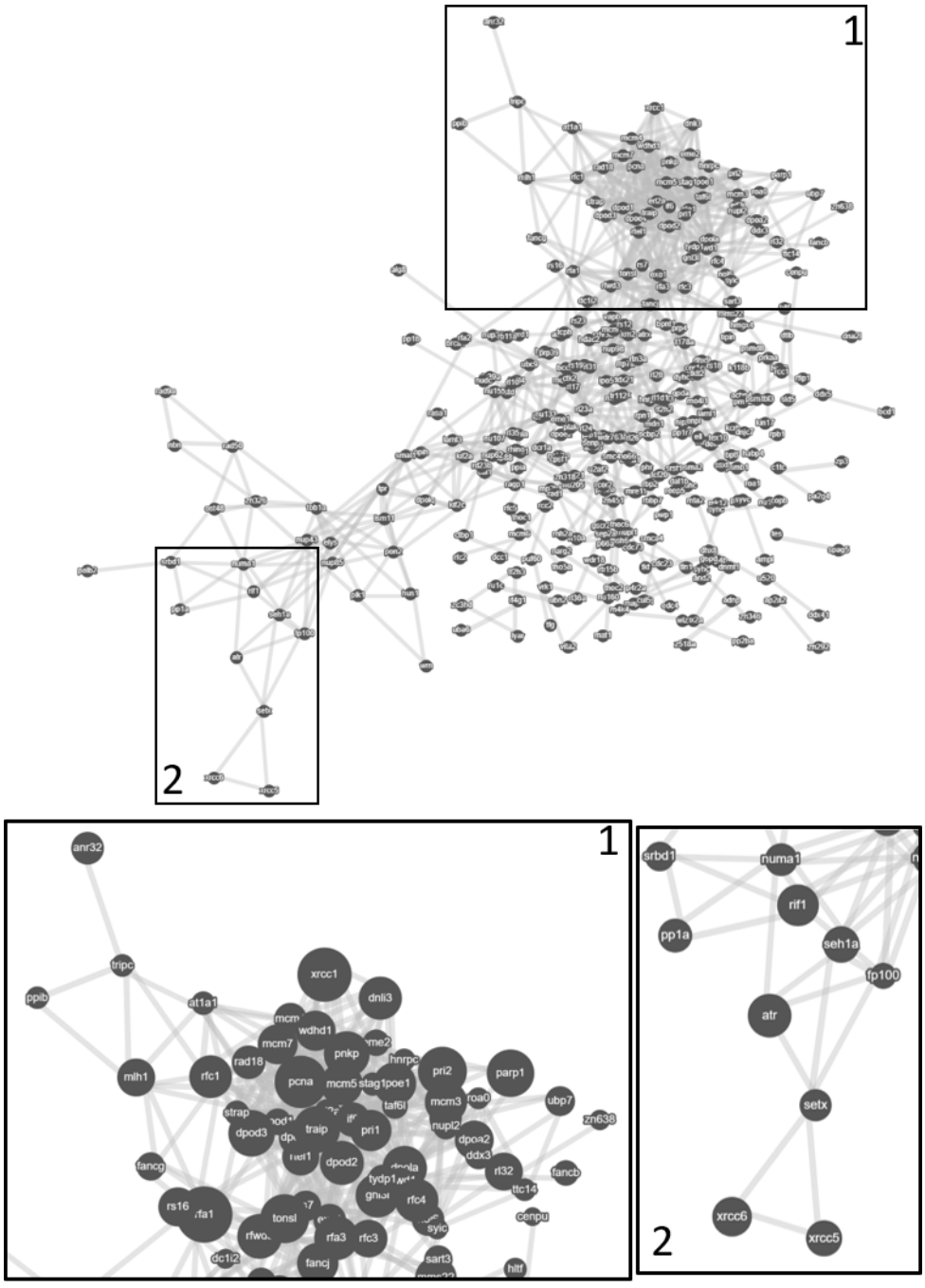

Figure 32: Doublestrand Break TOM Network. Visual representation of proteins loaded onto chromatin under Doublestrand Break conditions. Edges are filtered using a soft threshold of 70% and their width is proportional to their score. Node size represents the protein enrichment scores.

In contrast to all other networks looked at in this work the DSB network used experiment sets only collected using HSS meaning that this network depicts replication independent DSB repair. Proteins associated with replication initiation and origin licensing will be found in the set but replication firing could not occur. As can be seen in Figure 32 the Doublestrand Break network shows groups of proteins that are distinguishable by means of observation. The network itself feature a low number of nodes with diverse degrees. Looking at the top right (labeled **1**) group of proteins one can make out proteins associated with origin licensing (Section 1.2.2) and the formation of a pre-RC as well as the first polymerases with missing co-factors. Additionally, the E3 ubuquitin ligase TRAIP is closely connected to PCNA indicating the successful detection of DNA damage. In the bottom left of the network another high scoring group of proteins including ATR and the DSB repair factors XRCC5 and XRCC6 can be seen. This network represents mechanisms of DSB repair fairly well even in its filtered form for plotting. To see whether full DSB repair modules can be identified we used the known repair factors ATR, ATM, TRAIP, XRCC5 and XRCC6 as the query for our random walk algorithm and plotted the 30 highest scoring proteins in Figure 33.

As Cimprich, Cortez (2008) showed in their review on the function ATR in maintaining genome integrity the replication independent repair of DSBs works in two ways. The first mechanisms detects ssDNA by RPA binding which in turn leads to ATR/ATRIP recruitment as well as the loading of the 9-1-1 complex (Rad9-Hus1-Rad1) onto the first available dsDNA section after the DSB. This method is reflected in the lower part of the cluster seen on the right. The right side of the cluster represents the detection and repair of double stranded ends via ATM and the following recruitment of RAD50.

To further show how robust the DSB network is in representing protein recruitment profiles we used the MCM helicase subunits MCM2-7 as an input for random walk and plotted the 15 highest scoring proteins

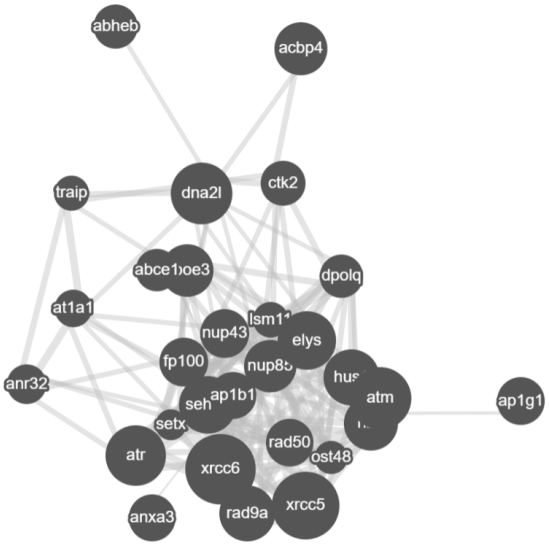

Figure 33: Random walk result for ATR, ATM and other DSB repair proteins under DSB conditions. Shown is the resulting cluster graph of the random walk algorithm based on the unfiltered DSB network.
Query: ATR, ATM, TRAIP, XRCC5, XRCC6; 30 nodes

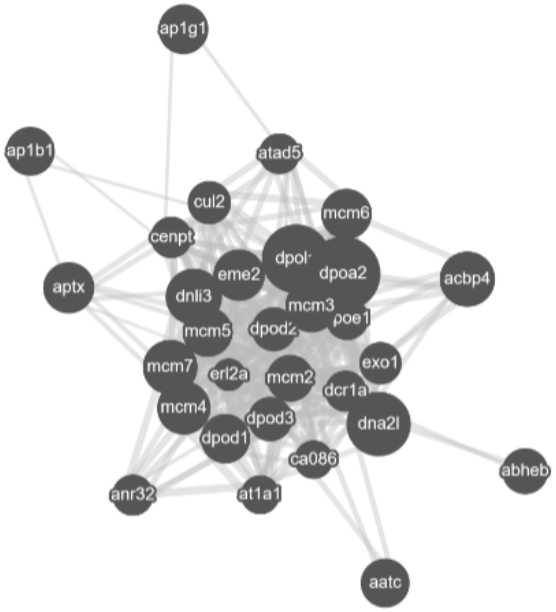

Figure 34: Random walk result for the MCM helicase. Shown is the resulting cluster graph of the Random walk algorithm based on the unfiltered DSB network.
Query: MCM2-7; 15 nodes

the result of which can be seen in Figure 34. It shows a closed cluster of all MCM subunits as well as the DNA polymerases (Polδ and Polα) expected on licensed chromatin. Additionally the SMC5/6 loading factor Slf1 (as ANR32; Räschle et al. (2015)) is depicted as behaving similar to MCM4, 5 and 7 as well as Polδ subunit 1. This result suggests that the newly deployed DSB TOM network represents the recruitment of DSB repair proteins as well as the processes of chromatin licensing and the formation of the pre-RC in non-replicating extracts.

3.3.3 Analysis of the Replication Fork Collapse network

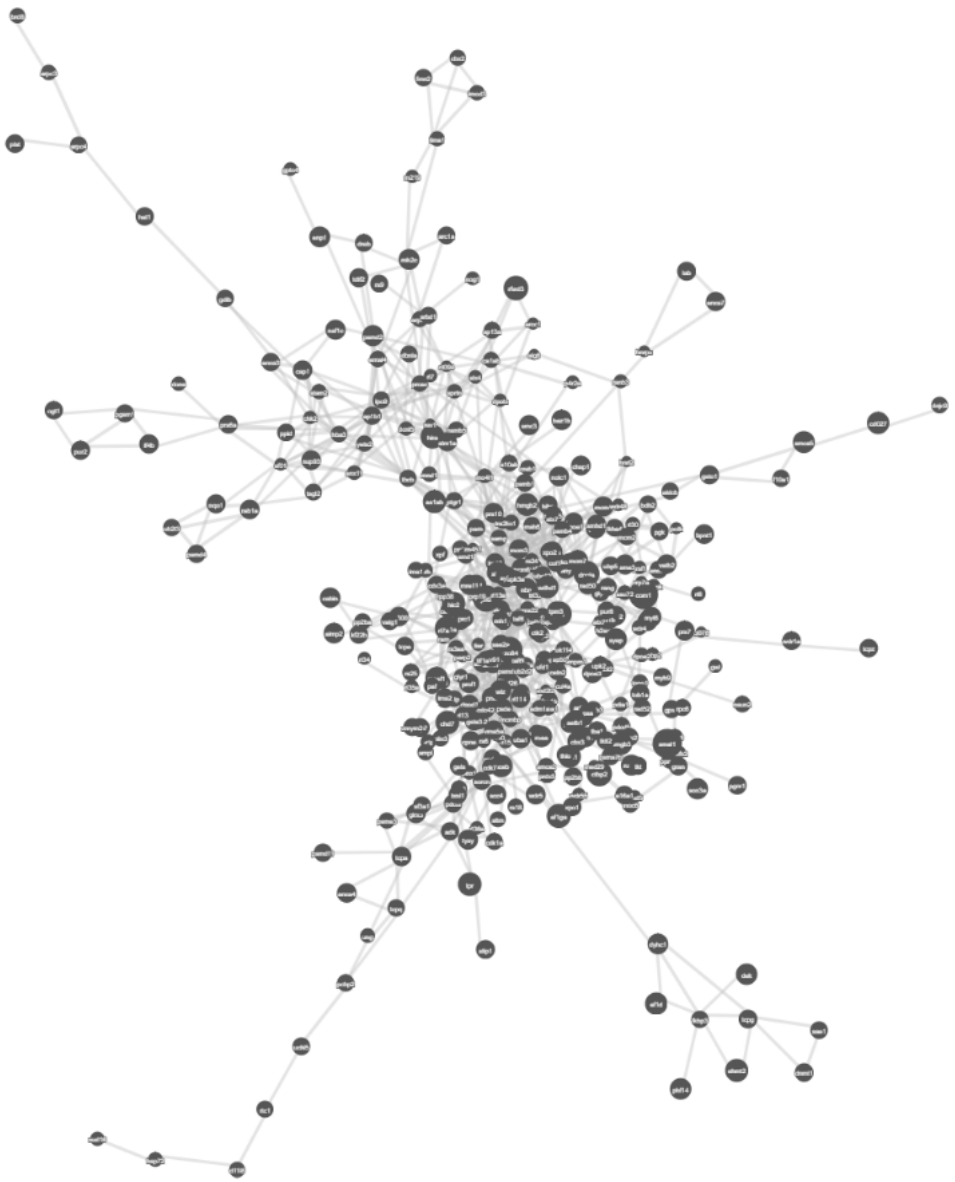

Figure 35: Fork Collapse TOM Network. Visual representation of proteins loaded onto chromatin during Replication Fork Collapse. Edges are filtered using a soft threshold of 70% and their width is proportional to their score. Node size in zoomed windows represents the protein enrichment scores.

During the replication of DNA the replisome encounters many obstacles such as DNA damages or sequences that are difficult to replicate. These obstacles can cause fork stalling that may lead to the collapse of the fork itself by replisome disassembly. In his review Cortez (2015) describes how the replication checkpoint prevents replication fork

stalling for example via the regulation of helicases like BLM, nucleases like DNA2 and EXO1 and translocases like SMARCAL1 via ATR. All of the mentioned enzymes fulfill roles in different repair mechanisms but all have in common that they are involved in fork regression or fork reversal (see Section. 1.3.1). The result of using these enzymes as a query to diffuse over the unfiltered Fork Collapse network can be seen in Figure 36. The highest node with the highest random walk score within the resulting cluster is ATR indicating that it connects most of the found nodes with another and therefore could play a crucial role in their recruitment or function. As is already known, direct targets of ATR are involved in fork reversal suggesting that the FC-TOM network represents the mechanism of fork collapse well. On the right side of the cluster one can see the translocase HUS1 (SMARCAL1) in close vicinity to the helicase BLM and ATRIP, the ATR interacting protein. These targets of ATR are closely connected to the high scoring repair proteins BRCA1, BRCA2 and APTX (XRCC1) that were also present and enriched in experiments investigating the replication-independent repair of DSBs (see Section 3.3.2). Given the fact that double- and singlestrand breaks are a common cause of replication fork stalling and also a result of proteasome dissociation (Cortez, 2015) the inclusion and high random walk- and enrichment scores for these factors is expected.

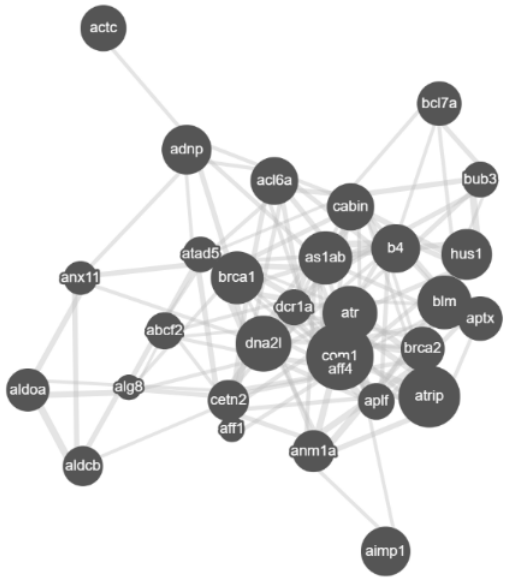

Figure 36: Random walk result for ATR, BLM, DNA2 and SMARCAL1 under Fork Collapse conditions. Shown is the resulting cluster graph of the Random walk algorithm based on the unfiltered FC network.
Query: ATR, BLM, DNA2(DNA2L), SMARCAL1 (HUS1); 30 nodes

3.3.4 Analysis of the Interstrand Crosslink network

The Interstrand Crosslink network is built on the person correlation based TOM of NPE/HSS using Psoralen-treated sperm chromatin. Under Psoralen treatment interstrand crosslinks are introduced at random into the chromatin template and mostly repaired via the Fanconi Anemia (FA) pathway. The mechanism of ICL bypass and repair in *Xenopus* egg extracts has been resolved in high detail by Räschle et al. (2015).

Replication of Psoralen-treated chromatin starts with the loading of all necessary replication proteins followed by triggering a transient checkpoint response. After encountering an ICL replicative DNA polymerases are unloaded while TLS polymerases as well as the entire Fanconi core complex are loaded. Additionally the nucleases XPX and FAN1 are loaded together with the FA core complex that is loaded by a ubiquitylated FANCD2/FANCI dimer (Räschle et al., 2015; Knipscheer et al., 2009; Räschle et al., 2008).

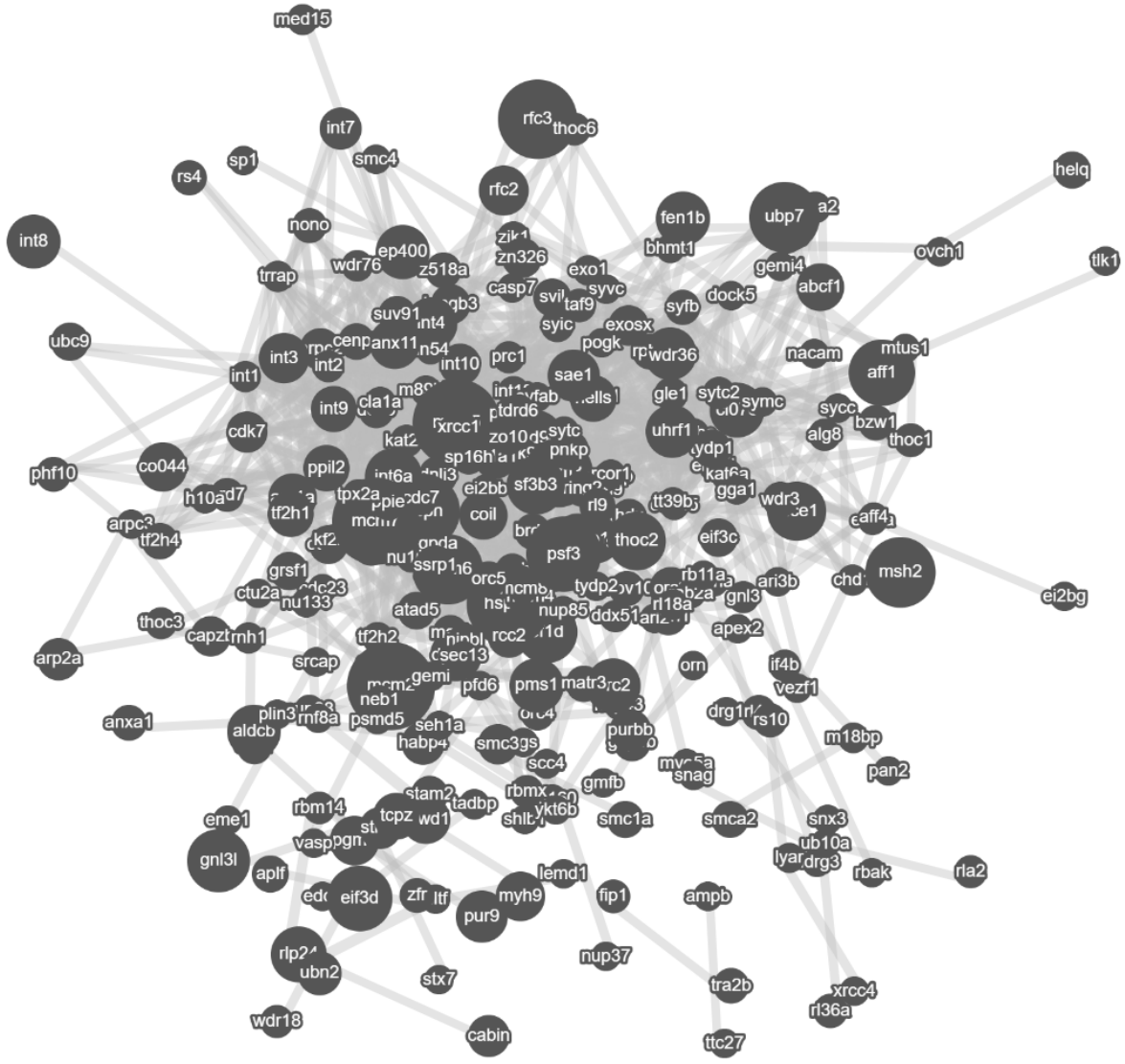

Figure 37: Interstrand Crosslink TOM Network. Visual representation of proteins loaded onto chromatin under Interstrand Crosslink conditions. Edges are filtered using a soft threshold of 70% and their width is proportional to their score. Node size in zoomed windows represents the protein enrichment scores.

Using the FA core complex as the query to diffuse over the ICL network we expected to find most repair enzymes as well as TLS polymerases in the resulting cluster. Surprisingly, of the FA loading complex dimer only FANCI could be detected. Otherwise all proteins of the FA core complex are closely connected to the helicase BLM as well as ATR and BRCC3. The latter is a subunit of an E3 ubiquitin ligase which is involved in the recruitment of BRCA1 to DNA lesions. BRCA1 is also found in the resulting cluster being a direct neighbor to FANCC and FANCA as well as ATRIP and ATR. Given the slightly shifted peaks of FA and its recruiters as well as the TLS polymerases and FA itself it is possible that the proteins expected to be included in the first cluster are not in the list of 50 highest scoring neighbor nodes we defined. Outputting a cluster consisting of 100 nodes for example will not be plotted because the server takes to long to answer the users request and the script runs into a hard coded timeout that can not be changed.

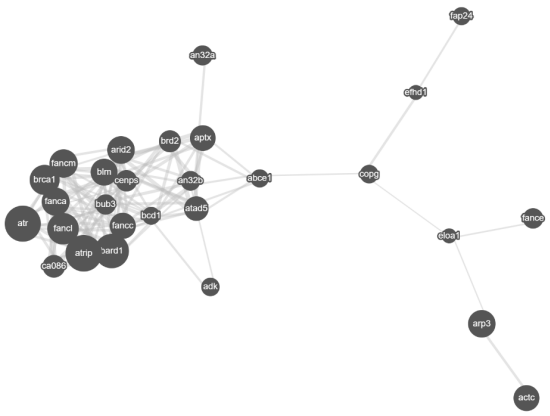

Figure 38: Random walk result for the Fanconi core complex under Interstrand Crosslink conditions. Shown is the resulting cluster graph of the random walk algorithm based on the unfiltered ICL network.
Query: GO:0043240; 30 nodes

To circumvent this problem we repeated the random walk using the FA recruiter FANCD2 and POL$\varkappa$, a TLS polymerase, as the query proteins and plotting the 50 highest scoring proteins. Approaching the search for ICL repair related proteins from the angle of known regulatory proteins resulted in the expected group of proteins.

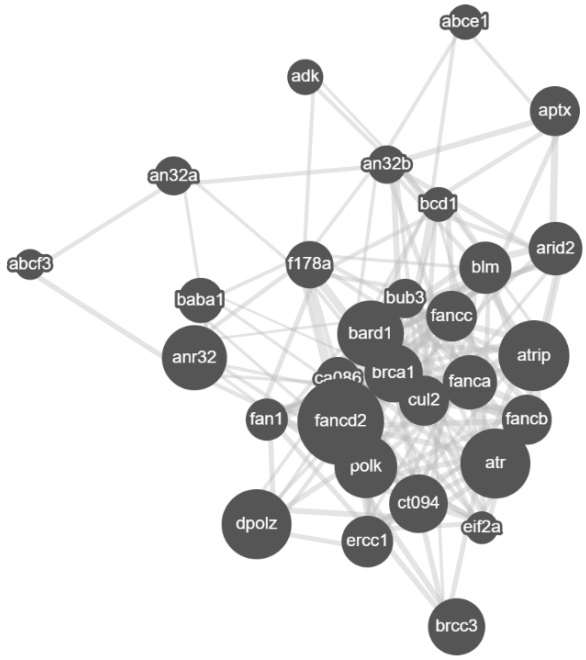

Figure 39: Random walk result for FANCD2 and POLχ under Interstrand Crosslink conditions. Shown is the resulting cluster graph of the random walk algorithm based on the unfiltered ICL network.
Query: FANCD2, POLχ; 30 nodes

Using FANCD2 and POLχ to diffuse over the ICL network yielded a cluster of highly enriched enzymes known to be involved in ICL repair. Aside from the DNA binding subunits of the FA core complex (FANCA and FANCB) the repair associated helicase BLM was found in direct neighborhood of FANCC together with ATR and some subunits of the BRCA1-associated genome surveillance complex (BASC) as it was the case for the diffused cluster using the FA core complex. Other enzymes involved in doublestrand break repair such as HUS1 and ATRIP are also part of the cluster which is to be expected considering that it is possible for ICLs to be bypassed using a combination of translesion synthesis and doublestrand break repair (see Section 1.3.2 for details). Another highly enriched protein included in the cluster is FAN1 which is a nuclease found to be associated with the FA core complex and FANCD2 thought to act together with MUS81 cut covalently crosslinked DNA (O'Donnell, Durocher, 2010). Another interesting protein in the cluster is DPOLZ (also known as REV3), the catalytic subunit of DNA polymerase zeta. It is known to be involved in DNA repair, especially in the replication-dependent resolution of doublestrand breaks and therefore it plays a major role in bypassing covalent ICLs (Räschle et al., 2008) as REV3 depleted *Xenopus* egg extracts have been shown to be ICL repair deficient (Räschle et al., 2008).

3.3.5 Analysis of the Replication Termination network

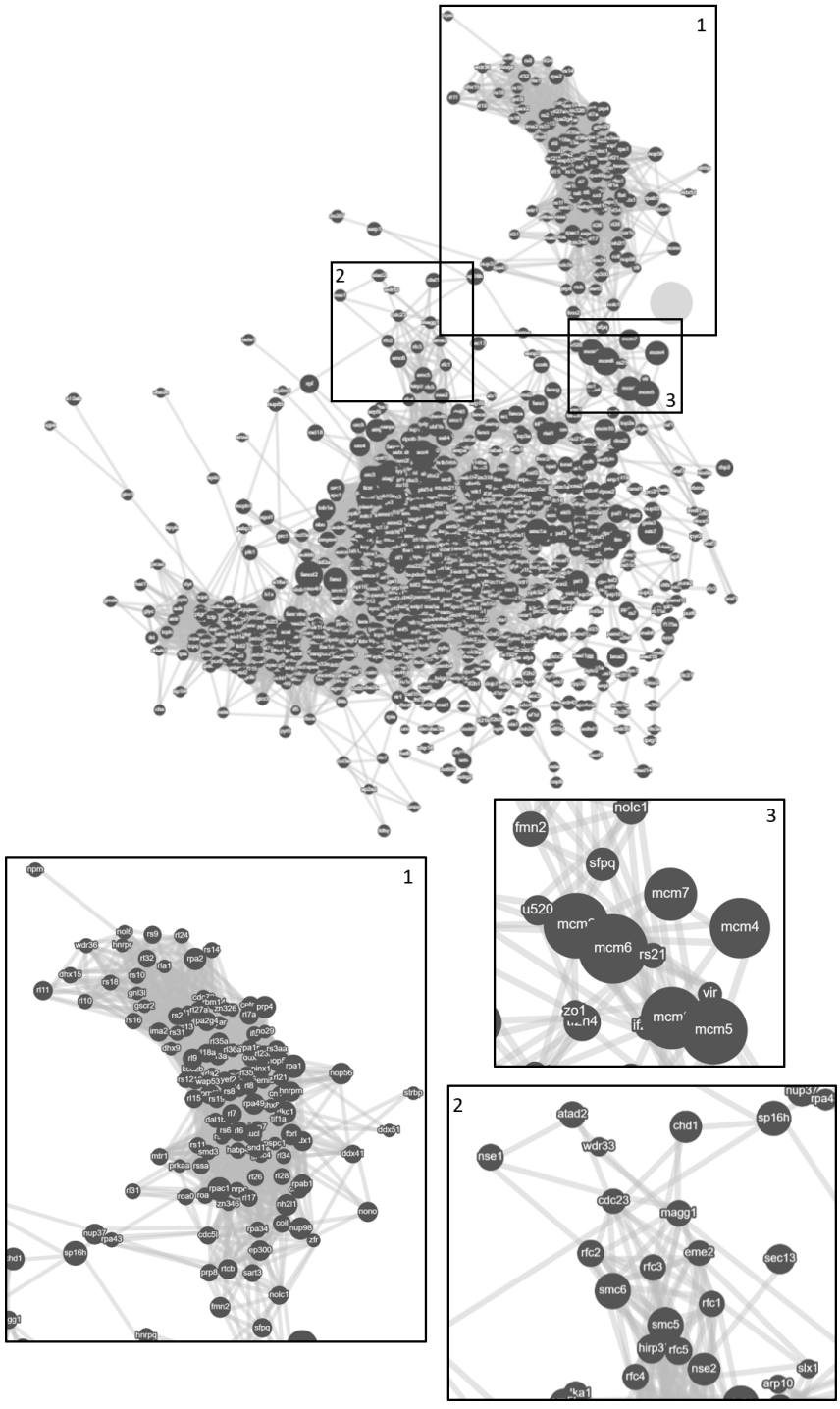

Figure 40: Replication Termination TOM Network. Visual representation of proteins loaded onto chromatin during Replication Termination. Edges are filtered using a soft threshold of 70% and their width is proportional to their score. Node size in zoomed windows represents the protein enrichment scores.

The network representation of chromatin binding profiles for repair and replication factors during Replication Termination differs greatly from the other damage specific networks due to the inclusion of only one comprehensive experiment looking at one defined time point after the start of replication in this collection. The single included experiment used the PP-MS system and a plasmid template capable of synchronous replication over all plasmid molecules enabled by blockage of converging forks using 16 *lacO* repeats and *lacI* without stalling them (Dewar et al., 2015). This means that the network representation does not show any internal batch effects and can indeed be considered scale free (see Figure 28) and its filtered visualization on the DRA is more accessible to the user than, for example, the ICL network.

As can be seen in Figure 40 there are distinct groups of enriched proteins visible in the network representation. Group 1 consists mostly of ribosomal proteins as well as subunits of the nuclear pore complex. These proteins are present in all PP-MS and especially CHROMASS experiment sets due to the method of isolating chromatin bound protein and their nuclear localization but are especially enriched in this set due to the low number of proteins involved in replication termination in general. The relative amount of proteins present in the sample *by chance* increases as the number and abundance of proteins loaded onto DNA decreases. Group 2 shows most subunits of the SMC5/6 complex (SMC5/6, NSE2,NSE1 and NSE3 (MAGG1) in close proximity to replication factor C (RFC1-5).

Dewar et al. (2015) showed that most the CMG helicase dissociates relatively late during termination while MCM7 unloads from DNA together with POLη even later than that. Given these findings we used the CMG helicase as our query to diffuse over the unfiltered RT network (Figure 41).

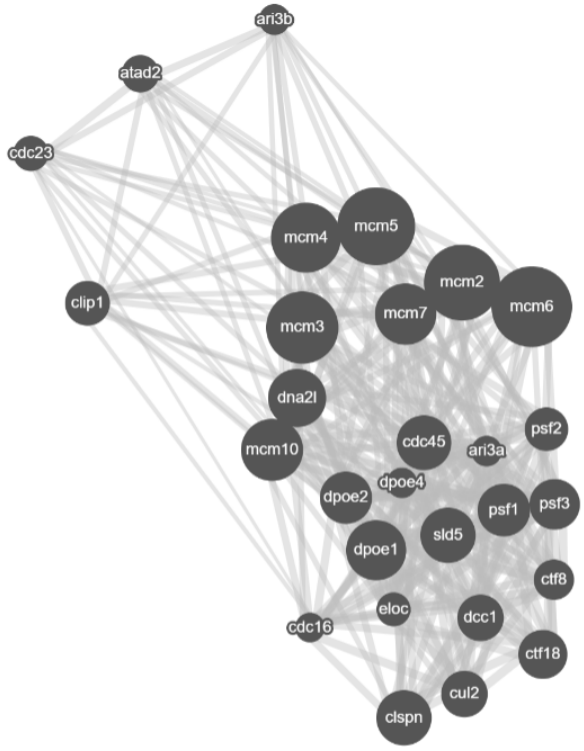

Figure 41: Random walk result for the CMG helicase during Replication Termination. Shown is the resulting cluster graph of the random walk algorithm based on the unfiltered RT network.
Query: MCM2-7, CDC45, SLD5, PSF1-3; 50 nodes

The resulting cluster includes all subunits of the CMG helicase and shows a high similarity between them. Additionally, all subunits of DNA polymerase ε are included which is expected due to its leading strand elongation function. Given the experimental setup of converging replication forks on a plasmid it is expected to find highly enriched leading and lagging strand DNA polymerases. CDC45 clusters closely with the GINS complex (SLD5 and PSF1-3) as well as the DNA replication initiation protein MCM10, possibly hinting on its proposed function as a fork stabilizing factor in *Xenopus* extracts (Chadha et al., 2016). Additionally, CUL2 and ELOC can be found in this cluster which are part of the E3 ligase that ubiquitylates MCM7. Missing from this cluster are the additional subunits ELOB and LLR1. The remaining included proteins are either known to be associated with already mentioned enzymes or artifacts of the number of nodes not being optimized (upper left side).
Given the findings of Maric et al. (2014), the disassembly of the CMG helicase consisting of CDC45 as well as the MCM and GINS complexes at the end of DNA replication is mediated by CDC48 (in our system known as p97) and the ubiquitin ligase SCF^{Dia2}.

SCF^{Dia2} specifically has been shown to ubiquitylate the MCM helicase on its MCM7 subunit, driving dissociation and disassembly of the CMG helicase complex in yeast. Using p97 or CDC48 and its indirect target MCM7 as our query to diffuse over the RT network yielded the results shown in Figure 42.

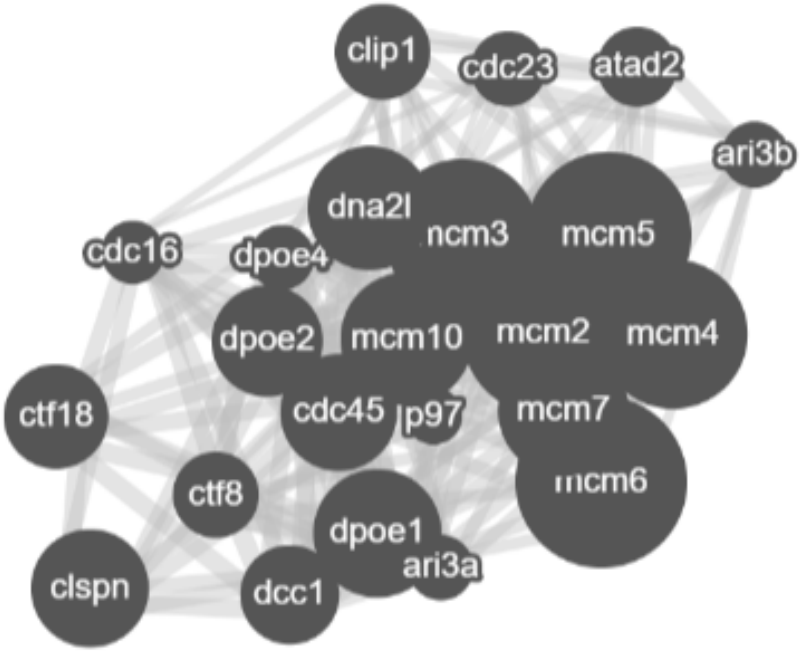

Figure 42: Random walk result for the CMG helicase during Replication Termination. Shown is the resulting cluster graph of the random walk algorithm based on the unfiltered RT network.
Query: MCM7, p97 (CDC48); 25 nodes

During termination of DNA replication one would expect p97 to be enriched and behave similar to its indirect target MCM7. In the resulting random walk cluster a clear relationship between MCM7 and CDC45 can be seen as well as high similarities of these two factors with other cell cycle controlling (CDC) enzymes. A direct connection between p97 and MCM7 or CDC45 could be detected using the TOM network under RT conditions indicating that the network represents replication termination mechanisms reasonably well.

4 Discussion

In this thesis we presented a user friendly web resource for the visualization of functional DNA repair modules as well as features for detailed mining of chromatin proteomic (MS) data collected using *Xenopus* cell free egg extract systems and the MS based CHROMASS and PP-MS methods described in the introduction to this work. A selection of published and unpublished data sets had been pre-processed using the program Perseus and the software environment for statistical computing R. Label-free quantification values calculated by the MS analysis software MaxQuant were standardized as Z-scores and grouped by experiments. A dimensionality reduction algorithm, namely t-SNE, was applied to the data to validate its integrity and comparability by ruling out possible batch effects from the combination of different experiments of such diversity. Although we could see some effects of this diversity, the majority of differences could be explained as effects of the use of different DNA templates (Figure 21). Due to this result we decided to process individual experiments grouped by the DNA repair mechanism they investigated. All further steps of this work were conducted using individual matrices for each DNA lesion type while a combination of all sets into one large repair network was ommited.
After validating data integrity and comparability pairwise Pearson's correlation was performed for each individual set of experiments investigating a specific DNA lesion. A soft-threshold was applied using a sigmoid adjacency function to filter the correlation matrices from which the topological overlap measures (TOM) were computed using the R package WGCNA.

The functional programming language F# was used to process the data further by creating new and optimizing existing functions to fit record types based on a user's input. Transformed data was then formatted as JSON strings and visualized client-sided via the use of JavaScript libraries to reduce the computational load of the web server. From this an already existing reference network based on data from all included experiments and a filtered list of 266 known DNA repair factors could be visualized. Details about this reference network can be found in Menges (2018).
To expand on this feature, reference networks based on the mentioned TOM matrices for each individual DNA repair pathway were added using the same functional framework. To lower the computational effort needed to plot the networks themselves, the list of possible node proteins was filtered using enrichment scores calculated individually from the p-values of statistical comparisons between experimental conditions within

each experiment set (Section 3.2.5). To automate this the R package *samr* was used which defines the Nullhypothosis of each comparison based on the input data and a non-linear regression coefficient. This set of enrichment scores builds the base for the **Protein Search** function of the DRA because these scores represent the involvement of each measured protein in the individual DNA repair mechanisms independent from the network representations.
The Enrichment Scores were also used to define the highest enriched proteins in comparison to their respective control over all experiments. This yielded a list of 1747 proteins that were used to filter the TOM networks visualized on the user-accessible front-end of the resource. The resulting networks can be seen in Figures 30, 32, 35, 37 and 40 respectively. When visualized using cytoscape.js the size of each node corresponds to the enrichment score calculated computed via the use of SAM-optimized t-Tests while edge width is defined by its weight or TOM.
Visualization of the filtered networks shows varying degrees of node cluster formation. While Figure 30 and 37 show a tendency for the nodes to group together in the middle of the graph with a few outliers, Figures 32, 35 and 40 show distinct node clusters representing known biological interactions. An exceptional example of this is the node cluster labeled as **1** in Figure 40 that includes small and large ribosomal subunit proteins as well as subunits of the nuclear pore. These proteins are clearly distanced from the remainder of the network which mainly contains DNA repair and replication factors that can for example be seen in the clusters labeled as **2** and **3** in the same graph. Thereby, TOM networks based on simpler experiment sets could be shown to be visualized in a way that allows the user to see relationships between certain groups of proteins without the addition of more computational steps. Through the option of mapping different enrichment scores on the network it is possible to see the behavior of certain proteins over all repair pathways and therefore compare the profiles of each network with one another. This shows the importance of complex visualization algorithms as it is not possible to draw such conclusions from the raw correlation or TOM matrix.

For networks where it is not possible to identify clusters by simple means of observation or if more detailed information about possibly filtered relationships between proteins is wanted, the two clustering methods mentioned above have been adapted to work on the newly added TOM networks. These allow the interactive drawing of data based subnetworks, or clusters using user defined proteins as a starting position. Menges (2018) already showed that the Neighborhood Search was capable of identifying the correct number of protein complexes when starting from a single protein. It was

also shown that the Neighborhood Search is highly error prone due to it only plotting direct neighbors. In addition to that the combination of multiple neighborhood searches was shown to be even more effective in identifying functional protein modules.
To improve on this an algorithm based on Google's PageRank (Taher Haveliwala et al., 2003) was implemented. This algorithm uses the principle of network propagation and has been shown to find the proteins missing from neighborhood search results for known protein complexes. The main benefit of this algorithm is its use of a list of proteins as the input query whereas neighbor hood searches only work on one protein at a time.
The performance of this algorithm given the input matrices was determined by the use of protein complexes known to be involved in the specific damage repair pathway as the input for the algorithm and diffuse over each damage set individually. The results of those searches were then compared with published interactions and repair models to validate that each network represents its respective repair pathway adequately. It was proposed that the random walk with restart algorithm could prove to be able to give novel insights to DNA repair pathways by amplifying weak signals through the use of custom scores such as somatic mutation frequencies (Leiserson et al., 2015). To this end, a collaboration with Prof. Dr. Maik Kschischo from the University of Applied Sciences in Koblenz was established with the goal of expanding the DNA Repair Atlas to be used for the identification of novel factors. Though each network represents all known repair and replication complexes well, the performance of the clustering algorithms in identifying all connected nodes varies with each network. This is most likely an artifact of the heterogeneous nature of the sets especially in regards to the make-up of individual treatments and conditions within each set. The DSB and ICL networks for example represent their respective repair pathways well with all proteins involved in DSB repair, or the Fanconi Anemia pathway respectively, could be identified using simple input queries (see Figures 33 and 39).
The same is true for the replication termination network, though here only one time point was included in the experiment set and some factors could not be identified due to the nature of the experimental methods. Additionally, the RT network included a lot of proteins that could be considered ”noise” in the context of chromatin proteomics solely due to their nuclear localization. In contrast to this, the DPC network (Figures 30) does not directly represent experimental findings regarding the repair of DPCs. Though it is generally possible to connect functionally similar proteins together, normalization and statistical artifacts caused by the comparison of *Xenopus* extracts at different time points lead to the inability to connect known regulatory factors to their targets reliably.

In this set, an untreated control was measured at one time point and compared to a time series of treated samples starting 5 minutes after the control. While this is absolutely not a problem for traditional methods of data analysis such time shifted controls can greatly influence the representation of protein relationships in a network visualization.

In conclusion, the newly generated TOM networks represent their respective repair pathways reasonably well but networks based on a high number of experimental conditions using synchronous time courses seem to perform better. Additionally, the implementation and usage of a suitable wavelet based correlation analysis (Heidari et al., 2016; Zhao et al., 2018) instead of simple Pearson's correlation coefficient could yield networks that retain more time-dependent information and therefore improve clustering.
The use of TOM as a measure for neighborhood similarity for the random walk with restart algorithm yielded similar if not better results in comparison to the combined correlation network. Using this measure and optimizing the input query could yield novel interactions between known proteins, although improvements should be made in the way missing values are handled before calculating the correlation matrix. The implementation of sophisticated filtering algorithms that adapt to each set of experiments and corresponding expectations could also reduce the number of identified similarities caused by artifacts of normalization and imputation.

Another addition made to the DNA Repair Atlas was a function that gives the user insight on the enrichment of his or her protein of interest over all included data sets. This function is based on the previously mentioned enrichment scores calculated using SAM-optimized t-Test results that are also used to style the final network plots on the front-end. A function has been implemented that takes a list of input proteins and returns a record type including the enrichment score for all five DNA lesion types included in the DRA. Using JavaScript libraries this information is then provided in a polar plot and gives first insight in the behavior of a protein of interest. From there, the user can start searching for interactors using the TOM network the protein of interest scores highest.

Finally, the ability to search for GO terms and UniProtIDs instead of the included protein/gene names was implemented. This function takes an input string and tries to match it against a list of known annotations such as GO terms and in some cases additional names. This feature is especially useful if one wants to see the enrichment

or raw abundance of a large group of proteins at once. It also allows the use of one GO term as an input query for the random walk with restart algorithm. The algorithm will then look for all proteins in the data base that are annotated with this GO term and include every match in the input list for the random walk. This allows the quick and efficient analysis of the behavior of a whole functional group in either the combined or one of the DNA lesion specific networks.

Lastly, the functional web-server built with Suave.IO to distribute the data in an easily modifiable and accessible way was adapted to fit the requirements for the newly implemented functions. The website itself was moved from a free tier Microsoft Azure server with shared resources to a virtual machine running locally and with full control for future administrators through the RHRK. Deployment is done simply by adding and modifying files in the Internet Information Service Control window inside the virtual machine via remote access from any computer within the local network of the TU Kaiserslautern. The final application can be accessed via the URL http://dnarepairatlas.bio.uni-kl.de/. Currently the resource is password protected until included data sets are published separately. The password can be requested by contacting the project lead Dr. Markus Räschle.

In conclusion the web resource delivers a combination of different approaches to visualize and analyze CHROMASS and PP-MS data sets in accessible ways. The user experience has been improved since the last revision and new features have been added that streamline the experience of using the DNA Repair Atlas. It is now possible to visualize complex relationships of repair proteins under specific DNA lesion conditions as well as relationships between repair proteins in DNA repair in general. The implementation of mutation frequencies to enable a random *weighted* walk with restart could enable the use of those network visualizations for the identification of novel repair factors with respect to their mutation rates in human cancer patients. Minor changes to the data and especially the application itself would be necessary to include such a feature that is currently in development with one of our collaborators. As the application running the web resource is written in a modular way, it is easy to add new data or functions and modify existing ones to for new requirements. Changes made in the structure of the plain text database especially improved the ability to add new data sets quickly and reliably.
The DNA Repair Atlas proved to be a reliable tool for the identification of protein interactions under specific DNA lesion conditions.

5 Acknowledgement

I'd like to thank the whole team of the department of Molecular Genetics at the Tu Kaiserslautern and especially Prof. Dr. Zuzana Storchová for enabling me to work in her lab. I want to thank Dr. Markus Räschle for introducing me to the DNA Repair Atlas and the ways of chromatin proteomics and of course for supervising this project and my thesis.
I also want to thank Jun.-Prof. Dr. Timo Mühlhaus for being the second reader of this thesis.

Another big "Thank You!" has to be provided to Paul Menges for helping me with F# and especially JavaScript and to Sushweta Sen for being an awesome teacher during my first experimental steps with the *Xenopus* systems. I want to thank Sushweta, Steffi, Paul, Kristina, Naren as well as Isabell and Robin and of course the rest of the lab crew for being just awesome people one can have a laugh with, even through tough times, be it personal or professional.
I thank all our Hiwis for their entertainment during lunch breaks and of course my favourite lab member, Nala the cute little fur ball.

Without the support of my friends and my family I would have never come this far considering where I started in terms of education. Thank you Max, Moritz, Laura, Rahel, Jörg, Emelie and especially Adriana for being there for me when I needed support, for pulling me through when I was close to throwing everything overboard and for just being there whenever I called - be it for a beer or to talk.

6 Literature

Alver Robert C., Chadha Gaganmeet Singh, Blow J. Julian. The contribution of dormant origins to genome stability: from cell biology to human genetics // DNA repair. 2014. 19. 182–189.

Amunugama Ravindra, Willcox Smaranda, Wu R. Alex, Abdullah Ummi B., El-Sagheer Afaf H., Brown Tom, McHugh Peter J., Griffith Jack D., Walter Johannes C. Replication Fork Reversal during DNA Interstrand Crosslink Repair Requires CMG Unloading // Cell reports. 2018. 23, 12. 3419–3428.

Aquiles Sanchez J., Wonsey D. R., Harris L., Morales J., Wangh L. J. Efficient plasmid DNA replication in Xenopus egg extracts does not depend on prior chromatin assembly // The Journal of biological chemistry. 1995. 270, 50. 29676–29681.

Bétous Rémy, Mason Aaron C., Rambo Robert P., Bansbach Carol E., Badu-Nkansah Akosua, Sirbu Bianca M., Eichman Brandt F., Cortez David. SMARCAL1 catalyzes fork regression and Holliday junction migration to maintain genome stability during DNA replication // Genes & development. 2012. 26, 2. 151–162.

Bhat Kamakoti P., Cortez David. RPA and RAD51: fork reversal, fork protection, and genome stability // Nature structural & molecular biology. 2018. 25, 6. 446–453.

Bhowmick Rahul, Minocherhomji Sheroy, Hickson Ian D. RAD52 Facilitates Mitotic DNA Synthesis Following Replication Stress // Molecular cell. 2016. 64, 6. 1117–1126.

Blow J. J., Sleeman A. M. Replication of purified DNA in Xenopus egg extract is dependent on nuclear assembly // Journal of cell science. 1990. 95 (Pt 3). 383–391.

Bönisch Clemens, Nieratschker Sonja M., Orfanos Nikos K., Hake Sandra B. Chromatin proteomics and epigenetic regulatory circuits // Expert review of proteomics. 2008. 5, 1. 105–119.

Broido Anna D., Clauset Aaron. Scale-free networks are rare // Nature communications. 2019. 10, 1. 1017.

treeClust: An R Package For Tree-Based Clustering Dissimilarities. // . 2015.

Campbell Neil A., Reece Jane B. Biology: A global approach. Boston: Pearson, 2015. 10. ed., global ed.

Carter Scott L., Brechbühler Christian M., Griffin Michael, Bond Andrew T. Gene co-expression network topology provides a framework for molecular characterization of cellular state // Bioinformatics (Oxford, England). 2004. 20, 14. 2242–2250.

Ceccaldi Raphael, Sarangi Prabha, D'Andrea Alan D. The Fanconi anaemia pathway: new players and new functions // Nature reviews. Molecular cell biology. 2016. 17, 6. 337–349.

Chadha Gaganmeet Singh, Gambus Agnieszka, Gillespie Peter J., Blow J. Julian. Xenopus Mcm10 is a CDK-substrate required for replication fork stability // Cell cycle (Georgetown, Tex.). 2016. 15, 16. 2183–2195.

Ciccia Alberto, Elledge Stephen J. The DNA damage response: making it safe to play with knives // Molecular cell. 2010. 40, 2. 179–204.

Cimprich Karlene A., Cortez David. ATR: an essential regulator of genome integrity // Nature reviews. Molecular cell biology. 2008. 9, 8. 616–627.

Cook Jeanette Gowen, Chasse Dawn A. D., Nevins Joseph R. The regulated association of Cdt1 with minichromosome maintenance proteins and Cdc6 in mammalian cells // The Journal of biological chemistry. 2004. 279, 10. 9625–9633.

Cortez David. Preventing replication fork collapse to maintain genome integrity // DNA repair. 2015. 32. 149–157.

Cortez David. Replication-Coupled DNA Repair // Molecular cell. 2019. 74, 5. 866–876.

Couch Frank B., Bansbach Carol E., Driscoll Robert, Luzwick Jessica W., Glick Gloria G., Bétous Rémy, Carroll Clinton M., Jung Sung Yun, Qin Jun, Cimprich Karlene A., Cortez David. ATR phosphorylates SMARCAL1 to prevent replication fork collapse // Genes & development. 2013. 27, 14. 1610–1623.

Cowen Lenore, Ideker Trey, Raphael Benjamin J., Sharan Roded. Network propagation: a universal amplifier of genetic associations // Nature reviews. Genetics. 2017. 18, 9. 551–562.

Cupello Steven, Richardson Christine, Yan Shan. Cell-free Xenopus egg extracts for studying DNA damage response pathways // The International journal of developmental biology. 2016. 60, 7-8-9. 229–236.

Dewar James M., Budzowska Magda, Walter Johannes C. The mechanism of DNA replication termination in vertebrates // Nature. 2015. 525, 7569. 345–350.

Dilley Robert L., Verma Priyanka, Cho Nam Woo, Winters Harrison D., Wondisford Anne R., Greenberg Roger A. Break-induced telomere synthesis underlies alternative telomere maintenance // Nature. 2016. 539, 7627. 54–58.

Dungrawala Huzefa, Rose Kristie L., Bhat Kamakoti P., Mohni Kareem N., Glick Gloria G., Couch Frank B., Cortez David. The Replication Checkpoint Prevents Two Types of Fork Collapse without Regulating Replisome Stability // Molecular cell. 2015. 59, 6. 998–1010.

Dytham Calvin. Choosing and using statistics: A biologist's guide. Chichester: Wiley-Blackwell, 2011. 3. ed.

Franke Lude, van Bakel Harm, Fokkens Like, Jong Edwin D. de, Egmont-Petersen Michael, Wijmenga Cisca. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes // American journal of human genetics. 2006. 78, 6. 1011–1025.

Gao Alan, Larsen Nicolai B., Sparks Justin L., Gallina Irene, Mann Matthias, Räschle Markus, Walter Johannes C., Duxin Julien P. Mechanism of replication-coupled DNA-protein crosslink proteolysis by SPRTN and the proteasome. 19. 2018.

Gillis Jesse, Pavlidis Paul. "Guilt by association" is the exception rather than the rule in gene networks // PLoS computational biology. 2012. 8, 3. e1002444.

Goodman Myron F., Woodgate Roger. Translesion DNA polymerases // Cold Spring Harbor perspectives in biology. 2013. 5, 10. a010363.

Gupta Dipika, Heinen Christopher D. The mismatch repair-dependent DNA damage response: Mechanisms and implications // DNA repair. 2019. 78. 60–69.

Gupta Surya, Turan Demet, Tavernier Jan, Martens Lennart. The online Tabloid Proteome: an annotated database of protein associations // Nucleic acids research. 2018. 46, D1. D581–D585.

Haahr Peter, Hoffmann Saskia, Tollenaere Maxim A. X., Ho Teresa, Toledo Luis Ignacio, Mann Matthias, Bekker-Jensen Simon, Räschle Markus, Mailand Niels. Activation of the ATR kinase by the RPA-binding protein ETAA1 // Nature cell biology. 2016. 18, 11. 1196–1207.

Hanada Katsuhiro, Budzowska Magda, Davies Sally L., van Drunen Ellen, Onizawa Hideo, Beverloo H. Berna, Maas Alex, Essers Jeroen, Hickson Ian D., Kanaar Roland. The structure-specific endonuclease Mus81 contributes to replication restart by generating double-strand DNA breaks // Nature structural & molecular biology. 2007. 14, 11. 1096–1104.

Heidari Zahra, Roe Daniel R., Galindo-Murillo Rodrigo, Ghasemi Jahan B., Cheatham Thomas E. Using Wavelet Analysis To Assist in Identification of Significant Events in Molecular Dynamics Simulations // Journal of chemical information and modeling. 2016. 56, 7. 1282–1291.

Horvath Steve. Weighted Network Analysis: Applications in Genomics and Systems Biology. New York, NY: Springer Science+Business Media LLC, 2011. 1.

Huang Jing, Liu Shuo, Bellani Marina A., Thazhathveetil Arun Kalliat, Ling Chen, Winter Johan P. de, Wang Yinsheng, Wang Weidong, Seidman Michael M. The DNA translocase FANCM/MHF promotes replication traverse of DNA interstrand crosslinks // Molecular cell. 2013. 52, 3. 434–446.

Huang Jing, Zhang Jing, Bellani Marina A., Pokharel Durga, Gichimu Julia, James Ryan C., Gali Himabindu, Ling Chen, Yan Zhijiang, Xu Dongyi, Chen Junjie, Meetei Amom Ruhikanta, Li Lei, Wang Weidong, Seidman Michael M. Remodeling of Interstrand Crosslink Proximal Replisomes Is Dependent on ATR, FANCM, and FANCD2 // Cell reports. 2019. 27, 6. 1794–1808.e5.

Iyer Ravi R., Pluciennik Anna, Burdett Vickers, Modrich Paul L. DNA mismatch repair: functions and mechanisms // Chemical reviews. 2006. 106, 2. 302–323.

Kermi Chames, Aze Antoine, Maiorano Domenico. Preserving Genome Integrity During the Early Embryonic DNA Replication Cycles // Genes. 2019. 10, 5.

Kile Andrew C., Chavez Diana A., Bacal Julien, Eldirany Sherif, Korzhnev Dmitry M., Bezsonova Irina, Eichman Brandt F., Cimprich Karlene A. HLTF's Ancient HIRAN Domain Binds 3' DNA Ends to Drive Replication Fork Reversal // Molecular cell. 2015. 58, 6. 1090–1100.

Klages-Mundt Naeh L., Li Lei. Formation and repair of DNA-protein crosslink damage // Science China. Life sciences. 2017. 60, 10. 1065–1076.

Knipscheer Puck, Räschle Markus, Smogorzewska Agata, Enoiu Milica, Ho The Vinh, Schärer Orlando D., Elledge Stephen J., Walter Johannes C. The Fanconi anemia pathway promotes replication-dependent DNA interstrand cross-link repair // Science (New York, N.Y.). 2009. 326, 5960. 1698–1701.

Kobak Dmitry, Berens Philipp. The art of using t-SNE for single-cell transcriptomics // Nature communications. 2019. 10, 1. 5416.

Kunkel Thomas A., Erie Dorothy A. DNA mismatch repair // Annual review of biochemistry. 2005. 74. 681–710.

Kustatscher Georg, Grabowski Piotr, Schrader Tina A., Passmore Josiah B., Schrader Michael, Rappsilber Juri. Co-regulation map of the human proteome enables identification of protein functions // Nature biotechnology. 2019. 37, 11. 1361–1371.

Langston Lance D., Mayle Ryan, Schauer Grant D., Yurieva Olga, Zhang Daniel, Yao Nina Y., Georgescu Roxana E., O'Donnell Mike E. Mcm10 promotes rapid isomerization of CMG-DNA for replisome bypass of lagging strand DNA blocks // eLife. 2017. 6.

Larsen Nicolai B., Gao Alan O., Sparks Justin L., Gallina Irene, Wu R. Alex, Mann Matthias, Räschle Markus, Walter Johannes C., Duxin Julien P. Replication-Coupled DNA-Protein Crosslink Repair by SPRTN and the Proteasome in Xenopus Egg Extracts // Molecular cell. 2019. 73, 3. 574–588.e7.

Lebofsky Ronald, Takahashi Tatsuro, Walter Johannes C. DNA replication in nucleus-free Xenopus egg extracts // Methods in molecular biology (Clifton, N.J.). 2009. 521. 229–252.

Leiserson Mark D. M., Vandin Fabio, Wu Hsin-Ta, Dobson Jason R., Eldridge Jonathan V., Thomas Jacob L., Papoutsaki Alexandra, Kim Younhun, Niu Beifang, McLellan Michael, Lawrence Michael S., Gonzalez-Perez Abel, Tamborero David, Cheng Yuwei, Ryslik Gregory A., Lopez-Bigas Nuria, Getz Gad, Ding Li, Raphael Benjamin J. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes // Nature genetics. 2015. 47, 2. 106–114.

Loeb Lawrence A., Monnat Raymond J. DNA polymerases and human disease // Nature reviews. Genetics. 2008. 9, 8. 594–604.

Lygerou Z., Nurse P. Cell cycle. License withheld–geminin blocks DNA replication // Science (New York, N.Y.). 2000. 290, 5500. 2271–2273.

Marians Kenneth J. Lesion Bypass and the Reactivation of Stalled Replication Forks // Annual review of biochemistry. 2018. 87. 217–238.

Maric Marija, Maculins Timurs, Piccoli Giacomo de, Labib Karim. Cdc48 and a ubiquitin ligase drive disassembly of the CMG helicase at the end of DNA replication // Science (New York, N.Y.). 2014. 346, 6208. 1253596.

Massey Thomas H., Jones Lesley. The central role of DNA damage and repair in CAG repeat diseases // Disease models & mechanisms. 2018. 11, 1.

Menck Carlos Fm, Munford Veridiana. DNA repair diseases: What do they tell us about cancer and aging? // Genetics and molecular biology. 2014. 37, 1 Suppl. 220–233.

Identification and visualization of DNA repair modules through network analysis of proteomic data. // . 2018.

Mijic Sofija, Zellweger Ralph, Chappidi Nagaraja, Berti Matteo, Jacobs Kurt, Mutreja Karun, Ursich Sebastian, Ray Chaudhuri Arnab, Nussenzweig Andre, Janscak Pavel, Lopes Massimo. Replication fork reversal triggers fork degradation in BRCA2-defective cells // Nature communications. 2017. 8, 1. 859.

Mutreja Karun, Krietsch Jana, Hess Jeannine, Ursich Sebastian, Berti Matteo, Roessler Fabienne K., Zellweger Ralph, Patra Malay, Gasser Gilles, Lopes Massimo. ATR-Mediated Global Fork Slowing and Reversal Assist Fork Traverse and Prevent Chromosomal Breakage at DNA Interstrand Cross-Links // Cell reports. 2018. 24, 10. 2629–2642.e5.

Neelsen Kai J., Lopes Massimo. Replication fork reversal in eukaryotes: from dead end to dynamic response // Nature reviews. Molecular cell biology. 2015. 16, 4. 207–220.

O'Donnell Lara, Durocher Daniel. DNA repair has a new FAN1 club // Molecular cell. 2010. 39, 2. 167–169.

Oldham Michael C., Langfelder Peter, Horvath Steve. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease // BMC systems biology. 2012. 6. 63.

Oliver S. Guilt-by-association goes global // Nature. 2000. 403, 6770. 601–603.

Peter Langfelder , Steve Horvath . WGCNA: an R package for weighted correlation network analysis // BMC bioinformatics. 2008. 1. 559.

Pommier Yves, Huang Shar-yin N., Gao Rui, Das Benu Brata, Murai Junko, Marchand Christophe. Tyrosyl-DNA-phosphodiesterases (TDP1 and TDP2) // DNA repair. 2014. 19. 114–129.

Räschle Markus, Knipscheer Puck, Knipsheer Puck, Enoiu Milica, Angelov Todor, Sun Jingchuan, Griffith Jack D., Ellenberger Tom E., Schärer Orlando D., Walter Johannes C. Mechanism of replication-coupled DNA interstrand crosslink repair // Cell. 2008. 134, 6. 969–980.

Räschle Markus, Smeenk Godelieve, Hansen Rebecca K., Temu Tikira, Oka Yasuyoshi, Hein Marco Y., Nagaraj Nagarjuna, Long David T., Walter Johannes C., Hofmann Kay, Storchova Zuzana, Cox Jürgen, Bekker-Jensen Simon, Mailand Niels, Mann Matthias. DNA repair. Proteomics reveals dynamic assembly of repair complexes during bypass of DNA cross-links // Science (New York, N.Y.). 2015. 348, 6234. 1253671.

Rechkoblit Olga, Gupta Yogesh K., Malik Radhika, Rajashankar Kanagalaghatta R., Johnson Robert E., Prakash Louise, Prakash Satya, Aggarwal Aneel K. Structure and mechanism of human PrimPol, a DNA polymerase with primase activity // Science advances. 2016. 2, 10. e1601317.

Sanchez J. A., Marek D., Wangh L. J. The efficiency and timing of plasmid DNA replication in Xenopus eggs: correlations to the extent of prior chromatin assembly // Journal of cell science. 1992. 103 (Pt 4). 907–918.

Semlow Daniel R., Zhang Jieqiong, Budzowska Magda, Drohat Alexander C., Walter Johannes C. Replication-Dependent Unhooking of DNA Interstrand Cross-Links by the NEIL3 Glycosylase // Cell. 2016. 167, 2. 498–511.e14.

Sparks Justin L., Chistol Gheorghe, Gao Alan O., Räschle Markus, Larsen Nicolai B., Mann Matthias, Duxin Julien P., Walter Johannes C. The CMG Helicase Bypasses DNA-Protein Cross-Links to Facilitate Their Repair // Cell. 2019. 176, 1-2. 167–181.e21.

Stingele Julian, Bellelli Roberto, Alte Ferdinand, Hewitt Graeme, Sarek Grzegorz, Maslen Sarah L., Tsutakawa Susan E., Borg Annabel, Kjær Svend, Tainer John A.,

Skehel J. Mark, Groll Michael, Boulton Simon J. Mechanism and Regulation of DNA-Protein Crosslink Repair by the DNA-Dependent Metalloprotease SPRTN // Molecular cell. 2016. 64, 4. 688–703.

Stingele Julian, Jentsch Stefan. DNA-protein crosslink repair // Nature reviews. Molecular cell biology. 2015. 16, 8. 455–460.

An Analytical Comparison of Approaches to Personalizing PageRank: Technical Report. // . 2003. 2003-35.

Teixeira-Silva Ana, Ait Saada Anissia, Hardy Julien, Iraqui Ismail, Nocente Marina Charlotte, Fréon Karine, Lambert Sarah A. E. The end-joining factor Ku acts in the end-resection of double strand break-free arrested replication forks // Nature communications. 2017. 8, 1. 1982.

Vaz Bruno, Popovic Marta, Newman Joseph A., Fielden John, Aitkenhead Hazel, Halder Swagata, Singh Abhay Narayan, Vendrell Iolanda, Fischer Roman, Torrecilla Ignacio, Drobnitzky Neele, Freire Raimundo, Amor David J., Lockhart Paul J., Kessler Benedikt M., McKenna Gillies W., Gileadi Opher, Ramadan Kristijan. Metalloprotease SPRTN/DVC1 Orchestrates Replication-Coupled DNA-Protein Crosslink Repair // Molecular cell. 2016. 64, 4. 704–719.

Weston Ria, Peeters Hanneke, Ahel Dragana. ZRANB3 is a structure-specific ATP-dependent endonuclease involved in replication stress response // Genes & development. 2012. 26, 14. 1558–1572.

Wu R. Alex, Semlow Daniel R., Kamimae-Lanning Ashley N., Kochenova Olga V., Chistol Gheorghe, Hodskinson Michael R., Amunugama Ravindra, Sparks Justin L., Wang Meng, Deng Lin, Mimoso Claudia A., Low Emily, Patel Ketan J., Walter Johannes C. TRAIP is a master regulator of DNA interstrand crosslink repair // Nature. 2019. 567, 7747. 267–272.

Yuan Jingsong, Ghosal Gargi, Chen Junjie. The annealing helicase HARP protects stalled replication forks // Genes & development. 2009. 23, 20. 2394–2399.

Zellweger Ralph, Dalcher Damian, Mutreja Karun, Berti Matteo, Schmid Jonas A., Herrador Raquel, Vindigni Alessandro, Lopes Massimo. Rad51-mediated replication fork reversal is a global response to genotoxic treatments in human cells // The Journal of cell biology. 2015. 208, 5. 563–579.

Zhang Bin, Horvath Steve. A general framework for weighted gene co-expression network analysis // Statistical applications in genetics and molecular biology. 2005. 4. Article17.

Zhao Yifan, Laguna Ramon C., Zhao Yitian, Liu Jimmy Jiang, He Xiongxiong, Yianni John, Sarrigiannis Ptolemaios G. A Wavelet-Based Correlation Analysis Framework to Study Cerebromuscular Activity in Essential Tremor // Complexity. 2018. 2018. 1–15.

van Dam Sipko, Võsa Urmo, van der Graaf Adriaan, Franke Lude, Magalhães João Pedro de. Gene co-expression analysis for functional classification and gene-disease predictions // Briefings in bioinformatics. 2018. 19, 4. 575–592.

van der Maaten Laurens, Hinton Geoffrey. Viualizing data using t-SNE // Journal of Machine Learning Research. 2008. 9. 2579–2605.