Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Maia Segura Wang, M.Sc.
born in San José, Costa Rica

Oral examination: 18 May 2015

Genome-wide approaches for identifying genes involved in the maintenance of genomic stability

Referees:
Dr. Lars Steinmetz
Prof. Dr. Robert Russell

Abstract

The maintenance of genomic stability and the repair of DNA damage are essential for the survival of all cells. Despite diverse pathways for repair of DNA lesions, different mutations can arise, ranging from Single Nucleotide Variants (SNVs) to larger Structural Variants (SVs). The processes that play a role in the formation of these alterations are not fully understood. In this thesis, I present two complementary approaches for accumulating genomic variants and for identifying pathways involved in the suppression of mutation formation using *Saccharomyces cerevisiae* (budding yeast) gene knockout strains.

First, using next-generation sequencing, I studied neutral variants through a mutation accumulation assay for up to 1800 generations. I used 47 yeast strains with known defects in DNA replication, repair and recombination pathways. In all strains, small insertions and deletions (indels) were more common than larger SVs (>50bp). Most mutations occurred in repetitive sequences, implicating replication based mechanisms and homologous recombination in the formation of genomic variants. Furthermore, the knockout of *MSH2* produced a hypermutable strain that acquired the highest number of indels. Moreover, the knockout of the genes *SWR1* and *ISW1*, involved in chromatin remodeling, resulted in strains with high number of deletions. These results suggest that defects in establishing a correct chromatin architecture may play a role in the formation of genomic variants.

I further performed a genome-wide screen for genes that suppress deletion formation under different drug treatments in the presence or absence of homologous repeats by using designed constructs. As expected, deletions occurred more often between repeats, in support of the frequent involvement of homologous recombination in the formation of chromosome rearrangements. In addition, I identified genes whose knockout led to increased levels of deletions. Among these, *IOC4* is of particular interest given that it belongs to the same chromatin remodeling complex as *ISW1*, identified in the neutral mutation accumulation assay. This provides further evidence that chromatin remodeling may be involved in preventing the occurrence of SVs. Furthermore, several meiosis-related mutants also showed increased levels of deletions, suggesting that meiosis proteins may have additional roles in the maintenance of genomic stability during vegetative growth. By performing additional experimental validations, I verified the higher vulnerability of meiosis gene knockouts to acquire deletions, especially in their diploid stages.

In the last chapter, I briefly describe the results of several side projects in which I applied computational methods learned through the above mentioned projects, to identify and characterize genomic rearrangements in different human cancers.

In summary, I have found that genome-wide approaches can provide interesting insights into the understanding of genomic variants in yeast and human cancers. In particular, given the evolutionary conservation of the ISWI chromatin remodeling complex and meiosis-related genes, the results presented here point to potentially novel functions of these proteins in the maintenance of genomic stability.

Zusammenfassung

Die Aufrechterhaltung der genomischen Stabilität sowie die Reparatur von DNA Schäden sind essentiell für das Überleben aller Zellen. Trotz unterschiedlicher Mechanismen für die Behebung von DNA Defekten, können verschiedene Arten von Mutationen, wie zum Beispiel Einzelnukleotid-Variationen (SNVs) oder Strukturvariationen (SVs), auftreten. Die Prozesse, die bei deren Entstehung eine Rolle spielen sind noch nicht komplett verstanden. In dieser Arbeit stelle ich zwei komplementäre Ansätze zur Anreicherung von genomischen Veränderungen sowie zur Identifikation von, für die Vermeidung von Mutationen wichtigen, Signalwegen vor. Für beide Ansätze verwendete ich *Saccharomyces cerevisiae* (Hefe) Gen-Knockout-Stämme.

Als erstes untersuchte ich, unter Verwendung von Hochdurchsatz-Sequenzierung, neutrale Mutationen anhand eines Mutations-Akkumulations-Experimentes über einen Zeitraum von bis zu 1800 Generationen. Dabei verwendete ich 47 Hefestämme, welche bekannte Defekte in DNA Replikations-, Reparatur-, und Rekombinationsmechanismen haben. In allen Stämmen waren Indels (kleine Insertionen und Deletionen) häufiger als große SVs (>50bp). Die meisten Mutationen traten in repetitiven Sequenzen auf, was auf eine Beteiligung replikations-basierender Mechanismen sowie homologer Rekombination bei der Entstehung genomischer Veränderungen deutet. Des Weiteren hatte der Knockout von *MSH2* einen hypermutablen Stamm zur Folge, welcher die meisten Indels akkumulierte. Außerdem zeigten die Knockouts von *SWR1* und *ISW1* eine große Anzahl von Deletionen. Beide Gene sind an Chromatin-Umstrukturierungen beteiligt. Diese Ergebnisse deuten darauf hin, dass Probleme beim Aufbau einer korrekten Chromatin-Architektur eine wichtige Rolle bei der Entstehung genomischer Veränderungen spielen könnten.

Als zweiten Ansatz habe ich ein genomweites Screening mit Hilfe eines speziell entworfenen DNA Konstruktes durchgeführt, um Gene zu identifizieren, welche die Entstehung von Deletionen unterdrücken. Hierfür wurden verschiedene Bedingungen, wie die Anwendung unterschiedlicher Wirkstoffe sowie die An- beziehungsweise Abwesenheit homologer Sequenzwiederholungen, getestet. Wie erwartet entstanden Deletionen häufiger zwischen Sequenzwiederholungen, welches die häufige Beteiligung homologer Rekombination bei dem Auftreten chromosomaler Veränderungen unterstreicht. Außerdem habe ich Gene identifiziert, deren Knockout zu einem erhöhten Level von Deletionen führte. Von diesen ist besonders *IOC4* interessant, da es zum gleichen Chromatin-Umstrukturierungs-Komplex wie *ISW1* gehört, welches im Mutations-Akkumulations-Experiment identifiziert wurde. Dies ist ein weiterer Hinweis darauf, dass die Chromatin-Umstrukturierung eine Rolle bei der Vermeidung von SVs spielen könnte. Des Weiteren zeigten Knockouts von in der Meiose beteiligte Genen ein verstärktes Auftreten von Deletionen, was darauf hindeutet, dass diese neben der Meiose auch bei der Erhaltung der genomischen Stabilität während des vegetativen Wachstums eine Rolle spielen könnten. Durch weitere experimentelle Validierungen konnte ich die erhöhte Anfälligkeit für

Deletionen bei Meiose-Gen-Knockout-Stämmen bestätigen, vor allem in ihrer diploiden Form.

Im letzten Kapitel dieser Arbeit gehe ich schließlich kurz auf mehrere Nebenprojekte ein, in denen ich bioinformatische Methoden, welche ich in den oben genannten Projekten gelernt hatte, angewandt habe, um genomische Veränderungen in verschiedenen Krebsarten von Menschen zu identifizieren und zu charakterisieren.

Zusammenfassend konnte ich zeigen, dass genomweite Ansätze interessante Einblicke in das Verständnis genomischer Veränderungen in Hefe und bei Krebserkrankungen geben können. Insbesondere in Anbetracht der evolutionären Konservierung des ISWI Chromatin-Umstrukturierungs-Komplexes und der in Meiose involvierten Gene weisen die hier präsentierten Ergebnisse auf mögliche neue Funktionen dieser Proteine beim Erhalt genomischer Stabilität hin.

Para ma, pa y las chicas,
que *siempre* me han apoyado...

Acknowledgments

My experience in EMBL will certainly be unforgettable and the best moments were possible due to the help and contributions of several people to whom I am really thankful.

First, I would like to thank my PhD supervisor Jan Korbel for giving me the opportunity to work in his lab. I appreciate his advise and his openness for scientific discussions. I also want to recognize the advice and suggestions from the members of my Thesis Advisory Committee, Lars Steinmetz, Christian Häring and Robert Russell.

It was of great benefit to be surrounded by so skillful people from which I could learn and discuss ideas. I am particularly thankful to Megumi, who introduced me to yeast genetics and had always very helpful suggestions and feedback. Special thanks to Balca, who was also open for discussions at all times, giving me really useful recommendations and ideas for the development of my yeast project.

I appreciate all the help I received from the computational people in the lab. I was able to learn a considerable amount of new things from them, especially from Thomas who was always available for helping and patient to answer my questions. His experience and knowledge, plus his willingness to share them, make him without a doubt one of the most valuable persons in the lab during my PhD.

I also want to thank Adrian, with whom I had a really enjoyable time discussing projects and learning a large number of wet lab procedures. His constant interest and ability to solve problems, his general skepticism and positive attitude were key to creating a very encouraging working environment (...even on Fridays).

I am also really happy to have shared the lab with Stephanie and Jelena. I have many things to thank them for, including sharing their ideas, support, constant feedback, and the large number of activities we did together, inside and outside of the lab. It was a lot of fun!

Many other people were very helpful as well during my time in EMBL. In this regard, I want to acknowledge the Genomics Core facility and the Lab Kitchen Staff for their high quality service and support.

I would like to thank my family, who has always backed up my ideas and decisions. Without them I would not have been able to pursue this degree and I appreciate their unconditional encouragement. And finally, I am very grateful to Thomas for his continuous motivation, understanding and optimism, KHK.

Table of Contents

List of Abbreviations

1000GP	1000 Genomes Project
5FOA	5-Fluoroorotic acid
alt-EJ	Alternative end joining
ARS	Autonomously replicating sequence
BIR	Break-induced replication
bp	Base pair(s)
Campt	Camptothecin
chr	Chromosome
CNS	Central nervous system
CNV	Copy number variant
DDR	DNA damage response
DelNoRep	Deletion construct without direct repeats
DelRep	Deletion construct with direct repeats
DKFZ	Deutsches Krebsforschungszentrum
DNA	Deoxyribonucleic acid
Doxo	Doxorubicin
DSB	Double strand break
DT	Downtag
EMBL	European Molecular Biology Laboratory
FDR	False discovery rate
FoSTeS	Fork stalling and template switching
GFP	Green fluorescent protein
GO	Gene ontology
GCR	Gross chromosomal rearrangements
HIP	Haploinsufficiency profiling
HR	Homologous recombination
HU	Hydroxyurea
Hyg	Hygromycin
Indel	Short insertion/deletion
kb	Kilobase(s)
KO	Knockout

LTR	Long terminal repeat
MA	Mutation accumulation
MB	Medulloblastoma
Mb	Megabase(s)
MEI	Mobile element insertion
MMBIR	Microhomology-mediated break-induced replication
MMEJ	Microhomology-mediated end joining
MMS	Methyl methanesulfonate
MMR	Mismatch repair
mRNA	Messenger RNA
NAHR	Non-allelic homologous recombination
NGE	Neighboring gene effect
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining
Noco	Nocodazole
ORF	Open reading frame
PCR	Polymerase chain reaction
qPCR	Quantitative PCR
RNA	Ribonucleic acid
SDSA	Synthesis-dependent strand annealing
SGA	Synthetic genetic array
SGD	*Saccharomyces* Genome Database
SMUFIN	Somatic mutation finder
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SSA	Single-strand annealing
ssDNA	Single stranded DNA
SV	Structural variant
tRNA	Transfer RNA
TSS	Transcription start site
UCSC	University of California, Santa Cruz
UT	Uptag
UTR	Untranslated region
WGS	Whole-genome sequencing
WT	Wild type

Chapter 1

Introduction

1.1 Motivation and outline of this thesis

The maintenance of genome integrity is an essential process for the survival of any organism. Throughout the life of a single cell, numerous endogenous and exogenous agents can damage DNA [Ciccia and Elledge, 2010]. This damage can impair essential processes, including DNA replication and transcription. If the damage is not repaired, it can give rise to mutations or large-scale genomic aberrations, which can be passed to the next generations. Therefore, to ensure genome stability, all organisms have evolved diverse mechanisms, generally known as the DNA damage response (DDR), for sensing and repairing damaged DNA and avoiding the inheritance of incorrect genetic information [Jackson and Bartek, 2009].

Nevertheless, DNA damage is inevitable and can result in changes ranging from Single Nucleotide Variants (SNVs) to larger Structural Variants (SVs). Many thousands of DNA lesions occur [Lindahl and Barnes, 2000] mainly because of byproducts of the normal cellular metabolism, as a consequence of radiation or other environmental factors, or even as intermediates of normal developmental processes such as those related to lymphocytes and germ cell formation [Jackson and Bartek, 2009]. Most of the time, these lesions are repaired, often causing no further damage. Defects in the mechanisms of detection or repair of DNA damage can cause alterations that eventually lead to uncontrolled cellular proliferation and the development of cancer and other diseases such as developmental disorders [Branzei and Foiani, 2010; Carvalho et al., 2010].

Even though these mutations often have negative effects on the phenotype, they are the ultimate source of variation. A major field of research in genetics is to understand how genetic variation occurs and how the underlying sequence variants give rise to certain phenotypic characteristics [Frazer et al., 2009]. A better comprehension of mutational processes is important to gain insights into the evolution of specific traits, *i.e.* why and how some phenotypes are adaptive, but also to understand disease processes, such as why some individuals are more susceptible to a particular disorder.

Given recent advances in tools for detection and analysis of genetic variation, such as next-generation sequencing technologies, we are now able to explore genomes more thoroughly than ever and to better understand the nature and impact of genetic variation. In this regard, model organisms, such as the budding yeast, *Saccharomyces cerevisiae*, are particularly useful for assessing the impact of molecular pathways on the maintenance of genome stability.

During the course of my PhD work, I was therefore interested in characterizing genetic variation in a wide range of genetic backgrounds. Using next-generation sequencing technologies and experimental approaches, my main aim was to improve the understanding of the processes involved in preserving genome integrity. For this purpose, I made use of a set of publicly available yeast deletion mutants, the Yeast Deletion Collection, each with a knockout of an open reading frame (ORF).

In the remainder of this chapter, I introduce basic concepts related to the content of this thesis, including several topics on genomic variation and the use of yeast as a model organism to study the maintenance of genomic stability.

Chapter 2 describes a mutation accumulation approach to study spontaneous mutations in a set of 47 knockout strains with defects in genes related to DNA repair, recombination, chromosome segregation and chromatin remodeling. In this approach, strains were passed through up to 90 recurrent single-cell-to-colony bottlenecks, consequently reducing the effects of selection and allowing neutral mutations to accumulate. By sequencing the strains at several time points, I investigated different characteristics of genomic variants, including their size, frequency, genome-wide distribution.

In **Chapter 3**, I describe an experimental approach using the budding yeast to study genomic variants in a broader collection of genetic backgrounds. I designed specific constructs to enrich for strains with defects in genes that are involved in preventing the formation of deletions. This led to the identification of several genes that may play a role in the maintenance of genomic stability.

These studies using the budding yeast allowed me to familiarize myself with the computational tools available for the identification and analysis of genomic sequencing data. Therefore, during my PhD I also participated in several side projects with the main aim of improving the understanding of SVs in human cancers. As part of these projects, I identified and performed experimental validations of SVs. A summary of my main contributions in these collaborations are mentioned briefly in **Chapter 4**.

Finally, in the last Chapter of this thesis I summarize my main results and conclusions, and give future perspectives on potential research directions. Detailed experimental procedures are described in **Appendix A**. Supplementary Figures and Tables are provided in **Appendix B**. A list of publications in which I was involved during my PhD is included in Appendix C.

1.2 Genomic variation

Even though members of the same species have essentially the same sets of genes, no two individuals, including closely related ones, are genetically identical. Genomic variation refers to these differences between DNA sequences within and among populations. While genomic differences between individuals tend to be small. They largely account for variation at the phenotypic level, from molecular differences in gene expression to distinctive characteristics in appearance, adaptation to environmental conditions, susceptibility to diseases and response to drugs [Wilson et al., 2001; Perry et al., 2007; The Wellcome Trust Case Control Consortium, 2007; Pickrell et al., 2010].

Genomic differences between individuals can occur at distinct scales, from changes in a single base pair (SNVs), to large genomic alterations, often called SVs. The latter can range from around 50bp to several megabases and can include whole chromosome gains or losses (aneuploidies). Additionally, between these two extremes, there are small insertions and deletions (indels) of up to 50bp in size. This size cutoff is mainly an operational definition. Earlier definitions set the size larger because sequencing technologies were not able to detect smaller variants as efficient as they do nowadays. The impact of genomic variants on phenotypic variation depends largely on their size and genomic location, and substantial investigations have been made to catalog variants in humans [The 1000 Genomes Project Consortium, 2010] as well as other organisms such as yeast [Cherry et al., 1998] in an effort to better understand their functional impact and effects on phenotype.

1.2.1 Single nucleotide variants (SNVs)

SNVs are the most common type of genomic variation between individuals. As an operational definition, SNVs are called single nucleotide polymorphisms (SNPs) if the frequency of the variant is 1% or greater in a population. As part of the 1000 Genomes Project (1000GP), 38 million SNPs have been discovered and genotyped in human genomes and they occur mainly as biallelic variants [The 1000 Genomes Project Consortium, 2012].

The distribution of SNVs in the genome is not homogeneous. Variants in coding regions are less frequent than in non-coding ones, mainly due to purifying selection acting against mutations with a negative effect on the phenotype as well as the relatively small size of the coding regions in humans. These variants in non-coding regions can affect regulatory elements, such as transcription factor binding sites or non-coding RNAs, or introns affecting the splicing of a gene. Additionally, SNV distribution is also influenced by the mutation and recombination rates along the genome [Nachman, 2001].

SNVs within coding regions can be synonymous if they do not change the sequence of the protein encoded by the gene, or nonsynonymous if they have an effect on the amino acid sequence of the protein. Nonsynonymous variants can be classified into missense and nonsense. The first type results in a change of amino acid, whereas the second one produces a premature stop codon at the variant site. Depending on the base change that an SNV causes, it can be referred to as a transition, if a purine is replaced by another purine, or if a pyrimidine is replaced by a pyrimidine. In contrast, if a purine is substituted by a pyrimidine, or the other way around, the variant is called a transversion. Transitions are in general more common than transversions, being the former twice as common as the latter in humans [Zhang, 2003].

SNVs can originate from errors during DNA synthesis by the DNA polymerase or after the mismatch repair processes. Environmentally induced DNA damage can also give rise to point mutations, not only by external conditions such as chemicals or radiation, but also by the cells' own environment. On the other hand, spontaneous mutations can occur due to the deamination of cytosine to uracil [Pfeifer, 2006]. In fact, the most common point mutation is the deamination of 5-methylcytosine resulting a C to T transition, which occurs predominantly at CpG islands and has been observed at high rates in human cancers, *e.g.* by Alexandrov et al. [2013].

1.2.2 Small insertions and deletions (indels)

Indels are also relatively frequent in the genome, making them the second most abundant type of variant in human genomes after SNVs [Mullaney et al., 2010]. However, their identification and cataloging has been more challenging than other types of variants. The 1000GP has detected around 2 million indels in human genomes [The 1000 Genomes Project Consortium, 2012; Montgomery et al., 2013], with an average of one every 8kb per individual. Indels are also abundant in other organisms, and they account for around 20% of all genetic variation in *Caenorabditis elegans* [Wicks et al., 2001] and *Drosophila melanogaster* [Berger et al., 2001]. On the other hand, in yeast, for example, where the genome is relatively stable, indels occur less frequently [Nishant et al., 2010]. Nevertheless, they still account for a high percentage of the variation in the yeast genome in comparison to other types of variants [Zhu et al., 2014].

Similar to the uneven distribution of SNVs in the genome, indels are also not randomly distributed. There are some hotspots of variation with increased indel occurrence compared to the chromosomal average [Mills et al., 2006]. These regions also contain higher numbers of SNVs. A set of smaller indel hotspots within genes was also identified, and similarly co-occurs with increased levels of SNVs [Mills et al., 2006]. There may be several reasons for these unusual regions of high genetic variation. For example, older segments of DNA with a longer evolutionary history have more time to accumulate mutations. Additionally, higher rates of homologous recombination and a lack of selective pressure if the variants have little impact on the fitness of the individual can result in increased levels of genetic variants [Montgomery et al., 2013].

Indels can arise due to slippage of the DNA polymerase during replication, especially in regions of highly repetitive sequences. This mechanism can explain almost 75% of indels in the human genome including the ones that occur in hotspots [Montgomery et al., 2013]. The remaining indels are most likely the result of a fork stalling and template switching (FoSTeS) mechanism which is associated with palindromic sequences and which is also related to the formation of larger structural variants [Montgomery et al., 2013].

1.2.3 Structural variants (SVs)

Larger rearrangements in the DNA sequence are termed SVs due to their overall effect on the structure of the genome. These include variants of different sizes. They can be unbalanced, such as deletions, duplications, insertions (both of novel sequences or of mobile elements, MEIs), which produce gains or losses of DNA and are therefore usually referred to as copy-number variants (CNVs). Additionally, there are also balanced large variants such as inversions and translocations, which do not change the overall content of DNA in the cells.

Due to their size, SVs account for a higher difference in nucleotides between individuals compared to SNVs and indels. In fact, in humans, up to 1% of sequence differences between two individuals are due to SVs compared to only around 0.1% for SNVs [Pang et al., 2010]. Consequently, these larger variants can have significant impact on phenotypic variation and evolution [Stankiewicz and Lupski, 2010; Zhang et al., 2009a].

Several studies have undertaken the task of identifying, cataloging and assessing the functional impact of SVs in human genomes [Pang et al., 2010; Conrad et al., 2010; Mills et al., 2011]. It has become clear that SVs, similar to SNVs, occur more frequently in some regions of the genome compared to others, which creates hotspots of recurrent variation [Mills et al., 2011]. Some of these regions also contain a higher amount of SNVs. Some factors that contribute to the clustering of mutations have been mentioned above, but additionally, they depend on the sequence context and the local genomic architecture [Stankiewicz and Lupski, 2002; Shaw and Lupski, 2004].

1.2.4 Mechanisms of formation of structural variants

There are diverse molecular mechanisms that can lead to structural variant formation. The natural processes by which cells acquire genetic variation are of biological importance. For example, DNA double strand breaks (DSBs) are required for proper meiotic recombination [Keeney and Neale, 2006] and for V(D)J recombination as part of the adaptive immune response [Jung et al., 2006].

Even though this variation can be beneficial, cells usually prevent DNA damage by a careful regulation of processes that involve DNA replication, recombination and repair. However, there may be some alterations that escape these controls and lead to the formation of mutations, including structural variants.

SV formation mechanisms have been operationally classified into different types. Although this classification is still partly in debate, to simplify this section, these mechanisms will be broadly categorized into two main groups: those that involve homologous sequences and those that are homology independent [Pâques and Haber, 1999; Hastings et al., 2009b]. Mechanisms that are homology dependent include those that require regions of sequence homologies that range from 50bp to 300bp, for *Escherichia coli* and mammals respectively [San Filippo et al., 2008; Liskay et al., 1987]. On the other hand, nonhomologous events can join DNA strands that are not complementary to each other, although in some cases they make use of very short sequences (microhomologies) to create the junctions [Pâques and Haber, 1999] (Figure 1.1).

In yeast, these homology independent mechanisms occur less often than the homology dependent ones. In fact, most of the DSBs in wild type yeast are repaired by homologous recombination. However, when this pathway is eliminated, other pathways are then activated for the fast repair of broken DNA strands, even though this process may be less efficient. Consequently, some yeast deletion mutants deficient in a repair process can still result in viable strains without severe phenotypes [Pâques and Haber, 1999].

Homology-mediated mechanisms. Homologous recombination is an important basis for several mechanisms that accurately repair damaged DNA using an identical sequence as a template. However, genomes contain repeated sequences that are not in the exact same chromosomal position in the homologous chromosome nor in the sister chromatid, but rather in different loci in the same or even in another chromosome [Hastings et al., 2009b]. This process, called non-allelic homologous recombination (NAHR), can be responsible for the formation of deletions, translocations, duplications and inversions and is believed to be a major source of rearrangements in cancer genomes [Hoeijmakers, 2001] (Figure 1.1A).

Homology dependent mechanisms include single-strand annealing (SSA), double-strand break repair (DSBR), synthesis-dependent strand annealing (SDSA) and break-induced replication (BIR).

If a DSB occurs between two flanking repeat regions, it can be repaired by homologous recombination in a process called single-strand annealing (SSA) [Sung and Klein, 2006] that leads to the deletion of a single copy of the repeated sequence (Figure 1.1B). In this case, the ends of the DSB are processed to create single stranded DNA (ssDNA) tails that can subsequently anneal to each other. Similarly, in the DSBR model, the broken chromosomes are also processed into ssDNA, which then invade the homologous

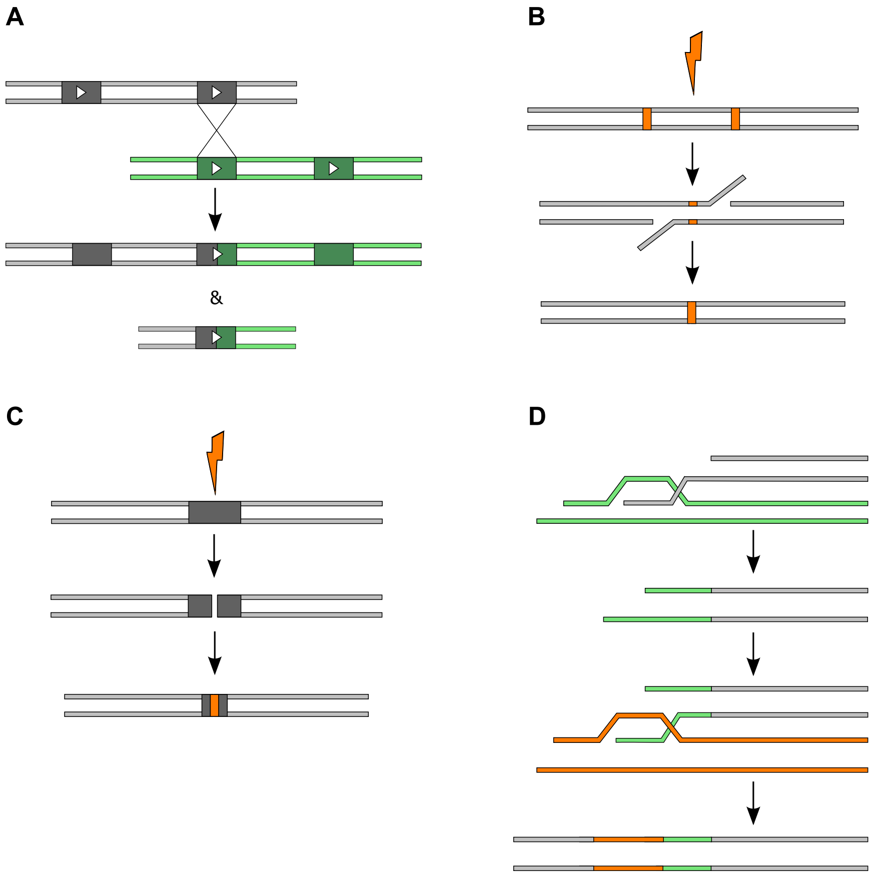

Figure 1.1: Mechanisms of formation of structural variants. **A** | Non-allelic homologous recombination (NAHR) can give rise to large deletions, translocations, inversions and duplications when recombination occurs between long sequence repeats (filled squares with white arrows). **B** | Single-strand annealing (SSA) depends on homologous sequences (orange squares) to repair DSBs. The removal of the single stranded flaps and the re-ligation of the broken ends creates a deletion of the sequence between the homologous sequences. **C** | Non-homologous end joining (NHEJ) is a non-homology-mediated mechanism to repair DNA DSBs. The rejoining of the DSB ends often leaves a repair "scar" in the form of small deletions or insertions. **D** | More complex mechanisms, like micro-homology mediated break-induced replication (MMBIR) and fork stalling and template switching (FoSTeS) can lead to the formation of simple and complex SVs with microhomologies at the breakpoints. They occur during replication and involve one or multiple rounds of single strand invasion into different replication forks. Modified from [Hastings et al., 2009a,b; Currall et al., 2013].

chromosome or an ectopic location to copy genetic information. This exchange of strands can result in a crossover with the exchange of segments of the two interacting chromosomes [Szostak et al., 1983].

On the other hand, in some cases, the DSBR process is not associated with a crossover. For this, the SDSA model has been proposed, in which the invading DNA strands and the newly synthesized one are separated from the template by a helicase and returned to the original broken molecule. This allows for the two ends to encounter and anneal to each other [Pâques and Haber, 1999], although it is possible that changes in copy numbers occur when the DNA template contains direct repeats [Hastings et al., 2009b].

Furthermore, if a DSB has only one end, *e.g.* when a replication fork breaks or collapses, the single end can then be repaired by the process of break-induced replication (BIR) [McEachern and Haber, 2006; Sung and Klein, 2006]. In this model, the single-stranded end is processed to form a ssDNA tail that invades a homologous sequence, and in this regard, it is similar to SDSA. However, it is independent of some proteins required for SDSA, such as *rad51* [Pâques and Haber, 1999]. Following the invasion, DNA is synthesized by copying information from the donor chromosome, and DNA synthesis continues to the end of the chromosome. If the donor is a repeated sequence located in a different chromosomal position, it can result in non reciprocal translocations, duplications or deletions [Hastings et al., 2009b].

Non-homology-mediated mechanisms Pathways of DNA repair that use limited or no homology also act to rejoin DNA molecules together. However, due to the lack of a homologous template, they are more likely to introduce errors. Non-homologous repair can be non-replicative or replicative. Among the non-replicative mechanisms, there are non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ) and breakage-fusion-bridge cycles.

NHEJ is one of the major ways to repair DSBs in eukaryotic cells, including V(D)J recombination and damage induced by ionizing radiation [Lieber, 2010]. NHEJ is used to rejoin DSB ends and proceeds by molecularly bringing the broken DNA ends together, then modifying the ends to make them compatible and finally ligating them [Gu et al., 2008; Lieber, 2010] (Figure 1.1C). The processing of the ends can lead to small deletions and insertions (1-5bp) due to the cleavage or addition of nucleotides to the ends. On the other hand, in MMEJ, short sequence homologies of 5-25bp are required to rejoin the DSB ends. This mechanism also leads to deletions of sequences between the microhomologies [McVey and Lee, 2008]. One main difference between MMEJ and NHEJ, is that the former is not dependent of classical repair factors required for NHEJ, such as DNA ligase IV, Ku70 and Ku80. Therefore, MMEJ as well as a few other methods independent of those classical factors are referred to as alternative end joining (alt-EJ) mechanisms [Lieber, 2010].

Breakage-fusion-bridge cycle is a mechanism implicated with the fusion of chromosome ends and plays a role in oncogene amplification in some cancers. Chromosomes that lack telomeres can fuse and create a dicentric chromosome. During anaphase the two centromeres are pulled to opposite poles of the cell, causing the breakage of the dicentric chromosome. At the same time, this process leads to the formation of new ends that also lack the telomeres, inducing their fusion and breakage once again, establishing a cycle until the ends are stabilized [McClintock, 1942].

DNA replication based mechanisms can also account for more complex chromosomal rearrangements. DNA replication can lead to the formation of microhomology junctions that, together with mechanisms similar to BIR, are involved in the formation of SVs with microhomologies at the breakpoints. Such a process has been termed microhomology mediated break-induced replication (MMBIR) [Hastings et al., 2009a]. A process that might be related to MMBIR, or is a subtype of this mechanism, is fork stalling and template switching (FoSTeS) (Figure 1.1D). It can occur when template switching happens between different replication forks [Zhang et al., 2009b]. Briefly, when replication forks stall, the 3' DNA end can change template to another ssDNA in a nearby replication fork, leading to the formation of SVs ranging in size from a few hundreds of base pairs to megabases. The junctions of the SVs typically show only microhomologies of 4-14bp, indicating that homologous recombination is likely not involved [Zhang et al., 2009b].

More recently, complex genomic rearrangements with multiple breakpoints have also been described and the chromothripsis model has been proposed [Stephens et al., 2011; Rausch et al., 2012a]. Chromothripsis, which was reported originally occurring in 2-3% of cancers, has also been observed in the germline [Kloosterman et al., 2011] and involves the acquisition of a large number of structural rearrangements in a single catastrophic event which may include the shattering of entire chromosomes. The mechanisms by which this is achieved are still not clear, but it may involve breakage-fusion-bridge cycles and potentially also NHEJ and alt-EJ repair mechanisms to rejoin the DNA ends.

1.3 Functional impact of genomic variation

Investigating the landscape of genomic variation is only one step towards understanding the effect of the variants on gene function and on the health of an individual. Large efforts have been made to determine the impact of mutations and the mechanisms by which variants occur. In general, the phenotypic impact of these variants depends on their size and their location, such that the larger the modified stretch of DNA, the higher the chance to affect an important region and to have higher damaging potential. Additionally, some variants occur in the germline, *i.e.* they are inherited from the parents, whereas others occur somatically, meaning that they are acquired after the formation of the zygote.

SNVs can have a direct effect on the phenotype if they change the amino acid sequence of a protein, if they cause a stop-gain or stop-loss or if they modify a splice site. Additionally, they can affect regulatory elements and promoter activities.

Although approximately 0.1% of genomic variation between the genomes of two individuals is due to point mutations, most of these variants are non deleterious. In humans, it has been reported that each individual carries around 2500 nonsynonymous variants at conserved positions [The 1000 Genomes Project Consortium, 2012]. However, some nonsynonymous SNVs can have a milder effect if the amino acid substitution is functionally similar to the original one. Additionally, most of these nonsynonymous variants are actually common in the population (with frequencies >0.5%), and it is therefore unlikely that they are of pathological importance [Frazer et al., 2009; The 1000 Genomes Project Consortium, 2012]. In highly conserved loci, the majority of the variants are present in very low frequencies, *i.e.* below 0.5%, because these sites of high evolutionary conservation are usually functionally important and can have negative consequences on the individual when disrupted [The 1000 Genomes Project Consortium, 2012].

Nevertheless, several SNVs are deleterious and are known to be responsible both for human disorders as well as other traits that are not pathological. SNVs can affect how an individual responds to drugs and the susceptibility to disease. Some of the best studied examples are mutations in *TP53* increasing the risk of cancer [Li and Fraumeni, 1969], or that can be directly related to the cause of an illness. Consistently, the cause of several monogenetic disorders has been mapped to single nucleotide mutations. This is the case of phenylketonuria [DiLella et al., 1986], Tay-Sachs disease [Myerowitz, 1997] and sickle-cell anemia [Marotta et al., 1977]. To predict the functional impact of SNVs, several algorithms have been developed, and the predictions have proved to be useful for characterizing the effects on the phenotype. In this regard, SIFT [Ng and Henikoff, 2003] and PolyPhen2 [Adzhubei et al., 2010] can predict the effect of amino acid substitutions on protein structure and function.

Furthermore, indels can affect the phenotype in similar ways as SNVs, modifying the amino acid composition of a protein or causing frame shifts which mostly result in a dysfunctional gene product. The functional impact of indels is also of great interest because they can alter phenotypic traits and are the cause of several diseases. For example, cystic fibrosis is one of the most common genetic disorders in humans and it is caused by indels in the *CFTR* gene [Collins et al., 1987]. Indels not occurring in gene coding sequences can also have severe effects. For example, insertions in the promoter of the *FMR1* gene can lead to Fragile X syndrome, but only if the size of the trinucleotide expansion reaches a certain threshold [Warren et al., 1987].

In the case of structural variants, the effects on the phenotype can be caused by the direct modification of the sequence of a gene or by disruption of regulatory elements. Additionally, positional effects can result from rearrangements that bring together genomic elements that are normally not interacting [Northcott et al., 2012a]. Copy number variants (CNVs), such as a duplication of a region, can lead to dosage effects, and translocations can create new fusion genes.

Several larger rearrangements are known to be the cause of human diseases. Long before the advent of sequencing technologies, through the use of microscopic techniques, many of these large rearrangements were identified and characterized, such as aneuploidies and fragile sites [Feuk, 2010]. More recently, many other smaller variants have been linked to different phenotypes. These include non-pathogenic traits, such as the copy number variation in the amylase gene (*AMY1*) [Perry et al., 2007], and other rearrangements associated with genomic disorders such as Williams-Beuren and Potocki-Lupski syndromes [Zhang et al., 2009a].

1.4 Structural variants in cancers genomes

Structural variants, as well as other types of somatic variations, have also been shown to play an important role in the development of different types of tumors [Pleasance et al., 2010a; Rausch et al., 2012a; Yang et al., 2013; Zack et al., 2013]. Cancer includes a diverse group of diseases, of over 100 different types with distinct characteristics. These include different risk factors, such as environmental conditions and genetic features.

Specific genomic abnormalities have been associated with particular cancers, and they encompass a wide range of rearrangements, including SVs and SNVs [Stratton et al., 2009]. In fact, some of the first recurrent rearrangements discovered in cancers were large events, like the translocation between chromosomes 9 and 22 (Philadelphia translocation) common in chronic myeloid leukaemia [Rowley, 1973] given that they were visible with staining techniques and microscopes.

With the advent of new sequencing technologies, the study of cancer genomics has been highly improved by the ability to sequence a high number of genomes in unprecedented coverage and timing. The increasing availability of data has allowed the characterization of the genomic landscapes of a wide variety of cancer types [Pleasance et al., 2010a; Stephens et al., 2012; Ho et al., 2013; Ojesina et al., 2014], revealing the large variation and high complexity of cancer genomes and the processes that govern cancer development [Berger et al., 2011].

Structural variants identified in human cancers include all simple types, like deletions, inversions, duplications and translocations [Campbell et al., 2008; Stephens et al., 2009; Yang et al., 2013]. It has been earlier proposed that cancers arise by an accumulation of somatic variants and that therefore its development follows a progressive model [Stratton et al., 2009]. In this model, the cells acquire consecutive mutations that can activate proto-oncogenes or inactivate tumor suppressor genes. Not all mutations are similarly important for tumorigenesis, rather there are some mutations that are “drivers”, they have an effect in tumor growth and are directly implicated in oncogenesis. Other less important mutations are “passangers” and may have no impact on the development of the tumor [Stratton et al., 2009].

As mentioned previously, more recent analyses of different cancer genomes have revealed that a large number of structural variants can arise by a different mechanism involving a one-step catastrophic event called chromothripsis [Stephens et al., 2011; Rausch et al., 2012a]. By this process, massive rearrangements are formed, occurring in only one or two chromosomes, in which several DNA fragments are lost and many others are rejoined in a random way. The derived chromosome shows variation in copy number changes that typically oscillates between two or three states [Korbel and Campbell, 2013].

The identification and annotation of structural variants in cancers has lead to the discovery of rearrangements that affect the coding regions of genes, either by removing part of the coding sequence or by creating the fusion of genes, such as the *TMPRSS2* fusion to the ETS transcription factors *ERG* or *ETV1* seen in prostate cancers[Tomlins et al., 2005]. Duplications can lead to changes in the gene dosage when the repeated units contain intact elements. High level amplifications are also characteristic of specific cancers and can lead to overexpression of oncogenes, *e.g.* the amplification of *MYC* in group 3 medulloblastomas [Bigner et al., 1990; Northcott et al., 2012b]. Furthermore, SVs can also affect the expression of a gene without causing direct damage to the coding region of that particular ORF. Such effects can be the result of changes in the location of regulatory elements, including enhancers and isolators, which lead to the missregulation of genes that were otherwise not regulated by these elements. An example of this case has been shown in medulloblastoma subgroups 3 and 4, in which by “enhancer hijacking” an enhancer element is brought to the proximity of the proto-oncogenes *GFI1* and *GFI1B* and leads to their activation [Northcott et al., 2014].

Even though great advances have been made in the understanding and interpretation of the effects of SVs on the phenotype, there are still many questions that remain open. The study of the functional impact of these large genomic variants is still ongoing and several challenges remain to be overcome. Therefore, many cancer genomic projects are being performed, with larger datasets (*e.g.* the Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative), better computational tools and more accurate matched clinical descriptions. The detection and understanding of the influence of SVs in the genome will probably improve by the development of new laboratory and computational techniques [Weischenfeldt et al., 2013]. For example, the increase in sequencing read lengths will help to understand more complex rearrangements or those occurring in repetitive areas of the genome that are at the moment difficult to ascertain [Huddleston et al., 2014].

1.5 Next-generation sequencing (NGS) methods and the identification of genomic variants

Last years witnessed major improvements in the technologies used for detection and characterization of genomic variants. Experimental and bioinformatic methods have been developed to allow fast, reproducible, and reliable identification and analysis of variants [Mardis, 2013]. About a decade ago, there were only a few dozen submicroscopic SVs detected which were discovered by microarray technologies and capillary-based DNA sequencing (Sanger sequencing). These techniques provided useful information and eventually led to the sequencing of the genomes of several model organisms, including yeast and mouse, and of the first human genome.

Microarrays initially allowed the identification of copy number differences mainly by means of array comparative genomic hybridization (array CGH) [Pinkel et al., 1998; Conrad et al., 2010] and by SNP microarrays [Cooper et al., 2008; McCarroll et al., 2008]. The array CGH method compares the signal ratios of two labeled samples, test and reference, obtained when hybridizing them to a chip containing defined probes (long oligonucleotides that align to known targets in the genome). Based on these ratios, copy number gains and losses can be estimated. Some of these platforms contain up to 42 million probes, which allow the detection of CNVs as short as 500bp, although they are not practical for large sets of samples [Iafrate et al., 2004; Conrad et al., 2010]. Similarly, SNP microarrays use hybridization methods with the advantage of designing probes that are allele specific and that increase CNV detection sensitivity. Although both methods have proved useful for the detection of CNVs, they both have certain limitations. For example, it is not possible to accurately define the breakpoints of events and they only allow larger and unbalance events to be detected [Alkan et al., 2011].

At the beginning of 2005, new sequencing technologies, commonly called next-generation sequencing technologies (NGS), became commercially available. Since then, they have revolutionized genomics studies and in particular, biomedical research, by achieving high-throughput and efficiency. The main improvements came from the amount of sequence that could be produced per run, the increase in the number bases that can be sequenced, the improved base-calling accuracy and the overall lower costs [Mardis, 2011]. Additionally, with these technological and experimental advances, bioinformatics tools and methods for data analysis have also seen great progress, allowing not only the detection of different types of rearrangements, including balanced events, but also the mapping of breakpoints at nucleotide resolution.

Several NGS platforms were developed in the last years, including SOLiD (Life Technologies), Roche/454 and Illumina platforms. Illumina's HiSeq2000/2500 is the most commonly used platform to date, with the ability to generate more than 300 million "reads" of 100bp (the resulting sequenced fragments) per lane. All experiments using NGS technologies in this thesis were done using the "paired-end" method and Illumina sequencing platforms, where both ends of linear DNA fragments of similar sizes are sequenced (Figure 1.2). Sequencing can also be done by using the "mate-pair" method, in which the sequencing is carried out at both ends of previously circularized DNA fragments. This circularization step of longer fragments brings together ends which are normally distant to each other [Metzker, 2010]. Therefore, the mate-pair method allows the spanning of larger genomic regions.

In summary, for these Illumina NGS approaches, the DNA sample to be sequenced is used to construct a collection of DNA fragments called a library. In the case of paired-end libraries, these fragments can be 200-500bp long. For mate-pair libraries the fragments are much larger, ranging from 3 to 6kb. The size of these fragments are usually referred to as "insert size". All the fragments have universal adapters covalently ligated in both sides. These adapters are used to PCR amplify the library products before sequencing and they are also used to hybridize the fragments to the flowcells [Mardis, 2013]. These flowcells are microfluidic channels that serve as a solid platform where the amplification for sequencing occurs. The sequencing reaction is a process of repeated steps, where in each cycle one fluorescently labeled nucleotide is incorporated. The nucleotides are blocked from further addition of bases, but after the detection step that identifies which base was added, a washing step cleaves the fluorescent labels [Metzker, 2010]. Therefore, the sequencing occurs in a nucleotide-by-nucleotide process.

After the reads are generated, they can be used for *de novo* assembly or they can be aligned to a reference genome. By determining differences of the sequences or the position and the orientation of the reads with respect to the reference, different types of genomic variants can be identified. These methods are described below and are summarized in Figure 1.3.

Read alignment and split reads. Detection of SNVs and indels can be done directly by mapping of the sequencing reads to a reference. Computational methods have been developed, such as SAMtools mpileup [Li et al., 2009] followed by BCFtools, to identify positions in the genome where multiple reads disagree. Since sequencing errors are in general random, if several reads disagree at the same position, it is possible that the mismatch corresponds to an SNV. In this regard, the amount of reads that cover each region of the genome, *i.e.* the sequencing coverage, is essential to correctly identifying SNVs. The same approach is also applied for the reliable identification of indels.

Additionally, when reads directly span a breakpoint of a variant, the mappers may not be able to align these reads as a whole. However, some tools can map only the beginning or the end of a read, *i.e.* the read is split. With this approach, all types of variants can be identified, with the advantage of detecting the breakpoint at nucleotide resolution.

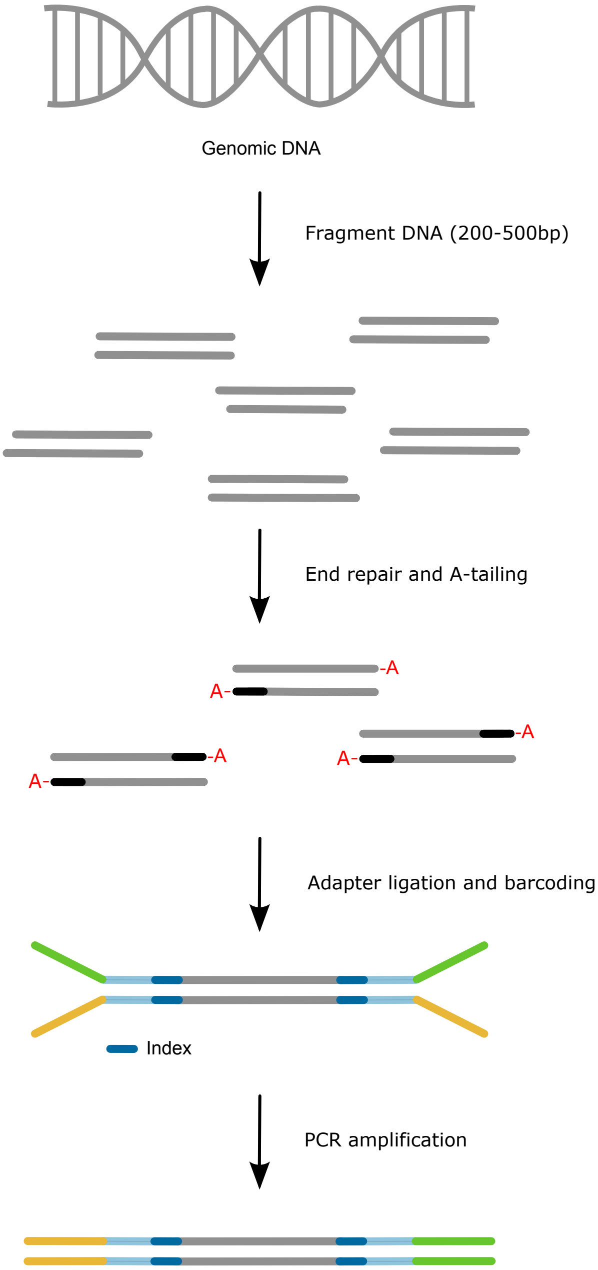

Figure 1.2: Library preparation for paired-end sequencing with an Illumina instrument. First, genomic DNA is sheared into fragments of about 500bp. The ends of these fragments are repaired to create blunt-ended dsDNA. Following the end-repair, an "A" nucleotide is added to the 3' end of the fragments to prevent the formation of concatemers in the following ligation steps. Then specific adapters for annealing to the sequencing flowcell are ligated to both ends of the fragments. For multiplexed library preparations, different molecular indexes can be added, in this case depicted in dark blue (Protocol following the paired-end Sample Preparation Guide, Illumina).

This method is computationally challenging since short sequences can align to multiple places in the genome. Therefore, to increase specificity, local alignment strategies can be used, and the position of the other read of a pair is used to reduce the search space [Ye et al., 2009]. For larger variants, another way of reducing complexity is to first identify the SVs by other methods, and then perform the split read search [Rausch et al., 2012b]. Similarly, higher sequencing coverage and longer reads are necessary to have better support for the variants. Specifically, the detection and validation of indels is more challenging compared to the detection of SNVs and SVs, especially because they occur more frequently in repetitive regions which are harder to accurately map to the reference genome.

Read depth. The read depth (also called sequencing coverage or depth of coverage) refers to the amount of sequenced reads that aligns to a specific region of the genome. These methods assume that there is a random distribution of the sequencing reads mapping to the genome. Deviations from this distribution suggest the existence of deletions or duplications (unbalanced SVs) [Alkan et al., 2011; Raphael, 2012]. Other variants that do not cause a change in the read depth, like translocations and inversions, cannot be identified by this method.

The general workflow following the alignment of the reads to a reference genome is to divide the genome into equally sized windows (*e.g.* 10kb) and to determine the number of reads mapped per window. By comparing the number of reads per window to the genome-wide read depth average, unbalanced events can be identified. Contiguous windows with a lower or higher than expected coverage indicate the occurrence of a deletion or a duplication respectively. For example, in the case of diploid organisms, if the coverage of a segment is reduced by half, it is assumed that there is a heterozygous deletion, and if it is reduced to zero, a homozygous deletion occurred. Similarly, if a region is duplicated, or amplified in the case that there are more than 2 copies, the coverage will in theory increase in proportion to the number of copies.

There are several factors that can cause deviations from the simple rules described above. Repetitive sequences in the genome and biases in the sequencing, such as GC-rich regions, can affect the read depth. The latter occurs mainly at the PCR step of the library preparation, where it has been shown that the PCR efficiency is lower in GC-rich parts of the genome [Quail et al., 2012]. This particular bias, however, can be corrected by normalizing the coverage by the GC content before making inferences about the copy number. One additional disadvantage of identifying SVs only based on read depth approaches is the inability to map the breakpoints at a nucleotide resolution. Regions with gains and losses are identified typically with a breakpoint within at least 1kb.

Despite these disadvantages, read depth based methods have been successfully used to map copy number variants in healthy and disease human genomes. For example, Sudmant et al. [2010] identified regions of human-specific expansions, *i.e.* highly amplified gene families, specifically in genes involved in brain development. Additionally, several copy number analyses have been carried out in cancer genomes [Campbell et al., 2008; Chiang et al., 2009]. There are several computational tools for the identification of CNVs based on read depth, for instance, CNVnator [Abyzov et al., 2011], Genome STRiP [Handsaker et al., 2011] and BIC-seq [Xi et al., 2010].

Paired-end sequencing and mapping. For paired-end mapping a paired-end or mate-pair sequencing protocol, such as the ones described above, are required. In both protocols the two ends of each DNA fragment are sequenced. The sequences are then aligned to a reference genome. In general, most reads will map concordantly with an expected distance between them equal to the fragment length (insert size). On the other hand, if the reads align discordantly, with an abnormal distance between them, with an unexpected orientation, or if they align to different chromosomes, they can be considered as evidence for the occurrence of an SV [Korbel et al., 2007]. Several computational tools have been developed for the identification of SVs based on paired-end mapping, such as DELLY [Rausch et al., 2012b] and Genome STRiP [Handsaker et al., 2011].

This approach has the advantage of allowing the identification of both balanced and unbalanced rearrangements. For example, if a discordant pair of reads aligns farther away than the mean insert size of the library, the occurrence of a deletion in the sample genome is suggested. Additionally, if the two reads of a pair align to different chromosomes, a translocation may have happened. Based on the orientation of the reads, inversions can also be identified through a similar logic (Figure 1.3). Since the length of a particular sequenced DNA fragment is not exactly known, the definition of discordantly mapping reads is done by comparing to the empirical distribution of fragment lengths obtained for a particular library [Korbel et al., 2009]. It is therefore desirable that the variance of the insert size distribution is small. This can be achieved by doing a tight size selection of the fragments around the target size during the library preparation.

On the other hand, discordantly mapping reads can also arise by sequencing errors or by misalignments due to, for example, repetitive sequences in the genome. To overcome these issues, clusters of discordant pairs that support the same event are usually required before calling an SV [Korbel et al., 2009; Rausch et al., 2012b]. Furthermore, some algorithms do not take into account reads that map to many different positions in the genome. One disadvantage of paired-end mapping is that the breakpoints of an SV cannot be determined at nucleotide resolution. Only an approximation is obtained, which can be a few base pairs to several kilobases away from the exact position.

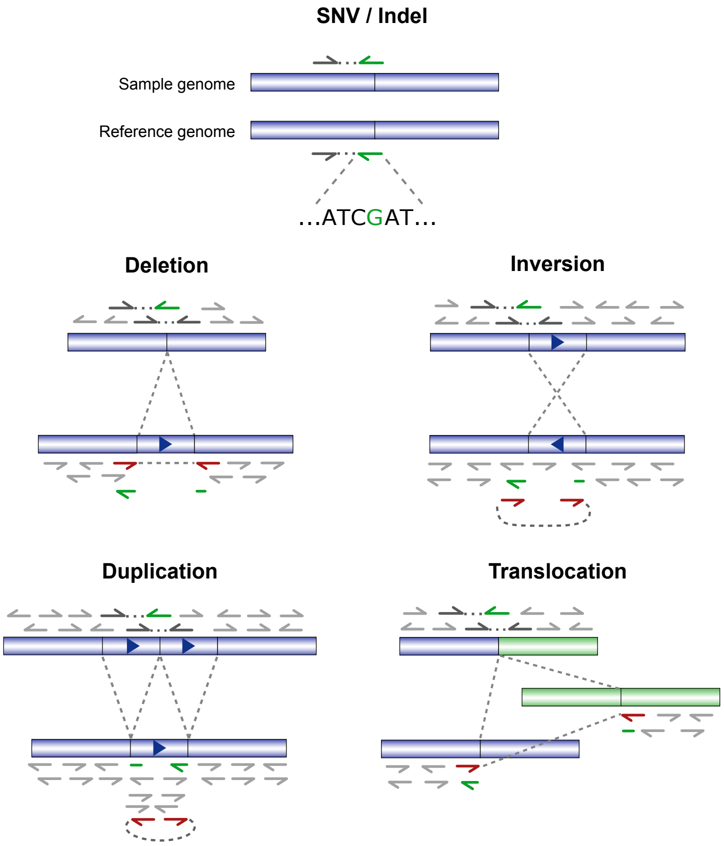

Figure 1.3: Methods to detect genomic variants based on next-generation sequencing data. Different approaches are used in combination to identify the sites where mutations occur. Paired-end mapping uses the information of discordantly mapping reads (red arrows) to infer SVs. Read depth methods are based on the sequence coverage (number of reads aligning to each region, gray arrows). Split reads refer to reads where the start maps to a region and the end maps to another region (dark green arrows). Modified from Weischenfeldt et al. [2013].

Assembly. Assembling the whole genome of an individual from scratch, *i.e.* without the use of a reference genome, is a very complex computational task, and it is even more challenging using short sequence reads [Nagarajan and Pop, 2013]. This means that millions of DNA sequences have to be put together into the right order to completely build a genome. In addition to the high number of fragments, other issues like the existence of repetitive sequences, errors in sequencing, and missing fragments, can contribute to the complexity. Because of these challenges, assembling whole genomes is at the moment very time consuming and computationally challenging. On the other hand, *de novo* local assembly is useful to better understand the derivative DNA molecules after the occurrence of complex rearrangements. The major advantage is that by assembling a region affected by structural variants, the exact positions of SVs can be determined at nucleotide resolution.

1.6 Yeast as a model organism

The baker's yeast or budding yeast, *Saccharomyces cerevisiae*, has been one of the most extensively studied organisms in molecular biology. In addition to its known traditional and commercial uses, the budding yeast is not only an excellent model organism but it has been used to establish completely new fields of biology, such as functional genomics and systems biology [Botstein and Fink, 2011]. In fact, the budding yeast has initiated and improved the molecular understanding of essential cellular processes such as DNA repair mechanisms, recombination, cell cycle, protein networks, DNA transcription [Pelechano et al., 2013] and many others [Barnett, 2003].

As an eukaryotic organism, *S. cerevisiae* has several advantages that make it suitable for biological research. Among its important features, yeast is not pathogenic, which makes it easy to handle in the laboratories [Sherman, 2002], and it is unicellular but still has similar fundamental cellular processes and cell constitution as higher eukaryotic cells. In addition, it is inexpensive to grow, can be grown on defined media facilitating the control of environmental conditions. Furthermore, it has a fast doubling time [Sherman, 2002], saving experimental time when compared to other model systems.

Genetic studies are facilitated due to the fact that yeast can grow stably in diploid and haploid states. In the haploid state, the effects of genetic modifications can be directly observed, in contrast to the masking of recessive mutations in the diploid state. On the other hand, the diploid cells can carry heterozygous mutations in essential genes. Furthermore, *S. cerevisiae* has a very efficient homologous recombination system, making the process of gene modification, such as knockouts and gene disruptions, efficient and simple. For all these reasons, the budding yeast is one of the most useful organisms to study the relation between genotype and phenotype in eukaryotic cells [Botstein and Fink, 2011].

Another advantage of working with yeast is the abundant biological information available. Its genome, fully sequenced and published since 1996 [Goffeau et al., 1996, 1997], was the first eukaryotic genome to become available. It is compact, with around 12Mb packed into 16 chromosomes. In fact, the genes comprise about 70% of the total genome [Goffeau et al., 1996] and most of them lack introns. Furthermore, *S. cerevisiae* has around 6000 well characterized genes, and only one sixth of them are essential [Giaever et al., 2002].

Perhaps one of the most important reasons why yeast became such a successful model system is the straightforward assessment of the relationship between gene structure and protein function [Botstein and Fink, 2011]. Until now the function of around 85% known protein coding genes in *S. cerevisiae* has been described [Botstein and Fink, 2011]. This is an advantage considering that around 31% of all potential protein coding genes in yeast have a mammalian homolog [Botstein et al., 1997; Foury, 1997]. Additionally, approximately 17% of genes belong to orthologous gene families that are related to diseases in humans [Botstein and Fink, 2011; Heinicke et al., 2007].

Consequently, it is not unexpected that a wealth of information that has become available on human genes was initially discovered by studying their yeast homologs. For example, the function of candidate genes that resulted from linkage studies of human diseases could be better understood by comparing them to yeast genes with high sequence homology [Botstein et al., 1997]. For example, genes causing nonpolyposis colon cancer could be identified by resemblance to *MSH2* and *MLH1* in yeast [Strand et al., 1993]. Similarly, genes for Werner's syndrome (*SGS1* in yeast) [Sinclair et al., 1997] and ataxia telangiectasia (*TEL1* in yeast) [Greenwell et al., 1995] could be identified. Today, a vast amount of knowledge, including what has been discovered so far about each yeast gene, is collected in the *Saccharomyces* Genome Database (SGD `http://www.yeastgenome.org/`) [Cherry et al., 1998]. This includes information about gene expression and functional analysis data [Ball et al., 2001], gene ontology [Dwight et al., 2002], the roles of gene products and their interaction with other proteins, and also organizes published literature about each gene [Christie et al., 2004].

One of the most important contributions to the understanding of the connection between genes and proteins was the production of yeast deletion mutants for every ORF of the genome. This became what is known as the Yeast Deletion Collection. By systematically deleting single genes and later on subjecting these mutants to a diverse range of experiments, many biological consequences of the lack of a gene's function were discovered [Botstein and Fink, 2011].

1.6.1 The Yeast Deletion Collection

As mentioned, one of the main advantages of using budding yeast as an experimental organism is the wealth of data and strains available. In addition to the publication of the complete genomic sequence of *S. cerevisiae* around 1986, a consortium of European and North American laboratories [Winzeler, 1999; Giaever and Nislow, 2014] established the *Saccharomyces* Genome Deletion Project. The main goal was to create a collection of yeast knockout (KO) mutant strains to identify, confirm and study the function of essential genes. Additionally, it became useful for analyzing molecular components of the cell involved in basic physiological and developmental pathways, and their interactions.

The collection is composed of >21,000 knockout strains, with deletions in most of the known ORFs, defined as a DNA sequence that could potentially encode a protein of more than 100 amino acids. It comprises homozygous and heterozygous diploid strains with deletions in around 6000 genes, including some putative ORFs and over a thousand essential genes, and an additional set of haploid strains of both mating types for over 4700 non-essential genes [Scherens and Goffeau, 2004].

Each knockout strain contains a cassette that carries two unique "molecular barcodes", which are 20bp long, to allow working with the pooled collection while still being able to identify the mutant strains. There is a great number of studies that have used these collections to research a broad range of biological questions. Novel protocols and big screens have been developed over the past years and have increased our understanding of biological functions, response to stress and mechanism of drug action, of phenotypes occurring under different physiological conditions, and of previously unknown genes involved in essential metabolic processes [Giaever and Nislow, 2014].

1.6.2 Design of the Yeast Deletion Collection

The yeast deletion strains were constructed using a PCR-based strategy [Baudin et al., 1993; Wach et al., 1994], in which a deletion cassette replaces each ORF (Figure 1.4). The cassette contains the KanMX4 marker, which confers G418 resistance, and it is required to select for the transformant strains. Additionally, it carries two unique 20bp sequences that function as molecular barcodes or tags that are commonly referred to as "uptag" and "downtag". These barcodes are surrounded by universal primer sequences, common to all the strains in the collection, making it very simple to amplify them from a pooled culture. Furthermore, the barcodes are different enough so that it is unlikely that PCR or sequencing errors can lead to an incorrect identification of the strains [Pierce et al., 2007].

This cassette is integrated into the genome and replaces the target ORFs from start to stop codon by mitotic recombination. To achieve this, the cassette is surrounded by two sequences of 45bp directly upstream and downstream of the selected ORF [Wach et al., 1994; Giaever et al., 2002; Pierce et al., 2007]. A total of over 50,000 oligonucleotides were needed for the PCR-mediated construction of the deletion cassettes [Giaever et al.,

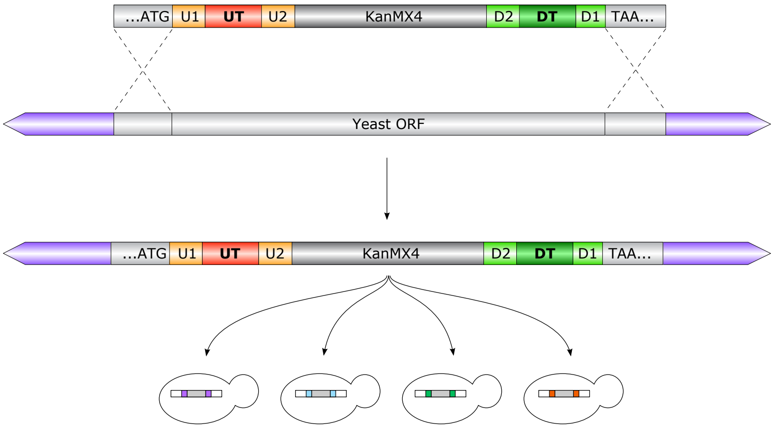

Figure 1.4: The yeast deletion collection. In the deletion strains of the yeast deletion collection each ORF has been substituted with a KanMX cassette, which integrates into the genome by homologous recombination. The cassette confers resistance to geneticin (G418) and carries two molecular barcodes, **UT** and **DT**, specific for each deletion strain. All UTs can be amplified with universal primers U1 and U2, while all DTs with primers D1 and D2. (UT: uptag, DT: downtag). Modified from Giaever and Nislow [2014].

2002]. Each deletion was then verified using a set of PCRs to amplify a sequence spanning the left and right junctions of the deletion cassette and the genomic location.

1.6.3 Main applications of the Yeast Deletion Collections

The yeast KO strains of the yeast deletion collection have been successfully used in many studies and in over 1000 genome-wide screens. In fact, the deletion project became a landmark and a model to develop many other genome-wide technologies (*e.g* the GFP collection [Ghaemmaghami et al., 2003]) and deletion collections in other organisms, such as in *Arabidopsis thaliana* [Alonso et al., 2003] and in *Escherichia coli* [Baba et al., 2006]. With no doubt, since the first results and description of the project were published in 1999 [Winzeler, 1999], the collection became widely used, contributing to many biological fields and having a deep impact on the yeast and genomics research communities [Giaever and Nislow, 2014]. Some of the most important research topics that have been addressed using the yeast KO strains are mentioned below.

Functional characterization of genes. The first studies to be published addressed the question of the functional impact of gene deletions [Winzeler, 1999; Giaever et al., 2002]. These studies investigated the phenotypic consequences of grown on stress conditions and revealed that around 18% of the genes were essential. The researchers were able to identify new genes required for growth in these conditions. They also found that around 15% of the mutants had growth problems on regular rich media, indicating that they are more vulnerable than others and present a slower growth rate when compared to the wild type.

With these studies it became evident that there is a low correlation between the genes that are required for survival in one condition and the genes with increased expression in the same condition. In fact, only less than 7% of the genes exhibited a significant increase in mRNA levels and were required for optimal growth [Birrell et al., 2002; Giaever et al., 2002]. This finding was surprising at that time, when there was little understanding of all the complex events that occur post-transcriptionally and that regulate protein expression at a translational level.

Furthermore, Giaever et al. [2002] observed the cell shape and size of the knockout strains upon growth in different conditions. They were able to identify deletion strains that had morphological defects compared to the ellipsoid shape of the wild type cells. As expected, these deletion strains, comprising around 15% of the mutants, were defective in genes related to cell growth, cell division and DNA synthesis.

These initial studies validated the suitability of the deletion collection as a valuable tool for functional genomics experiments. Many large scale phenotypic screens have been done so far including the assessment of cell growth and cell size (*e.g.* [Deutschbauer et al., 2002; Jorgensen et al., 2002]), mating and membrane trafficking.

Identification of novel drug targets and mechanisms of action of toxic compounds. Several studies have focused on identifying novel protein targets for multiple drugs. The yeast deletion collection has been used to identify members of pathways that are altered upon the application of different chemicals and to understand the global functions of the targets of certain drugs, such as beomycin [Aouida et al., 2004] and rapamycin [Chan et al., 2000].

Drug induced haploinsufficiency profiling (HIP) is one of the methods developed based on the yeast KO collections. It takes advantage of the fact that if a heterozygous deletion strain is sensitive to a specific drug, then the drug may act on the product of that locus [Smith et al., 2010a]. Therefore, by screening the entire heterozygous deletion collection, several drug targets have been identified [Giaever et al., 2004]. Additionally, Lum et al. [2004] assessed the effects of 78 compounds on the heterozygous collection and were able to verify the targets of previously known compounds and to propose mechanisms of action of several of those drugs. Several of these compounds are clinically relevant, confirming the utility of these types of screens [Lum et al., 2004].

Gene and protein interaction networks. Additional technologies based on the yeast KO strains were also developed to study more complex gene functions. One important development was the Synthetic Genetic Array (SGA) analysis [Tong et al., 2001]. With this method, the impact of double mutations could be assessed in a systematic way. Crossing one mutant with an array of around 4700 other deletion mutant strains, and evaluating the progeny generated by meiosis allowed the identification of the functional relationship between genes. This approach proved very useful for the generation of interaction networks and for the production of a increasingly better map of gene function.

Many insights have been obtained from the study of yeast interaction networks, from the identification of novel interacting components to a better understanding of the properties of networks and of how genes regulate themselves and other cellular processes [Boone et al., 2007]. Earlier functional studies in the deletion mutants mentioned above revealed that only a few genes were essential, implying that this biological system is robust to change and that there is a buffering machinery preventing drastic variation [Hartman et al., 2001]. Several studies have addressed this topic by using synthetic lethality approaches, in which the combination of mutations in two genes causes a reduced fitness or cell death, which then is an indication of the interaction between these genes [Davierwala et al., 2005].

Evolutionary studies. The process of adaptive evolution driven by the accumulation of beneficial mutations has also been studied through the use of the deletion collections. It has been shown that mutant strains with growth defects due to the lack of a specific gene can regain fitness to a level comparable to the wild type strain if they are left to grow for several generations. This allows for the accumulation of compensatory mutations that give rise to genetic divergence and consequently diverse phenotypes across different environments [Szamecz et al., 2014].

On the other hand, these mutations arising to compensate for defects in a certain pathway are advantageous in the particular environment in which they were positively selected. If the growth conditions change, mutations that were beneficial at some point can become deleterious.

The relative advantage of different alleles of a particular gene can be opposite in different environmental conditions, an effect known as pleiotropy. By studying the collection of deletion mutants, Qian et al. [2012] were able to identify genetic mechanisms that are related to the pleiotropic effects of mutations. They demonstrated that hundreds of genes expressed by yeast cells under certain growth conditions were actually harmful, but could become useful in other conditions. These studies can be applied to higher organisms as well, in areas such as senescence or neurodegenerative disorders, in which the harmful effect of mutations can be seen only in later stages of life, but that might be beneficial in earlier development [Carter and Nguyen, 2011; Qian et al., 2012].

Studies of adaptive evolution have contributed to our knowledge of how mutations become fixed in the populations [Kao and Sherlock, 2008; Lang et al., 2013]. Importantly, the dynamics of how some mutations survive even if they are not fully beneficial, while others are lost, by for example genetic hitchhiking, closely resembles the underlying processes that act in tumor cells during cancer progression [Nik-Zainal et al., 2012b].

Human diseases. Since many of the fundamental processes of the cell are conserved from *S. cerevisiae* to higher eukaryotes [Botstein et al., 1997], it is compelling to perform experiments in budding yeast to understand mechanistic aspects of human cells. In this regard, studies of DNA repair mechanisms have facilitated the discovery of genes involved in these processes, and functional analogies between human and yeast genes can often be made [Foury, 1997]. For example, Ooi et al. [2001] performed a screen for identifying components of the NHEJ pathway. They were able to find a new gene, NEJ1, which interacts with LIF1, homologous to the mammalian XRCC4 protein, an important factor in guarding against cancer [Ooi et al., 2001].

Other important screens include the identification of genes involved in human mitochondrial diseases [Steinmetz et al., 2002] and a screen to identify genes and pathways relevant for the development of Huntington's and Parkinson's diseases [Willingham et al., 2003]. In the latter study, 4850 yeast strains were transformed with constructs that expressed a mutant huntingtin fragment and α-synuclein, the major components of the inclusion bodies that are a feature of these two diseases respectively [Willingham et al., 2003]. It was possible to identify several yeast genes, with human orthologs, that are relevant for these diseases.

The availability of the yeast deletion collections has been certainly a great advantage for biological research, and will probably continue to be, especially with the availability of cheaper and better genome sequencing technologies.

1.6.4 Uncertainties and caveats of the Yeast Deletion Collections

Despite the many advantages of yeast assays and the yeast deletion collections, there are also a few limitations that need to be considered. While there are many shared processes between yeast and higher eukaryotes, making the interpretation of results in some cases easily extrapolated from one organism to another, not all cellular mechanisms can be taken as equal. It is therefore useful to always verify the findings in other models. Moreover, big screens are subject to a higher chance of finding false positives. It is therefore recommended that any results should be validated by other experiments [Scherens and Goffeau, 2004].

Furthermore, drug assays are restricted by the amount of each compound needed to cause a response in the cell. Often very high concentrations are needed, due to reasons such as the low permeability of the cell wall to these substances, which makes the experiments more expensive and difficult to perform [Smith et al., 2010b].

One further potential source of error in studies using the deletion collections is the presence of wrong annotations and the existence of a neighboring gene effect (NGE) [Ben-Shitrit et al., 2012]. It was recently demonstrated that the phenotype in one deletion mutant can be influenced by the effect of the deletion on a neighboring gene, instead of being caused only by removing the particular ORF. The NGE may account for approximately 10% of all deleted genes, and therefore it can have some consequences on the interpretation of gene function and genetic interactions [Ben-Shitrit et al., 2012]. Therefore, verifying results by different experiments and techniques is highly recommended.

1.7 Mutation rates in the budding yeast

Mutations are the ultimate source of genetic variation. Even when the replication and transmission of genetic material is very accurate, all cells acquire spontaneous mutations during DNA replication or by environmental DNA damage, a process that is biologically and evolutionary important. Most mutations are deleterious, *i.e.* they affect negatively the fitness of the individuals. However, there are some that can confer an advantage allowing for adaptive evolution [Baer et al., 2007]. The rates at which mutations occur and the specific location where they arise are therefore important factors in the evolution of a species and how it is able to adapt to new conditions. In yeast, it has been estimated that there are around 10^{-9} single nucleotide substitutions per base per cell division [Lynch et al., 2008], whereas chromosomal structural variants occur less often [Nishant et al., 2010].

Several experimental studies have reported that mutation rates vary across the genome. For example, microsatellite sequences or polynucleotide runs [Hawk et al., 2005; Verstrepen et al., 2005], as well as genes related to the immune response [Papavasiliou and Schatz, 2002] have higher mutation rates than other genomic regions. Apart from this variation within the genome, mutations rates can vary due to environmental conditions. On one hand, the environment can be directly mutagenic, or it can also stress the cells making them more prone to mutations. Examples of direct mutagens are ultraviolet and ionizing radiation, alkylating agents, and crosslinking agents [Hoeijmakers, 2001]. Agents that can cause elevated mutation rates without damaging the DNA directly are for example those that inhibit the mismatch repair process. For instance, it has been reported that yeast growing in high cadmium concentrations show hypermutability due to the inhibition of mismatch repair [Jin et al., 2003].

Traditionally, mutations rates were estimated using genetic reporter assays. These experiments can unfortunately be applied only to model organisms, and their results can be biased because only mutations having a phenotypic effect can be scored and phenotypically silent mutations are omitted. For example, the canavanine resistance assay Can^{r} allows detection of different types of mutations when these inactivate the arginine permease activity of this gene [Forsburg, 2001]. In these experiments, the wild type cells are sensitive to the toxic arginine analogue, canavanine. Mutations in the *CAN1* gene prevent the cells to import canavanine, making then resistant to the drug. Additionally, some methods for detecting larger rearrangements have been developed.

For example, Chen and Kolodner [1999] measured the rates of gross chromosomal rearrangements by scoring the simultaneous loss of the *CAN1* and *URA3* markers at one end of chromosome V.

The current knowledge of mutation rates is still incomplete due to the difficulty of studying significant number of events [Zhu et al., 2014]. However, with novel sequencing techniques and increasing amount of genomic data available, more general inferences of the mutation rate and its variation across the genome have been made [Ellegren et al., 2003; Lang and Murray, 2008, 2011].

Since mutation rates are in general very low, the estimation of events per generation is difficult [Nishant et al., 2010]. One method to increase the number of events to a level that is high enough to make useful inferences is through a mutation accumulation (MA) assay. In this type of experiment the population is passed through many generations of very sharp bottlenecks. By this method, only highly deleterious mutations, *e.g.* those that hamper growth and survival, are selected against. Otherwise, there is no selective pressure applied to the population, allowing for the accumulation of neutral mutations. MA assays, combined with NGS technologies, have provided an experimental set up for the study of larger amounts of mutations in a relatively unbiased way.

Several MA experiments have been performed in eukaryotic model organisms, for example, in *Arabidopsis thaliana* [Ossowski et al., 2010], in the nematode *Caenorhabditis elegans* [Denver et al., 2012], in *Drosophila melanogaster* [Haag-Liautard et al., 2007] and also in *Saccharomyces cerevisiae* in the haploid [Lynch et al., 2008] and diploid [Nishant et al., 2010] states. Specifically, in yeast these MA studies have revealed that its genome is relatively stable, with few mutations accumulating over several generations in wild type strains in both vegetative and meiosis stages [Nishant et al., 2010]. In a MA experiment, Zhu et al. [2014] identified a total of 867 SNVs, 26 indels, 3 CNVs and 31 aneuploidies after studying 145 strains grown through more than 2000 generations each.

One exception of the low mutation rates generally observed are the mutator strains, *i.e.* strains that have increased spontaneous mutation rates. Usually these mutator phenotypes arise by acquiring mutations in one of the numerous genes that prevent errors to occur during replication, repair, recombination or processes that ensure the correct chromosome segregation.

In the budding yeast there are several mutations, such as in *MSH2* [Huang et al., 2003] a mismatch repair gene, *MRE11* [Chen and Kolodner, 1999] a gene involved in recombination and in the replication related gen *RAD27* [Chen and Kolodner, 1999] which have been reported to exhibit mutator phenotypes. Serero et al. [2014] investigated the genome-wide mutational spectra of MA lines of wild type and nine mutator strains. These authors found that there are diverse processes that govern the accumulation of mutations in the different strains. Each mutator strain had complex patterns of mutational spectra that included point mutations as well as structural variants and whole chromosome gains and losses. It is very interesting and important to study the genome-wide mutational patterns that arise in these deletion mutants because most of these genes have conserved orthologs in humans and have been associated to genetic diseases and higher susceptibility to certain cancers.

Apart from the limitations mentioned above, MA and reporter gene experiments have provided estimates of mutation rates [Lynch et al., 2008] and insights into patterns of mutations. However, due to the low amounts of mutational events, little is known about the general mechanisms of mutation formation, of the variation of these mutations in the genome and of the effects of gene deletions over the general occurrence of mutations. Furthermore, most studies of mutations in the budding yeast have focused only on point mutations. Therefore, there is little knowledge of the general rates and patterns of structural variants and indels in the budding yeast.

Chapter 2

Genome-wide mutational landscapes of yeast deletion mutants

2.1 Motivation

Even though *Saccharomyces cerevisiae* is one of the most studied model organisms, a genome-wide characterization of the types, distribution, and frequency of mutations, particularly of structural variants, is still not complete. Due to limitations in the costs and experimental techniques available, genetic variation studies had typically used single reporter assays, *e.g.* [Forsburg, 2001], or focused only on a particular part of the genome, *e.g.* chrV [Chen and Kolodner, 1999]. Additional studies have identified genes involved in the maintenance of genomic stability by analyzing specific types of DNA sequences, such as ectopic retrotransposition of Ty1 elements [Scheifele et al., 2009].

These assays have helped to identify key players in DNA repair pathways and the maintenance of genomic stability. The impact of these discoveries has been particularly influential due to the fact that yeast shares many homologies with human genes, and several pathways are conserved. In fact, many of the genes implicated in preserving genomic integrity are also involved in human diseases, including cancer [Yuen et al., 2007]. However, previous work revealed that mutation rates are not uniform along the genome, therefore focusing on specific regions does not cover the full spectrum of mutations and the mechanisms that can lead to their formation.

As a consequence, genome-wide studies are required to better understand the causes and consequences of genomic instability. In this regard, two very recent studies have started to describe the large variation existing in the mutational landscapes of yeast strains. Serero et al. [2014] used mutator strains to characterize the acquisition of mutations in strains with defects in different genome maintenance processes using mutation accumulation experiments. The study revealed a very diverse and complex mutational spectrum for each deletion strain, indicating that the mutational processes might be very dynamic given certain gene defects. The main limitation of this study was that only 9 different deletion mutants were studied. In addition, Zhu et al. [2014] described the mutational spectrum and mutation rates in 145 yeast deletion strains. The authors were able to identify a relatively large number of SNVs. However, their main goal was to estimate mutation rates, rather than analyzing the context in which mutations occur and linking the particular deficiencies of each mutant with their mutation spectrum.

One of the main findings of these studies is that different deletion mutants can have very diverse mutational spectra [Serero et al., 2014].It is still not fully understood how different types of mutations arise and how often they occur under different environmental conditions.

Thorough characterization of genomic variation is crucial to understand the relationship between genotype and phenotype since genetic variants are known to underlie complex diseases [Feuk et al., 2006]. Given the countless advantages of yeast as an experimental model, as mentioned in Chapter 1.6, I chose this organism as a model to investigate the impact of defects in main pathways of DNA repair, replication, and recombination on genomic instability. Specifically, by using yeast deletion mutants, I was able to characterize the mutational landscape of 47 haploid deletion mutants under no, or very mild, selective pressure. The results from this study aim to better understand the impact of particular defects in specific pathways and the mechanisms involved in maintaining genome stability.

2.1.1 Contribution

The experiments in this chapter were initially started by Megumi Onishi-Seebacher, a former postdoc in Jan Korbel's group, who performed the mutation accumulation assays up to bottleneck 30. I performed all the remaining bottleneck passages, as well as the sequencing experiments and validations described in the following sections. Additionally, I carried out all data analyses for the mutation identification and annotation presented in this chapter.

2.2 A mutation accumulation approach for studying neutral mutations in yeast

Structural variants, indels, and SNVs can arise by a diverse set of mechanisms including errors in the processes of DNA repair, recombination and replication, or through the insertion of transposable elements. They can arise in the presence (*e.g.* non-allelic homologous recombination NAHR) or absence (non-homologous end joining NHEJ, fork-stalling and template switching FoSTeS, microhomology-mediated break induce replication MMBIR) of homologous sequences around the breakpoints [Hastings et al., 2009a]. In more catastrophic forms of chromosomal rearrangements, *e.g.* chromothripsis, as observed in different cancers types [Stephens et al., 2011; Rausch et al., 2012a]

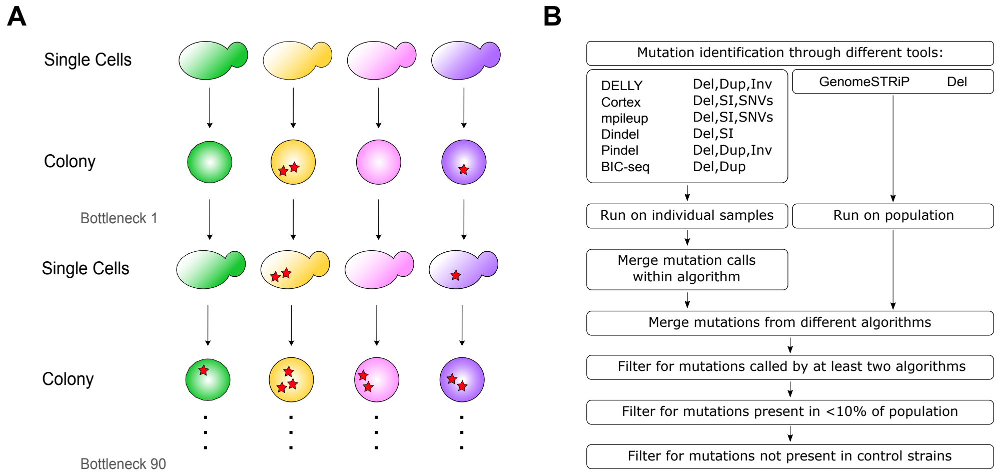
A
Single Cells
Colony
Bottleneck 1
Single Cells
Colony
Bottleneck 90
B
Mutation identification through different tools:
DELLY Del,Dup,Inv
Cortex Del,SI,SNVs
mpileup Del,SI,SNVs
Dindel Del,SI
Pindel Del,Dup,Inv
BIC-seq Del,Dup
GenomeSTRiP Del
Run on individual samples
Run on population
Merge mutation calls within algorithm
Merge mutations from different algorithms
Filter for mutations called by at least two algorithms
Filter for mutations present in <10% of population
Filter for mutations not present in control strains

Figure 2.1: Mutation accumulation assay and mutation identification pipeline. **A** | Mutation accumulation through recurrent single-cell-to-colony bottlenecks which consisted of multiple rounds of randomly choosing an individual colony, streaking it out on a YPAD plate to separate it into individual cells, letting it grow until new colonies were formed, and choosing again another colony. Yeast deletion strains in this study were passed through a total of 90 bottlenecks in which they accumulate mainly neutral mutations (red stars). **B** | Mutation identification pipeline using a set of computational tools which are based on different algorithms for the discovery of SVs, indels and SNVs. A more detailed explanation is described in the methods Section A.1.4. (Del: deletions, Dup: tandem duplications, Inv: inversions, SI: short insertions).

and in the germline [Kloosterman et al., 2011], many different mechanisms might be involved.

For this study we selected a set of 47 knockout strains from the Yeast Deletion Collection, each with a deletion of one ORF. These deleted genes belonged to a wide range of pathways, including DNA damage checkpoints and chromatin remodeling, with many having human homologs. The complete list of KO genes and their broad functional categories is shown in Table B.1. We also included the wild type strain (BY4741) and two knockout controls *his1* and *trp5* predicted not to have effects on genome stability.

The yeast genome has been shown to be relatively stable, acquiring low number of mutations over several generations [Nishant et al., 2010]. Therefore, in order to obtain a larger number of mutations for analyses, we performed a mutation accumulation assay to increase the number of neutral mutations over time. For this, each of the deletion strains was propagated through a series of single-cell-to-colony bottlenecks (Figure 2.1A). This was done for a total of up to 90 bottlenecks, producing mutation accumulation lines of around 1800 generations (given that 48h of clonal growth is 20 generations per bottleneck).

By using this approach, there is only natural selection against very damaging mutations that can interfere with growth or survival. Therefore, we were able to study the mutational landscapes of deletion mutants under no selective pressure, *i.e.* we characterized the natural variation in these strains.

Two mutation accumulation (MA) lines were propagated for each deletion strain. For each strain, one of the two MA lines was sequenced at the beginning of the experiment (hereafter b0, for bottleneck at time point 0). After 30 bottlenecks, *i.e.* around 600 generations, 22 strains were sequenced. Similarly, 93 lines were sequenced after 60 bottlenecks, including both MA lines per strain, and 21 strains were sequenced after 90 bottlenecks, for a total of 184 deletion strains. Given the relatively small yeast genome, the sequencing was done by multiplexing up to 55 different lines per sequencing lane, as explained in detail in Section A.1.3. All sequenced strains had around 20× coverage, for an overall total of 3460× coverage.

2.3 Mutational landscape of yeast deletion strains

2.3.1 Mutation identification

Due to the difficulty in detecting mutations in the genome, and in particular SVs, a combination of different methods was used. These included *de novo* assembly, read depth, paired-end mapping [Korbel et al., 2007] and split read [Mills et al., 2006] methods which are commonly used for the detection of genomic variants (details in Section 1.5). Several computational tools have been developed for this purpose. Most of them focus on predicting mutations by using only one or a few of these detection methods but they differ in the sensitivities, specificities, length and types of variants detected.

Since there is great advantage in combining several tools for the discovery of mutations, we made use of several computational approaches (Figure 2.1B) covering the mentioned detection methods to have a more comprehensive list of variants (Further details of methods are explained in Section A.1.4). However, to avoid a large number of false-positives and redundant calls, merging and filtering of the detected variants was performed as explained in Section A.1.4.

After sequencing the 47 different deletion mutants at several time points, we were able to identify a total of 9888 variants, including deletions, short insertions, tandem duplications and SNVs. The total number of *de novo* mutations (not present in the strains at b0) are shown in Table 2.1. The most common mutations were SNVs followed by deletions. Only a small number of tandem duplications were observed.

2.3.2 Higher accumulation of short insertions and deletions

In general, deletions were the most abundant SVs, with size ranging from 1 to 10,299bp. However, they mainly corresponded to short events, smaller than 5bp. In fact, only 3% of identified deletions had a size larger than 50bp. The insertions detected were also short, with sizes smaller than 20bp. The size distributions of deletions and short

Table 2.1: Total number of *de novo* variants found in 47 different deletion strains after 90 bottlenecks.

Mutation type	Total
Deletions	3786
Short insertions	1348
Inversions	87
Tandem duplications	39
SNVs	4628
Total	9888

insertions are shown in Figure 2.2A. It is noticeable that a big proportion of deletions and short insertions are 3bp long, likely because of the less deleterious effects of these non-frameshift mutations. There is also an additional small peak around 300bp that corresponds to the deletion of solo-LTR elements. These retrotransposon-derived elements constitute around 3% of the yeast genome [Goffeau et al., 1996], and their deletion is facilitated by surrounding small homologous sequences.

Tandem duplications corresponded to events of 20-2460bp, with a median size of 493bp. Furthermore, the inversion sizes ranged from 52bp to 11kb. However the existence of larger variants can not be ruled out, as the detection of larger inversions was limited by the fact that Pindel can only identify events ranging from approximately 50-10kb.

I randomly selected a set of 35 indels for validation by PCR and Sanger sequencing. Out of the indels, 33 loci could be amplified and sequenced. In total 28 calls were positively validated and 5 were classified as false calls accounting for a false discovery rate of 15%.

Figure 2.2B shows a strong positive correlation between chromosome length and number of indels (r^2=0.71, Spearman P<0.001), suggesting that at a genome-wide scale there may not be a great variation in mutation rates among chromosomes.

The total number of deletions per MA line is shown in the Supplementary Table B.2 and Supplementary Figure B.1A. Deletion strain *msh2* had the largest number of both deletions and short insertions, which was expected given the important role of this protein in mismatch repair (MMR) (Figure 2.3). There was a significant higher number of deletions accumulated in *msh2* than the rest of the deletion mutants (one-tailed *t*-test, P=0.005), suggesting that the *msh2* strain acquired a hypermutation state in which it gained many more mutations than the rest of the strains. This was also confirmed by the fact that this deletion mutant also acquired the highest number of SNVs as described below.

The next 9 deletion strains with the highest number of deletions are shown in Figure 2.3A. Among these strains there was lower variation in the number of events iden-

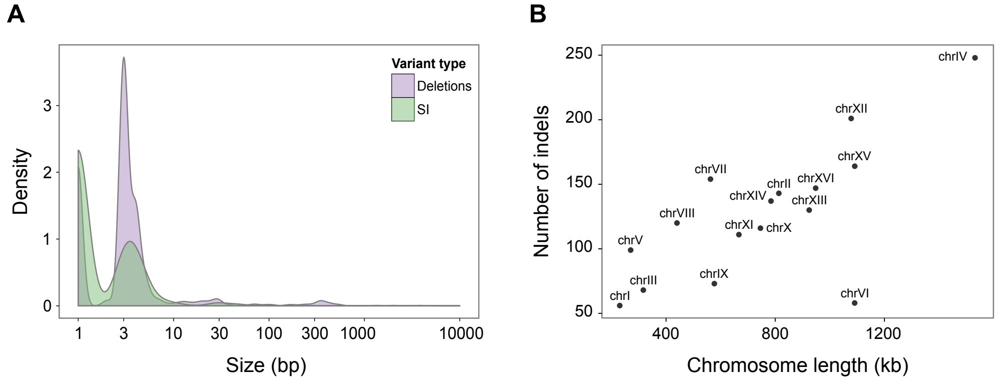

Figure 2.2: Sizes of *de novo* deletions and short insertions and their distribution per chromosome. **A** | Size distribution of the overall set of *de novo* deletions and short insertions discovered in the 47 yeast deletion mutant strains after a total of 90 bottlenecks. **B** | Total number of deletions and short insertions per chromosome.

tified (79±21, median±SD), compared to the difference between them and the *msh2* strain.

Additional mutant strains with high numbers of short deletions were *isw1*, *swr1* and *sgo1* (Figure 2.3A). These mutants are defective in chromosome segregation or chromatin remodeling pathways. Swr1 and Isw1 both form parts of different chromatin remodeling complexes [Smolle et al., 2012]. These results were therefore not expected due to the less clear connection between chromatin remodeling and the formation of short deletions.

Similarly, since Sgo1 is a spindle checkpoint component involved in meiotic chromosome cohesion [Indjeian et al., 2005], its link to the formation of short deletions may not be direct. However, as it is described below, this strain became aneuploid between b30 and b60 (Figure 2.7), and there may be a connection between aneuploidy and genomic instability [Pavelka et al., 2010], although the order of events is still not fully understood.

The total number of short insertions for all MA lines is shown in Supplementary Table B.2 and Supplementary Figure B.1B. The *msh2* strain was the one with the highest amount of short insertions (Figure 2.3B). However, the total number of events was lower than deletions, and there was not a significantly higher number of SI between *msh2* and the other mutant strains (one-tailed *t*-test, P=0.168). The second strain with the highest accumulation of SI was *rad27*. In this case, the difference between the first and second strains was not as striking as for deletions.

The mutant strains *isw1* and *sgo1*, which ranked high in mutants accumulating deletions, were also among the top ten strains with SI. However, different strains also showed increased numbers of SI. These include the mutants *mms4* and *srs2*. Mms4 is part of an endonuclease complex that cleaves branched DNA. *mms4* mutants have deficiencies in recombination and DNA repair. Together with Mus81 it is involved in

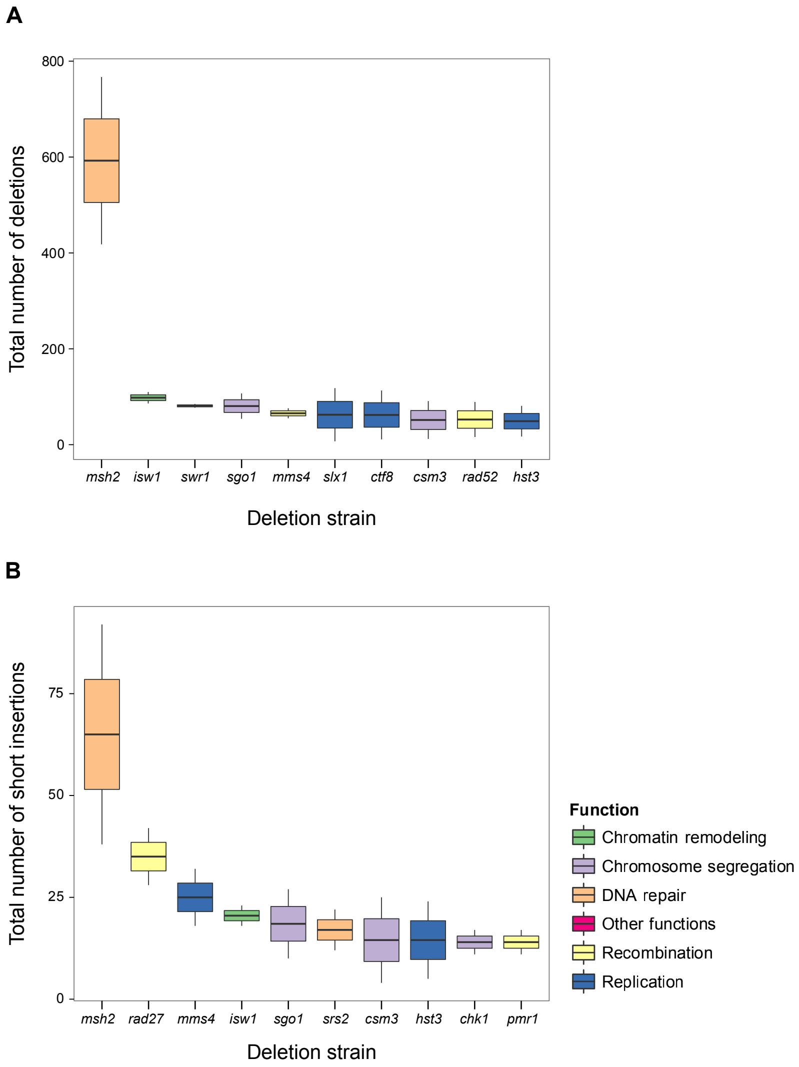

Figure 2.3: The 10 strains with the highest number of deletions and short insertions classified into broad functional categories. The total number of *de novo* mutations occurring in less than 10% of the strains is shown for the two mutation accumulation lines. **A** | Total number of *de novo* deletions. **B** | Total number of *de novo* short insertions. The *msh2* deletion mutant is a mutator strain and accumulated the highest number of both mutation types.

mismatch repair and plays a role in homologous recombinational repair [Odagiri et al., 2003]. Similarly, Srs2 is also highly involved in DNA repair and checkpoint recovery, and its deletion leads to genome instability.

From these MA experiments one interesting finding is that there are differences in the types and frequencies of mutations among deletion mutants, in which some strains acquire only, or predominantly, short deletions while others mostly short insertions. For example, *swr1* mutants have mainly short deletions (one-tailed *t*-test, P=0.004). This significant difference (with a significance level P<0.05) was also true for 7 other deletion strains, *rsc1*, *rad24*, *ard1*, *ctf4*, *dot1*, *htz1* and *rad18*. In contrast, the other deletion mutants included in this study had comparable amounts of both types of indels.

2.3.3 Functional annotation of SVs and indels

I then investigated the overlap of all mutations identified with multiple genetic features available for the yeast genome. The positions for a diverse set of genetic features were obtained from multiple published datasets and were used to annotate the variants detected in the yeast deletion mutants. I assessed the overlap with genes (ensGene from UCSC) to investigate the impact of the accumulated mutations on the yeast genome. Additionally, to further understand potential mechanisms that are involved in the formation of SVs and indels, I also performed an analysis to overlap mutations with recombination and crossover hotspots [Mancera et al., 2008], 3'UTR and 5'UTR [Nagalakshmi et al., 2008], nucleosome positioning based on H2AZ [Albert et al., 2007], origin of replication sites, DSB hotspots [Pan et al., 2011], TATA elements [Rhee and Pugh, 2012], transposable elements from the SGD, transcription start sites (TSS) [Zhang and Dietrich, 2005] and simple repeats from the SGD.

Around 26% of the mutations overlapped genes (Figure 2.4). The deletion mutants were classified into broad functional categories and the overlap of each category was assessed separately. There was no significant difference in the percentage of overlap with genes among the different functional categories (Kruskal-Wallis rank sum test, P=0.4329). However, I observed a depletion of variants overlapping genes when compared to a randomly generated list of mutations distributed along the genome (Wilcoxon rank-sum test, P<0.0001).

Among the affected genes, no particular GO term was significantly enriched as assessed by gene function analysis using the PANTHER classification system. In fact, 32.9% of genes belonged to unknown biological processes. This is in agreement with no selection being applied to through the accumulation of mutations.

No significant differences between the strains from different categories was found in the percentages of overlap with the above mentioned genomic features (Kruskal-Wallis rank sum test, P-values ranging from 0.439-0.569). On the other hand, there was a

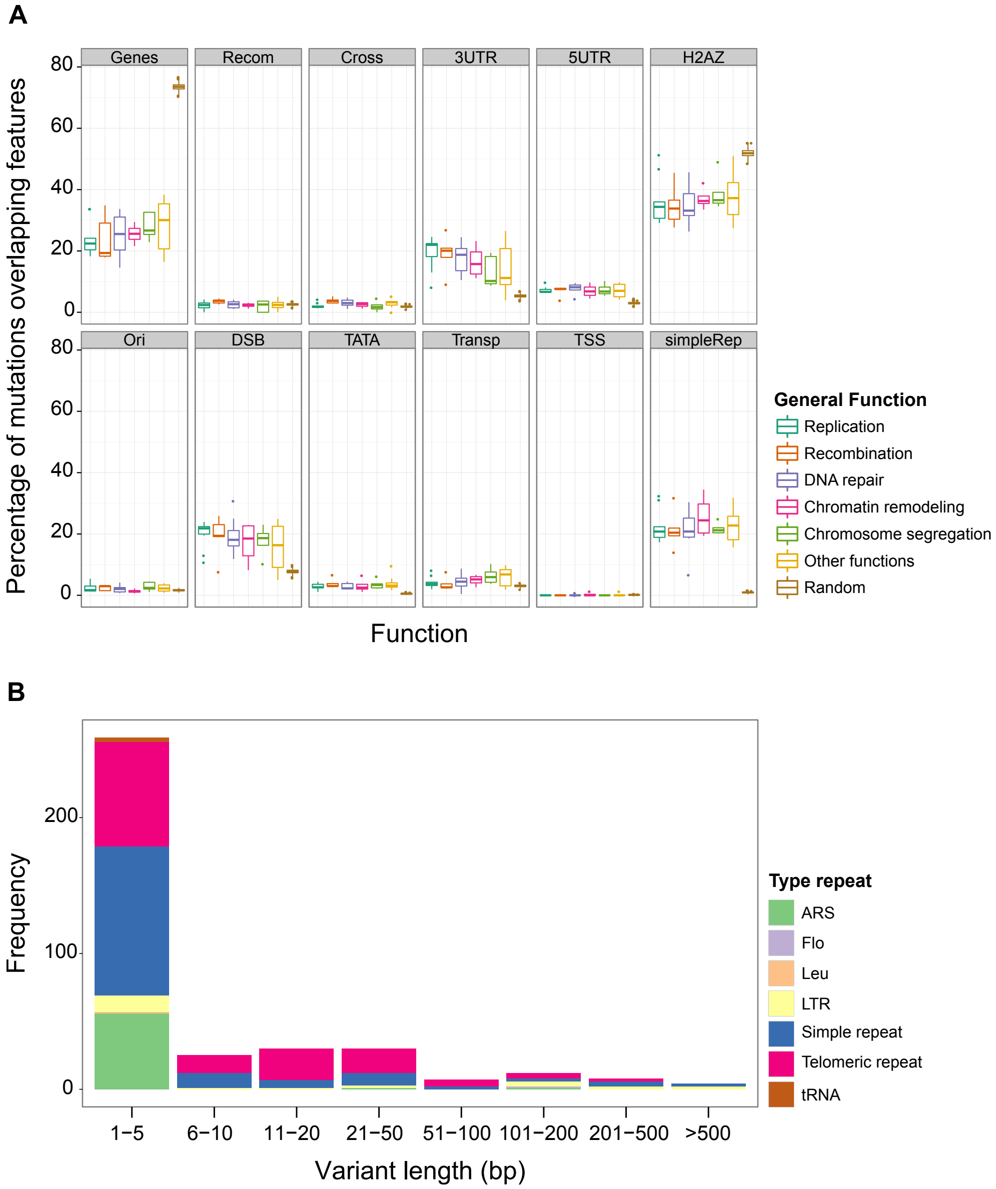

Figure 2.4: Overlap of SVs and indels with different features in the genome. **A** | Total *de novo* SVs and indels overlapped with different features in the genome. **B** | Most SVs and indels overlap with repetitive elements, in particular simple repeats and telomeric repeats. All strains were divided into broad functional categories and compared to a randomly distributed set of mutations. (Recom: recombination hotspots, Cross: sites of crossovers, 3UTR and 5UTR: 3' and 5'UTR of genes, H2AZ: nucleosome positioning, Ori: origins of replication, DSB: sites of double strand breaks, TATA: TATA elements, Transp: transposable elements, TSS: transcription start sites, simpleRep: simple repeats. ARS: autonomously replicating sequence, Flo: flocculin gene family sequence, Leu: leucine-rich repeats, LTR: long terminal repeat, tRNA: tRNA-derived repetitive elements. See text for data sources).

significantly higher number of mutations in our set overlapping 3’UTRs, 5’UTRs, DSB hotspots, transposable elements and simple repeats, as compared to the random set of variants (Methods Section A.1.4). There was also a significantly lower number of mutations overlapping sites of nucleosome positioning (P<0.0001).

Additionally, 22% of the identified mutations were found to overlap simple repeats and 5% transposable elements. Given that there are different types of simple repeats in the yeast genome, I then investigated which types of these repeats more commonly overlapped with mutations. Repetitive sequences such as telomeric repeats, autonomously replicating sequences or transposon related repeats were the most frequently mutated elements (Figure 2.4). In this regard, the presence of complete or fragmented Ty elements, retrotransposons that are very abundant in yeast, seemingly contribute to the formation of indels and SVs, presumably by mechanisms involving HR.

2.3.4 Single nucleotide variants

By applying the filtering steps depicted in Figure 2.5A, we were able to identify a total of 4628 SNVs following the WGS of the 47 strains. The studied strains also accumulated different amounts of SNVs through the generations, with a significant increase in b60 and b90 compared to b30 (one-tailed t-test, P<0.001) (Figure 2.5C). The total number of SNVs per MA line is shown in the Supplementary Table B.2.

Similar to the SVs and indels, the *msh2* deletion mutant accumulated the highest number of point mutations, confirming its hypermutable state. Other strains had a high number of SNVs, but not a high number of indels, suggesting that the formation of different types of mutations may be regulated by diverse mechanisms, and that different pathways are involved. However, some strains accumulated even lower numbers of SNVs as compared to the wild type strain, BY4741 (Supplementary Figure B.1), indicating that not all knockouts are heavily damaging or that there are coping mechanisms that allow the cells to maintain genome stability in spite of defects in particular pathways.

The types of nucleotide substitutions for the top 10 strains are shown in Figure 2.5D. The *msh2* mutants had a high number of G>A and G>T mutations. Other strains with high number of SNVs included *rad52*, a gene essential for homologous recombination, and two deletion strains involved in stabilizing damaged or stalled replication forks, namely *rtt107* and *ctf4*.

Mutation rates per base per generation in each deletion strain ranged from as low as 1.35×10^{-9} in *slx1* to 13.19×10^{-9} in *msh2* (median=3.33×10^{-9}) (Figure 2.6A). As a comparison, the wild type strain showed a mean mutation rate of 4.13×10^{-9}. These estimates were in agreement with previously reported genome-wide mutation rates per base per generation [Lynch et al., 2008]. On the other hand, lower mutation rate estimates have also been published for a wild type strain [Lang and Murray, 2008].

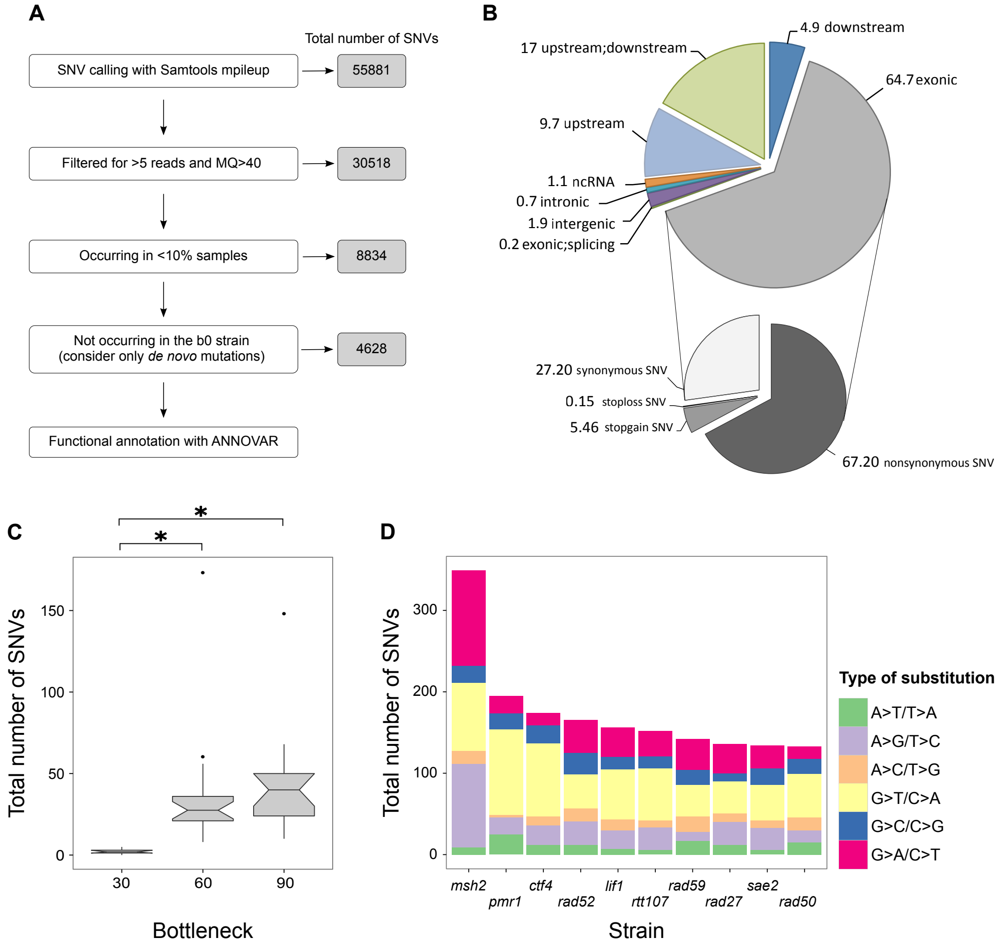

Figure 2.5: SNV discovery and filtering pipeline and total number of mutations. **A** | Discovery and filtering pipeline to generate a collection of *de novo* SNVs in the MA lines (MQ: Mapping quality). **B** | Functional annotation of SNVs. Larger pie chart shows the percentages of all SNVs at exonic, intronic and intergenic regions. The smaller pie chart depicts the percentages of the different categories for exonic variants (down and upstream indicate variants that are located 1kb from the the transcription start or end site respectively; splicing refers to variants within 2bp of a splicing site; and exonic are variants that overlap a coding exon [Wang et al., 2010]). **C** | Mean number of SNVs per bottleneck accumulated in the 47 yeast deletion strains (* indicates one-tailed *t*-test $P<0.001$). **D** | Total number of *de novo* SNVs and the types of nucleotide substitutions in the 10 strains with the highest number of SNVs. The sum of the total number of *de novo* mutations occurring in the two mutation accumulation lines is shown.

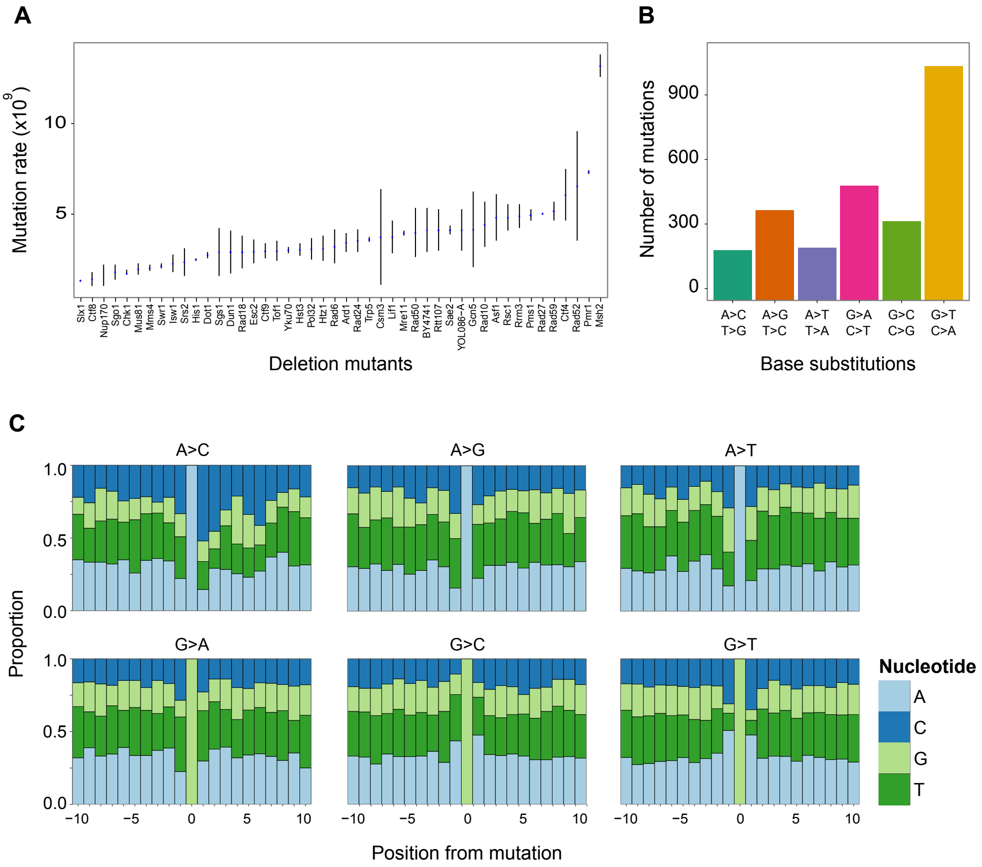

Figure 2.6: Mutation rates and types of single nucleotide substitutions. **A** | Mutation rates of the 47 yeast deletion mutants estimated by the accumulation of mutations in 60 bottlenecks. **B** | Number of mutations in each base substitution class. **C** | Across the genome, somatic variants occur mainly at G-C base pairs. The 0 marks the position of the substitution, 10bp upstream and downstream are shown.

However, these calculations were done by evaluating mutations in only two loci, and both loci differed significantly, indicating that there is variation in the mutation rates along the genome.

Considering the overall set of SNVs in all the strains, we identified G>T transversions as the most common type of SNV (37.7%) (Figure 2.6B). The second most prevalent change was G>A (18%). This pattern represents an overview of all the mutant strains together. We found that this high amount of transversions (G>T) was different to wild type strains where C>T mutations are the predominant type [Zhu et al., 2014]. However, it is important to keep in mind, that even if this is the general trend, the pattern was not uniform among all the strains, as mentioned before for the top 10 strains.

Across the genome, somatic variants occurred mainly at G and C base pairs, corresponding to mutations from G>T and C>T. By looking at the surrounding (±10bp) sequences from each point mutation class, we could identify that 50.8% of these G>C substitutions were followed by an A and 31% by C (Figure 2.6C). The second most common type of substitutions was G>A, in with the neighboring bases were mainly A and T.

Furthermore, we performed functional annotation of SNVs using ANNOVAR [Wang et al., 2010]. In total we found that around 64% of the SNVs were exonic, out of which 67% were nonsynonymous substitutions (Figure 2.5B), similar to the expected 75% of nonsynonymous changes if the mutations were random. We also detected 334 SNVs causing stop gain mutations, some in important genes involved in the maintenance of genome stability. For example, a stop gain was observed in *MLH2*, a protein involved in mismatch repair, in the *ctf9* deletion strain. Additionally, mutations in *MEC1* and *MEC3*, genes which have essential roles in DNA replication, repair and telomere maintenance, were observed in *pms1* and *mus81* deletion strains. This accumulation of mutations in strains with a knockout background makes it difficult to unravel the specific contributions to the mutational landscape of the original deletion and newly acquired mutations.

2.4 Aneuploidy and the accumulation of mutations

We also identified several deletion mutants that had entire chromosome gains giving rise to aneuploid strains (an example is shown in Figure 2.7A). In total we found 15 MA lines with 1 chromosome gain, 3 strains with 2 chromosome gains, and 2 strains with 3 chromosomes gained. Aneuploidies were detected only in strains in the 60th or 90th bottlenecks, but not earlier (Table 2.2). Chromosome I was gained more often than other chromosomes in most of the aneuploidy strains, which is in agreement with the fact that this is the shortest of all chromosomes in yeast, with 230kb, and it contains the least number of genes. However, the number of other chromosomes gained did not correlate with their lengths.

Strains with chromosome gains had higher number of SNVs (one-tailed *t*-test, P=0.01) compared to euploid strains (Figure 2.7B). In particular, *ctf4* (involved in chromatin cohesion), *rtt107* (important for replication fork repair) and *rad27* showed the highest numbers of SNVs among the aneuploid strains. Although there was no significant increase in the numbers of deletions for the aneuploid strains in general, *rtt107* and *sgo1* did show higher number indel events compared to the euploid strains (Figure 2.7C).

2.4.1 Validation of aneuploidies

I performed validations of aneuploidies, specifically for the gains of chromosomes I, II and VI, by qPCR. A total of 13 chromosome gains were tested. We followed a similar approach as described by Pavelka et al. [2010], in which for each chromosome tested, two regions (one in each arm) were chosen to increase the confidence of the aneuploidy

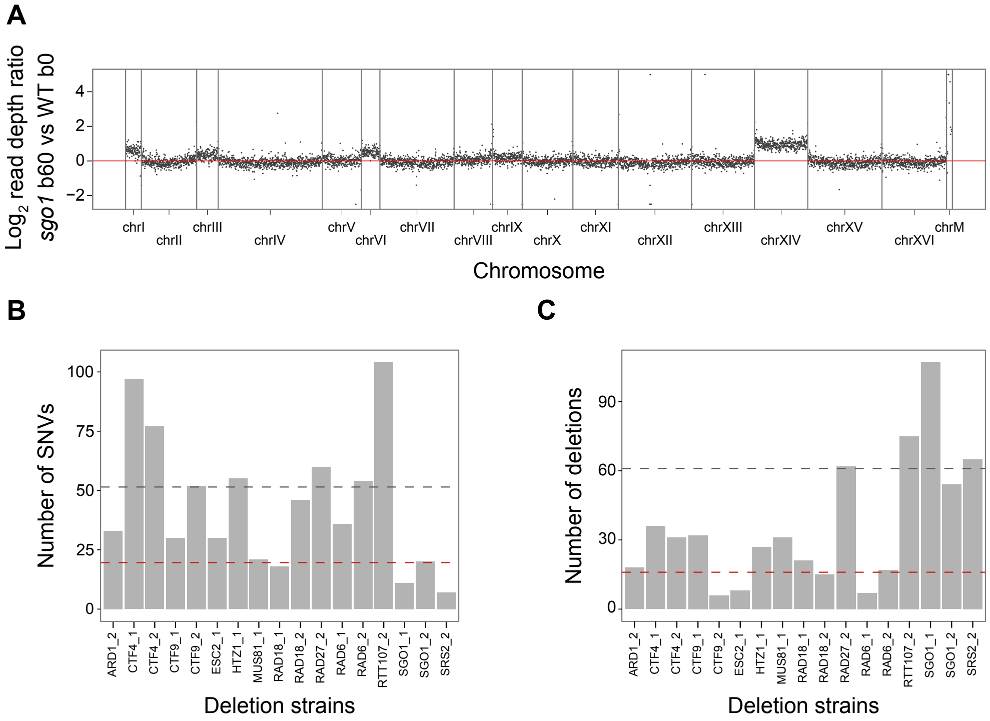

Figure 2.7: Several yeast deletion strains became aneuploid at the later bottlenecks and carried higher number of deletions and SNVs. **A** | Example of whole genome read depth plots for one of the *sgo1* mutation accumulation lines at bottleneck 60 (b60) with a gain of chromosome XIV compared to the wild type (WT) control at the beginning of the experiment (b0). This chromosome gain was absent in the b0 *sgo1* strain as well. **B** | Number of SNVs in the aneuploid strains. **C** | Number of deletions in the aneuploid strains. (For plots B and C the grey dashed line marks the mean count in all euploid strains and the red dashed line marks the mean count in the control wild type strain; "_1" and "_2" refer to the two mutation accumulation lines).

Table 2.2: Number of MA lines with chromosome gains after 60 and 90 bottlenecks. No aneuploidies were observed after 30 bottlenecks.

Number of chromosomes gained	Number of MA lines	b60	b90
1	15	13	2
2	3	3	0
3	2	1	1
Total	20	17	3

detection. Initially, the primers from these authors were used in efficiency testing. However, due to low efficiencies for the primers in chrII and chrVI, new primers were designed as described in Section A.1.6. All primer sequences used are also shown in the Supplementary Table B.4. Primer efficiencies for the tested chromosomes were high, ranging from 93% to 97%, confirming that they were usable for the validations. Example standard curves for chrI are shown in Figure 2.8A and B, showing a good agreement between the C_t and the quantities of the standards (r^2, Spearman P<0.001). Similarly, good primer efficiencies were obtained for the other two chromosomes and the selected internal reference locus *SPT15*(Supplementary Figure B.2).

Aneuploidies were supported by a significant correlation between the fold change predictions obtained with the left and right arms of chrI (r^2=0.956, P<0.0001) as shown in Figure 2.8C. Similarly, there was as good correlation between both arms of chrII (r^2=0.926, P<0.0001), and of chrVI (r^2=0.937, P<0.0001).

Additionally, significant correlation between the chromosome copy number predictions based on the qPCR results and the read depth based approach (for chrI r^2=0.953, Spearman P<0.0001, Figure 2.8D) also confirmed the chromosome gains. Based on these values, we were able to confirm aneuploidies for a total of 13 chromosome gains, specifically for 6 samples with gain of chrI, 3 samples with gain of chrII and 4 samples with a gain of chrVI.

2.5 Discussion

Based on next-generation sequencing technologies, I generated a catalog of genomic variants, including deletions, short insertions, duplications and SNVs for 47 yeast deletion strains. This allowed us to make inferences of the possible impact of deficiencies in genome maintenance pathways and of broad mechanisms involved in the formation of mutations in the yeast genome. I also estimated mutations rates and described the most common nucleotide substitution classes arising under a neutral accumulation assay.

I showed that in spite of the long term accumulation of mutations, the total amount of rearrangements per strain was in general low. This result was not completely unex-

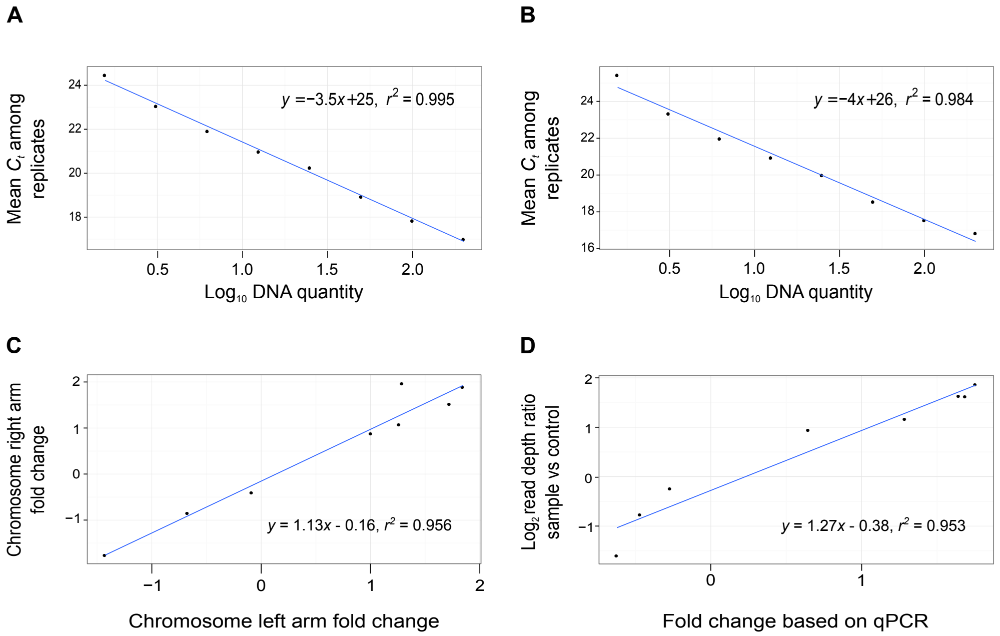

Figure 2.8: Validation of aneuploidies in chrI by qPCR. **A**, **B** | Standard curves for primer efficiency estimation for regions in the left and right arms of chrI respectively. **C** | Correlation between the fold change of the left and the right arms of chrI estimated by qPCR. There is good agreement in the predicted fold change between both arms. **D** | Correlation between the predicted fold changes based on qPCR and copy number changes detected by read depth ratios between each sample and the wild type control at b0.

pected given the reported stability of the yeast genome. In fact, Nishant et al. [2010] have shown that during both vegetative and meiotic growth, yeast strains remain with very few mutations after over 1700 generations. They used wild type strains to study the accumulation of mutations during these asexual and sexual growth phases. Therefore, their mutation rate estimates reflect the stability of this particular strain.

Similarly, Lynch et al. [2008] observed a total of 33 single base pair substitutions, 1bp deletion and one 3bp inversion after performing similar MA experiments in wild type strains. Regarding larger scale variants (>10kb) they reported only 11 inversions and 4 deletions. Taken together, their results also support the idea of a very stable yeast genome, robust to changes, at least in the wild type strain.

In this regard, our MA approach, using deletion mutants from the Yeast Deletion Collection, proved to be a good experimental method of accumulating neutral mutations in numbers high enough to be able to observe more general patterns of mutations. We identified higher numbers of short deletions and insertions, which in general overlapped with repetitive elements. Additionally, differences in the mutational landscapes of the deletion strains were observed and although not all can be fully understood, some patterns may reflect more directly the deficiencies in the particular pathway associated with the deleted gene.

For example, we identified *msh2* as being the deletion strain with the highest number of deletions, short insertions and SNVs. In addition to its role in MMR, which can explain the large number of SNVs accumulated, Msh2 can inhibit DNA recombination between sequences that are not 100% identical [Selva et al., 1997]. Therefore, the lack of this protein can lead to increased numbers of deletions due to higher recombination between direct repeats, even if they are slightly dissimilar.

The deletion of *MSH2* has been shown to make the cells become hypermutable, creating mutator strains [Huang et al., 2003]. These strains have significantly higher rates of mutation due to the deficiencies in important genome maintenance processes. Our results are also supported by observations of human samples with mutations in the homolog gene, *MSH2*. Interestingly, germline mutations of *MSH2* account for a significant predisposition to cancer [Dowty et al., 2013]. They constitute one of the main causes of Lynch syndrome [Rahner et al., 2007; Bonadona et al., 2011], an autosomal-dominant cancer predisposition syndrome. Mutations in human *MSH2* can also lead to a hypermutation state in which the cells have an increased rates of mutations, for example in prostate cancer [Pritchard et al., 2014]. In this state, the tumor cells also showed signs of microsatellite instability.

The reasons for the high number of short insertions in the *rad27* deletion mutant together with the deletions in this mutant may be also explained by the roles that Rad27 plays in the cells. This protein is involved in maintaining genome stability by processing Okasaki fragments, base excision repair and preventing the expansion of repeats [Reagan et al., 1995]. Consequently, strains deficient in this protein are expected to accumulate a high number of point mutations and microsatellite instability.

However, Rad27 can also participate in restricting recombination between homologous sequences. Interestingly, Negritto et al. [2001] have shown that the shorter the distance between homologous repeats, the more likely it is to be lost in *rad27* mutants. Therefore, deletions in this strain may be explained by deficiencies in homologous recombination. On the other hand, *rad27* mutants can exhibit higher amounts of short insertions and duplications flanked also by direct repeats due to its role in repairing DSBs during replication [Tishkoff et al., 1997]. *RAD27* is the homolog of the human *FEN-1* gene, and has an exonuclease role in removing “flaps” of DNA (short sections of ssDNA) during replication. Mutations in this gene have been associated with cancer [Zheng et al., 2007]. And finally, Rad27 can inhibit Ty1 mobility [Sundararajan et al., 2003], likely preventing the accumulation of deletions, which as shown in this study, overlap frequently with transposable elements.

Isw1 and Swr1 play roles in chromatin remodeling [Mizuguchi et al., 2004; Smolle et al., 2012]. Swr1 is known to be recruited to DNA double strand breaks and has been recently implicated in facilitating NHEJ at sites of DSBs by contributing to the mobility of the DNA DSBs to the periphery of the nucleus, where it gets stabilized [Horigome et al., 2014]. Being among the strains with high mutation numbers implicates that deficiencies in chromatin remodeling pathways are important sources of mutations. Furthermore, the role of Sgo1 in preventing the formation of short deletions is not completely clear, since its main role is related to accurate segregation of chromosomes through its sensing of mitotic chromosome tension [Indjeian et al., 2005]. However, its high number of mutations may be related to the fact that this strain became aneuploid in the later generations, as discussed below.

Since the yeast genome is very dense, with only less than 30% being intergenic [Goffeau et al., 1996], it is not surprising that we observed 65% of exonic mutations. Similar to our findings, Lynch et al. [2008] also reported around 27% of mutations being intergenic. The amount of nonsynonymous SNVs observed, close to 75%, which is the expected if mutations were randomly distributed, is in agreement with the neutral accumulation of mutations. This is especially true given that our mutation accumulation assay did not involve any selection process and therefore the mutations acquired are mostly neutral, even if they occur in exons. In fact, Giaever et al. [2002] observed that only 15% of knockout strains had slower growth rates under normal environments, pointing out that some genes may have more important roles only under stressful conditions.

Our results indicate that repetitive regions are more prone to have mutations, in agreement with other studies Nishant et al. [2010]. These results suggest that replication slippage and homologous recombination play an important role in the formation of indels. Nishant et al. [2010] similarly found that the majority of variants occurred at subtelomeric regions. In addition to the high content of repetitive sequences in these regions, the lower gene content indicates that mutations there may have less phenotypic impact [Nishant et al., 2010; Wellinger and Zakian, 2012]. Another possibility is that the number of false positive variant predictions may be higher in repetitive regions due to mapping artifacts.

The observed pattern of nucleotide substitutions (Figure 2.6B) is very similar to the one described for small-cell lung cancer, with G>T mutations being the most prevalent substitutions [Pleasance et al., 2010b; Lee et al., 2010; Pfeifer and Hainaut, 2003]. These mutations are usually the result of the conversion of guanine to 8-oxo-guanine, a lesion commonly induced by oxidative agents found in tobacco [Paz-Elizur et al., 2003; Feng et al., 2006]. Although there is obviously no direct connection between this type of mutagenesis and the yeast deletion mutants, there are indeed some tobacco agents that inhibit DNA repair [Feng et al., 2006] and there are some yeast deletion strains in our set that are deficient in this pathway as well. The G>A mutations, constituting the second most common substitutions, are probably the result of the spontaneous deamination of cytosine [Duncan and Miller, 1980; Maki, 2002].

The results for SNV sequence context are in agreement with other studies in yeast strains. Zhu et al. [2014] also found higher mutations rates at C and G positions. In fact, the mutation rates at these positions were also dependent on the neighboring sequence. The highest mutation rates was seen for cGg and cGa environments (where the big letters indicate the site of mutation and the small letters the preceding and following one respectively), comparable to our observation of G mutations frequently followed by an A.

Furthermore, I observed that some strains became aneuploid at later generations and that these strains carried higher numbers of mutations compared to the euploid strains. For example, the deletion of *RTT107* caused whole chromosome gains together with a high accumulation of indels and SNVs. This protein contains BRCT (BRCA1 C-terminal) domains that function in the recruiting of signaling and repair factors upon DNA damage [Mohammad and Yaffe, 2009]. Additionally, Rtt107 can interact with the SMC5/5 complex which is involved in the maintenance of sister chromatid cohesion and by this prevents errors during DNA repair. In fact, Rtt107 is recruited to the sites of DSBs and is important for the repair through the sister chromatid recombination pathway [Ullal et al., 2011; Leung et al., 2011]. Therefore, defects caused by the deletion of this protein can certainly lead to both whole chromosome gains and high indel formation rates.

Moreover, the *sgo1* deletion mutant also became aneuploid at later generations and showed a high accumulation of short deletions, consistent with its role in correct chromosome segregation and in keeping with previous reports that aneuploidy can drive genomic instability in yeast [Sheltzer et al., 2011]. Interestingly, a similar phenomenon has been described for human cancers. The inactivation of *STAG2* gene, involved in the correct separation of sister chromatids, results in chromosomal instability [Solomon et al., 2011] and may be related to tumorigenesis.

The direction of the relationship between aneuploidy and the accumulation of mutations in our study is not clear yet, whether the aneuploidy was an event preceding the accumulation of mutations or if it was the other way around. However, our results are in agreement with earlier reports showing that aneuploidy may be responsible for genomic instability, even though the gain or loss of a chromosome is usually detrimental to the cells [Sheltzer et al., 2011; Zhu et al., 2012]. Nevertheless, the heterogeneity provided by the increased instability may be beneficial in more selective environments.

As mentioned before, under neutral mutation accumulation, the total number of events scored was higher as compared to other studies. However, our study is still limited by the total number of mutations detected. A larger collection of mutations is required to identify patterns like kataegis, where by studying over 100,000 events, it was possible to identify the clustering of substitutions occurring in breast and other cancer types [Nik-Zainal et al., 2012a; Alexandrov et al., 2013; Chen et al., 2014].

Another limitation of this project is the fact that we chose to study 47 deletion mutants with known defects in the maintenance of genome stability. Therefore, all our results relate to mechanisms involving those pathways, such as DNA repair by homologous recombination or mismatch repair. The constraint of studying a small set of strains was in part overcome by setting up a genome-wide approach, which will be described in the following chapter.

The results of this section indicate that neutral mutational processes are complex and involve a combination of added effects instead of being the result of one single gene knockout. The mutational landscapes could, in some of the cases, be attributed to the particular deficiency caused by the knockout of a gene, such as the case of *msh2* mutants. In others, the patterns of mutations may be the contribution of a deficiency in a specific pathway plus acquiring some other slightly damaging (although not deleterious) mutations. These processes resemble those occurring in higher eukaryotes and can be used to understand, for example, aspects of disease progression, evolution of tumors or sites with higher risks of mutations.

Chapter 3

Genome-wide screen for genes that suppress deletion formation

3.1 Motivation

Despite the extensive work carried out in budding yeast assessing mutation rates and the processes involved in the maintenance of genome stability, most studies have focused on SNVs and short deletion and insertions [Lynch et al., 2008; Zhu et al., 2014], but not larger variants. Indeed, our study on the mutation accumulation lines presented in Chapter 2 indicated that these types of short mutations are the most abundant in the genome, under neutral growth conditions.

Moreover, several studies have reported that the disruption of pathways involved in DNA repair, recombination and replication can lead to an increase in the rates of SV formation (commonly referred to as gross chromosomal rearrangements, GCR) [Myung et al., 2001b,a; Smith et al., 2004; Motegi and Myung, 2007; Kanellis et al., 2007]. However, due to the low rates of occurrence of these types of mutations, some studies focus on strains with defects in known pathways of genome maintenance [Myung et al., 2001b]. Others enhance the levels of SV formation by the introduction of a *PIF1* deletion, known to significantly increase the levels of GCR, in addition to the specific KO [Myung et al., 2001a; Smith et al., 2004].

From our previous results in yeast deletion mutants under a neutral accumulation assay, we found that mutation frequencies and types of mutations acquired could vary depending on the specific gene defect each strain carries. These results indicate that distinct disrupted pathways can lead to the formation of different mutation types, probably through a combination of mechanisms. These findings motivated us to find an independent but complementary approach to study a wider range of deletion mutants with the aim of identifying strains that have not been previously linked to increased levels of mutation formation.

We were interested in investigating larger structural variants. SVs may have been missed in previous studies either because the experimental methods did not detect them, or because these larger variants occur in very low rates impeding their detection. Therefore, we designed an assay in order to identify KO mutants that are more prone to form deletions. To avoid an *a priori* selection of genes to be studied, the methods and results I describe here are based on screening a pooled collection of around 5000 homozygous deletion strains, in a series of bulk experiments, where all the deletion mutants could be tested simultaneously.

The advantages of this method, which has no biases of strain selection and which can be performed in a pool of deletion mutants, proved useful for the identification of genes not previously known to be involved in genome maintenance.

3.1.1 Contribution

The experimental set up presented in this chapter was designed in collaboration with Megumi Onishi-Seebacher, a former postdoc in the group of Jan Korbel. I performed all wet lab experiments, data analyses and validations described here.

3.2 An experimental approach for identifying and enriching for strains that acquire deletions

Using the deletion strains from the Yeast Deletion Collection, we designed an approach to identify genes that suppress the formation of deletions. By optimizing a bulk transformation assay, I inserted specifically designed constructs into most of the deletion strains and performed experiments using the pooled collection of mutants in a high-throughput way. I also tested the effects of several chemical agents causing direct DNA damage or replication stress on the formation of deletions.

3.2.1 Design of a construct to detect deletions

With the main goal of identifying strains with increased genome instability, we designed a construct in a way that allows the selection of mutant strains that acquire deletions. The construct is used to identify strains with defects in particular pathways that are more prone to form deletions (Figure 3.1A).

The construct carries a *URA3* gene to select for positive transformation by growth on media lacking uracil and the hygromycin phosphotransferase gene (*HPH*) to confer hygromycin resistance. More specifically, the *HPH* gene is interrupted by a modified actin intron which contains the *URA3* gene inside, making it longer than the original one. Important sequences required for the splicing of the intron were kept in place, preserving the distance between the 5' splice site and the branchpoint [Thompson-Jager and Domdey, 1987].

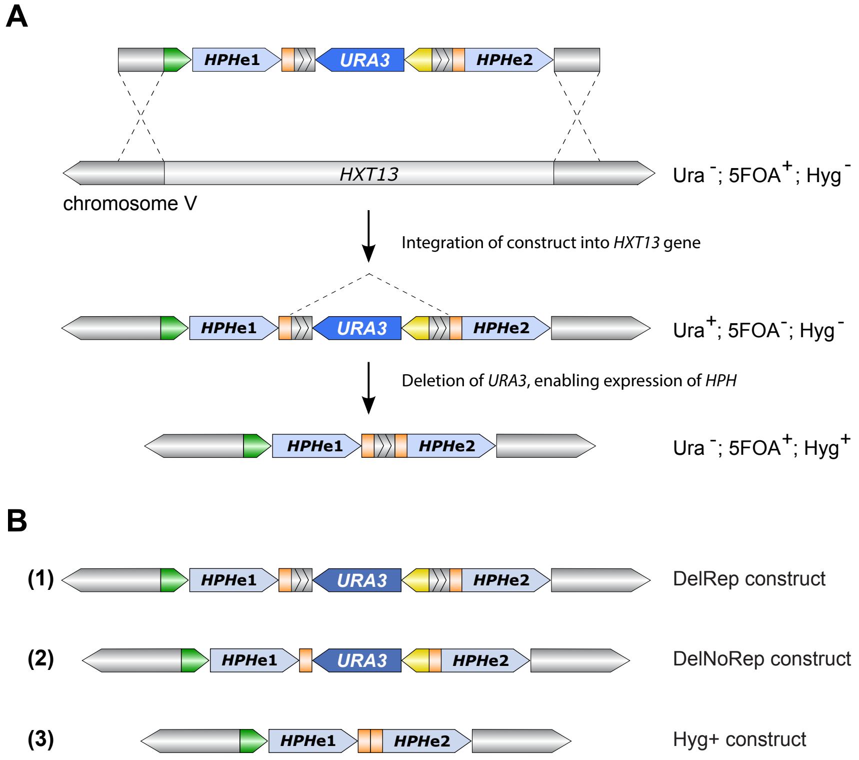
A
HPHe1
URA3
HPHe2
HXT13
chromosome V
Ura $^-$; 5FOA $^+$; Hyg $^-$
Integration of construct into *HXT13* gene
HPHe1
URA3
HPHe2
Ura $^+$; 5FOA $^-$; Hyg $^-$
Deletion of *URA3*, enabling expression of *HPH*
HPHe1
HPHe2
Ura $^-$; 5FOA $^+$; Hyg $^+$
B
(1)
HPHe1
URA3
HPHe2
DelRep construct
(2)
HPHe1
URA3
HPHe2
DelNoRep construct
(3)
HPHe1
HPHe2
Hyg+ construct

Figure 3.1: Constructs designed to detect deletions. **A** | The constructs contain the *HPH* gene that confers resistance to hygromycin. This gene is divided into exon1 (*HPH*e1) and exon2 (*HPH*e1) by a modified actin intron that contains the *URA3* gene with its own promoter. For integration into the yeast genome, the left and right sides of the constructs have 40bp homologous sequences up and downstream of the *HXT13* gene in chrV. Positive transformation of the constructs is selected in Ura- medium. Upon deletions occurring in the constructs (dashed lines) the strains become hygromycin and 5FOA (5-fluoroorotic acid) resistant. **B** | (1) The **DelRep** version of the construct has the *URA3* gene surrounded by direct repeats (grey arrows), 30bp long, belonging to a human *Alu* sequence and by actin intron splice sites (orange boxes). (2) The **DelNoRep** version of the construct lacks the direct repeats. (3) A **Hyg+** control construct lacks the *URA3* gene and has constitutive hygromycin resistance. (Green arrow: *ADH* promoter, Yellow arrow: *URA3* promoter).

The rationale behind this design is that most introns in the budding yeast genome are approximately 100bp long. There are some as short as 60bp and some as long as 600bp, the longest ones mainly found in ribosomal genes [Spingola et al., 1999; Kupfer et al., 2004]. Longer versions of the actin intron are either spliced in a more inefficient manner, or are not spliced at all [Klinz and Gallwitz, 1985]. The actin intron, which is 308bp long [Ng and Abelson, 1980], has been shown to be spliced when inserted into other genes in the yeast genome [Yu and Gabriel, 1999; Weigand and Suess, 2007], making it a useful tool as an artificial intron. Hygromycin resistance is only obtained upon deletions that remove or shorten the length of this intron to a size that can be spliced. Therefore, by screening for hygromycin resistance, we were able to detect strains with deletions in the construct.

Since the rate of occurrence of large chromosomal rearrangements is expected to be low, two versions of the construct were designed (Figure 3.1B). The two constructs differ only in that one contains two direct repeats flanking the *URA3* gene, in order to increase the rate at which these mutations occur. These repeats correspond to 30bp from the human *Alu* element consensus sequence [Voineagu et al., 2008] to avoid potential homologous recombination between these sequences and other regions of the yeast genome. In the following sections I refer to these two constructs as “DelRep” (construct with direct repeats) and “DelNoRep” (construct without direct repeats).

Additionally, the two versions of the construct are surrounded by two 40bp homologous sequences for insertion by homologous recombination into the nonessential gene *HTX13*, located in the left arm of chromosome V. This locus has been used to test for chromosomal rearrangements in other studies [Chen and Kolodner, 1999; Myung et al., 2001b], where it was shown that the disruption of this locus has no or little impact on the fitness of the cells. Some deletion strains may have deficiencies in homologous recombination pathways. Therefore, they may be excluded from the final set of strains that can be transformed with the construct, since its insertion is dependent on the HR mechanism.

As a control, another construct with the same structure as the deletion constructs was generated in which the hygromycin resistant gene was not interrupted by the long intron (Figure 3.1B). Hygromycin will be expressed after insertion in the yeast strains, without the need of rearrangements occurring in the cassette. Therefore, this control construct (hereafter called “Hyg+” construct) confers constitutive hygromycin resistance to the cells.

3.2.2 General experimental workflow

To be able to assess a wide range of deletion strains simultaneously and in an unbiased fashion, we optimized a bulk transformation experiment in which we used a pool of the yeast deletion mutants from the Yeast Deletion Collection to introduce the constructs described in the previous section. A general workflow of the concepts of the full experiment is shown in Figure 3.2. As mentioned in Section 1.6.2, each strain in this collection has molecular barcodes (uptag and downtag) that can be used to identify them (Figure 1.4). Therefore, the composition of this initial pool was assessed by amplifying the barcodes of the deletion strains in a single PCR reaction and using next-generation sequencing methods to sequence the pool of amplicons in order to detect the barcodes (Figure 3.2A). All procedures are described in more detail in Section A.2.

Several independent bulk transformation experiments using aliquots of this original pool were used to introduce the three different types of constructs (DelRep, DelNoRep and Hyg+) (Figure 3.2B). I obtained a total of 20,000 transformants per construct. Considering that there are around 5000 strains in the collection, this number of transformants allowed us to cover each strain almost 4 times to have a higher probability of being represented in the pools of transformed strains.

After performing these transformations I confirmed the insertion of the constructs by PCRs (Figure 3.3A). I successfully amplified the uptags and downtags in 10 independent PCRs for each transformed pool (an example is shown in Figure 3.3B) and the products were mixed together. These PCR amplicons were then used for the sequencing of the tags. Three replicates were performed for each construct. Based on this, I confirmed the presence of, on average, 4852 deletion strains in the original pool, meaning that the majority of the 5083 deletion strains from the Yeast Deletion Collection could be detected by the PCR amplification of the tags. In the transformed pools, a mean of 76% of all strains from the original pool could be identified (Figure 3.3C). The total number of strains identified in each transformed pool is shown in the Supplementary Table B.7.

The strains recovered after the transformation of the different constructs were similar in the different pools. We observed an overlap of 90% between the pools transformed with DelRep and DelNoRep, and approximately 85% overlap between these two pools and the Hyg+ transformed one (Figure 3.3D). Given that these sets had a large representation of the deletion collection strains and that the overlap between them was also very high, this gave us a set of transformed strains with which we performed all the following experiments.

Additionally, we tested whether the amplification of both the uptags and downtags was similar and that we could recover equal numbers and types of strains with both tags. Indeed, we saw a high correlation between the number of reads per strain recovered with the up and downtags for all the transformed pools (for example, for the original pool r^2=0.78, Spearman P<0.001, Supplementary Figure B.3).

With the transformed pools, we then performed several experiments to identify the strains with increased formation of deletions. To increase the rates of mutations, the pools were treated with different stressors after which the selection for those strains

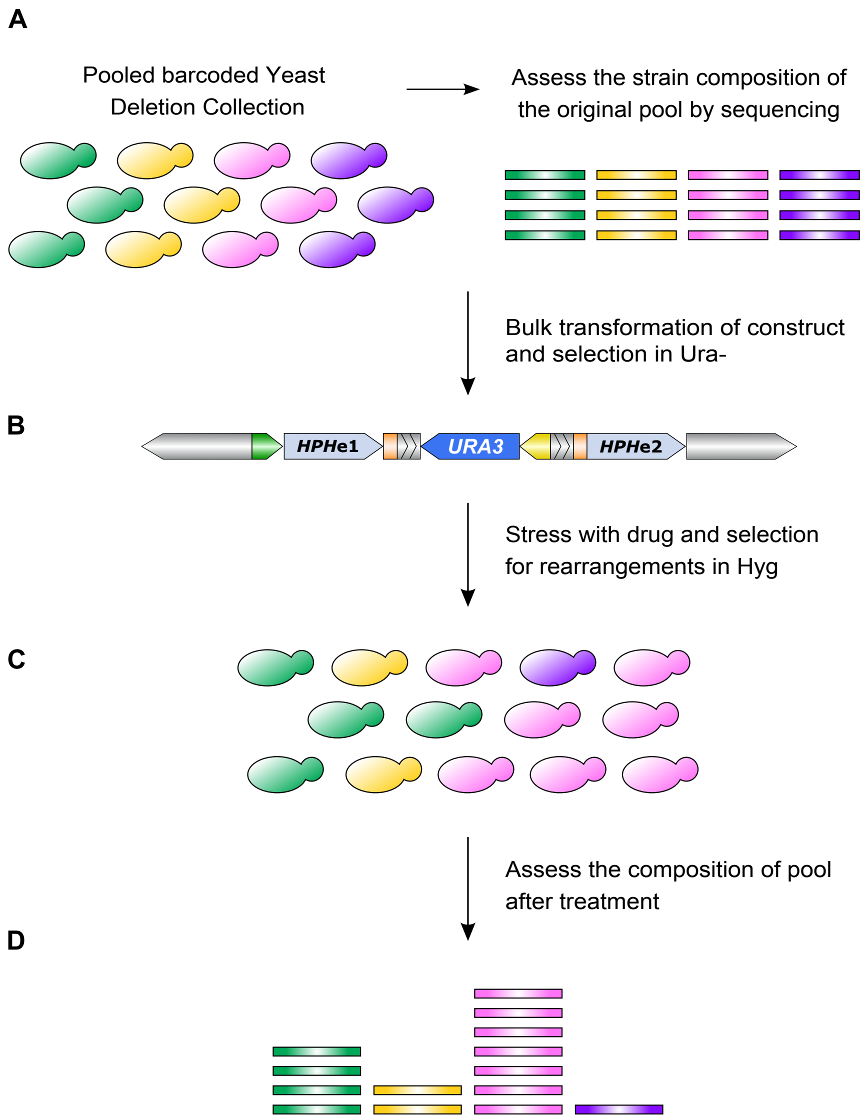
A
Pooled barcoded Yeast Deletion Collection
Assess the strain composition of the original pool by sequencing
Bulk transformation of construct and selection in Ura-
B
HPHe1
URA3
HPHe2
Stress with drug and selection for rearrangements in Hyg
C
Assess the composition of pool after treatment
D

Figure 3.2: Experimental workflow to identify genes that suppress the formation of deletions. **A** | The experiment starts with a pool of all the homozygous deletion strains from the Yeast Deletion Collection. The strain composition of this original pool is assessed at the beginning of the experiment by amplifying and sequencing the barcodes. **B** | Using the pooled deletion strains, the designed constructs are introduced in a bulk transformation step. **C** | The transformed pools of deletion strains are then subjected to different drugs to induce the formation of rearrangements, and the strains that acquire mutations in the construct are selected. **D** | Finally, the composition of the pools after selection (containing the strains that acquired deletions) are assessed by the sequencing of the barcodes.

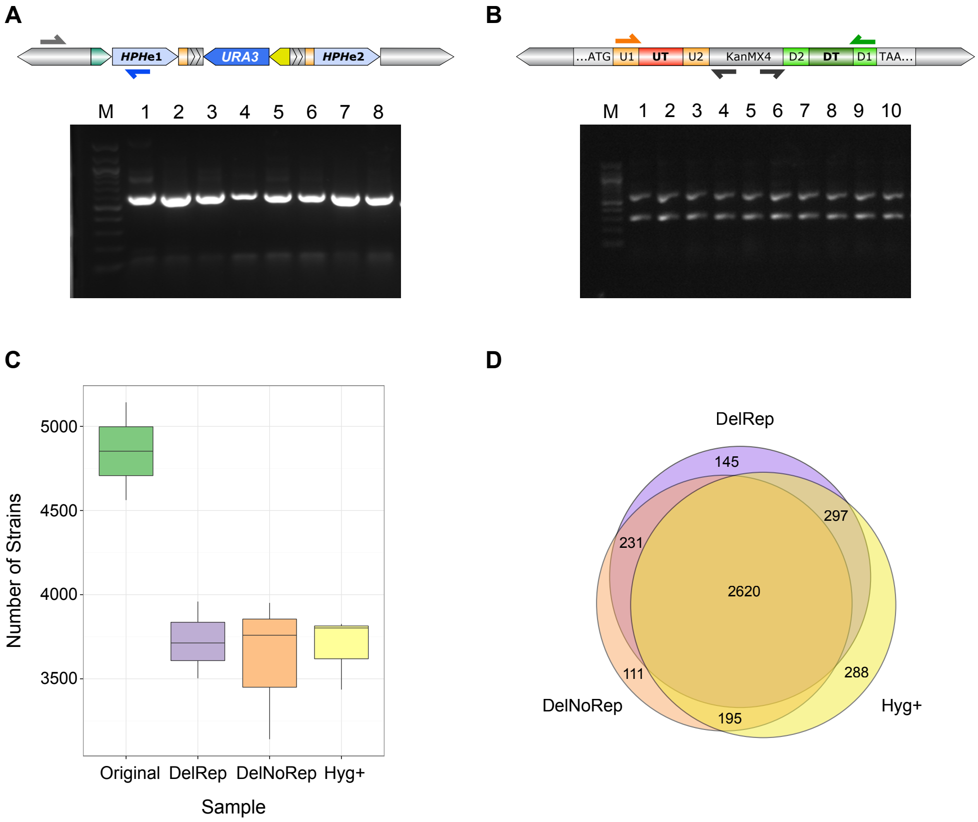

Figure 3.3: PCR validation of construct insertion and amplification of the tags. **A** | Verification of the insertion of the construct into the *HXT13* gene in chrV. Eight single different colonies are shown (M: 100bp marker). **B** | Amplification of the uptags and downtags to assess the composition of the pools of deletion strains was done using universal primers internal to the KanMX cassette and at the ends of the tags. Both tags were amplified in a single PCR reaction. PCRs for ten different pools are shown (M: 100bp marker). **C** | Number of strains detected after sequencing of the original pool of deletion strains and after transformation with DelRep, DelNoRep and Hyg+ constructs. **D** | Shared strains between the pools transformed with the DelRep, DelNoRep and Hyg+ constructs before treatment (only strains present in all replicates of each construct transformation are shown).

that acquire deletions was done on a hygromycin-containing medium (Figure 3.2C). The composition of these final sets of strains was then assessed similarly as before, by sequencing the barcodes and comparing them to the control (Figure 3.2D).

The stressors chosen are all known to cause either directly or indirectly DNA damage: hydroxyurea (HU), methyl methanesulfonate (MMS), doxorubicin (Doxo) and camptothecin (Campt). HU can cause replication fork arrest [Bianchi et al., 1986], which has been linked to an increased amount of DSBs [Arnaudeau et al., 2001; Petermann et al., 2010; Arlt et al., 2011] leading to chromosome rearrangements. MMS is a DNA alkylating agent that can lead to DNA DSBs [Chang et al., 2002] and has been shown to be a carcinogen [Beranek, 1990]. Doxorubicin and camptothecin are both inhibitors of DNA topoisomerases (I and II respectively) [Liu et al., 2006; Patel et al., 1997]. Doxorubicin can also intercalate in the DNA resulting in the formation of DNA DSBs [Patel et al., 1997].

3.2.3 Experiment design

To identify strains with increased levels of deletion formation, we grew independent aliquots of the transformed pools with different drug treatments or without applying any stress (YPAD control). For consistency, each transformed pool, namely the strains containing the deletion construct with repeats, without repeats and the control with the construct conferring constitutive Hyg resistance were all treated similarly, with the same drug concentrations (Supplementary Table B.6), for the same time period and plated on the same conditions.

A schematic of the experiments performed is shown in Figure 3.4. We compared the strains that were enriched in the treated transformed samples to the respective treated pools transformed with the Hyg+ control. To control for differences in the growth rates of the strains under the influence of stressors we chose to use the Hyg+ control as a baseline. In other words, all the strains carrying the Hyg+ construct have the possibility to be discovered in the final set of strains after treatment and selection because they do not need to acquire deletions to become hygromycin resistant. Therefore, the absence of some strains in the final Hyg+ pool is due to either strains being too sensitive to the drug, strains having too slow growth rates or due to random chance. Consequently, we were able to use this pool to normalize the strains in the other pools transformed with the DelRep or DelNoRep constructs. For all treatments and pools, I included three replicates, and the entire experiment was performed in duplicate. The pools grown in YPAD were used to account for the genotype effect, *i.e.* the specific gene KO in these strains had an influence in the susceptibility to acquire deletions in the construct independent of being stressed by a drug.

Before identifying the enriched strains, we assessed the correlation between different replicates by comparing the number of reads that supported the presence of a particular

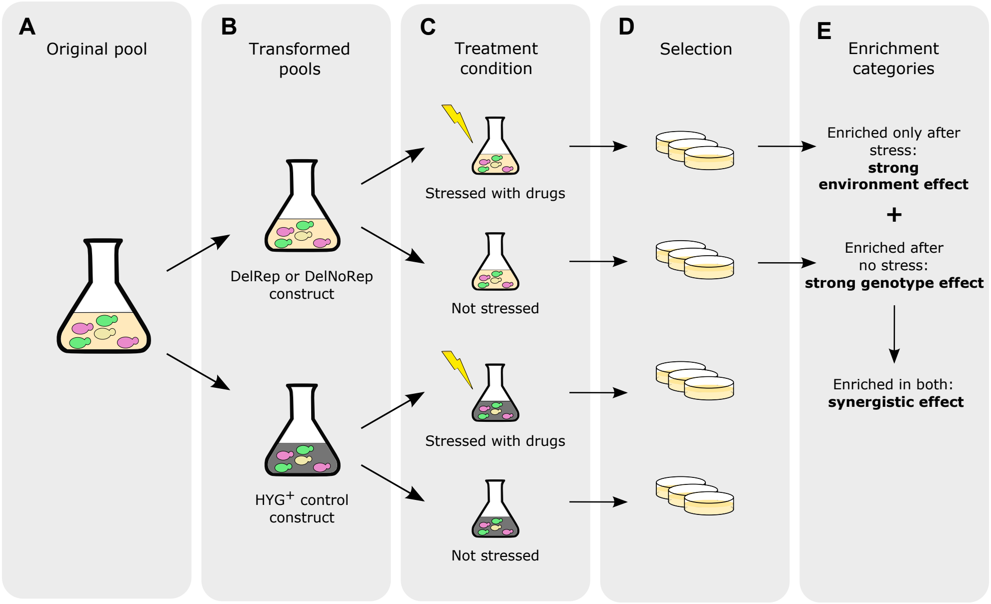
A
Original pool
B
Transformed pools
DelRep or DelNoRep construct
HYG^{+} control construct
C
Treatment condition
Stressed with drugs
Not stressed
Stressed with drugs
Not stressed
D
Selection
E
Enrichment categories
Enriched only after stress: strong environment effect
+
Enriched after no stress: strong genotype effect
Enriched in both: synergistic effect

Figure 3.4: Experimental design. **A** | A pool from the yeast deletion collection was used as a starting set of strains. **B** | From the original pool three pools were derived. One transformed with the deletion construct with repeats, one with the construct without repeats, and one transformed with a construct that confers constitutive hygromycin resistance (Hyg+). **C** | Aliquots of the transformed pools were used for all experiments. In each experiment the samples were treated with specific drugs or grown on rich media (YPAD control) without any stress. **D** | After the treatment or growth on YPAD, strains that acquired deletions were identified by selecting for hygromycin resistance. **E** | For the strains enriched only after drug treatment, environmental effects (drugs) were considered to have a stronger effect than the KO of a gene. The genotype (the specific gene KO) has a stronger effect in the strains that were already enriched after growth without drug stressor. Strains that were shared between these groups are influenced by their genotype as well as by the environment (treatment), *i.e.* a synergistic effect.

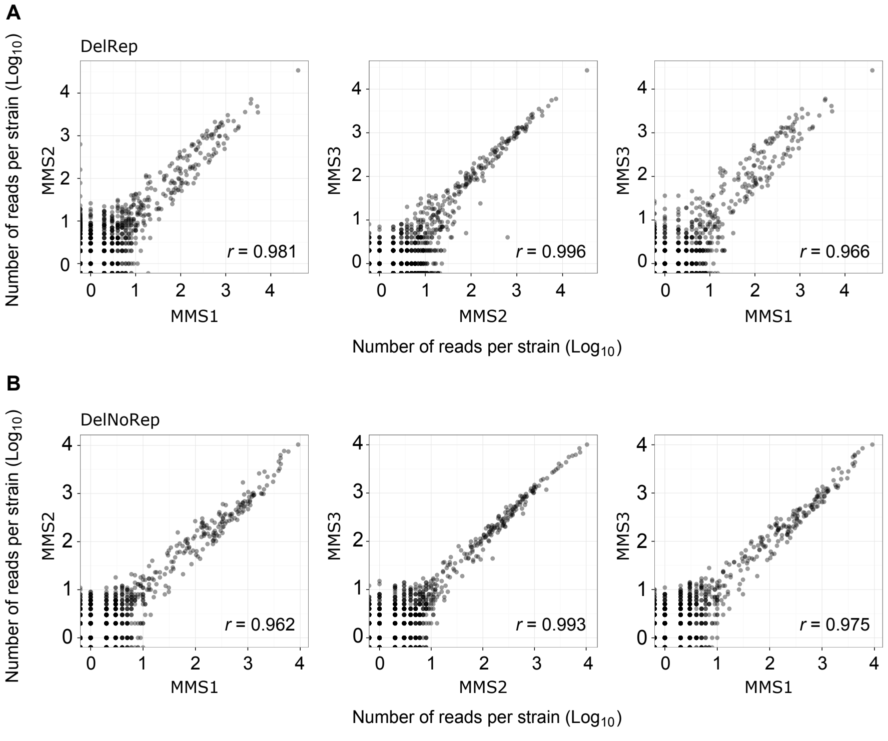

Figure 3.5: Correlation between MMS replicates for the pools transformed with DelRep and DelNoRep constructs. **A** | Correlation between the three replicates for the transformed pool with the DelRep construct and stressed with MMS. **B** | Correlation between the three replicates for the transformed pool with the DelNoRep construct and stressed with MMS. (r: Pearson correlation coefficient).

strain. We confirmed that all replicates within a treatment were highly correlated (Pearson correlation coefficients ranging from 0.737 to 0.999) and that we were able to identify similar sets of strains in each replicated experiment. Figure 3.5 shows as example the correlation between the replicates for the transformed pools stressed with MMS. Other correlations for the pools transformed with the DelRep and DelNoRep constructs are shown in the Supplementary Figures B.4 and B.5 respectively.

3.3 Deletions can occur in the absence of stressors

We first assessed the strains that were enriched in the pool without stressors compared to the control pool carrying the Hyg+ construct. Only strains enriched 2-fold when compared to the Hyg+ control and with more than 10 supporting reads, were considered. The strains enriched after growth on YPAD, represent the KO mutants that were able to acquire deletions in the construct in the absence of a drug treatment. We identified total of 250 strains that belonged to this group (Figure 3.6), indicating that some KO strains had increased rates of deletion formation even in the absence of an external stress.

When looking at the functions of the strains in this category, we could identify that

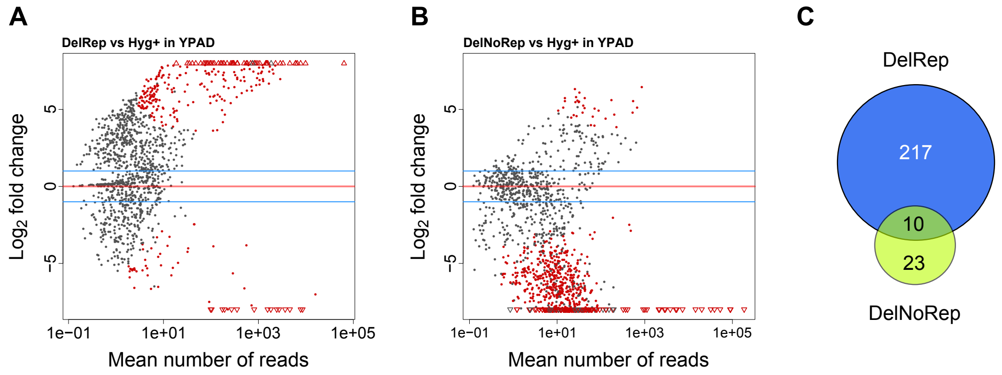

Figure 3.6: Strains enriched at least 2 fold in the pools transformed with the DelRep and DelNoRep constructs after growth without any stressor. **A** | Strains significantly enriched in the pool transformed with the DelRep construct, *i.e.* strains that acquired deletions between direct repeats. **B** | Strains significantly enriched in the pool transformed with the DelNoRep construct, *i.e.* strains that acquired deletions without the presence of direct repeats. **C** | Ten enriched strains were shared between the DelRep and DelNoRep pools.

they belonged to a very wide range of biological processes, mainly strains with defects in metabolic processes, synthesis and processing of cellular products. However, no significant enrichment for a particular functional category was found.

3.4 Deletion formation between direct repeats is more common than in the absence of repeats

The number of strains that acquired deletions in the absence of stressors was higher in the pools transformed with the DelRep construct than in the DelNoRep construct, implying that the formation of deletions between direct homologous repeats is more common than in the absence of these repeats (Figure 3.6). This result was not completely surprising given the known important role of homologous recombination in the formation of SVs. Only 10 strains were shared between the DelRep and the DelNoRep pools, indicating that there are different mechanisms behind the formation of deletions in the presence and absence of direct repeats.

In addition to the strains that acquired deletions in the absence of stressors, we were able to identify a set of KO mutants that acquired deletions after the treatment with different drugs. Similar to the pools without treatment, a higher number of enriched strains was identified in the DelRep pool as compared to the DelNoRep pool (Figure 3.8A).

There was a high overlap between the enriched strains after each drug stressor and the strains enriched under no treatment. In fact, on average 65.5% (SD ±11.5) of the strains that acquired deletions in the DelNoRep construct were shared with the pool of strains that acquired deletions without the influence of stressors. Similarly, on average 82% (SD ±5.1) of the strains that gained deletions between the direct repeats under drug treatment were also detected in the pools grown without any stress.

Interestingly, for strains that were detected after treatment as well as without treatment, a significantly higher fold change enrichment was observed compare to those strains that were only enriched after a specific drug treatment (Figure 3.7A), indicating potential synergistic effects of the environment and the gene KO. This trend was only seen for strains acquiring deletions between direct repeats. This suggests that overall there are some KO strains that are more prone to gain deletions between homologous repeats, and this can occur in the absence of stressors because of the particular deficiencies caused by the gene KO. At the same time, this effect can be amplified by the stress caused by chemical agents. Therefore, those strains are even more likely to form deletions in the presence of a drug stress than the strains where the genotype has a smaller effect.

Furthermore, for the highly enriched KO strains (taking the top 25 percentile) in both treated and untreated conditions, their fold increase under the drug stress was always higher compared to the increase in no stress conditions (using the respective Hyg+ controls as normalizers in both cases) (Figure 3.7B). Taken together, these results are additional support for the presence of synergistic effects suggesting that those strains that gain deletions in the absence of any stressor are also the ones that are more prone to form deletions when subjected to growth in stressful conditions.

I assessed the functions of the KO strains where both genotype and environmental effects play a role in the formation of deletions (See Supplementary Table B.10 for the complete set of strains). Remarkably, five KO genes, *RDH54*, *MMS2*, *IRC20*, *IOC4*, were directly related to DNA repair and recombination pathways such as synthesis-dependent strand annealing and mismatch repair. They were not uniquely detected upon a specific drug treatment, but rather enriched after stress with several drugs. Interestingly, Ioc4 associates with Isw1 to form the chromatin remodeling complex Isw1p [Vary et al., 2003; Pinskaya et al., 2009; Smolle et al., 2012], one of the two versions of the ISWI complexes. Interestingly, the *isw1* mutant was the strain with the second highest number of deletions in the mutation accumulation assay described in Chapter 2 (Figure 2.3). Additionally, two meiosis related genes were also identified among the highest enriched strains, namely *MUM2* and *GMC2*.

3.5 Several strains are enriched only after growth in the presence of stressors

A set of KO strains were only enriched after treatment with a drug. In total we were able to detect 176 strains that were enriched upon stress, after filtering out those detected in the absence of stressors. Independent of the drug treatment we observed always a higher number of strains acquiring deletions between homologous repeats (Fisher's-exact test, P=3.5×10^{-5}) (Figure 3.8A).

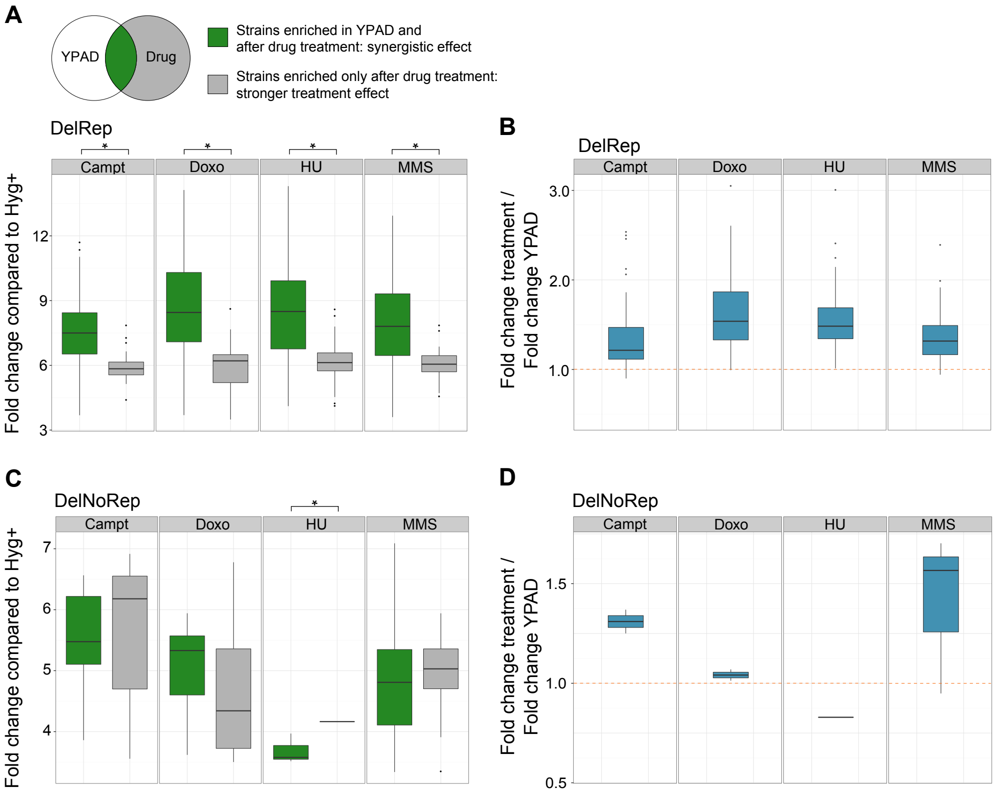

Figure 3.7: Fold enrichment (FDR<10%) is higher for strains detected both before and after treatment. **A** | Fold enrichment of strains that acquired deletions between direct repeats in the DelRep construct. **B** | Comparison of the fold changes after treatment and before treatment (YPAD), both normalized by the Hyg+ control, for the strains with the DelRep construct. **C** | Fold enrichment of strains that acquired deletions in the DelNoRep construct without direct repeats. **D** | Comparison of the fold changes after treatment and before treatment (YPAD) for the strains with DelNoRep construct. (*:P<0.01).

The overlap between the strains detected after treatment with different drugs was on average 45.8% (SD ±30.1) for the strains carrying the DelRep construct and 38.8% (SD ±51.0) for the ones with the DelNoRep construct (Figure 3.8B,C). As seen from the variation in these overlaps, there were cases with no strains shared between the pools treated with different drugs, whereas other pools shared all the same strains. For example, of the 16 strains highly enriched in the DelRep pools after treatment with doxorubicin, all were also enriched after the treatment with camptothecin. A similar trend was seen for the pools transformed with the DelNoRep construct, in which 3 out of 4 strains enriched after treatment with doxorubicin were also enriched after camptothecin treatment.

3.5.1 Meiosis related genes are enriched among the KO strains with deletions between direct repeats upon drug stress

The strains significantly enriched only after drug treatment are shown in the Supplementary Tables B.8 and B.9 (only top 10 strains shown). The majority of the strains that acquired deletions without direct repeats belonged to diverse biological processes, including cellular carbohydrate metabolic processes and protein folding. No significant Gene Ontology (GO) term was enriched among these strains.

However, among the strains that acquired deletions between homologous repeats, I could identify several KO strains that belonged to DNA repair pathways. The top ten strains that acquired deletions between the homologous repeats (DelRep construct) are listed in the Supplementary Table B.9. Interestingly, also several of the top enriched strains belonged to processes related to meiotic recombination. From the top ten strains in every drug stressor, an overall total of 5 deletion mutants are meiosis related, with camptothecin treatment being the one with the highest number of meiosis genes detected. Some of these KO strains were also shared among the treatments, *e.g. zip2* was detected after camptothecin and after HU treatments, and *spo73* was detected after camptothecin and doxorubicin treatments. Furthermore, one meiosis related gene was also observed among the top strains with the DelNoRep construct (Supplementary Table B.8).

By performing a GO term enrichment analysis on strains that acquired deletion between the homologous repeats, I observed a significant enrichment for genes related to the response to stress, DNA metabolic processes, DNA repair and strains related to meiosis pathways (Figure 3.9). Other terms enriched include processes involved in protein post-translational modification, such as glycosylation and acylation.

The enrichment of meiosis related genes was not expected and is of particular interest because in this assay strains were undergoing only vegetative growth. Among the genes enriched, *MSH4* and *ZIP2*, both involved in meiosis, are known to co-localize, supporting the functional significance of their enrichment under the same growth condition.

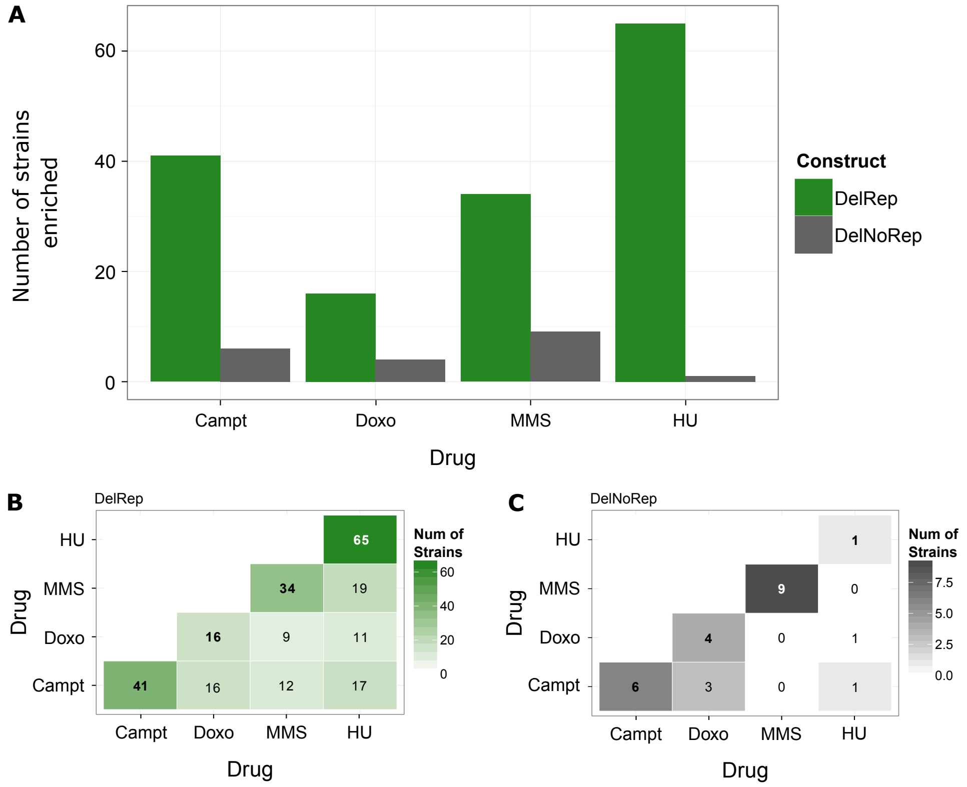

Figure 3.8: Number of enriched strains per treatment for the pools transformed with DelRep and DelNoRep constructs. **A** | Total number of enriched strains per treatment. Only strains enriched 2-fold when compared to the Hyg+ control, with more than 10 supporting reads and not present in the YPAD control, are shown. **B** | Total number of shared strains enriched in the treated pools transformed with the DelRep construct. The diagonal shows the total number of strains enriched per treatment. **C** | Total number of shared strains enriched in the treated pools transformed with the DelNoRep construct. The diagonal shows the total number of strains enriched per treatment.

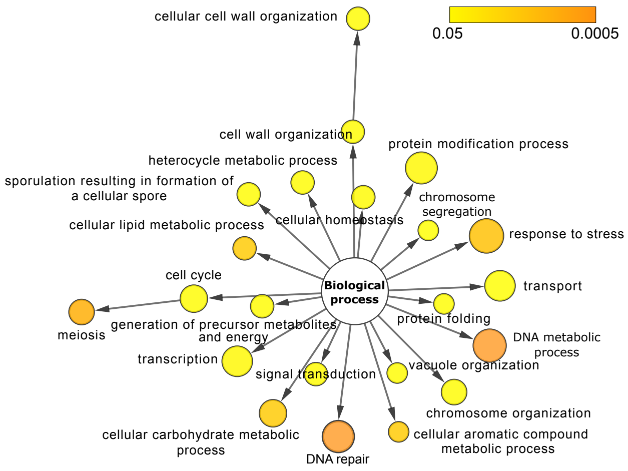
cellular cell wall organization
0.05
0.0005
cell wall organization
heterocycle metabolic process
protein modification process
sporulation resulting in formation of a cellular spore
chromosome segregation
cellular homeostasis
response to stress
cellular lipid metabolic process
cell cycle
Biological process
transport
meiosis
generation of precursor metabolites and energy
protein folding
DNA metabolic process
transcription
signal transduction
vacuole organization
chromosome organization
cellular carbohydrate metabolic process
DNA repair
cellular aromatic compound metabolic process

Figure 3.9: Main GO terms enriched in the strains that acquired deletion between homologous repeats. The color represents the significance and the size the number of strains in the set that belong to each term. (The GOSlim Yeast set was used for annotation with Cytoscape [Shannon et al., 2003]).

They have been previously reported to form discrete foci in the meiotic chromosomes [Novak et al., 2001].

Based on these results, and given that the meiosis related genes have not previously been related to the suppression of deletion formation, we selected the deletion mutants highlighted in Table B.9 to perform experimental validations on their increased rates of deletion formation (Section 3.7).

3.6 Enriched strains have higher growth rates than a set of KO strains known to be involved in DNA repair

It was unexpected that among the strains enriched in our experiments, only few belonged to the known major pathways of DNA repair and genome maintenance. Therefore, I assessed whether genome maintenance genes have comparably lower growth rates and higher vulnerability to the drug stressors than the strains that were enriched in our screen. For this analysis, I used previously published growth rate values in rich medium for the strains enriched with and without stressors [Steinmetz et al., 2002]. As a comparison, I used 44 (excluding the wild type and the control strains) mutant strains from the mutation accumulation assay described in Chapter 2 (Table B.1). These mutant strains were selected specifically due to their involvement in pathways required for genome stability.

Indeed, the genes identified in this screen had overall higher growth rates than those involved in genome maintenance pathways (Wilcoxon rank-sum test, P<0.0001). Similarly, the strains enriched without stress, had also higher growth rates than genome maintenance related strains (Wilcoxon rank-sum test, P<0.0001). There were no dif-

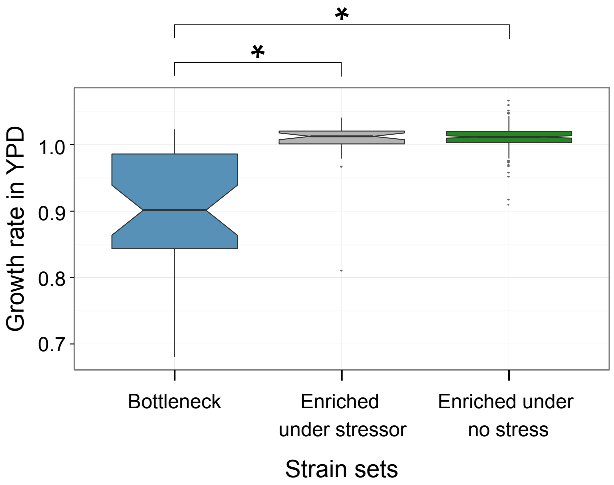

Figure 3.10: Growth rates in rich medium for the strains in different groups. “Bottleneck strains” correspond to the mutation accumulation assay, including the strains shown in Table B.1 with known defects in genome maintenance pathway. “Enriched under stressor” correspond the top ten strains that acquired deletions between direct repeats (DelRep construct) and that were enriched after drug treatment. “Enriched under no stressor” are the strains carrying the DelRep construct that were detected after growth without drug treatment. (*:P<0.0001).

ferences in the growth rates between the treated and untreated strains (Wilcoxon rank-sum test, P=0.674). This suggests that in our enrichment assay we were mainly able to recover strains that had no strong growth deficiencies compared to strains with known defects in genome maintenance (Figure 3.10).

3.7 Experimental validation of candidate genes

3.7.1 The KO strains were tested individually for increased rates of deletion formation

Due to reported concerns about the original strains in the Yeast Deletion Collection, including the existence of aneuploidies in several strains or additional mutations other than the specific KO genes [Hughes et al., 2000; Lehner et al., 2007; Ben-Shitrit et al., 2012], we decided to newly generate the individual KO strains for the candidate genes. Starting from the wild type strains of both mating types (BY4341 and BY4342) I created haploid and diploid KO strains using the PCR deletion strategy used to create the original deletion mutants (Detailed methods are described in Section A.2.11). Given the less direct connection between meiosis genes and the suppression of deletion formation, we chose to perform experimental validations initially for the KO mutants *msh4*, *zip2*, *eno1* and *apn2*. As mentioned before, these strains were identified as having increased rates of deletion formation between direct homologous repeats and after the treatment with different drugs.

In addition to generating KO strains for the candidate genes, I also created the KO strain *trp5* as a negative control, and *rad52* as a positive control. The former is a gene coding for the tryptophan synthetase, which is not related to DNA repair or recombination pathways, and the latter is known to be important for DNA repair by homologous recombination [Lisby et al., 2001]. Each newly generated KO strain, in the haploid and diploid stages, was additionally transformed with the DelRep construct and subjected to the same experimental workflow as described in Figure 3.4. A wild type haploid (*MAT*a) and a diploid strains, untransformed, were also included in the experiments.

The KO strains have higher deletion formation than the control strains

The formation of deletions in the construct was again monitored by selecting for, and quantifying, hygromycin resistant colonies. The newly generated KO strains exhibited significantly higher levels of deletion formation when compared to the wild type strain and to the negative control for both the diploid and the haploid strains (Figure 3.11A).

The *msh4* deletion mutant showed the highest increase in the number of resistant colonies, in both the haploid and diploid stages, when compared to the other KO strains, even higher than the *rad52* positive control (Figure 3.11B). On the other hand, even though *apn2* and *zip2* showed an increase in deletions compared to the *trp5* negative control and to the wild type, it was not significantly higher than for the *rad52* positive control (Figure 3.11B and Supplementary Table B.11).

Interestingly, no significant differences were found between rates of deletion formation among the different drug treatments, except for the *msh4* deletion mutant (Figure 3.11B). Furthermore, most KO strains also acquired deletions when grown in a rich medium (YPAD) in the absence of any drug stress. In fact, for some knockout strains we could see similar fold increases in the number of strains that acquired deletions in the presence and absence of a drug (Figure 3.11B). These results suggest that the KO strains themselves had higher rates of deletion formation independent of the treatment used, indicating a stronger genotype effect compared to the influence of the chemical agent.

KO diploids have higher rates of deletion formation than their corresponding haploids.

Furthermore, the diploid KO strains showed higher deletion levels than their haploid stages (Figure 3.11A). In both growth conditions, with and without drug stress, the difference between diploid and haploid strains was significant. No difference between diploids and haploids was observed for the wild type and negative control strains (Figure 3.11A).

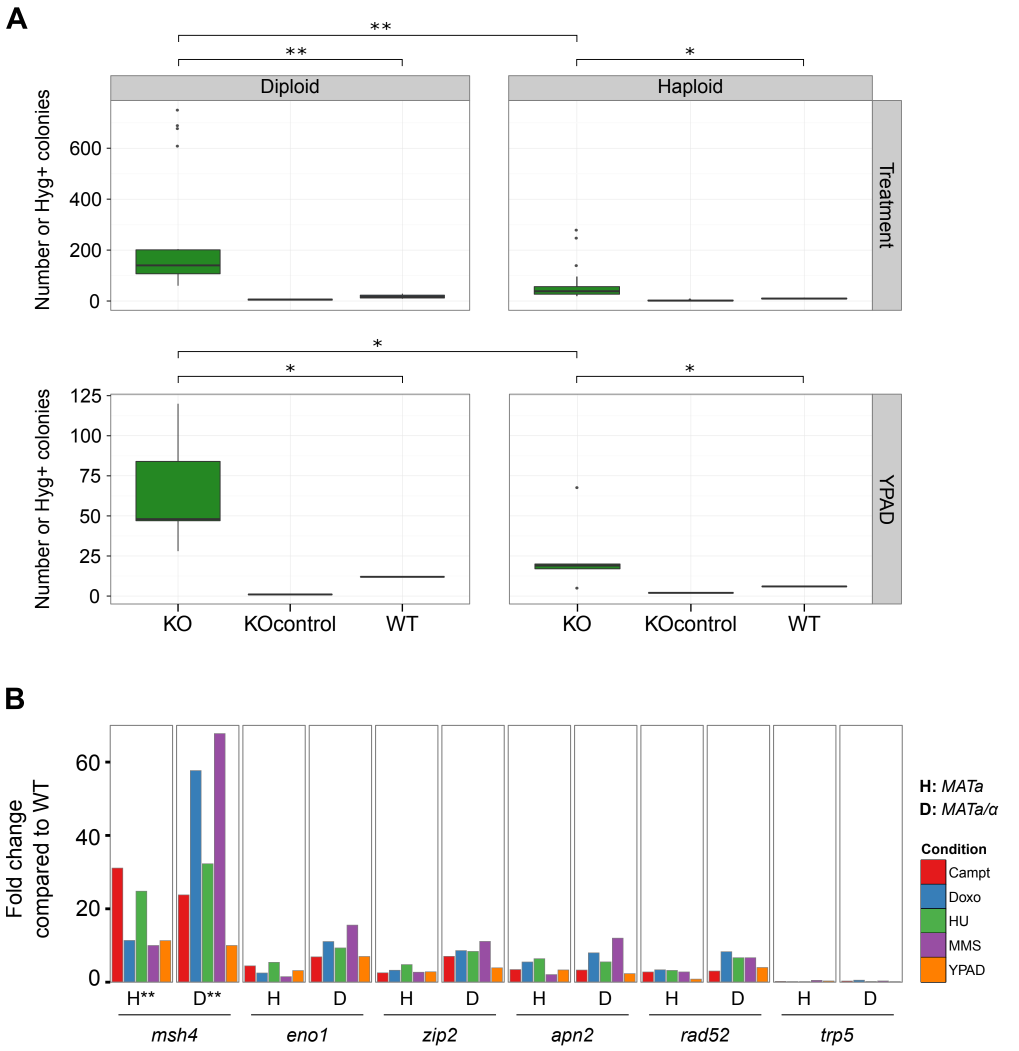

Figure 3.11: Number of KO colonies that acquired deletions between direct repeats in the DelRep construct after growth under different drug treatments or under no stress (YPAD). **A** | Comparison between the number of strains with deletions between direct repeats in the DelRep construct (hygromycin resistant) in diploids and haploids under treatment and no treatment (YPAD). **B** | Fold increase in the number of colonies that gained deletions in the KO strains compared to the WT control. (*:P<0.01; **:P<0.001).

The difference between diploids and haploids was less pronounced in the *msh4* mutant, where the haploid strains had also a significant increase in the formation of deletions (Supplementary Table B.11).

3.8 Discussion

I presented here a genome-wide screen for strains that have an increased tendency to acquire deletions. With the use of specifically designed constructs inserted into a non-essential region in chromosome V, we were able to identify strains that are more prone to acquire deletions in the presence of direct homologous repeats and in the absence of them. Several of these strains represent mutants that have not previously been linked with the suppression of deletion formation.

Homologous repeats are often involved in the formation of deletions. DNA double-strand breaks can occur due to the natural metabolism of the cells or as a consequence of external factors, such as chemical agents or irradiation. When DSBs occur, their repair is achieved by different mechanisms, some of which involve homologous recombination (HR) [Pâques and Haber, 1999]. However, defects in DSB repair pathways can lead to the formation of genomic rearrangements. Homologous recombination between repeated sequences has long been recognized as a mechanism of deletion formation [Eichler, 1998]. For example, in humans, *Alu*-rich regions show increased genomic instability [Calabretta et al., 1982], and several diseases are caused by deletions mediated by flanking repetitive elements [Ledbetter et al., 1981; Yen et al., 1990].

The majority of the identified candidate strains acquired deletions surrounded by direct repeats. Some of the strains had increased deletion formation even under no stress. In yeast, naturally occurring repeats can be found at recombination hotspots [St. Charles and Petes, 2013; Song et al., 2014]. In addition, solo-LTRs (the long terminal repeats at the ends of LTR-retrotransposons) in the yeast genome are associated with regions of slow replication fork progression, and these sites are more susceptible to recombinogenic lesions that can result in the formation of rearrangements [Song et al., 2014]. These results support our observation that without the influence of drug treatment, a higher number of deletions were formed between the direct repeats.

Solo-LTRs represent sequences that are reminiscent of older retrotransposition events [Carr et al., 2012; Neuvéglise et al., 2002]. Even though their sequences are not 100% identical, there is high homology between them. In budding yeast it has been shown that recombination between non identical sequences, homeologous recombination, can occur even between sequences that share only 70% identity [Mézard et al., 1992]. The recombination between homeologous sequences depends on the existence of shorter stretches of high identity within them that allow the pairing of the DNA sequences

[Mézard et al., 1992]. The homologous sequences that are found in the designed construct are 100% identical, but given that they derive from a human sequence, they do not show high stretches of identity with other loci in the yeast genome.

The sequence configuration of the DelRep construct used in this screen, with direct repeats separated by 1kb, is not completely uncommon in the yeast genome. For example, there are 383 long terminal repeats (LTR) annotated in the yeast genome (*Saccharomyces* Genome Database). From these LTRs, more than 15% have sequences around 300bp long and are separated from each other between 1kb and 500bp, mimicking the genomic conformation that the constructs have. Therefore, the results obtained with our constructs are a good model for processes that can occur in yeast.

Furthermore, our results suggest that deficiencies caused by the knockout of a gene may be enough to give rise to such deletions. Spontaneous HR events can be initiated by DSBs, but additionally by other processes inherent to replication, such as single strand breaks (SSBs), fork stalling and collapse [Lettier et al., 2006; St. Charles and Petes, 2013]. In fact, different DNA lesions can trigger HR, such as pyrimidine dimers produced by UV irradiation [Lettier et al., 2006]. If left unrepaired, for example in the case of defects in nucleotide excision repair pathways, these lesions can lead to replication fork stalling exposing regions of ssDNA and resulting in genomic rearrangements. These findings support the presence of genes involved in recombination or base excision repair (*e.g. RAD34* and *RDH54*) in our top candidate lists. The deletion mutant *shu2* has also been identified in a genome-wide screen for genes that suppress gross chromosomal rearrangements Smith et al. [2004].

Mutation of chromatin remodeling proteins may increase the rates of deletion formations. Among the strains that acquired deletion without treatment, a particularly interesting one is *ioc4*. Remarkably, Ioc4 belongs to the chromatin remodeling complex Isw1p, which is additionally composed of Ioc2 and Isw1 [Vary et al., 2003; Maltby et al., 2012]. Consistently, in the mutation accumulation assay described in Chapter 2, the mutant strain *isw1* showed the highest number of deletions after *msh2*. These results, derived from independent experiments, support the role of this complex in the suppression of deletion formation.

Isw1p belongs to the ISWI family of remodeling enzymes that is conserved from yeast to humans [Clapier and Cairns, 2009; Smolle et al., 2012]. Growing evidence indicates that chromatin remodeling complexes are implicated in the DNA damage response [Klochendler-Yeivin et al., 2006; Wilson and Roberts, 2011] particularly for oncogenic mutations in the SWI/SNF family [Wilson and Roberts, 2011; Shain and Pollack, 2013; Hohmann and Vakoc, 2014]. Additionally, the ISWI family is also emerging as a player in the DNA damage response. The human homolog of Isw1, SNF2H, has been recently shown to accumulate at sites of DNA breaks and prevents genomic instability by promoting correct HR repair [Toiber et al., 2013; Vidi et al., 2014]. The mechanisms by which mutations in these chromatin remodeling families lead to genomic instability are not clear yet. However, one possibility is the recent finding that SNF2H can reduce the barriers imposed by heterochromatin during the repair of heterochromatic DSBs. Its interaction with ACF1 induces heterochromatic relaxation by respacing nucleosomes [Klement et al., 2014].

Treatment with different drug stressors can result in similar outcomes. The induction of replication stress or direct DNA DSBs by environmental pressures, such as treatment with chemical compounds, is also known to increase the recombination between homologous repeats [Bishop and Schiestl, 2000; Iraqui et al., 2012]. However, it was unexpected to observe overlaps between strains enriched after the different drug treatments used in our screen. Each of the stressors used has potentially a different effect on the cells, via diverse mechanisms. Therefore, the mutant strains that were prone to form deletions under each growth condition were expected to differ as well. Nonetheless, all these stressors can impede the progression of the replication fork.

For instance, MMS is a DNA alkylating agent that methylates N3-deoxyadenine, a lesion that can inhibit DNA synthesis and reduces the rate of fork progression [Chang et al., 2002]. On the other hand, camptothecin acts differently, by inhibiting DNA topoisomerase I [Liu et al., 2006]. However, interestingly, camptothecin has also been shown to cause stalling and collapsing of replication forks in yeast [Regairaz et al., 2011] and mammalian cells [Saleh-Gohari et al., 2005]. Furthermore, hydroxyurea can similarly lead to the stalling of replication forks by inhibiting the ribonucleotide reductase. This leads to replication stress by limiting the pools of nucleotides available and this way preventing the progression of the replication fork [Saintigny and Lopez, 2002; Petermann et al., 2010; Arlt et al., 2011]. Consequently, and as explained above, the defects in replication caused by different drug treatments could be responsible for an increase formation of deletions in similar mutant strains, especially if these strains have already deficiencies in the suppression of mutations due to their genotypes.

Defects in meiosis related genes may also lead to the formation of deletions during vegetative growth. I identified several yeast strains with deficiencies in meiosis related pathways that showed increased deletion formation. Yeast cells can enter meiosis only when the environmental conditions meet very specific criteria. The lack of at least one essential growth nutrient, such as nitrogen, and the absence of a fermentable carbon source, such as glucose, are required to arrest the cells in G1. Even very low concentrations of glucose can inhibit the initiation of sporulation [Honigberg and Purnapatre, 2003]. Furthermore, a non-fermentable carbon source should become available, to be metabolized by respiration, which in turn act as a signal for sporulation [Hirschberg and Simchen, 1977; Jambhekar and Amon, 2009; Piekarska et al., 2010].

All media and cultures used in this study contained all required nutrients and enough glucose to ensure their vegetative growth.

Non allelic homologous recombination was itself initially regarded as a meiosis specific process [Carr and Lambert, 2013]. Nevertheless, HR can also play a role in fork protection and fork restart after replication is blocked, making HR a pathway ensuring DNA replication at the expense of genomic rearrangements [Lambert et al., 2010; Lambert and Carr, 2013]. Similarly, proteins that have been characterized for their functions during meiosis may act in other processes related to the DNA damage response in vegetative growth as well. In agreement with this, Tkach et al. [2012] identified several meiotic processes involved in the processing of DNA DSBs, significantly enriched in the class of proteins that re-localizes to nuclear foci upon DNA replication stress. This suggests that some meiosis proteins may be involved in DNA repair pathways outside of meiosis. Furthermore, it is interesting to point out that the human homolog of *MSH4* (whose mutant strain gained deletion between direct repeats only after treatment with several stressors), *hMSH4*, has recently been implicated in other functions in mitotic cells, including double strand break repair [Her et al., 2003; Chu et al., 2013]. This is in agreement with our validation experiments, where *msh4* was the KO strain with the highest formation of deletions.

Diploid strains tend to acquire more deletions than their corresponding haploids. An additional finding derived from the validation experiments was that the KO strains tend to have more deletions in their diploid state than in the haploid. In mammals, hyperploid cells have higher frequencies of chromosome missegregation and chromosomal rearrangements [Fujiwara et al., 2005]. Furthermore, in yeast diploids have been shown to be more resistant to UV irradiation [Snow, 1967], EMS (an alkylating agent) [Mable and Otto, 2001] and fluconazole (an antifungal agent) [Anderson et al., 2004] than the corresponding haploids, meaning that they can tolerate a larger number of mutations. Perhaps this is a consequence of diploids still having a functional copy of the damaged gene, providing them with an advantage to survive in spite of genomic rearrangements.

In fact, some yeast KO mutants are lethal in a haploid state and can only exist in a heterozygous state. Interestingly, Lada et al. [2013] induced mutations in yeast by the base analog 6-hydroxylaminopurine (HAP) and by an ectopic DNA editing cytosine deaminase, resulting in a significant increase in the number of mutations accumulated in the diploid strains. In contrast, haploids acquired an order of magnitude less mutations. Most of the mutations found in the diploids were recessive and did not have fitness effects when heterozygous. [Lada et al., 2013]. Taken together, these results support the fact that diploids may be able to cope better with mutations and were therefore found to acquire more deletions in our screen.

Although this experiment monitors the occurrence of mutations at one specific locus, this may reflect processes influencing other parts of the genome as well. By using pools of deletion mutants, we were able to perform a genome-wide screen for genes that, when mutated, increase deletion formation. Similar experiments could be also implemented to study the formation of other structural variants, since the mechanisms that give rise to different types of mutation are also distinct. Further studies will be necessary to fully understand the functional impact of each candidate gene at a genome-wide level and the precise mechanisms by which they preserve genomic stability.

Chapter 4

Detection and validation of structural variants in cancer genomes

4.1 Motivation

As part of my PhD, in addition to the experiments and analyses of genomic variants in *Saccharomyces cerevisiae* presented in the previous chapters, I was involved in multiple side projects. By performing the data analyses of the yeast related projects I got acquainted with the tools and pipelines for the analysis of next-generation sequencing DNA data, especially those involving the calling, filtering and annotation of structural variants. For this reason, and to further extend my knowledge in this field, I participated in additional projects on human data, specifically on cancer samples, some resulting in co-authorships as described below. This chapter collects the analyses in these projects and it therefore contains information on different topics, not necessarily highly related to each other, but with several technical procedures in common.

4.1.1 Contribution

My main contributions in these projects were mostly on technical aspects of the data analyses. I performed all analyses and results presented in each section of this chapter.

4.2 Validation of structural variants identified by the reference free Somatic MUtation FINder (SMUFIN)

My contribution to this project is part of the following published paper:

Moncunill V., Gonzalez S., Beà S., Andrieux L. O., Salaverria I., Royo C., Martinez L., Puiggròs M., Segura-Wang M., Stütz A. M. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature Biotechnology*, 32(11), October 2014.

As it was mentioned in the introduction, the comprehensive identification and characterization of structural variants still faces several difficulties. One of these limitations is the need to compare the sequence of a sample against a reference genome. However, the reference genome is still incomplete, in particular in highly repetitive regions. Additionally, it is known that sequencing reads derived from rearranged loci are often not properly aligned to the reference genome and are discarded, leading to loss of information about the variants. Furthermore, since the tumor sample is aligned to the reference genome independently from the control sample, further filtering steps are needed to remove germline variants and to keep the somatic ones. Additional advantages of reference-free based SV detection methods are that the variants are identified at nucleotide resolution and that non-reference insertions can also be detected.

As part of a collaborative project, I performed validations for SVs detected using SMUFIN, a reference-free variant detection tool that identifies rearrangements by direct comparison of tumor and normal sequencing reads [Moncunill et al., 2014]. An initial set of 60 SVs detected with SMUFIN in a medulloblastoma sample affected by chromothripsis [Rausch et al., 2012a] was randomly selected for PCR and Sanger sequencing validations. The results of this first assessment were key to identifying some issues with the detection algorithm, in particular affecting the identification of interchromosomal events. Therefore, modifications in the tool were implemented that improved its specificity.

I then performed further validations on new sets of predicted breakpoints. Several variants identified by SMUFIN were previously called by other reference based methods but not at nucleotide resolution. Of these, a set of 39 SVs were tested, of which 36 (92%) were positively confirmed. Additionally, another set of 27 variants not previously identified in the same sample were also assessed by PCR and Sanger Sequencing. Of these, 25 (92.5%) were validated. With these results we were able to confirm the detection capabilities of SMUFIN and its ability to identify variants at nucleotide resolution, even in highly rearranged tumor samples.

4.3 Identification and characterization of structural variants in a large cohort of medulloblastoma patients

Medulloblastoma constitutes the most common malignant brain tumor in children [Dolecek et al., 2012] and is the major cause of cancer-related mortality in childhood. Based on transcriptional profiling with microarrays and genetic data, different subtypes of medulloblastoma have been defined. The current consensus is the classification into 4 distinct subgroups [Kool et al., 2008; Northcott et al., 2011; Taylor et al., 2012; Northcott et al., 2012a]. Apart from their different clinical characteristics, they also exhibit specific cytogenetic traits, mutations and gene expression profiles [Northcott et al., 2011, 2012a].

Several next-generation sequencing studies of medulloblastomas have been carried out in the last years, revealing a large number of previously unknown mutations, including structural variants, with several of these alterations being subgroup-specific [Jones et al., 2012; Northcott et al., 2012b, 2014]. More complex rearrangements, like chromothripsis have also been observed among these tumors [Rausch et al., 2012a]. However, for medulloblastomas, there is still a lack of integration of the different types of data available. For example, there is not a complete understanding of the contribution of point mutations and structural variants, especially in terms of defining driver and passenger variants. Additionally, the roles of the epigenome and transcriptome in the development of the tumors are yet to be further characterized.

With the goal of further understanding and integrating different types of data, currently we are analyzing the largest cohort of medulloblastoma samples for which whole-genome, exome, expression and methylation data is available. My role in this project is the detection and characterization of the SVs in paired-end WGS, which currently comprises a total of 454 tumor-control sample pairs.

I am analyzing all samples with a uniform computational pipeline for the discovery, annotation and filtering of variants. This workflow is being developed in the frame of the Pan-Cancer Analysis of Whole Genomes initiative by Joachim Weischenfeldt of the Korbel Lab. It involves the calling of SVs with DELLY [Rausch et al., 2012b], the filtering and functional annotation of somatic variants and the removal of low confidence calls by comparing to the 1000GP and other cancer sample sets.

I have analyzed 360 tumor-control pairs, and continue to call SVs in the remaining samples. Currently we are able to identify the main known SVs that have been described in medulloblastoma, in addition to other variants that are not recurrent or occur only in less than 5 samples. Given that the data set is large, once I finish the SV detection for all samples, I will assess the existence of new recurrent variants, especially for the subgroups 3 and 4, for which only few clear oncogenic drivers have been identified. Additional analyses will include the identification of breakpoint clusters and potential fusion genes. This data will be then integrated with the other types of genomic information available for these samples, in collaboration with researchers at the German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ), with the aim of generating a comprehensive mutational landscape for medulloblastomas. Studying the underlying genetic basis of these tumors will eventually help improve patient stratification, treatment and prognosis [Northcott et al., 2012a].

Additionally, some of these medulloblastoma samples exhibit chromothripsis, massive genomic rearrangements that result from a single catastrophic event [Stephens et al., 2011; Rausch et al., 2012a]. The structural variants that I identified in these samples have been used in a study that describes a reproducible experimental approach for the generation of complex rearrangements and that links these shattering events to hyperploidy.

My contribution to this project is included in the following manuscript:

Mardin, B., Drainas, A., Waszak, S., Weischenfeldt, J., Isokane, M., Stütz, A., Buccitelli, C., Segura-Wang, M., Northcott, P., Pfister, S. et al. A novel cell-based model system links chromothripsis with hyperploidy. (in review in *Nature Methods*).

4.4 Identification of structural variants at specific loci in small cell lung cancer

My results of this project have been included in the following manuscript:

George, J., Peifer, M., Cun, Y., Leenders, F., Müller, C., Dahmen, I., Schaub, P., Bosco, G., Pinther, B., Lu, X., Seidel, D., Fernandez-Cuesta, L., Sage, J., Lim, J., Jahchan, N., Park, K., Yang, D., Vaka, D., Torres, A., Karnezis, A., Korbel, J., Segura-Wang, M., Menon, R. et al. Comprehensive genomic characterization of small cell lung cancer. (in review in *Nature*).

I analyzed mate pair WGS data for 7 samples of small cell lung cancer (SCLC) from a study that included a total of 110 samples. The main variants observed in this large set of samples were universal inactivating mutations in both *TP53* and *RB1* detected in all samples except for 2 that exhibited chromothripsis. Additionally, 20% of the samples had mutations in genes from the *NOTCH* family.

However, for these 7 samples there was no clear structural variant detected in *NOTCH* genes or *RB1*. Therefore, mate pair sequencing was used to assess the presence of larger rearrangements, particularly at these loci. Although with the data available I was not able to confirm the presence of variants in these specific regions, I identified 2 samples with rearrangements in *CREBBP*, one with a deletion and another one with a translocation. This histone acetyl transferase is involved in the coactivation of several transcription factors, and has been also found mutated in other cancers [Mullighan et al., 2011]. Based on further analyses performed by our collaborators, it was possible to find that mutations in *CREBBP* and *NOTCH* family genes were mainly mutually exclusive in SCLC.

4.5 Identification of structural variants in embryonal tumors with abundant neuropil and true rosettes

Embryonal tumors with abundant neuropil and true rosettes (ETANTRs), named for their histological characteristics, are rare neoplasms of the central nervous system

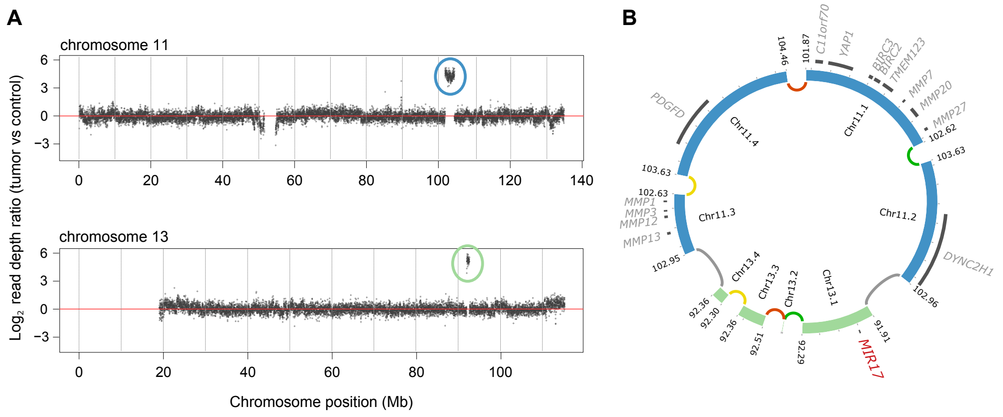

Figure 4.1: Focal amplifications in chromosomes 11 and 13 from an ETANTR sample form a double minute chromosome. **A** | Read depth profiles for chromosome 11 and chromosome 13 in one ETANTR sample. High level amplifications are marked with circles. **B** | The amplified regions in chromosome 11 and 13 are joined together forming a double minute chromosome. The detected SVs are shown by colored arches: grey: translocations, red: tandem duplication, yellow: inversion 5'-5', green: inversions 3'-3'. (Circular plot created using Circos [Krzywinski et al., 2009]).

(CNS) [Korshunov et al., 2014]. ETANTRs typically carry known amplifications in chromosome 19q13.42. Additionally, they show high expression of *LIN28*, which leads to the repression of the let-7 miRNA family and consequently to an upregulation of the targets of let-7, including *MYC*, *MYCN*, *RAS* and *CDK6* [Korshunov et al., 2012].

It has been recently discovered that the amplification in chromosome 19 in ETMR (a broader classification of embryonal tumors with multilayered rosettes that likely includes ETANTRs) causes the fusion of the *TTYH1* gene with the C19MC miRNA cluster, leading to the overexpression of these miRNAs. As a consequence, the gene *DNMT3B* is also overexpressed [Kleinman et al., 2014]. This gene is normally active only during early stages of the neural tube development. Its overexpression at later ages makes it an oncogene involved in ETMR, and probably ETANTR, tumorigenesis [Kleinman et al., 2014].

As part of a collaboration with researchers at the DKFZ, I analyzed mate-pair whole-genome sequencing data of 13 ETANTR tumor-normal pairs with the aim of confirming the high frequency of the chromosome 19 amplification and to identify other potential rearrangements in these tumors.

Out of the 13 samples, 8 carried the know amplification with support for the fusion of *TTYH1* and C19MC. However, many more rearrangements in this region of chromosome 19 were detected in addition to the amplification. Several samples showed deletions and inversions in the same locus.

Interestingly, from the other 5 samples where no amplification in 19q13.42 was detected, two samples had rearrangements involving additional miRNA clusters. One sample showed a high level amplifications in chr11 and chr13 (Figure 4.1A). Based on paired-end mapping I was able to infer that the amplified regions form a double minute chromosome containing the miR-17/92 cluster (Figure 4.1B). This cluster comprises at least 6 miRNAs involved in cell cycle and proliferation and has been shown to be amplified in other tumors [Mogilyansky and Rigoutsos, 2013].

An additional sample without the typical amplification contained a translocation between chromosomes 22 and 19, creating a fusion between the known C19MC miRNA cluster and *MIRLET7BHG*. The latter has been also shown to be associated with the development of other cancers [Peter, 2009; Saleh et al., 2011]. Interestingly, this miRNA belongs to the let-7 family, which acts as tumor suppressors [Viswanathan et al., 2008].

Although we have not yet found the direct link between these additional rearrangements, it is interesting that in both cases they involve miRNAs that have been linked to cell proliferation and tumor progression. At the moment, further research on expression and methylation data for these samples is being conducted by investigators at the DKFZ to further characterize and understand if these or other variants have some functional impact on the development of ETANTRs.

Chapter 5

Conclusions and future perspectives

In this thesis I described two complementary approaches for studying genomic variants in *Saccharomyces cerevisiae* and for identifying genes that prevent the formation of chromosome rearrangements. Using knockout strains I could identify specific genes related to the maintenance of genomic stability that when deleted, increase the occurrence of different types of mutations. Together, these approaches allowed us to have a broader overview of genomic features that lead to the formation of mutations, such as the presence of homologous repeats. The results from this work suggest a connection between chromatin remodeling the maintenance of genomic stability. Additionally, we found support for the involvement of meiosis related genes in the suppression of deletion formation.

In Chapter 2 I presented the results of a neutral mutation accumulation (MA) assay in a set of 47 selected yeast knockout strains with defects in known genome maintenance pathways. These strains were propagated through single-cell-to-colony bottlenecks and their whole genome was sequenced at different time points. I generated a comprehensive catalog of mutations acquired in these strains and analyzed their frequency, distribution along the genome as well as the genomic features that were related to the formation of mutations.

Chapter 3 described a genome-wide screen for genes involved in the suppression of deletion formation. I used a pooled collection of more than 5000 different KO strains to assess their tendency to acquire deletions with and without different chemical stressors. I designed specific constructs tailored to monitor the occurrence of deletions, in the presence and absence of homologous repeats, and integrated them in the genome of the KO strains. Additionally, through experimental validations, I was able to confirm the frequent generation of deletions in candidate strains, including two with KO of meiosis related genes, even in the absence of additional external stressors and especially in their diploid stages.

In the following sections I summarize the most important conclusions derived from these approaches. I further discuss limitations and additional experiments that can be carried out to pursue a better understanding of the processes that maintain genomic stability.

Under neutral mutation accumulation, yeast KO strains acquired more short deletions and insertions (1-5bp long) compared to large rearrangements. Some strains gained more deletions while others tend to have more short insertions, supporting that there are different processes involved in the formation of mutations. The *msh2* deletion strain became hypermutable and acquired the highest number of mutations compared to the other strains studied. *MSH2* is known to be involved in mismatch repair and mutations in this gene increase the predisposition to cancer significantly [Dowty et al., 2013]. It also plays a role in DNA recombination between homologous sequences, making cells with mutations in this gene more prone to acquire SNVs and further rearrangements.

Given that the fraction of the genome that is changed by an SV is larger than that accounted for by an SNV, SVs may have stronger phenotypic consequences [Conrad et al., 2010; Mills et al., 2011]. Even though no selection process was implemented in the MA assay, mutations with effects on growth or survival, likely including large SVs, were not highly represented in the sequenced strains.

Short deletions and insertions occur more frequently in repetitive regions and between homologous repeats. By performing functional annotation of the variants detected in the MA strains, I could confirm their frequent occurrence in repetitive regions. These results suggest that replication slippage and homologous recombination are important processes in the formation of mutations.

Repetitive sequences have been shown to be involved in genomic instability [Calabretta et al., 1982; Bzymek and Lovett, 2001]. Repeat instability is also the basis of several diseases, exemplifying the importance of studying repetitive DNA and the mechanisms behind its mutagenic role [López Castel et al., 2010]. Homologous sequences can act as sources for homologous recombination, causing deletions or duplications of the sequence between the repeats. Additionally, highly repetitive regions can adopt different secondary structures, inducing problems during replication and consequently giving rise to repeat expansions, chromosomal fragility and gross chromosomal rearrangements [Voineagu et al., 2009].

My results from the genome-wide screen described in Chapter 3 further support the role of homologous sequences in the formation of deletions. I identified a larger number of yeast KO strains that acquired deletions between repeats than in the absence of them.

Certain KO strains are more prone to acquire deletions, even in the absence of external stress. Using constructs designed to monitor the occurrence of deletions, I identified strains with increased deletion formation in the presence and absence of external stress factors. Interestingly, there were more than 200 KO strains that acquired deletions in the absence of additional external stress. This suggests that the KO of these genes may have a strong negative effect on the suppression of mutations. Indeed in this group we found mutant strains for genes involved in DNA recombination and repair, synthesis-dependent strand annealing and mismatch repair. Under the presence of drug stressors causing replication stalling or direct DNA damage, these strains showed a higher fold enrichment compared to strains that gained deletions only after stress. In this latter group, the strains are likely functional under normal environmental conditions, but may have defects in coping with stressful growth situations. The use of designed constructs proved a useful method for screening for mutant strains that suppress the formation of deletions.

Chromatin remodeling proteins may be involved in suppressing the formation of deletions. In addition to mismatch repair genes, I found that the knockout of chromatin remodeling genes, namely *ISW1* and *SWR1* may also play a role in preventing the formation of rearrangements, in particular of deletions. In the neutral MA assay *isw1* and *swr1* were found among the knockout strains with the highest numbers of deletions. Isw1 and Swr1 are both part of protein complexes involved in chromatin remodeling [Mizuguchi et al., 2004; Smolle et al., 2012].

Chromatin architecture plays an important role in the control of gene expression by regulating the access of proteins to DNA [Cairns, 2009; Clapier and Cairns, 2009]. There is increasing evidence that chromatin remodelers are also important for DNA repair. For example, they are required for the phosphorylation of serine 129 of histone H2A in yeast (and serine 139 of histone H2A.X in vertebrates) upon DNA damage. This phosphorylation step leads to the recruitment of other chromatin remodelers that facilitate the access of repair proteins to the broken DNA ends [Redon et al., 2003; Clapier and Cairns, 2009; Xu et al., 2012].

Additionally, the exchange of H2A to H2A.Z (Htz1 in yeast) on nucleosomes at DSBs is required for the loading of DNA repair proteins [Xu et al., 2012]. The exchange is mediated by the SWR1 complex. Consistently, the high number of deletions that accumulated in the *swr1* KO strain in the MA assay supports its importance in preventing rearrangements, in line with previous reports demonstrating the involvement of H2A in the DNA damage response [Morillo-Huesca et al., 2010; Papamichos-Chronakis et al., 2011].

The ISWI complex also belongs to a family of chromatin remodelers with different subunit composition and biochemical activities than the SWR1 family [Clapier and Cairns, 2009; Wilson and Roberts, 2011]. Much less is known of its role in maintaining genomic instability compared to the SWR1 and SWI/SNF families. Interestingly, our results from the genome-wide screen, where the *ioc4* strain was enriched for deletions without any additional stress, further support the involvement of the ISWI complex in maintaining genomic stability. This complex is composed of three proteins, including

Isw1 and Ioc4 and acts in the repositioning of nucleosomes during transcription. Additionally, *isw1* mutants show strong defects in the chromatin organization [Gkikopoulos et al., 2011].

Taken together, these results suggest a potential role of the ISWI complex in the suppression of deletion formation. The detailed molecular mechanisms that cause genome rearrangements in strains having defects or a lack of proteins from the ISWI complex are, to my knowledge, yet to be characterized.

Meiosis related genes may also play a role in the prevention of deletion formation during vegetative growth. Among the strains that gained deletions after the treatment with different stressors I found a significant enrichment for genes involved in meiosis. The roles of these genes during vegetative growth are not fully described and their role in genome maintenance is not known. However, there are proteins that are important components of the meiotic machinery, but have further been implicated in the DNA damage response in other stages of the cell cycle, such as the Mre11 complex [Haber, 1998]. Given that meiotic proteins are involved in the creation and repair of DSBs, it is possible that these functions are also used in additional processes. Initial evidence exists for the involvement of hMSH4 in DSB repair in mitotic cells [Her et al., 2003; Chu et al., 2013]. Consistently, as discussed in Chapter 3, its yeast homolog mutant strain *msh4*, currently known to be involved in meiosis related processes, showed high number of deletions.

Limitations. The detection and study of genomic rearrangements faces some limitations despite the recent advances in sequencing technologies and computational tools. Some types of variants are particularly difficult to ascertain. We are aware that the detection of indels in the genome is still hampered by the difficulties of mapping reads in repetitive regions. Furthermore, repeated sequences create ambiguities in the alignments that can lead to false positive calls [Treangen and Salzberg, 2012]. We used stringent filtering criteria to reduce false positives caused by alignment artifacts, such as removing variant predictions that were present in 90% of the strains. Additionally, we performed experimental validations to confirm the computational calls. However, with this filtering process we cannot rule out that some real mutations were removed from the final catalog, or that a few false positives still remained.

Another important consideration is that the identification of mutations leading to a certain phenotype is not always a simple task. In the MA approach, each KO strain carried a deletion of a particular ORF. However, through the accumulation of mutations, other important genes for the maintenance of genomic stability may have been altered. This complicates the assignment of a phenotype only to the gene initially deleted. Therefore, the results derived from our approach could also be caused by multiple factors influencing DNA repair and recombination processes.

Through the MA assay, I identified mainly short deletions and insertions. Because of their smaller size and their occurrence in repetitive regions, where there are commonly less genes, they tend to cause less phenotypic impact. In the MA approach no selection method was used. However, only mutations not affecting survival or growth can be recovered. Therefore, we cannot discard the possibility that we were not able to observe a higher number of larger variants because they may have caused larger fitness defects compared to the smaller variants.

Furthermore, in the genome-wide screen for genes that suppress deletion formation, we scored deletions occurring at a specific locus in the genome by using designed constructs inserted by homologous recombination. Even though a large number of strains was transformed, strains with strong defects in this process were likely not represented in the final pools. Additionally, given the differences in the growth rates of the KO strains, several of them may have been absent from the final pools due to their slower growth rate and not related to their propensity to formation of deletions.

Since the distribution of genetic variants varies along the genome [Nachman, 2001; Mills et al., 2006], the results from one locus may not reflect the formation of deletions in the entire genome. However, since we compare the formation of deletions in all strains based on this locus, our results can be used as a starting point to carry out further experiments on the strains with increased levels of deletions. Such experiments can be, for example, the sequencing of the whole genomes to confirm the increased frequency of deletions in other loci.

Future perspectives. Even though the budding yeast has been extensively studied, the phenotypic consequences of different types of mutations are not fully understood. Systematic analyses of the deletion collection to identify genes that prevent the formation of mutations is still of interest given the diverse pathways involved in these processes [Huang et al., 2003]. The advances in sequencing technologies will continue to improve the characterization of genomic variants and their formation mechanisms. However, further improvements in these technologies, especially in read lengths, are needed to increase our ability to identify more complex rearrangements and variants at repeated sequences.

Further studies can be performed using similar techniques as the ones described in this thesis. For example, the constructs can be modified to test for the formation of other types of rearrangements, such as inversions and duplications. Other modifications to the constructs can include the substitution of the direct repeats with inverted repeats. Moreover, the understanding of the mechanisms of how genomic variants arise is improved by analyzing the sequences around their breakpoints. Therefore, it is of interest to investigate the resulting sequences in the constructs, after acquiring rearrangements, to assess for differences between strains that may provide insights into the formation mechanisms. The use of additional strains from other Yeast Deletion Collections may be also beneficial. For example, using a heterozygous collection allows the screening of essential genes, not included in the homozygous collection used in this study.

Additionally, the identification of genes involved in genomic stability is the first step into understanding the molecular processes that operate in the cells. In this regard, further experiments are required to determine the underlying mechanisms involved and the specific roles that each protein plays. In order to test for potential indirect effects of the candidate KO genes in our screen, one possibility is to assess transcriptional changes (for example, through an RNAseq experiment) in the strains before and after stress with drugs. This may reveal protein changes associated with higher levels of deletion formation. Due to the high evolutionary conservation of genome maintenance mechanisms, further studies in the budding yeast such as the ones presented in this thesis, will contribute to the global understanding of these processes in higher eukaryotes.

Appendix A

Methods

A.1 General protocols for Chapter 2

A.1.1 Mutation accumulation assay

In order to study the rates and types of spontaneous mutations in the yeast genome, a single-cell-to-colony bottleneck mutation accumulation assay was used. For this study we selected a set of 47 haploid deletion mutants from the yeast deletion collection (BY4741 *MAT*a his3Δ1 leu2Δ0 met15Δ0 ura3Δ0), with a range of defects in DNA replication, repair and recombination. The complete list of mutant strains used and their general function is shown in Table B.1.

Two colonies of each deletion strain were passed through a series of colony-to-single-cell bottlenecks that consisted on selecting individual colonies and streaking them out on YPAD plates to separate them into individual cells. These cells were grown into colonies for around 2 days at 30°C after which one colony was selected again for isolation. This process was repeated for a total of 90 bottlenecks (Figure 2.1). Frozen stocks were saved after bottlenecks 30, 60 and 90. This process leads to an accumulation of neutral mutations that can then be related to the effect of the disruption of specific genes that play a role in the maintenance of genome stability. The saved stocks at bottlenecks 30, 60 and 90, in addition to the original stocks at time point 0, were used for whole-genome sequencing as described below.

A.1.2 DNA extraction from yeast on 96 well plates

For extracting DNA from the yeast strains for sequencing, the PrepEase Genomic DNA Isolation Kit (Affymetrix) was used, with some modifications. Specifically, the Spheroplast and Enzyme Solutions were replaced by Buffer Y1 (Qiagen's Yeast Lysis Buffer: 1M Sorbitol, 100mM EDTA, 14mM β-mercaptoethanol) and Zymolyase (Seikagaku Corporation) treatment.

Two colonies of each strain, *i.e.* the two lines grown for up to 90 generations, were subject to paired-end whole-genome sequencing (WGS) at the first generation, and then after 30, 60 and 90 generations. For WGS, a small aliquot from the saved stocks was streaked out for single colonies on YPAD plates. A single colony of each mutation accumulation line (2 per deletion mutant strain) was inoculated into a single 4ml liquid YPAD culture in a glass tube, or 4×1ml YPAD cultures in 2ml deep well plates, and let grow overnight in a rotor or shaker at 200rpm and 30°C.

After 12-14h of growth the cultures were centrifuged for 2min at 3000rpm and the supernatant was removed. If the cultures were grown in deep well plates, the cells of all cultures were merged into a single plate after centrifugation. The cells were washed once with 450µl water. To make spheroplasts, 200µl of Buffer Y1, 10µl of Zymolyase (1000U/ml) and 0.5µl RNaseA (10mg/ml) were added per well with a multichannel pipet. The plate was covered with a plastic seal and placed in a shaker at 25rpm, at 37°C for 1-2h. After this, 200µl of water were added and the solutions were mixed shortly on a vortex.

The spheroplasts were collected by centrifugation at 6000rpm for 5min. The supernatant was discarded and 120µl of Homogenization Buffer (PrepEase kit) were added and mixed with a multichannel pipet to resuspend the pellet completely. 120µl of chloroform and 400µl of Protein Precipitation Buffer (PrepEase kit) were added to each well and mixed with a pipet. The solutions were centrifuged at 4000rpm for 45min. From the upper aqueous phase, 450µl were transferred to a new halfdeep well plate containing 340µl of isopropanol per well, mixed and let stand for 15min at -21°C. These solutions were again centrifuged at 4000rpm for 1h. The supernatant was decanted, 500µl of cold 70% ethanol were added to wash the DNA and centrifuged at 6000rpm for 10min.

After the supernatant was removed, the pellets were dried for 5min at 37°C. The DNA was resuspended in 200µl of TE and shake at 37°C for 15min to completely disolve the pellet. Typically, the concentration of at least 10 samples was measured with Qubit dsDNA BR Assay (Life Technologies).

A.1.3 Whole-genome sequencing of yeast strains

Before preparing the libraries for paired-end WGS, 5µg of yeast genomic DNA from each strain was sheared using a Covaris S2 system (Covaris Inc.) set to generate 500bp fragments: 90s, duty cycle of 5%, intensity of 3 and 200 cycles per burst. To purify the sheared products, the 120µl solution was dried down completely in a vacuum at 45°C and then resuspended in 30µl of EB. The DNA was then purified with 1.8x Ampure XP beads (Agencourt) following the default protocol. The final elution was done with 34µl of EB.

Library preparation was carried out using the NEBNext DNA Sample Preparation kit (New England BioLabs) with 32µl of the sheared and purified DNA and following the protocol of the manufacturer. After each step the products were purified using Ampure XP beads. Size selection was performed by either loading the products on 2% agarose gels and cutting the respective products at around 500bp, or by one step of Ampure XP bead purification adapting the amount of beads to achieve the desired fragment size. After the size selection, the products were quantified using Qubit dsDNA HS Assay (Life Technologies). A representative set of products was also loaded on a Bioanalyzer DNA 1000 chip (Agilent Technologies, Inc.) to assess the size distribution of the fragments.

All strains were sequenced with 101bp paired-ends using an Illumina HiSeq 2000 and aiming for a coverage of 20×. The samples were multiplexed using a set of 55 different molecular barcodes 6pb long (Table B.5), most of them described in Wilkening et al. [2013]. These barcodes have an equilibrated base composition in the first 2 positions, allowing for better clustering, and differ from each other by at least 3bp preventing the confusion of samples by sequencing errors [Wilkening et al., 2013].

A.1.4 Data analysis

This section describes in detail the pipelines depicted in Figures 2.1 and 2.5.

Sequence alignment. For mutation identification, the sequences were aligned against the S288c *Saccharomyces cerevisiae* reference genome (sacCer3) using the software Novoalign version V2.07.06 (`http://www.novocraft.com/`) with the parameter -r Random (random placement of reads that align to multiple locations in the genome).

Identification of the mutational spectrum of each deletion strain. Different types of mutations were identified with a set of computational tools as follows. To detect deletions, tandem duplications and inversions, the DELLY tool (v0.0.11) developed by Tobias Rausch in the Korbel Lab [Rausch et al., 2012b] was used, requiring a minimum paired-end mapping quality (-q) of 20 and a minimum flanking sequence (-m) of 5. Deletions and tandem duplications were also detected by Pindel (v0.2.4s) [Ye et al., 2009] with maximum size set to 7, *i.e.* a maximum SV length of 517.8kb. The same type of variants were identified by a third algorithm, BIC-Seq [Xi et al., 2010], with the advantage that this is a read depth based algorithm and therefore adds a different way of discovering these types of rearrangements.

Shorter insertions and deletions (indels) were additionally identified by using Dindel (v1.01) [Albers et al., 2011] and mpileup from Samtools (v0.1.18) [Li et al., 2009]. Dindel was run using in addition the predicted calls from Pindel as candidate regions. These methods are able to call only variants of around less than 50bp, in contrast to the tools described before, which can identify longer variants.

Cortex (v1.0.5.3), a tool that uses a *de novo* assembly approach, which in higher eukaryotes can be a difficult computational task. However, given the advantage of the small yeast genome, this method provides a way of detecting genomic variants without comparing to the reference genome. The parameters used were -kmer_size 31 -mem_width 100 and -mem_height 22.

All algorithms described were applied to each individual sample. On the other hand, the population based method GenomeSTRiP [Handsaker et al., 2011] was run on all samples simultaneously, with a minimum mapping quality of 20. By using the information of all the genomes, the quality of the calls and the power to infer them are increased.

The SNVs identification pipeline included calling with Samtools mpileup [Li et al., 2009] followed by bcftools view, and filtering by using the varFilter tools from vcfutils.pl using a maximum read depth (-D) of 50,000. To obtain high quality SNVs they were required to have at least 5 supporting pairs and a mapping quality higher than 40. Since SNV calling can have artifacts that affect most of the samples at the same positions, only variants called in less than 10% of the samples were considered for further analyses. *De novo* mutations were obtained by removing all variants occurring the the b0 strains (the original strains before the bottleneck assay).

For the sequence context analyses, 10bp up and downstream of each SNV were extracted and for each position the number of each nucleotide was counted. Equivalent nucleotide substitutions, like G>T and C>A, were combined.

Merging of mutation calls. After collecting the variants by applying the tools described in the previous section, the calls were merged with the goal of reducing the redundancy in the final set of calls. For merging, first the calls per individual were combined using the in house tool imerge (developed by Tobias Rausch). For this, a confidence interval around the predicted breakpoints was defined taking into account the accuracy of variant identification for each tool. Specifically, intervals for variants detected a nucleotide resolution were defined as 10bp to each side of the breakpoint. For DELLY calls without a breakpoint, the intervals were 20bp outside the variant and 50bp into the variant. Given that variant discovery in BIC-seq is based on read depth, the breakpoint prediction is more inaccurate and therefore the intervals were defined as 1000bp outwards and 400bp outwards. For Cortex, intervals of only 1bp around the breakpoints were allowed. For GenomeSTRiP, this step was not necessary. A variant was consider to be the same as another one if the defined intervals of both start and end breakpoints overlapped. These events were therefore merged into single variants.

Then the lists of variants detected by each algorithm were filtered for good mapping quality (>20) and at least 3 supporting read pairs. After this step, confidence intervals were defined once again for each list, using the same criteria as above and 50bp outwards and 100bp inwards for GenomeSTRiP. Using these intervals, all variant calls among tools were merged together using imerge. Only those calls detected by at least two algorithms were considered. Additionally, given that there are regions where some artifacts can happen, or where the reference genome is different from all other samples, only those variants that were present in less than 10% of the samples were kept. This set was then called "unique" variants. On the other hand, since in this project there are strains sequenced at different time points, mutations detected in the later generations that were already present in previous generations were also removed, resulting in a final set of "*de novo*" variants.

Estimation of mutation rates. Estimation of the mutation rates per base per generation was based on the number of SNVs per strain. The rates were calculated as $N/n\times t\times bp$, with N=total number of SNVs at b60; n=number of mutation accumulation lines (two in this study); t=mean number of generations; and bp=yeast haploid genome size (12,162,995bp). The mean number of generations was estimated considering 20 generations per bottleneck (after each single-cell-to-colony bottleneck the cells were allowed to form colonies for 48h), and 60 bottlenecks, accounting for 1200 generations.

Functional annotation of mutations. The annotation of indels and SNVs was performed using the ANNOVAR software [Wang et al., 2010]), using sacCer3 as the reference genome and the Ensemble gene annotations for *Saccharomyces cerevisiae.*

Additionally, all deletions and insertions independent of size were overlapped with a set of different genomic features, using published data, to analyze the location of the variants and their possible functional effects. For this, I compared the overlap of the variants with genes (ensGene from UCSC), recombination hotspots [Mancera et al., 2008], crossovers [Mancera et al., 2008], 3'UTR and 5'UTR [Nagalakshmi et al., 2008], nucleosome positioning based on H2AZ [Albert et al., 2007], origin of replication sites, DSB hotspots [Pan et al., 2011], TATA elements [Rhee and Pugh, 2012], transposable elements from the SGD, transcription start sites (TSS) [Zhang and Dietrich, 2005] and simple repeats from the SGD.

For comparison, a permutation approach was used to assess the significant values of the overlaps. The set of discovered mutations from out strains was distributed along the whole yeast genome and the overlap with the mentioned features was assessed as described in the previous paragraph. This process was iterated 1000 times, and the percent of random mutations overlapping each of the defined genomic features was estimated. Wilcoxon rank-sum tests were used to compare the overlaps between the detected mutations with different features and the randomly positioned mutations.

SNVs and indels were also annotated using ANNOVAR [Wang et al., 2010] with gene annotations downloaded from UCSC (`ftp://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/database/`). The variants were classified as exonic or intronic. For exonic variants, their functional consequences were also predicted and classified as nonsynonymous SNV, synonymous SNV, frameshift insertion, frameshift deletion, nonframeshift insertion, nonframeshift deletion, and frameshift or nonframeshift substitution.

Identification of chromosome aneuploidies. Chromosome aneuploidies were identified by estimating chromosome copy number based on the read depth. To estimate the read depth, the yeast genome was divided into non overlapping 10bp windows. For each strain, the number of reads aligning to a specific window were counted using an in-house tool "cov" (developed by Tobias Rausch). To infer copy number changes between the initial wild type (WT-b0) strain and the deletion strains at all time points (b0, b30, b60, b90), the read counts of each sample were normalized for differences between the coverages. For this normalization step, the read depth ratio between the WT-b0 strain and each of the mutant strains was estimated. The median of these ratios was multiplied by the counts of each time point per strain. Then the $\log_2$ ratios of the normalized read counts were calculated per window and subjected to GC normalization based on the relation between the GC content and the $\log_2$ read depth ratio of each sample pair (WT-b0 versus any of other generations).

Based on the $\log_2$ read depth ratio estimations for each sample pair, chromosome gains were identified as those having ratios larger than 0.7, instead of the expected $\log_2$ ratio of one for a gain of a chromosome (or part of it) in a haploid genome. This was chosen because there were some cases where the sequenced colony might not have been completely pure, giving a ratio lower than one, although still sufficiently distinct to the b0 strain and to other chromosomes, making it possible to identify chromosome copy changes. An aneuploidy was then scored if the chromosome gain was seen in time points b30, b60 or b90, but absent in the b0 of the corresponding strain. This means, that chromosome gains occurred through the mutation accumulation assay.

Plots and statistical tests. All plots and statistical analyses in this section were done using the software environment R, version 3.1.0 [R Core Team, 2014].

A.1.5 Validation of genetic variants by PCR and Sanger sequencing

To validate the computationally predicted mutation calls, PCRs were performed aiming to have either a difference in size between a positive and a negative SV call, or a difference in the ability to produce an amplicon. In the latter, a lack of a PCR product is indicative of an absence of the SV. To validate SNVs and indels, the primers were designed to amplify a product where at least one of the primers was close enough to the prediction that it could be used to perform Sanger sequencing. This way the presence of the mutation could be validated by looking a the sequence directly. When possible, the fragments were sequenced from both directions. Considering that some of the SVs were not predicted at a nucleotide resolution, the primers were placed 500bp away from the breakpoints when there were no repetitive sequences in the region. Otherwise the primers were placed farther away to avoid the amplification of unspecific products.

All primers were designed using the Primer3Plus software [Untergasser et al., 2007], and amplicons were generated using the SequalPrep Long PCR Kit (Invitrogen). Reactions were done following the specifications of the manufacturer, with total volumes of 25µl, annealing temperatures of 56-54°C and adapting the elongation steps depending on the expected product size. The PCR products were loaded on 1% agarose gels and visualized with the SYBR Safe Dye (Invitrogen) and using a 100bp or 1kb DNA ladder (NEB) depending on the expected product size. When required, the products were cut out of the gel, purified with QIAquick Gel Extraction Kit (Qiagen), and eluted in 30µl EB buffer. The products were sent to GATC Biotech (Germany) for Sanger sequencing. All sequences were analyzed with the Sequence Scanner Software (Applied Biosystems). For these validations, the DNA used was the same as the one used for the WGS experiments, and obtained as described in Section A.1.2.

A.1.6 Verification of aneuploidies by qPCRs

To confirm the presence of aneuploidies in the deletion strains that arose during the bottleneck mutation accumulation assay, qPCR assays were designed. Two primer pairs per aneuploid chromosome were chosen, binding to the left and the right arms of each chromosome. Some primer sequences were obtained from Pavelka et al. [2010] whereas others were newly designed following the methods from these same authors and using the Primer3Plus software [Untergasser et al., 2007] with default qPCR settings (except reducing the maximum length of mononucleotide repeats to 3). As template sequences, yeast intergenic regions not overlapping repetitive elements were downloaded from the *Saccharomyces* Genome Database (`http://downloads.yeastgenome.org/sequence/S288C_reference/intergenic/NotFeature.fasta.gz`, version from the 03-Feb-2011). The amplicon lengths were chosen to be between 75-200bp. Primer sequences can be found in the Supplementary Table B.4.

qPCRs experiments were performed using the SYBR Green PCR Master Mix Kit and following the standard protocol (Applied Biosystems) and 20µl total reaction volumes. All assays were set up in 96 well plates and included three replicates of each sample, the standards and non-template controls. For the samples, a total of 0.1ng of DNA was used for a total concentration of 0.02ng/µl per reaction, and final primer concentrations of 0.2pM/ul. A master mix with the primers and the SYBR Green mix was prepared and aliquoted into each well.

To estimate the fold change in chromosome copy number a relative quantification based on the $\Delta\Delta C_t$ method was used. The C_t (cycle threshold) refers to the number of cycles at which the fluorescent signal produced by the DNA-binding dye in the SYBR Green is detectable. In this assay, the relative amount of an amplicon is estimated in relation to a reference locus in the same sample and then this is compared to the respective ratio in a control sample. By doing the normalization to a reference locus, differences in the amount of sample loaded into the assay can be compensated. In all experiments the WT-b0 strain was used as the control, and the *SPT15* gene was used as the internal reference locus.

First, the C_t of the target chromosome was normalized to the C_t of the reference locus for both, the test and control samples: $\Delta C_{t(test)} = C_{t(testTarget)} - C_{t(testRef)}$ and $\Delta C_{t(control)} = C_{t(controlTarget)} - C_{t(controlRef)}$. Then the ΔC_t of the test sample was normalized against the ΔC_t of the control sample: $\Delta\Delta C_t = \Delta C_{t(test)} - \Delta C_{t(control)}$. Finally, the fold change (or relative copy number) of the target chromosome in the test sample compared to the same locus in the control sample and normalized by the copy number of a reference locus was estimated as $2^{-\Delta\Delta C_t}$.

To estimate the primer efficiencies, standard curves of the C_t values versus the $\log_1 0$ of the quantities of template were produced using 8 serial two-fold dilutions from 2ng to 0.016ng for the WT-b0 control strain. A regression line was fitted and the amplification efficiencies were calculated as $10^{(-1/slope)}$ (Table B.4). Examples of the primer efficiencies for both arms of the assessed chromosomes and the reference locus are shown in Figures 2.8 and B.2. Spearman rank correlation tests were used to calculate the *P*-values for the relation between read depth base predictions and qPCR fold change estimates.

A.2 General protocols for Chapter 3

A.2.1 Construct design

The deletion construct used in this study was design using SnapGene Viewer v2.5 [SnapGene] and the plasmid editing software ApE [Davis, 2012].

A.2.2 Construct synthesis

Due to its complex sequence composition, the deletion construct was synthesized by GENEWIZ, Inc. Custom Gene Synthesis service and was received in a pUC57-Amp plasmid. It was then cloned and modified as detailed below.

From the synthesized deletion construct (DelRep), another construct was derived by removing the direct repeats (DelNoRep) present at the boundaries of the *URA3* gene (Figure 3.1). To remove the direct repeats, restriction digestions with enzymes PsiI (NEB) and NaeI (NEB) for the left side repeat and enzymes PmeI (NEB) and SnaBI (NEB) for the right side, were performed, independently for each repeat. For each restriction digestion, 1µl of the original purified plasmid with the construct was used, in addition to 2µl of 10x CutSmart Buffer (NEB) and 1µl of each enzyme. The reaction was incubated at 37°C for 1h. After the first restriction digestion, the open plasmid was re-ligated using the Quick Ligation Kit (NEB), with 10µl of restriction digestion reaction, 10µl of 2x Buffer and 1µl of Ligase enzyme. The mix was incubated for 5min. at room temperature. Once the left repeat was removed, 2µl of the quick ligation reaction were used to transform 50µl of JM109 cells for cloning of the plasmid (for more details about cloning see Section A.2.3). This process was repeated to perform the deletion of the repeat on the right side of the construct, using as a starting plasmid the one lacking the left repeat.

Confirmation of the deletion of both left and right homologous repeats in the deletion construct was performed by looking at the size differences of the products on a 1% agarose gel (Section A.1.5). Additionally, Sanger sequencing of the area surrounding the repeats was used to confirm de deletions.

The two deletion constructs, one carrying the direct repeats and the one without it, were cloned (Section A.2.3) and used to amplify the constructs. The specific deletion constructs were amplified then by PCR using primers M13FW and M13RV (the sequences are displayed in Table B.3). The PCR products were then separated on a 1% agarose gel (Section A.1.5), purified using the QIAquick Gel Extraction Kit (Qiagen), and used for the bulk transformation as described below.

Furthermore, a control construct (Hyg+), conferring constitutive hygromycin resistance was also generated by removing the *URA3* gene interrupting the *HPH* gene allowing its expression and conferring resistance to the strains. Starting from the DelRep construct in pUC57-Amp plasmid, a restriction digestion using blunt enzymes PsiI and PmeI (NEB) was performed to excise the *URA3* gene. Then the plasmid was re-ligated using the Quick Ligation Kit (NEB) as explained before. The size of the new construct was confirmed in a 1% agarose gel, and its sequence was confirmed by Sanger sequencing.

For transformation into the pool of yeast homozygous deletion mutants, all constructs were linearized by restriction digestion using enzymes HindIIIHF and SacI (NEB).

A.2.3 Construct cloning

Each plasmid containing the deletion constructs was transformed into JM109 competent bacteria for cloning. The bacteria were thawed on ice for approximately 5min. mixing the cells by gentle flicking. For each transformation reaction, 10ng of the plasmid and 50µl of competent bacteria were added to 1.5ml reaction tubes. The mix was incubated on ice for 20min. The cells were then heat shocked at 42°C for 45s. in a water bath without shaking. The tubes were then immediately return to ice for 2min. After this, 250µl of S.O.C. medium (Invitrogen) at room temperature were added to each reaction tube. The cells were then incubated for 1.5h at 37°C with shaking.

To select for transformants, 100µl of the cells were plated in duplicate on LB + Amp plates. The plates were incubated overnight at 37°C. The next day, the plates were screened for colonies and four colonies for each plasmid were selected. These colonies were then inoculated into 5ml LB + Amp cultures and incubated at 37°C overnight. The following day, the plasmids were purified using the Mini Prep plasmid Purification Kit (Qiagen).

A.2.4 Strains used

A homogeneous pool of the homozygous yeast deletion collection was obtained from the Steinmetz lab at EMBL. All strains were in a S288c background. The distribution of the strains was assessed by amplification of the uptags and downtags as explained below.

The pool was composed of 5083 different mutant strains, and the growth rates of these mutants is known to differ due to their deficiencies in different cellular processes. Therefore, to avoid an imbalance in strain constitution of the pool due to growth, the incubation times and the timing of the experiments was always optimized to be as short as possible.

A.2.5 Yeast bulk transformation

All transformation were done using the high-efficiency Lithium Acetate (LiAc), single stranded carrier DNA and Polyethylene Glycol 3350 (PEG) method [Gietz and Schiestl, 2007; Knop et al., 1999], with a few modifications. In summary a 50µl aliquot of the pooled homozygous yeast deletion collection, at an OD_{600} of 50 was inoculated into a 5ml YPAD culture and incubated overnight on a rotor at 30°C.

The next day, the overnight culture was diluted to get an OD_{600} between 0.2 and 0.3. Usually this meant diluting 2ml of the overnight culture in 50ml YPAD. The dilutions were grown at 30°C at 180rpm for around 2.5h to reach and OD_{600} of 0.5-0.7. Once this OD was reached, the cultures were centrifuged at 2500g for 3min and the supernatant removed. The cells were then washed by resuspending them in 50ml of water and repeating the centrifugation step using the same conditions. The supernatant was discarded, and the cells were once again resuspended in 1ml of water.

From these cells, 100µl were transferred to a new tube and used for each transformation. For each construct to be transformed, 20 independent transformations were done. The 100µl aliquot was centrifuged and the supernatant removed. Then, 34µl of the construct were added to the cells and mixed by pipetting, corresponding to around 800ng to 1µg of linearized construct DNA (prepared as explained in Section A.2.2), and incubated for 10min at room temperature.

A transformation mix was prepared using 240µl of PEG 50%, 36µl of LiAc 1M and 50µl of single stranded DNA (salmon testis DNA denatured by heating at 95°C for 10min and placed immediately on ice). This mix of 326µl was added to each of the 100µl grown yeast cells. The transformation tubes were incubated for 70min at 42°C. This time was optimized for high transformation efficiency of the yeast pool by performing several transformations with varying heat shock times and estimating the transformation efficiency.

After the heat shock, the mix was centrifuged at 2500g for 30s. The supernatant was removed and 500µl of YPAD were added to each transformed cells. These cultures were incubated at 30°C and 550rpm for recovering the strains for 2h before plating in synthetic complete medium lacking uracil (SC-URA). For the selection of transformed strains, 250µl of the recovered culture were plated in SC-URA on 10cm petri dishes. Two plates were done for each transformation experiment. The plates were incubated at 30°C for 4 days.

After this incubation time, the colonies present on each plate were counted and picked manually. All transformed strains with the same construct were combined in sets of around 2000 colonies and stored in YPAD glycerol stock and stored at -80°C. In total, 20,000 colonies were picked for each construct transformation to cover, in theory, each ORF in the homozygous yeast deletion collection approximately 5 times.

A.2.6 Confirmation of construct insertion

The insertion of the construct into the *HXT13* gene in chrV was verified by PCRs, positioning a primer into the cassette and a primer ouside of the cassette in both sides of the insertion were verified. Primers LeftOutConstruct, LeftIntConstruct, RightOutConstruct and RightIntConstruct in Table B.3 were used, yielding products of 531bp for the left side and 535bp for the right side. The PCRs were performed using the LongAmp Taq PCR kit (NEB), with an initial denaturation step of 94°C for 1min, followed by 35 cycles of 20s at 94°C, 30s at 54°C and 1min at 65°C, and a final elongation step of 10min at 65°C, ending on hold at 10°C. Presence of a PCR product of the correct size was confirmed in a 1% agarose gel (Section A.1.5).

A.2.7 Inducing replication stress to the pooled deletion collection

To increase mutation rates and to get some insights into the mechanisms that give rise to a higher mutation formation, the transformed pools were treated with several drugs that cause different stresses to the cells. The drugs applied were: hydroxyurea (HU), methyl methanesulfonate (MMS), doxorubicin (Doxo) and camptothecin (Campt) (all from Sigma) to induce replication stress or DNA damage.

The induction of replication stress and DNA damage was done in deep well plates of 2ml. Using 100µl of the corresponding transformed pooled strains, a dilution to 200µl was done with YPAD. These cultures were treated with the drugs and concentrations depicted in Table B.6 and were grown overnight at 30°C. A non treated control was always included. For MMS, the treatment time was reduced to 1h because cell viability is lower in this drug. A non treated control, grown for 1h, was also included while treating cells with MMS to have a similarly grown culture since the composition of the pool changes with the growing timings.

After treatment, the cultures were centrifuged and the supernatant removed. The cells were then washed 2 times by adding 500µl of YPAD, and centrifuged, and the supernatant was discarded. The cells were then let to recover by resuspending them in 500µl of YPAD and incubating for 2.5h at 30°C.

Selection of the strains that acquired rearrangements in the construct was then done by plating the cultures in YPAD + Hyg plates (with a hygromycin concentration of 200µg/ml) and letting them grow at 30°C for three days. After this, all the Hyg^+ colonies were picked from the plate and stored at -80°C for subsequent experiments.

For confirmation of Hyg resistance, some plates were also replica plated in YPAD + 5FOA (5-fluoroorotic acid). The selectable marker *URA3* allows for positive and negative selection. The presence of the *URA3* gene makes strains not able to grow on 5FOA. However, if the strains have lost the expression of this gene, and in this particular case, if they have deletions in this gene that allow the expression of the Hyg resistance gene, then they are also able to grow in 5FOA plates. Therefore, this method was used to double select the strains.

A.2.8 Amplification of the tags from the transformed pooled deletion collection

For each enrichment and selection experiment, all Hyg^+ colonies grown on YPAD + Hyg plates, *i.e.* all strains that acquired a rearrangement in the construct and therefore became hygromycin resistant, were picked from plates, and glycerol stocks were made for storage at -80°C. From these stocks, 10µl were used for PCR amplification of the tags. The aliquot was diluted in 20µl of SDS (0.2%) and incubated at 95°C for 5min to break the cells, after which the cells were centrifuged and 2µl of the supernatant were used for the PCR.

The PCRs to amplify the uptags and downtags were done for the pooled yeast cultures using primers U1+KanB and D1+KanC respectively (see Table B.3 for the specific sequences). PCR amplification using primers U1+KanB yields a product of 299bp containing the uptags, and D1+KanC yields a 624bp product. Both tags were amplified in a single 20µl PCR reaction using the SequalPrep Long PCR Kit (Invitrogen), with

2min at 94°C, 10 cycles of 20s at 94°C, 40s at 56°C, and 1min at 68°C, followed by another 25 cycles of 20s at 94°C, 40s at 54°C, and 1.5min at 68°C, and a final step of 10min at 72°. 2µl of the PCR products were loaded on a 1% agarose gel for verification (Section A.1.5), and upon successful amplification, the PCRs were then purified using 1.8µl of AMPure XP beads (Agencourt) per 1µl of PCR product.

A.2.9 Sequencing of tags from the pooled deletion collection

The purified PCR products were used for library preparation and multiplex sequencing using the procedures and barcodes described in Section A.1.3. The protocol was slightly modified, skipping the size selection step of the library preparation because the amplicons had already sizes small enough to reach the barcode sequence from one side. The sequencing was done on Illumina HiSeq 2000 or MiSeq instruments with paired ends of 101bp or 150bp respectively.

A.2.10 Data analysis

Identification of strains through the sequences of the barcodes. The sequencing reads were then used for barcode identification. In theory, each deletion strain contains one unique 20bp sequence in each of the amplicons. Based on some Sanger sequencing experiments, I realized that some of the mutants contain truncated barcodes, or sequences that extend beyond the theoretical 20bp. The deletion collection has been shown to have some inconsistencies, and therefore for the identification of the strains through the barcodes we used strict criteria to perform the correct annotation.

Given that the PCR amplicons containing the uptags and downtags of all deletion strains have the same sequence except for the mentioned 20bp, for each sequencing read, we could remove the shared sequences using the trimming tool cutadapt [Martin, 2011]. It finds and removes sequences (adapters) from the 3' and 5' ends of the reads. By using the parameter -a the sequences in the 3' end of the barcode were removed, and by using the parameter -g, the sequences at the 5' end of the barcode were trimmed. As queries, sequences of 20bp directly up and downstream of the barcode were used. Since the reads may contain sequencing errors, the allowed maximum error rate (-e) was set to 0.15. This accounts for up to 3 deletions, insertions or mismatches per adapter. The resulting sequences were kept for following analyses if they were between 15-21bp long.

The trimmed sequences, that corresponded to the tags, were used to identify the strains. For this, the list of all matched barcodes and strains was obtained from the *Saccharomyces* Genome Deletion Project website (`http://www-sequence.stanford.edu/group/yeast_deletion_project/strain_homozygous_diploid.txt`), and used to search for our sequences allowing up to 3 inconsistencies, including mismatches, deletions or insertions. Only barcodes that could be assigned uniquely to a strain name were used for the analysis. The number of reads per strain were quantified and only strains supported by at least 10 sequencing reads, and identified by both up and downtags, were considered to test for enrichments.

Assessing the strain composition of the transformed pools. After performing the bulk transformation of the constructs, the strain composition of the pooled cultures transformed with each of the constructs (the deletion construct with and without direct repeats and the control construct that confers constitutive Hyg resistance) was assessed by sequencing the barcodes and identifying the strains as described above. Additionally, the original pool before the transformation was done, was also sequenced as a control to assess how many strains could be identified by the PCR of the barcodes. The percentage of strains out of a total of 5083 possible strains was estimated.

Identification of significantly enriched strains. To identify Hyg resistant strains that were significantly enriched after applying each stressor or after growth in YPAD, the R package DESeq2 [Love et al., 2014] was used. Differences in sequencing coverage for each experiment are accounted for in the DESeq package, so that the read count data can be used directly for analysis. We assessed if there were differences in the composition of the Hyg resistant strain pools by using the read count data in the form of a matrix with rows being all strains detected and columns being each treatment. A significant enrichment of at least 2 fold was required ($\log_2$ fold change lfcThreshold=1). Differential analysis was performed considering the pool of strains with constitutive Hyg resistance as control.

Strains enriched in the pool without treatment were considered as having a stronger genotype effect, *i.e.* the KO of a gene had an influence on the ability to suppress the formation of deletions. The strains that were enriched only after drug treatment were considered to have a higher influence from the environment, *i.e.* the stressor effect had a stronger impact on the formation of deletions. Strains that were present before and after treatment had an additive effect of the influence of their genotype and the drug stressor.

Growth rates comparison. The growth rates in rich medium for all yeast homozygous deletion strains were downloaded from `http://www-deletion.stanford.edu/YDPM/index.html`. The results of two time course experiments are available, but are highly correlated. Therefore the measures of growth base in the first time course were used for the comparisons. The growth rates were estimated by a linear regression fit to intensity values obtained from growth experiments [Steinmetz et al., 2002].

The growth rates of the strains from the bottleneck mutation accumulation assay (Table B.1) were compared to the growth rates of the top ten strains detected in the enrichment assay carrying the DelRep construct and treated with drugs (Table B.9). Additionally, the growth rates of the strains detected after no stressor were also used for comparison. Wilcoxon rank-sum tests were used to assess the differences between the growth rates of these pools of strains.

Pathway analysis of the enriched strains. Gene ontology enrichment analysis was performed with the AmiGO 2 GO term enrichment tool based on the PANTHER Classification System [Mi et al., 2013] The PANTHER database maintains up to date GO annotations. The estimations of significance for the enrichment analysis are computed based on the background number of genes that belong to a specific GO term and the number of strains belonging to that pathway the sample set.

Additionally, GOSlim terms were used to create Figure 3.9, in which the GO terms are parental categories, *i.e.* broader groups that comprise several GO terms. The plot and the estimation of the significance of enrichment were done using Cytoscape [Shannon et al., 2003] and the plugin BiNGO [Maere et al., 2005], with a custom reference set of strains used as background. For this reference set, only the strains detected in the original pool of deletion mutants were included.

Plots and statistical tests. All plots and statistical analyses in this section were done using the software environment R, version 3.1.0 [R Core Team, 2014], unless otherwise specified.

A.2.11 Verification of candidate genes

Creation of KO strains. Due to several issues and concerns about the original yeast deletion collection, *e.g.* Ben-Shitrit et al. [2012], the deletion strains for the candidate genes were generated starting with the BY4341 (MATa) and the BY4342 ($MAT\alpha$) wildtype haploid strains. For this aim, a PCR deletion strategy was used [Baudin et al., 1993; Wach et al., 1994], similar to the procedure used to create the original deletion strains, where each desired ORF is substituted from its start and stop codons with the KanMX4 cassette. In this case, the unique barcodes incorporated in the original experiment were not included, as the strains were grown afterwards independently. Once strains of both mating types were created, homozygous diploid deletion mutants were also generated as described in the following section.

The KanMX4 cassette was amplified from the pFA6a-KanMX4 plasmid (kindly provided by the Steinmetz Lab, EMBL) with extended U2 and D2 primers, each including 40bp homologous to the upstream and downstream sequences of the yeast ORFs respectively (full sequences are provided in `http://www-sequence.stanford.edu/group/yeast_deletion_project/strain_homozygous_diploid.txt`). The cassette was amplified using the SequalPrep Long PCR Kit (Invitrogen) as described in Section A.2.8, with extension times of 2:30min for the first 10 cycles and and 2:45min for the remaining 25 cycles. PCR products were loaded on a 1% agarose gel (Section A.1.5) and the bands were cut out and purified with the QIAquick Gel Extraction Kit (QIAGEN) following the protocol from the provider. Concentrations of the purified products were measured with Qubit dsDNA BR Assay (Life Technologies).

The amplified KanMX4 cassette containing the homologous sequences were used in independent transformation assays, using the Gietz and Schiestl [2007] protocol as described in Section A.2.5. After the transformation, the strains were plated on YPAD + G418 plates (with a geneticin concentration of 200ug/ml). The plates were incubated at 30°C for 2 days and replica plated to new YPAD + G418 plates to remove false positives. After plate incubation for 2 more days, three colonies were confirmed by PCR as described in Section A.2.11, inoculated in a 5ml YPAD and grown overnight. Glycerol stocks were made for each strain and stored at -80°C.

Strain confirmation. PCRs using primers outside of the KanMX4 cassette 200-400bp upstream and downstream of each ORF (primers A and D), and a pair of primers within the cassette (primers KanB and KanC in Table B.3), were done as described by [Winzeler, 1999]. Therefore, upon successful deletion of the desired ORF by replacing it with the KanMX4 cassette, the PCRs using primers A for each specific ORF and KanB, or primers D and KanC, should both give products. PCRs were performed using the primers listed in the *Saccharomyces* Genome Deletion Project website (`http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html`) and the LongAmp Taq PCR kit (NEB) as described in Section A.2.6. All products had sizes of around 500-800bp.

Creating homozygous diploid KO strains. Diploids were created by mating the KO strains of each mating type. One *MAT*a colony KO was streaked onto a YPAD plate. On top of the streaked colony, another *MAT*α colony of the same KO was streaked out. The plates were left to grow at 30°C for two days after which a small amount of the combined strains was streaked out for single colonies into a new YPAD plate. From the growing colonies, two were selected for PCR assessment of the mating type to confirm the ploidy as explained below.

Assessment of yeast mating type by PCR. To verify the mating type of each newly created strain, a PCR-based assessment method was used following Huxley et al. [1990]. In summary, it is possible to amplify regions of the *MAT* locus that clearly distinguish between three possible mating types, namely a, α, or the a/α diploid. For the PCR, one single colony was dissolved in 20µl SDS 0.2% and incubated at 95°C for 5min. Then 1µl of the supernatant was used for each reaction. Three primers are used in each reaction (Table B.3). Two are forward primers: primer *MAT*a binds to a sequences within the a-specific region of the *MAT*a locus. Similarly, the *MAT*α primer binds to a α-specific sequence. The reverse primer aligns to a sequence directly downstream of the *MAT* locus. By using the three primers in a single reaction, the *MAT*a strains with generate a PCR product of 544bp, the *MAT*α strains with give a 404bp product. The diploid strains will show both bands.

Induction of replication stress in individual KO strains. Once the individual KO strains containing the constructs were created, they were treated independently with the same drugs used in the pooled screen to confirm the effect on the formation of deletions. For this step, one colony of each strain was inoculated to a 2ml YPAD culture and grown overnight at 30°C. The next day, several aliquots of around 9.25×10^5 cells of the grown strain were each diluted to a total of 100µl YPAD culture. Each 100µl culture was then treated with the appropriate amounts of drug to achieve the concentrations shown in Table B.6.

Identification and quantification of strains with rearrangements in the construct. The detection of strains with rearrangements in the construct that confer hygromycin resistance was done by plating around 3×10^4 of the treated cells on YPAD + Hyg plates. The plates were incubated at 30°C for 3 days after which the colonies on each plate were counted. To be able to quantify the differences between the treatments and the KO strains, the same amounts of cells were plated.

Appendix B

Supplementary figures and tables

B.1 Supplementary figures and tables for Chapter 2

Table B.1: Yeast deletion mutant strains used in the neutral mutation accumulation experiment grouped into broad functional categories.

Deleted yeast gene	General functions	Main functional category
MRE11	DSB repair by recombination	Recombination
RAD52	DSB repair by recombination	
SGS1	Illegitimate recombination suppression	
RAD59	Recombination, single-strand annealing	
ESC2	Transcriptional Silencing, Recombination, DNA Damage Checkpoint	
MSH2	Mismatch repair	DNA Repair
LIF1	Non-homologous end-joining	
SAE2	Hairpin DNA processing	
RTT107	Replication fork repair	
SRS2	DNA repair, genome stability	
RAD50	DSB repair, NHEJ	
MMS4	Recombination and repair	
RAD6	Histone Modifying, DSB repair, Checkpoint Control	
PMS1	Mismatch Repair	
RAD10	NER (Nucleotide Excision Repair), DSB repair	
MUS81	Replication fork restart, DNA repair	
RAD18	Postreplication repair, Ubiquitin ligase	Chromosome segregation
CHK1	Chromosome Segregation (DNA damage checkpoint)	
SGO1	Chromosome Segregation, Spindle Checkpoint	
CTF19	Chromosome Segregation	
CIN8	Chromosome Segregation	
NUP170	Chromosome Segregation, Nuclear Pore Complex	
RSC1	Chromatin Remodeling	Chromatin remodeling
ASF1	Chromatin Remodeling, Ty1 Transposition	
ISW1	Chromatin Remodeling	
SPT2	Chromatin Remodeling	
SWR1	Chromatin Remodeling	

continued on next page

Table B.1: continued from previous page

Deleted yeast gene	General function	Main functional category
CTF8	DNA damage replication checkpoint, Sister Chromatid Cohesion	Replication
CSM3	DNA damage replication checkpoint, meiotic chromosome segregation	
TOF1	DNA replication checkpoint	
RAD27	Okazaki fragment processing	
POL32	Error-prone DNA synthesis	
RRM3	Ty1 Transposition, Replication Fork Stalling	
CTF4	DNA Replication, Sister Chromatid Cohesion	
SLX1	Replication fork restart	
DUN1	DNA damage replication checkpoint, post-replicative DNA repair	
HST3	Transcriptional Silencing	Other functions
ARD1	Transcriptional Silencing, N-terminal acetylation	
RAD24	DNA damage checkpoint signaling	
DOT1	Histone Modifying, DNA damage response	
SET1	Histone Modifying	
PMR1	Protein sorting	
YKU70	Telomere maintenance and NHEJ	
YOL086W-A	Genome stability maintenance, homolog of Fanconi Anemia Complex	
GCN5	Transcription regulation	
BY4741	Parental strain wild type	
HIS1	Histidine biosynthesis - control	
TRP5	Tryptophan biosynthesis - control	

Table B.2: Total number of *de novo* mutations accumulated in different yeast deletion mutants. Two replicates are shown, corresponding to the two MA lines per deletion strain analyzed. (SI: short insertions, Tandem dup: tandem duplications).

Deleted yeast gene	Replicate	Deletions	SI	Tandem dup.	SNVs
ARD1	1	15	5	0	48
ARD1	2	18	8	0	33
ASF1	1	10	7	0	81
ASF1	2	13	6	1	47
BY4741	1	17	4	0	41
BY4741	2	15	5	0	77
CHK1	1	29	17	0	17
CHK1	2	53	11	1	12
CSM3	1	12	4	0	91
CSM3	2	91	25	2	10
CTF4	1	36	10	0	97
CTF4	2	31	13	1	77
CTF8	1	11	0	0	30
CTF8	2	113	21	10	20
CTF9	1	32	10	0	30
CTF9	2	6	2	0	52
DOT1	1	35	16	1	23
DOT1	2	30	6	1	15
DUN1	1	16	5	1	11
DUN1	2	6	2	1	49
ESC2	1	8	1	0	30
ESC2	2	11	8	0	40
GCN5	1	14	4	3	83
GCN5	2	30	14	1	23
HIS1	1	195	16	2	48
HIS1	2	26	12	4	29
HST3	1	17	5	1	42
HST3	2	81	24	4	32
HTZ1	1	39	12	2	59
HTZ1	2	47	13	1	29
ISW1	1	86	18	0	31
ISW1	2	110	23	2	22
LIF1	1	87	13	2	82
LIF1	2	26	4	1	74
MMS4	1	55	18	6	20
MMS4	2	76	32	3	26
MRE11	1	117	26	6	51
MRE11	2	38	12	2	67
MSH2	1	1002	146	0	175
MSH2	2	418	38	0	174
MUS81	1	31	12	1	21
MUS81	2	1	2	0	26
NUP170	1	73	11	4	10
NUP170	2	20	4	0	25
PMR1	1	18	17	0	91
PMR1	2	26	11	3	104

continued on next page

Table B.2: continued from previous page

Deleted yeast gene	Replicate	Deletions	SI	Tandem dup.	SNVs
PMS1	1	9	10	1	65
PMS1	2	17	10	3	48
POL32	1	57	4	1	37
POL32	2	58	9	3	37
RAD10	1	26	14	3	39
RAD10	2	12	6	4	66
RAD18	1	21	6	1	18
RAD18	2	15	2	3	46
RAD24	1	28	10	2	31
RAD24	2	28	11	0	37
RAD27	1	123	67	2	76
RAD27	2	84	85	6	60
RAD50	1	104	24	7	87
RAD50	2	14	8	2	46
RAD52	1	169	29	5	91
RAD52	2	88	19	13	75
RAD59	1	74	17	1	76
RAD59	2	30	6	0	67
RAD6	1	7	3	0	36
RAD6	2	17	2	2	54
RRM3	1	27	7	0	48
RRM3	2	19	4	1	61
RSC1	1	38	10	3	53
RSC1	2	40	12	5	70
RTT107	1	96	19	3	48
RTT107	2	75	20	3	104
SAE2	1	122	13	1	68
SAE2	2	23	8	0	66
SGO1	1	107	27	1	11
SGO1	2	54	10	2	20
SGS1	1	42	5	0	32
SGS1	2	24	8	4	52
SLX1	1	118	26	2	11
SLX1	2	7	1	0	35
SRS2	1	31	12	8	31
SRS2	2	65	22	3	7
SWR1	1	77	9	1	21
SWR1	2	85	14	1	30
TOF1	1	73	14	5	61
TOF1	2	76	14	3	52
TRP5	1	100	9	0	99
TRP5	2	128	21	3	42
YKU70	1	51	20	10	38
YKU70	2	31	5	1	66
YOL086-A	1	22	14	4	68
YOL086-A	2	18	4	0	41

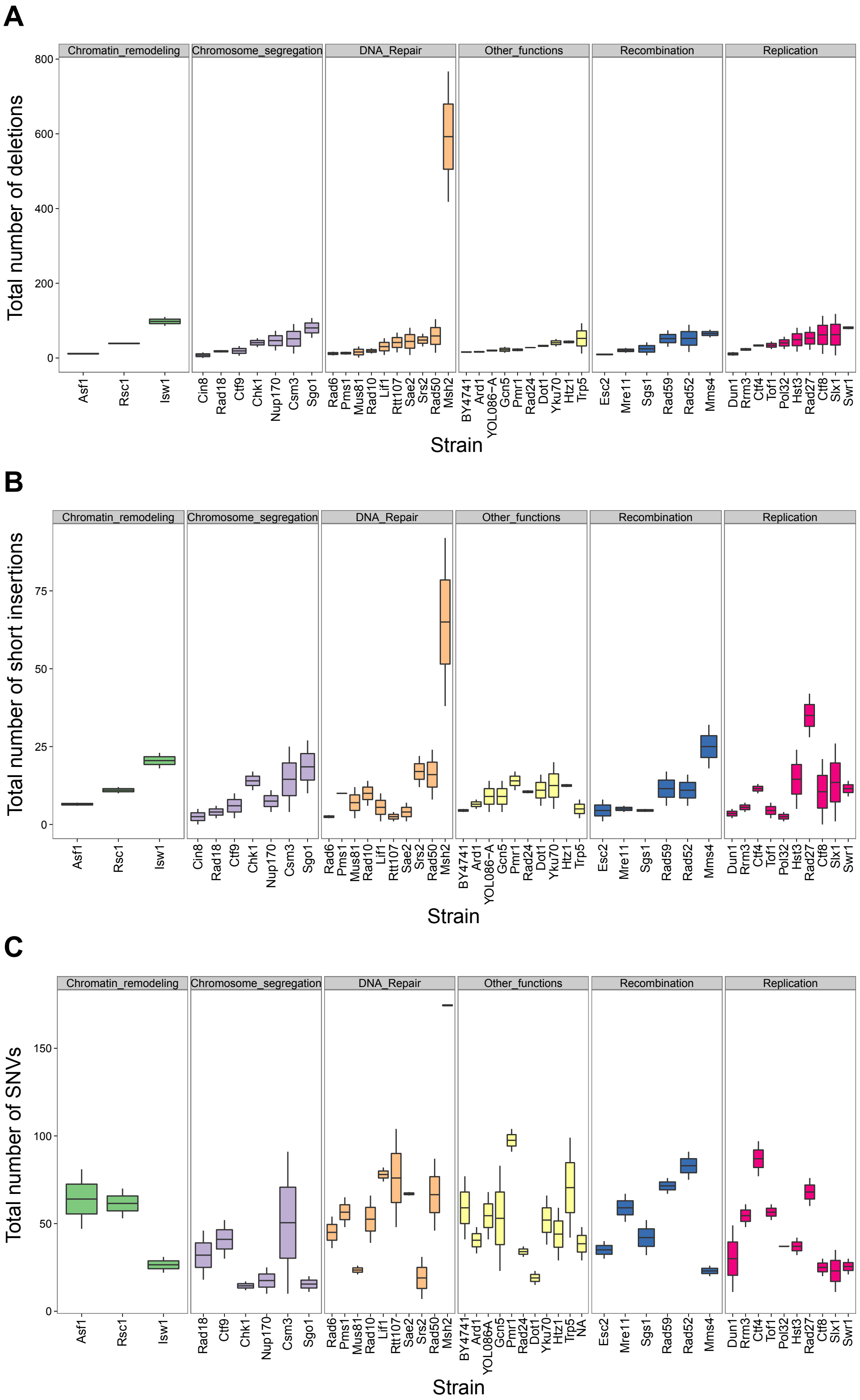

Figure B.1: Number of **A** | deletions, **B** | short insertions and **C** | SNVs in each yeast deletion mutant classified into broad functional categories. For comparison, the strains are ordered by the number of deletions they acquired, and the same order is preserved in the other plots. It is noticeable that strains with the highest number of deletions are not necessarily the ones with the highest number of short insertions or SNVs. The total number of *de novo* mutations occurring in less than 10% of the strains is shown.

Table B.3: Primers used in this study.

Name	Sequence 5'-3'	Function	Source
M13FW	TGTAAAACGACGGCCAGT	Amplify construct from plasmid	Universal plasmid primer
M13RV	CAGGAAACAGCTATGACC	Amplify construct from plasmid	Universal plasmid primer
ConstructSeqPrimerFW	CGTTGTTCCAGAGCTGATGA	Test for rearrangements in construct	This study
ConstructSeqPrimerRV	ATAGGTCAGGCTCTCGCTGA	Test for rearrangements in construct	This study
LeftOutConstruct	ACGGATATTCAGAACCCAATGA	Confirmation of construct transformation	This study
LeftIntConstruct	GTAACTGGAAGGAAGGCCGT	Confirmation of construct transformation	This study
RightIntConstruct	CATCCGGAGCTTGCAGGATC	Confirmation of construct transformation	This study
RightOutConstruct	TTCACTCCACCCCGCTTTAC	Confirmation of construct transformation	This study
U1	GATGTCCACGAGGTCTCT	Amplify the uptag	[Giaever et al., 2002]
D1	CGGTGTCGGTCTCGTAG	Amplify the downtag	[Giaever et al., 2002]
KanB	CTGCAGCGAGGAGCCGTAAT	Amplify the uptag	[Giaever et al., 2002]
KanC	TGATTTTGATGACGAGCGTAAT	Amplify the downtag	[Giaever et al., 2002]
*MAT*a	AGTCACATCAAGATCGTTTATGG	Mating type assessment	[Huxley et al., 1990]
MATα	GCACGGAATATGGGACTACTTCG	Mating type assessment	[Huxley et al., 1990]
*MAT*a/α	ACTCCACTTCAAGTAAGAGTTTG	Mating type assessment	[Huxley et al., 1990]
GFPinternalRV	CGTTCTTCTGCTTGTCGGCCATGA	GFP tagging confirmation	This study
GFPinternalFW	TATGCTGTTATCGATTTGGGAT	GFP tagging confirmation	This study

Table B.4: Primers used for qPCR validations of aneuploidies found in the deletion strains and their efficiencies.

Name	Sequence 5'-3' Forward	Sequence 5'-3' Reverse	Amplicon size	Slope[1]	Source
chrI Left arm	ACAGCTTCTAAACGTTCCGTGTGC	GCGGTGTGTGGATGATGGTTTCAT	113	-3.477	[Pavelka et al., 2010]
chrI Right arm	GCACTTGATCCATGTAGCCATACTCG	TTCGGGTGACCCTTATGGCATTCT	90	-3.971	[Pavelka et al., 2010]
chrII Left arm	TGGCATAGAACTTGGCCTTC	GAGCCAGGAAGAAATACACTGC	94	-3.624	This study
chrII Right arm	GCCAATGGATTAAGGGTGAC	ACAGAACAATAAAGGACGTTGC	131	-3.508	This study
chrVI Left arm	GCGTGCCGGGTAATAAATC	AGCTTCCCTAGAGGTAGAATGC	110	-3.289	This study
chrVI Right arm	GCTTCAATGGAAGTTGAAGTGC	CACCCAATCTTCACAGATCTTC	92	-3.479	This study
SPT15 reference	TGCGCTACATGCCCGTAA	CGCATGATGACAGCAGCAA	60	-3.517	[Pavelka et al., 2010]

[1] The slope was estimated via a regression line of the standard concentrations versus the C_t values and was used to calculate primer efficiencies.

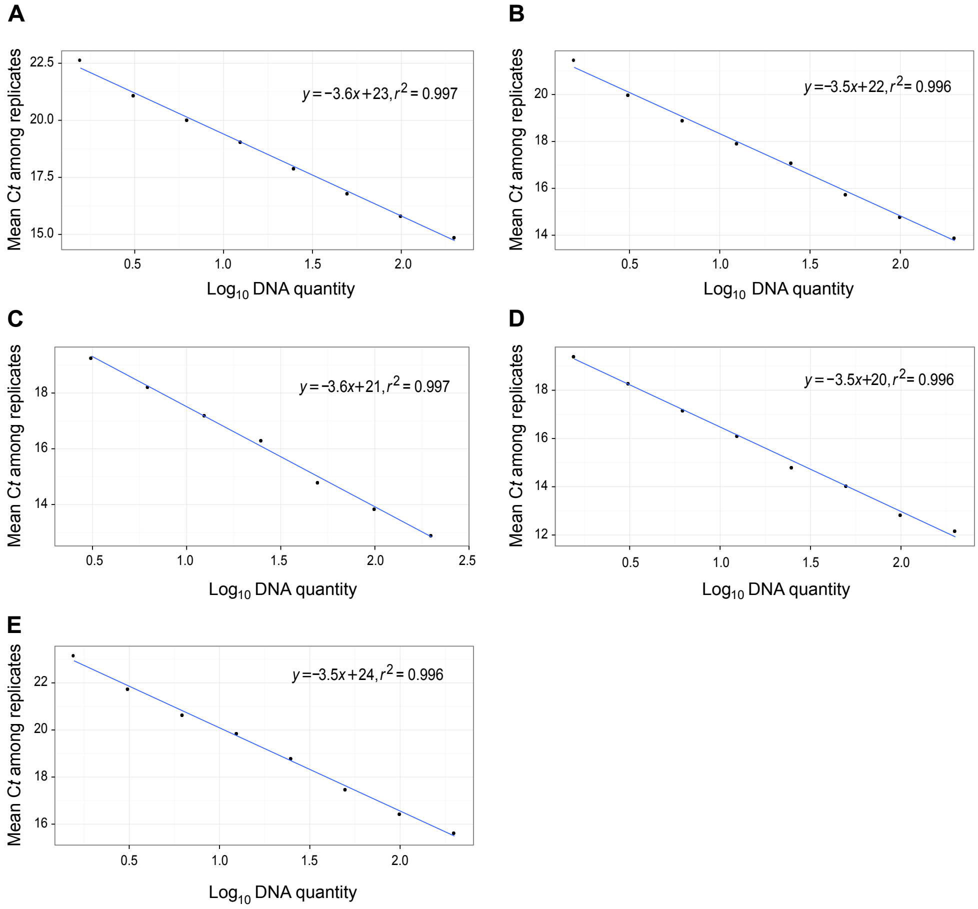

Figure B.2: Primer efficiencies for qPCR validation of aneuploidies. **A**, **B** | Standard curves for primer efficiency estimation for regions in the left and right arms of chrII respectively. **C**, **D** | Standard curves for primer efficiency estimation for regions in the left and right arms of chrVI respectively. **E** | Standard curves for primer efficiency estimation for SPT15, the internal reference locus used for calibration.

Table B.5: List of barcodes for multiplexed paired-end sequencing used in this study (WW barcodes were previously described in [Wilkening et al., 2013]).

Name	codePE2	codePE1
WWmp1	AGCGCT	AGCGCT
WWmp2	AGTCTT	AAGACT
WWmp3	TCGCTT	AAGCGA
WWmp4	TGACAT	ATGTCA
WWmp5	CACTGT	ACAGTG
WWmp6	GAGAGT	ACTCTC
WWmp7	GATGCT	AGCATC
WWmp8	CAAGTT	AACTTG
WWmp9	GGAACT	AGTTCC
WWmp10	CCGTAT	ATACGG
WWmp11	ATAGAT	ATCTAT
WWmp12	GCTCAT	ATGAGC
WWmp13	ATCGTG	CACGAT
WWmp14	TGAGTG	CACTCA
WWmp15	CGCCTG	CAGGCG
WWmp16	GCGTGG	CCACGC
WWmp17	TTGCGG	CCGCAA
WWmp18	CTAAGG	CCTTAG
WWmp19	ATTCCG	CGGAAT
WWmp20	CGTACG	CGTACG
WWmp21	AGCTAG	CTAGCT
WWmp22	GTATAG	CTATAC
WWmp23	TCTGAG	CTCAGA
WWmp24	TACAAG	CTTGTA
WWmp25	TCCGTC	GACGGA
WWmp26	CCACTC	GAGTGG
WWmp27	TATATC	GATATA
WWmp28	CTGATC	GATCAG
WWmp29	CCTTGC	GCAAGG
WWmp30	AAGTGC	GCACTT
WWmp31	TGGAGC	GCTCCA
WWmp32	TGTGCC	GGCACA
WWmp33	CAGGCC	GGCCTG
WWmp34	GCTACC	GGTAGC
WWmp35	CTCTAC	GTAGAG
WWmp36	GGCCAC	GTGGCC
WWmp37	CGAAAC	GTTTCG
WWmp38	GCTGTA	TACAGC
WWmp39	ATTATA	TATAAT
WWmp40	GAATGA	TCATTC
WWmp41	TCGGGA	TCCCGA
WWmp42	TGCCGA	TCGGCA
WWmp43	CATTCA	TGAATG
WWmp44	ATGGCA	TGCCAT
WWmp45	CCAGCA	TGCTGG
WWmp46	GCCTAA	TTAGGC
WWmp47	TTCGAA	TTCGAA
WWmp48	GGAGAA	TTCTCC
mp13	GGGGTT	AACCCC
mp19	GTTTGT	ACAAAC
mp21	AAAATG	CATTTT
mp5	ACATCG	CGATGT
mp49	AAACCT	AGGTTT
mp50	CTTTTG	CAAAAG
mp51	GGACGG	CCGTCC
mp52	GGTTTC	GAAACC
mp53	TTTCAC	GTGAAA

B.2 Supplementary figures and tables for Chapter 3

Table B.6: Drug concentrations used to induce mutations in the transformed pooled deletion collection.

Drug	Working Concentration
Hydroxyurea (HU)	50µM
Doxorubicin (Doxo)	25µM
Camptothecin (Campt)	10µM
Methyl methanesulfonate (MMS)	0.10%

Table B.7: Total number of reads sequenced for the original pool of homozygous yeast deletion mutant strains (before the transformation of any construct), and the pools after transformation of the DelRep, DelNoRep and Hyg+ constructs. The percentage of strains in each transformed pool compared to the original pool was calculated taken into account the strains that are shared between replicates.

Pool	Replicate	Total number of reads	Total number of strains identified	Percentage of strains in the original pool
Original pool	1	871666	5143	
	2	626098	4562	
DelRep construct	1	643728	3713	76.76%
	2	781458	3959	
	3	792034	3503	
DelNoRep construct	1	901580	3759	74.54%
	2	906772	3951	
	3	861048	3141	
Hyg+ construct	1	763190	3802	76.00%
	2	972044	3826	
	3	891776	3436	

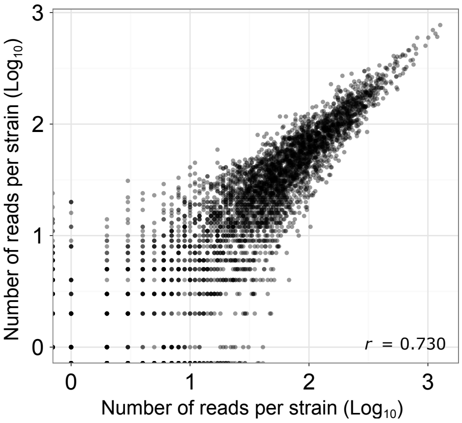

Figure B.3: Correlation between the number of reads per strain detected with the uptags and downtags in the original pool of yeast KO strains (r: Pearson correlation coefficient).

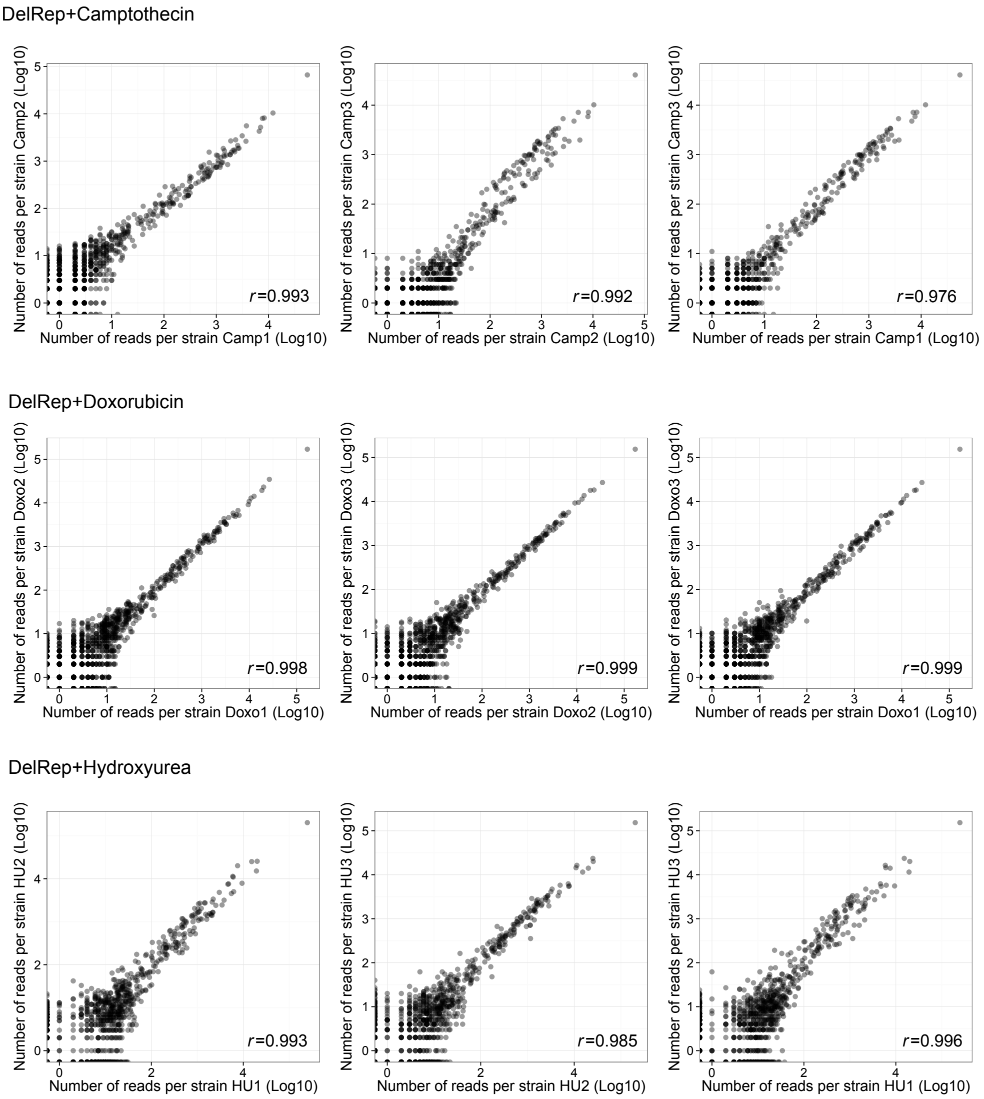

Figure B.4: Correlation between replicate experiments of the pools transformed with the DelRep construct, after the treatment with drugs (r: Pearson correlation coefficient).

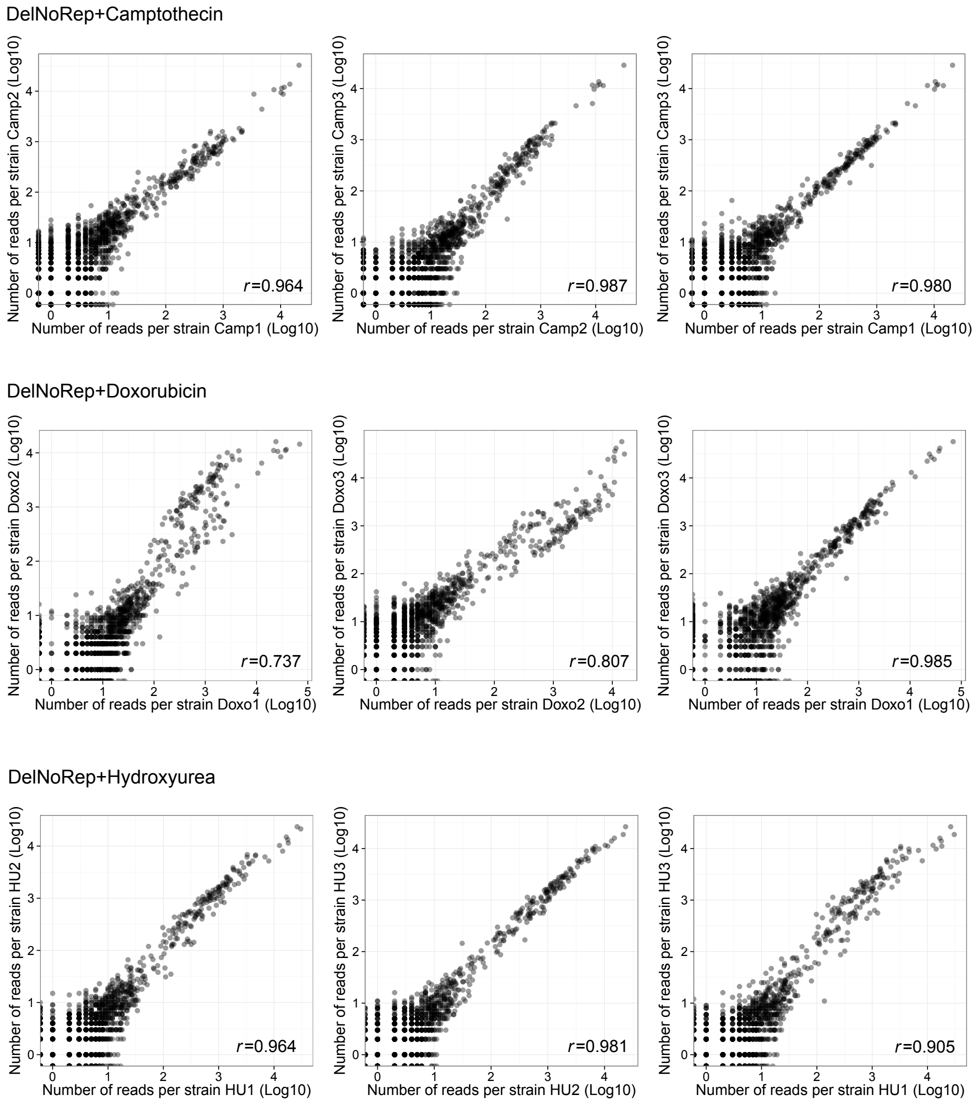

Figure B.5: Correlation between replicate experiments of the pools transformed with the DelNoRep construct after the treatment with drugs (*r*: Pearson correlation coefficient).

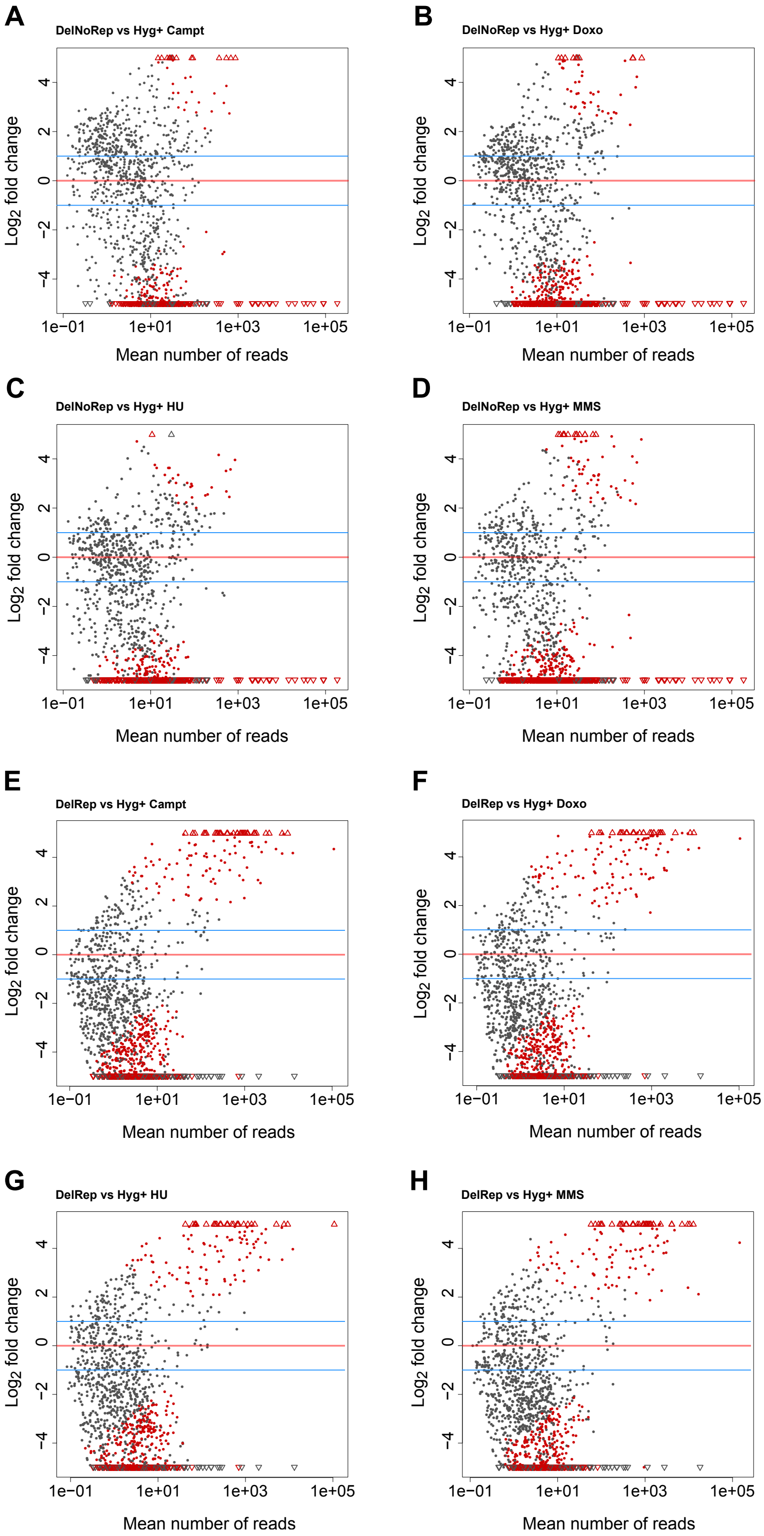

Figure B.6: Strains enriched after different treatments in the pools transformed with the DelRep and DelNoRep constructs. The blue lines delimit the 2-fold change threshold. **A** | DelNoRep+camptothecin (Campt). **B** | DelNoRep+doxorubicin (Doxo). **C** | DelNoRep+hydroxyurea (HU). **D** | DelNoRep+methyl methanesulfonate (MMS). **E** | DelRep+camptothecin. **F** | DelRep+doxorubicin. **G** | DelRep+hydroxyurea. **H** | DelRep+methyl methanesulfonate.

Table B.8: Significantly enriched strains in the pools transformed with the DelNoRep construct after different drug treatments. Highlighted are the strains selected for experimental validations (Gene Name indicates the particular genes that are deleted in these strains. padj: Benjamini-Hochberg adjusted *p*-value).

Strain	Fold Change	padj	Gene Name	Description
Campt				
YDR252W	6.92	6.14E-05	*BTT1*	Heterotrimeric nascent polypeptide-associated complex beta3 subunit
YKL047W	6.66	0.046	*ANR2*	Putative protein of unknown function
YDR316W	6.22	0.036	*OMS1*	Protein integral to the mitochondrial membrane
YFL027C	6.14	0.055	*GYP8*	GTPase-activating protein for yeast Rab family members
YJR128W	4.22	0.092	-	Dubious open reading frame
YOR374W	3.56	0.091	*ALD4*	Mitochondrial aldehyde dehydrogenase
Doxo				
YKL047W	6.78	0.027	*ANR2*	Putative protein of unknown function
YDR252W	4.88	0.001	*BTT1*	Heterotrimeric nascent polypeptide-associated complex beta3 subunit
YBL002W	3.80	0.005	*HTB2*	Histone H2B
YOR374W	3.50	0.077	*ALD4*	Mitochondrial aldehyde dehydrogenase
HU				
YDR252W	4.17	0.021	*BTT1*	Heterotrimeric nascent polypeptide-associated complex beta3 subunit
MMS				
YJR021C	5.94	0.056	***REC107***	Protein involved in early stages of meiotic recombination
YER005W	5.68	0.086	*YND1*	Apyrase with wide substrate specificity
YLR285W	5.36	0.005	*NNT1*	S-adenosylmethionine-dependent methyltransferase
YDR336W	5.34	0.017	-	Putative protein of unknown function
YKL205W	5.03	0.025	*LOS1*	Nuclear pore protein
YBR116C	4.73	0.030	-	Dubious open reading frame
YDL121C	4.70	0.082	-	Putative protein of unknown function
YDL233W	3.91	0.081	*MFG1*	Regulator of filamentous growth
YHR155W	3.35	0.093	*YSP1*	Mitochondrial protein

Table B.9: Top ten significantly enriched strains in the pools transformed with the DelRep construct after different drug treatments. Highlighted are the strains selected for experimental validations (Gene Name indicates the particular genes that are deleted in these strains. padj: Benjamini-Hochberg adjusted *P*-value. 11 strains are shown for HU to include *REC114*).

Strain	Fold Change	padj	Gene Name	Description
Campt				
YFL003C	7.86	0.001	***MSH4***	Protein involved in meiotic recombination; required for normal levels of crossing over, colocalizes with Zip2p to discrete foci on meiotic chromosomes
YGL226C-A	7.28	0.004	*OST5*	Zeta subunit of the oligosaccharyltransferase complex of the ER lumen
YGL248W	7.05	0.006	*PDE1*	Low-affinity cyclic AMP phosphodiesterase
YGR254W	6.64	0.018	***ENO1***	Enolase I, a phosphopyruvate hydratase
YDR370C	6.51	0.013	*DXO1*	mRNA 5'-end-capping quality-control protein
YER046W	6.44	0.020	***SPO73***	Meiosis-specific protein of unknown function
YGL249W	6.41	0.028	***ZIP2***	Meiosis-specific protein; involved in normal synaptonemal complex formation and pairing between homologous chromosomes during meiosis
YBR169C	6.30	0.025	*SSE2*	Member of the heat shock protein 70 (HSP70) family
YGR100W	6.21	0.049	*MDR1*	Cytoplasmic GTPase-activating protein
YLR131C	6.20	0.062	***ACE2***	Transcription factor required for septum destruction after cytokinesis
Doxo				
YBR217W	8.62	1.27E-04	*ATG12*	Ubiquitin-like modifier involved in autophagy and the Cvt pathway
YER046W	7.67	0.002	***SPO73***	Meiosis-specific protein of unknown function
YGR015C	7.54	0.002	-	Putative protein of unknown function
YBR169C	7.39	0.005	*SSE2*	Member of the heat shock protein 70 (HSP70) family
YDL093W	7.15	0.005	*PMT5*	Protein O-mannosyltransferase
YDR312W	7.12	0.008	*SSF2*	Protein required for ribosomal large subunit maturation
YDR503C	7.06	0.009	*LPP1*	Lipid phosphate phosphatase
YGR100W	7.02	0.009	*MDR1*	Cytoplasmic GTPase-activating protein
YOL158C	7.00	0.007	*ENB1*	Endosomal ferric enterobactin transporter
YBL055C	6.90	0.011	-	3'-5' exonuclease and endonuclease with a possible role in apoptosis
HU				
YDR314C	8.60	2.70E-05	***RAD34***	Protein involved in nucleotide excision repair (NER)
YJR082C	8.30	4.49E-05	*EAF6*	Subunit of the NuA4 acetyltransferase complex
YJL171C	7.80	0.001	-	GPI-anchored cell wall protein of unknown function
YGL249W	7.48	0.003	***ZIP2***	Meiosis-specific protein; involved in normal synaptonemal complex formation and pairing between homologous chromosomes during meiosis

continued on next page

Table B.9: continued from previous page

Strain	Fold Change	padj	Gene Name	Description
YBL019W	7.19	0.003	***APN2***	Class II abasic (AP) endonuclease involved in repair of DNA damage
YBR233W	7.17	0.004	*PBP2*	RNA binding protein; involved in the regulation of telomere position effect and telomere length
YDL110C	7.04	0.004	*TMA17*	ATPase dedicated chaperone that adapts proteasome assembly to stress
YDR078C	6.98	0.006	***SHU2***	Component of the Shu complex, which promotes error-free DNA repair
YDR421W	6.97	0.014	*ARO80*	Zinc finger transcriptional activator of the Zn2Cys6 family
YGR238C	6.96	0.010	*KEL2*	Protein that negatively regulates mitotic exit
YMR133W	6.94	0.013	***REC114***	Protein involved in early stages of meiotic recombination
MMS				
YDR497C	14.13	0.001	*ITR1*	Myo-inositol transporter; member of the sugar transporter superfamily
YLR047C	17.61	0.002	*FRE8*	Protein with sequence similarity to iron/copper reductase
YLR131C	4.46	0.015	***ACE2***	Transcription factor required for septum destruction after cytokinesis
YJL083W	3.30	0.028	*TAX4*	EH domain-containing protein
YKL061W	3.90	0.037	*BLI1*	Subunit of the BLOC-1 complex involved in endosomal maturation
YBR169C	2.62	0.025	*SSE2*	Member of the heat shock protein 70 (HSP70) family
YBL052C	4.13	0.029	*SAS3*	Histone acetyltransferase catalytic subunit of NuA3 complex
YGL257C	5.74	0.025	*MNT2*	Mannosyltransferase
YLR246W	3.21	0.051	*ERF2*	Subunit of a palmitoyltransferase
YLR456W	7.83	0.025	-	Putative pyridoxal 5'-phosphate synthase

Table B.10: Strains with the highest fold enrichments detected under both stress and no stress growth conditions. In bold, genes that are related to DNA repair and genome maintenance pathways. Underlined are genes uniquely detected in each drug.

Strain	Gene Name	Fold enrich. Drug	Fold enrich. YPAD	Description
Campt				
YBR222C	*PCS60*	11.4	5.5	Oxalyl-CoA synthetase
YBR073W	***RDH54***	11.0	5.2	DNA-dependent ATPase; DNA recombination/repair translocase, supercoils DNA and promotes DNA strand opening
YGR121C	*MEP1*	10.0	7.8	Ammonium permease
YIR013C	*GAT4*	9.9	6.3	Protein containing GATA family zinc finger motifs
YMR186W	*HSC82*	9.8	8.8	Cytoplasmic chaperone of the Hsp90 famil
YKL174C	*TPO5*	9.7	7.0	Protein involved in excretion of putrescine and spermidine
YBR066C	*NRG2*	9.6	5.2	Transcriptional repressor
YGL087C	***MMS2***	9.6	7.5	Ubiquitin-conjugating enzyme variant; involved in error-free postreplication repair
YOL136C	*PFK27*	9.5	9.2	6-PhosphoFructo-2-Kinase
YIL038C	*NOT3*	9.4	6.0	Subunit of CCR4-NOT global transcriptional regulator
YDR034C	*LYS14*	9.3	7.1	Transcriptional activator involved in regulating lysine biosynthesis
YLR247C	***IRC20***	9.3	8.1	E3 ubiquitin ligase and putative helicase; involved in synthesis-dependent strand annealing-mediated homologous recombination
YMR038C	*CCS1*	9.3	8.6	Copper chaperone for superoxide dismutase Sod1p; protein abundance increases in response to DNA replication stress
YAR042W	*SWH1*	9.2	3.6	Protein similar to mammalian oxysterol-binding protein
YDL230W	*PTP1*	9.2	7.1	Phosphotyrosine-specific protein phosphatase
YML074C	*FPR3*	9.2	8.6	Nucleolar peptidyl-prolyl cis-trans isomerase (PPIase)
YBL002W	*HTB2*	9.1	3.7	Histone H2B; core histone protein required for chromatin assembly and chromosome function
YDL182W	*LYS20*	9.1	7.0	Homocitrate synthase isozyme
YIL120W	*QDR1*	9.0	6.2	Multidrug transporter of the major facilitator superfamily
YLR445W	***GMC2***	8.9	8.4	Protein involved in meiotic crossing over
YJR001W	*AVT1*	8.8	7.8	Vacuolar transporter
YDR465C	*RMT2*	8.8	7.3	Arginine N5 methyltransferase; relative distribution to the nucleus increases upon DNA replication stress
YGL032C	*AGA2*	8.7	7.3	Adhesion subunit of a-agglutinin of a-cells
YHR155W	*YSP1*	8.7	5.8	Mitochondrial protein
YPL103C	*FMP30*	8.7	9.7	Protein with a role in maintaining mitochondrial morphology
YGL255W	*ZRT1*	8.7	7.6	High-affinity zinc transporter of the plasma membrane
YBR057C	***MUM2***	8.5	5.0	Protein essential for meiotic DNA replication and sporulation
YNL253W	*TEX1*	8.4	9.2	Protein involved in mRNA export
YMR044W	***IOC4***	8.4	8.7	Member of a complex (Isw1b) with Isw1p and Ioc2p
Doxo				
YOL136C	*PFK27*	14.1	9.2	6-phosphofructo-2-kinase; catalyzes synthesis of fructose-2,6-bisphosphate
YBR222C	*PCS60*	13.8	5.5	Oxalyl-CoA synthetase
YBR057C	***MUM2***	13.1	5.0	Protein essential for meiotic DNA replication and sporulation

continued on next page

Table B.10: continued from previous page

Strain	Gene Name	Fold enrich. Drug	Fold enrich. YPAD	Description
YDL182W	*LYS20*	12.8	6.6	Homocitrate synthase isozyme; catalyzes the condensation of acetyl-CoA and alpha-ketoglutarate to form homocitrate
YMR038C	*CCS1*	12.7	8.6	Copper chaperone for superoxide dismutase Sod1p; protein abundance increases in response to DNA replication stress
YDR465C	*RMT2*	12.2	7.3	Arginine N5 methyltransferase; relative distribution to the nucleus increases upon DNA replication stress
YBR216C	*YBP1*	12.2	5.4	Protein involved in cellular response to oxidative stress
YLR445W	***GMC2***	12.1	8.3	Protein involved in meiotic crossing over; component of the Synaptonemal Complex (SC)
YEL012W	*UBC8*	11.9	5.9	Ubiquitin-conjugating enzyme that regulates gluconeogenesis
YDR345C	*HXT3*	11.8	7.2	Low affinity glucose transporter of the major facilitator superfamily
YOR371C	*GPB1*	11.5	9.6	Multistep regulator of cAMP-PKA signaling
YMR186W	*HSC82*	11.4	8.7	Cytoplasmic chaperone of the Hsp90 family
YBR066C	*NRG2*	11.4	5.1	Transcriptional repressor
YBR258C	*SHG1*	11.3	5.5	Subunit of the COMPASS (Set1C) complex; COMPASS methylates histone H3
YGR121C	*MEP1*	11.2	7.5	Ammonium permease
YBL002W	*HTB2*	11.1	3.6	Histone H2B; core histone protein required for chromatin assembly and chromosome function
YAR042W	*SWH1*	11.0	11.1	Protein similar to mammalian oxysterol-binding protein
YGL087C	***MMS2***	10.8	7.3	Ubiquitin-conjugating enzyme variant; involved in error-free postreplication repair
YGR260W	*TNA1*	10.8	7.6	High affinity nicotinic acid plasma membrane permease
YBR073W	***RDH54***	10.8	5.2	DNA-dependent ATPase; DNA recombination/repair translocase, supercoils DNA and promotes DNA strand opening
YPL103C	*FMP30*	10.7	9.7	Protein with a role in maintaining mitochondrial morphology
YER037W	*PHM8*	10.5	6.0	Lysophosphatidic acid (LPA) phosphatase, nucleotidase
YKL137W	*CMC1*	10.5	6.6	Copper-binding protein of the mitochondrial intermembrane space
YDR357C	*CNL1*	10.5	7.3	Subunit of the BLOC-1 complex involved in endosomal maturation
YGR068C	*ART5*	10.4	7.3	Protein proposed to regulate endocytosis of plasma membrane proteins
YDR034C	*LYS14*	10.4	7.2	Transcriptional activator involved in regulating lysine biosynthesis
YCL026C-A	*FRM2*	10.4	5.6	Type II nitroreductase, using NADH as reductant; involved in the oxidative stress response
YMR101C	*SRT1*	10.4	8.6	Cis-prenyltransferase
YDL230W	*PTP1*	10.3	7.0	Phosphotyrosine-specific protein phosphatase
HU				
YMR038C	*CCS1*	13.5	8.6	Copper chaperone for superoxide dismutase Sod1p; protein abundance increases in response to DNA replication stress
YOL136C	*PFK27*	13.0	9.2	6-PhosphoFructo-2-Kinase
YMR186W	*HSC82*	13.0	8.8	Cytoplasmic chaperone of the Hsp90 family
YNL253W	*TEX1*	12.9	9.2	Protein involved in mRNA export
YJR004C	*SAG1*	12.2	7.8	Alpha-agglutinin of alpha-cells
YBR057C	***MUM2***	12.1	5.0	Protein essential for meiotic DNA replication and sporulation
YBR222C	*PCS60*	11.7	5.5	Oxalyl-CoA synthetase

continued on next page

Table B.10: continued from previous page

Strain	Gene Name	Fold enrich.	Fold enrich.	Description
		Drug	YPAD	
YIR013C	*GAT4*	11.6	6.5	Protein containing GATA family zinc finger motifs
YDR163W	*CWC15*	11.6	7.3	Non-essential protein involved in pre-mRNA splicing
YBR066C	*NRG2*	11.5	5.1	Transcriptional repressor
YDL188C	*PPH22*	11.5	7.3	Catalytic subunit of protein phosphatase 2A
YML100W	*TSL1*	11.4	8.6	Large subunit of trehalose 6-phosphate synthase/phosphatase complex; mutant has aneuploidy tolerance; protein abundance increases in response to DNA replication stress
YGR121C	*MEP1*	11.3	7.5	Ammonium permease
YDR345C	*HXT3*	11.0	7.3	Low affinity glucose transporter of the major facilitator superfamily
YAR042W	*SWH1*	10.9	3.6	Protein similar to mammalian oxysterol-binding protein
YLR247C	***IRC20***	10.8	8.1	E3 ubiquitin ligase and putative helicase; involved in synthesis-dependent strand annealing-mediated homologous recombination
YPL103C	*FMP30*	10.8	9.7	Protein with a role in maintaining mitochondrial morphology
YDL230W	*PTP1*	10.8	7.3	Phosphotyrosine-specific protein phosphatase
YOR371C	*GPB1*	10.7	9.6	Multistep regulator of cAMP-PKA signaling
YCL026C-A	*FRM2*	10.7	5.6	Type II nitroreductase, using NADH as reductant; involved in the oxidative stress response
YGR260W	*TNA1*	10.6	7.6	High affinity nicotinic acid plasma membrane permease
YGL087C	***MMS2***	10.6	7.4	Ubiquitin-conjugating enzyme variant; involved in error-free postreplication repair
YIL038C	*NOT3*	10.5	6.2	Subunit of CCR4-NOT global transcriptional regulator
YPR129W	*SCD6*	10.5	9.9	Repressor of translation initiation
YDR357C	*CNL1*	10.5	7.4	Subunit of the BLOC-1 complex involved in endosomal maturation
YPL070W	*MUK1*	10.4	9.7	Guanine nucleotide exchange factor (GEF)
YGL234W	*ADE5,7*	10.4	7.5	Enzyme of the de novo purine nucleotide biosynthetic pathway
YNL082W	***PMS1***	10.4	8.8	ATP-binding protein required for mismatch repair; required for both mitosis and meiosis
YKL137W	*CMC1*	10.1	7.2	Copper-binding protein of the mitochondrial intermembrane space
YIL120W	*QDR1*	10.1	6.4	Multidrug transporter of the major facilitator superfamily
YHR155W	*YSP1*	10.1	6.0	Mitochondrial protein
MMS				
YBR066C	*NRG2*	12.9	5.4	Transcriptional repressor
YMR186W	*HSC82*	12.2	9.0	Cytoplasmic chaperone of the Hsp90 family
YAR042W	*SWH1*	12.1	11.1	Protein similar to mammalian oxysterol-binding protein
YIL038C	*NOT3*	11.9	7.0	Subunit of CCR4-NOT global transcriptional regulator
YKL137W	*CMC1*	11.3	7.6	Copper-binding protein of the mitochondrial intermembrane space
YOR371C	*GPB1*	11.3	9.7	Multistep regulator of cAMP-PKA signaling
YLR028C	*ADE16*	11.0	8.3	Enzyme of de novo purine biosynthesis
YCL026C-A	*FRM2*	11.0	5.9	Type II nitroreductase, using NADH as reductant; involved in the oxidative stress response
YBR258C	*SHG1*	10.8	5.8	Subunit of the COMPASS (Set1C) complex
YBR222C	*PCS60*	10.8	5.6	Oxalyl-CoA synthetase
YNL082W	***PMS1***	10.7	9.1	ATP-binding protein required for mismatch repair; required for both mitosis and meiosis
YLR247C	***IRC20***	10.6	8.3	E3 ubiquitin ligase and putative helicase; involved in synthesis-dependent strand annealing-mediated homologous recombination

continued on next page

Table B.10: continued from previous page

Strain	Gene Name	Fold enrich. Drug	Fold enrich. YPAD	Description
YER037W	*PHM8*	10.6	6.5	Lysophosphatidic acid (LPA) phosphatase, nucleotidase
YDR465C	*RMT2*	10.4	7.9	Arginine N5 methyltransferase; relative distribution to the nucleus increases upon DNA replication stress
YBR057C	***MUM2***	10.2	5.1	Protein essential for meiotic DNA replication and sporulation
YDR357C	*CNL1*	10.2	7.8	Subunit of the BLOC-1 complex involved in endosomal maturation
YIR013C	*GAT4*	10.2	7.3	Protein containing GATA family zinc finger motifs
YOL136C	*PFK27*	10.1	9.2	6-PhosphoFructo-2-Kinase
YKL174C	*TPO5*	10.1	7.6	Protein involved in excretion of putrescine and spermidine
YML100W	*TSL1*	10.1	8.7	Large subunit of trehalose 6-phosphate synthase/phosphatase complex; mutant has aneuploidy tolerance; protein abundance increases in response to DNA replication stress
YDR345C	*HXT3*	9.9	7.7	Low affinity glucose transporter of the major facilitator superfamily
YHR044C	***DOG1***	9.9	6.6	2-deoxyglucose-6-phosphate phosphatase
YMR101C	*SRT1*	9.8	8.8	Cis-prenyltransferase
YMR038C	*CCS1*	9.6	8.7	Copper chaperone for superoxide dismutase Sod1p; protein abundance increases in response to DNA replication stress
YDR163W	*CWC15*	9.3	7.6	Non-essential protein involved in pre-mRNA splicing

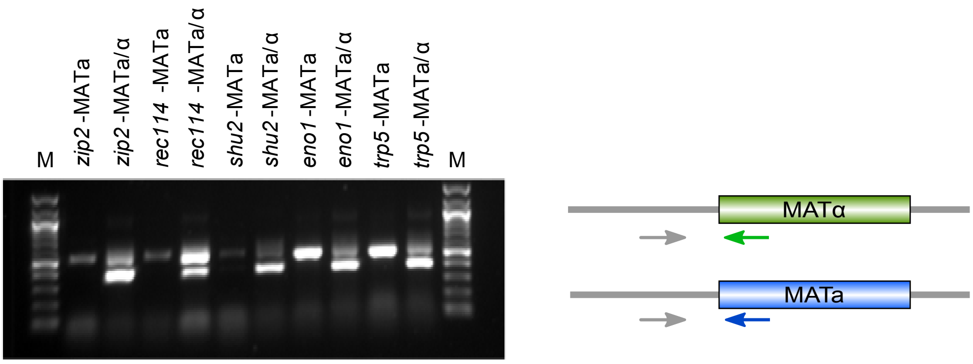

Figure B.7: PCR validations of mating type for newly generated KO strains for the validation experiments (M: 100bp marker). The PCRs were done using three primers in the same reaction: two primers that anneal to a specific region of *MAT*α and *MAT*a respectively and one that anneals in a common region of both mating types and it is directed towards the *MAT* locus [Huxley et al., 1990].

Table B.11: Number of hygromycin resistant colonies recovered from each KO mutant under different growth conditions. These KO strains, in their haploid and diploid stages were created newly for this experiment to have uniform backgrounds. YPAD refers to growth in rich medium without any drug stress. WT: wild type

Treatment	*msh4*		*eno1*		*zip2*		*apn2*		*rad52*		*trp5*		WT	
	*MAT*a	*MAT*a/α	*MAT*a	*MAT*a/α	*MAT*a	*MAT*a/α	*MAT*a	*MAT*a/α	*MAT*a	*MAT*a/α	*MAT*a	*MAT*a/α	*MAT*a	*MAT*a/α
YPAD	68	120	19	84	17	47	20	28	5	48	2	1	6	12
HU	248	678	54	196	48	176	64	116	32	140	1	4	10	21
Campt	280	690	40	200	23	204	31	96	25	88	2	9	9	29
MMS	140	610	21	140	38	100	29	108	39	60	7	3	14	9
Doxo	91	750	20	144	26	112	44	104	27	108	1	7	8	13

Appendix C

List of publications

Moncunill V., Gonzalez S., Beà S., Andrieux L. O., Salaverria I., Royo C., Martinez L., Puiggròs M., Segura-Wang M., Stütz A. M. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature Biotechnology*, 32(11), October 2014.

Mardin, B., Drainas, A., Waszak, S., Weischenfeldt, J., Isokane, M., Stütz, A., Buccitelli, C., Segura-Wang, M., Northcott, P., Pfister, S. et al. A novel cell-based model system links chromothripsis with hyperploidy. (in review in *Nature Methods*).

George, J., Peifer, M., Cun, Y., Leenders, F., Müller, C., Dahmen, I., Schaub, P., Bosco, G., Pinther, B., Lu, X., Seidel, D., Fernandez-Cuesta, L., Sage, J., Lim, J., Jahchan, N., Park, K., Yang, D., Vaka, D., Torres, A., Karnezis, A., Korbel, J., Segura-Wang, M., Menon, R. et al. Comprehensive genomic characterization of small cell lung cancer. (in review in *Nature*).

Bibliography

Abyzov A., Urban A. E., Snyder M., and Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21(6):974–84, June 2011. Cited on page 17.

Adzhubei I. A., Schmidt S., Peshkin L., Ramensky V. E., Gerasimova A., Bork P., Kondrashov A. S., and Sunyaev S. R. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–9, April 2010. Cited on page 10.

Albers C. A., Lunter G., MacArthur D. G., McVean G., Ouwehand W. H., and Durbin R. Dindel: accurate indel calls from short-read data. *Genome research*, 21(6):961–73, June 2011. Cited on page 87.

Albert I., Mavrich T. N., Tomsho L. P., Qi J., Zanton S. J., Schuster S. C., and Pugh B. F. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature*, 446(7135):572–6, March 2007. Cited on pages 36 and 89.

Alexandrov L. B., Nik-Zainal S., Wedge D. C., Aparicio S. A. J. R., Behjati S., Biankin A. V., Bignell G. R., Bolli N., Borg A., Borresen-Dale A. L. et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–21, August 2013. Cited on pages 4 and 47.

Alkan C., Coe B. P., and Eichler E. E. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–76, May 2011. Cited on pages 13 and 16.

Alonso J. M., Stepanova A. N., Leisse T. J., Kim C. J., Chen H., Shinn P., Stevenson D. K., Zimmerman J., Barajas P., Cheuk R. et al. Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, 301(5633):653–7, August 2003. Cited on page 22.

Anderson J. B., Sirjusingh C., and Ricker N. Haploidy, diploidy and evolution of antifungal drug resistance in Saccharomyces cerevisiae. *Genetics*, 168(4):1915–23, December 2004. Cited on page 71.

Aouida M., Tounekti O., Leduc A., Belhadj O., Mir L., and Ramotar D. Isolation and characterization of Saccharomyces cerevisiae mutants with enhanced resistance to the anticancer drug bleomycin. *Current genetics*, 45(5):265–72, May 2004. Cited on page 23.

Arlt M. F., Ozdemir A. C., Birkeland S. R., Wilson T. E., and Glover T. W. Hydroxyurea induces de novo copy number variants in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 108(42):17360–5, October 2011. Cited on pages 56 and 70.

Arnaudeau C., Lundin C., and Helleday T. DNA double-strand breaks associated with replication forks are predominantly repaired by homologous recombination involving an exchange mechanism in mammalian cells. *Journal of molecular biology*, 307(5):1235–45, April 2001. Cited on page 56.

Baba T., Ara T., Hasegawa M., Takai Y., Okumura Y., Baba M., Datsenko K. A., Tomita M., Wanner B. L., and Mori H. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology*, 2:2006.0008, January 2006. Cited on page 22.

Baer C. F., Miyamoto M. M., and Denver D. R. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics*, 8(8):619–31, August 2007. Cited on page 26.

Ball C. A., Jin H., Sherlock G., Weng S., Matese J. C., Andrada R., Binkley G., Dolinski K., Dwight S. S., Harris M. A. et al. Saccharomyces Genome Database provides tools to survey gene expression and functional analysis data. *Nucleic acids research*, 29(1):80–81, January 2001. Cited on page 20.

Barnett J. A. Beginnings of microbiology and biochemistry: the contribution of yeast research. *Microbiology*, 149(3):557–567, March 2003. Cited on page 19.

Baudin A., Ozier-Kalogeropoulos O., Denouel A., Lacroute F., and Cullin C. A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucleic acids research*, 21(14):3329–30, July 1993. Cited on pages 21 and 99.

Ben-Shitrit T., Yosef N., Shemesh K., Sharan R., Ruppin E., and Kupiec M. Systematic identification of gene annotation errors in the widely used yeast mutation collections. *Nature methods*, 9(4):373–8, April 2012. Cited on pages 25, 26, 65, and 99.

Beranek D. T. Distribution of methyl and ethyl adducts following alkylation with monofunctional alkylating agents. *Mutation Research*, 231(1):11–30, July 1990. Cited on page 56.

Berger J., Suzuki T., Senti K. A., Stubbs J., Schaffner G., and Dickson B. J. Genetic mapping with SNP markers in Drosophila. *Nature genetics*, 29(4):475–81, December 2001. Cited on page 4.

Berger M. F., Lawrence M. S., Demichelis F., Drier Y., Cibulskis K., Sivachenko A. Y., Sboner A., Esgueva R., Pflueger D., Sougnez C. et al. The genomic complexity of primary human prostate cancer. *Nature*, 470(7333):214–20, February 2011. Cited on page 11.

Bianchi V., Pontis E., and Reichard P. Changes of deoxyribonucleoside triphosphate pools induced by hydroxyurea and their relation to DNA synthesis. *The Journal of biological chemistry*, 261:16037–16042, December 1986. Cited on page 56.

Bigner S. H., Friedman H. S., Vogelstein B., Oakes V. V. J., and Bigner D. D. Amplification of the c-myc gene in human medulloblastoma cell lines and xenografts. *Cancer research*, 50: 2347–2351, April 1990. Cited on page 12.

Birrell G. W., Brown J. A., Wu H. I., Giaever G., Chu A. M., Davis R. W., and Brown J. M. Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13):8778–83, June 2002. Cited on page 23.

Bishop A. J. R. and Schiestl R. H. Homologous recombination as a mechanism for genome rearrangements: environmental and genetic effects. *Human molecular genetics*, 9(16):2427–2434, 2000. Cited on page 70.

Bonadona V., Bonaiti B., Olschwang S., Grandjouan S., Huiart L., Longy M., Guimbaud R., Buecher B., Bignon Y., Caron O. et al. Cancer risks associated with germline mutations in MLH1, MSH2 and MSH6 Genes in Lynch Syndrome. *JAMA*, 305(22):2304–2310, June 2011. Cited on page 45.

Boone C., Bussey H., and Andrews B. J. Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, 8(6):437–49, June 2007. Cited on page 24.

Botstein D., Chervitz S. A., and Cherry J. M. Yeast as a model organism. *Science*, 277(5330): 1259–60, August 1997. Cited on pages 20 and 25.

Botstein D. and Fink G. R. Yeast: an experimental organism for 21st Century biology. *Genetics*, 189(3):695–704, November 2011. Cited on pages 19 and 20.

Branzei D. and Foiani M. Maintaining genome stability at the replication fork. *Nature Reviews Molecular Cell Biology*, 11(3):208–19, March 2010. Cited on page 1.

Bzymek M. and Lovett S. T. Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15):8319–25, July 2001. Cited on page 80.

Cairns B. R. The logic of chromatin architecture and remodelling at promoters. *Nature*, 461 (7261):193–8, September 2009. Cited on page 81.

Calabretta B., Robberson D., Barrera-Saldana A., Lambrou T., and Saunders G. Genomic instability in a region of human DNA enriched in Alu repeat sequences. *Nature*, 296:219–225, March 1982. Cited on pages 68 and 80.

Campbell P. J., Stephens P. J., Pleasance E. D., O'Meara S., Li H., Santarius T., Stebbings L. A., Leroy C., Edkins S., Hardy C. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature genetics*, 40(6):722–9, June 2008. Cited on pages 11 and 17.

Carr A. M. and Lambert S. Replication stress-induced genome instability: the dark side of replication maintenance by homologous recombination. *Journal of molecular biology*, 425 (23):4733–44, November 2013. Cited on page 71.

Carr M., Bensasson D., and Bergman C. M. Evolutionary genomics of transposable elements in Saccharomyces cerevisiae. *PloS one*, 7(11):e50978, January 2012. Cited on page 68.

Carter A. J. R. and Nguyen A. Q. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. *BMC medical genetics*, 12(1):160, January 2011. Cited on page 24.

Carvalho C. M. B., Zhang F., and Lupski J. R. Evolution in health and medicine Sackler colloquium: Genomic disorders: a window into human gene and genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 107 Suppl:1765–71, January 2010. Cited on page 1.

Chan T. F., Carvalho J., Riles L., and Zheng X. F. A chemical genomics approach toward understanding the global functions of the target of rapamycin protein (TOR). *Proceedings of the*

National Academy of Sciences of the United States of America, 97(24):13227–32, November 2000. Cited on page 23.

Chang M., Bellaoui M., Boone C., and Brown G. W. A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):16934–9, December 2002. Cited on pages 56 and 70.

Chen C. and Kolodner R. D. Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants. *Nature genetics*, 23(1):81–5, September 1999. Cited on pages 27, 29, and 52.

Chen X., Bahrami A., Pappo A., Easton J., Dalton J., Hedlund E., Ellison D., Shurtleff S., Wu G., Wei L. et al. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell reports*, 7(1):104–12, April 2014. Cited on page 47.

Cherry J. M., Adler C., Ball C., Chervitz S. A., Dwight S. S., Hester E. T., Jia Y., Juvik G., Roe T., Schroeder M. et al. SGD: Saccharomyces Genome Database. *Nucleic acids research*, 26(1):73–79, January 1998. Cited on pages 3 and 20.

Chiang D. Y., Getz G., Jaffe D. B., O'Kelly M. J. T., Zhao X., Carter S. L., Russ C., Nusbaum C., Meyerson M., and Lander E. S. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods*, 6(1):99–103, January 2009. Cited on page 17.

Christie K. R., Weng S., Balakrishnan R., Costanzo M. C., Dolinski K., Dwight S. S., Engel S. R., Feierbach B., Fisk D. G., Hirschman J. E. et al. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic acids research*, 32:D311–4, January 2004. Cited on page 20.

Chu Y.-L., Wu X., Xu Y., and Her C. MutS homologue hMSH4: interaction with eIF3f and a role in NHEJ-mediated DSB repair. *Molecular cancer*, 12(1):51, January 2013. Cited on pages 71 and 82.

Ciccia A. and Elledge S. J. The DNA damage response: making it safe to play with knives. *Molecular cell*, 40(2):179–204, October 2010. Cited on page 1.

Clapier C. R. and Cairns B. R. The biology of chromatin remodeling complexes. *Annual review of biochemistry*, 78:273–304, January 2009. Cited on pages 69 and 81.

Collins F., Drumm M., Cole J., Lockwood W., Vande Woude G., and Iannuzzi M. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science*, 235:1046–1049, February 1987. Cited on page 10.

Conrad D. F., Pinto D., Redon R., Feuk L., Gokcumen O., Zhang Y., Aerts J., Andrews T. D., Barnes C., Campbell P. et al. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–12, April 2010. Cited on pages 5, 13, and 80.

Cooper G. M., Zerr T., Kidd J. M., Eichler E. E., and Nickerson D. A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature genetics*, 40(10): 1199–203, October 2008. Cited on page 13.

Currall B. B., Chiang C., Talkowski M. E., and Morton C. C. Mechanisms for structural variation in the human genome. *Current genetic medicine reports*, 1(2):81–90, June 2013. Cited on page 7.

Davierwala A. P., Haynes J., Li Z., Brost R. L., Robinson M. D., Yu L., Mnaimneh S., Ding H., Zhu H., Chen Y. et al. The synthetic genetic interaction spectrum of essential genes. *Nature genetics*, 37(10):1147–52, October 2005. Cited on page 24.

Davis W. M. ApE-A plasmid Editor., 2012. Cited on page 92.

Denver D. R., Wilhelm L. J., Howe D. K., Gafner K., Dolan P. C., and Baer C. F. Variation in base-substitution mutation in experimental and natural lineages of Caenorhabditis nematodes. *Genome biology and evolution*, 4(4):513–22, January 2012. Cited on page 27.

Deutschbauer A. M., Williams R. M., Chu A. M., and Davis R. W. Parallel phenotypic analysis of sporulation and postgermination growth in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24):15530–5, November 2002. Cited on page 23.

DiLella A., Kwok S., Ledley F., Marvit J., and Woo S. Molecular structure and polymorphic map of the human phenylalanine hydroxylase gene. *Biochemistry*, 25(4):743–749, February 1986. Cited on page 10.

Dolecek T. A., Propp J. M., Stroup N. E., and Kruchko C. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2005-2009. *Neuro-oncology*, 14 Suppl 5:v1–49, November 2012. Cited on page 74.

Dowty J. G., Win A. K., Buchanan D. D., Lindor N. M., Macrae F. A., Clendenning M., Antill Y. C., Thibodeau S. N., Casey G., Gallinger S. et al. Cancer Risks for MLH1 and MSH2 Mutation Carriers. *Human Mutation*, 34(3):490–497, March 2013. Cited on pages 45 and 80.

Duncan B. and Miller J. Mutagenic deamination of cytosine residues in DNA. *Nature*, 287 (5782):560–561, October 1980. Cited on page 46.

Dwight S. S., Harris M. A., Dolinski K., Ball C. A., Binkley G., Christie K. R., Fisk D. G., Issel-tarver L., Schroeder M., Sherlock G. et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic acids research*, 30(1):69–72, January 2002. Cited on page 20.

Eichler E. E. Masquerading repeats: paralogous pitfalls of the human genome. *Genome research*, 8:758–762, October 1998. Cited on page 68.

Ellegren H., Smith N. G., and Webster M. T. Mutation rate variation in the mammalian genome. *Current Opinion in Genetics & Development*, 13(6):562–568, December 2003. Cited on page 27.

Feng Z., Hu W., Hu Y., and Tang M. Acrolein is a major cigarette-related lung cancer agent: Preferential binding at p53 mutational hotspots and inhibition of DNA repair. *Proceedings of the National Academy of Sciences of the United States of America*, 103(42):15404–9, October 2006. Cited on page 46.

Feuk L. Inversion variants in the human genome: role in disease and genome architecture. *Genome medicine*, 2(2):11, January 2010. Cited on page 11.

Feuk L., Carson A. R., and Scherer S. W. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, February 2006. Cited on page 30.

Forsburg S. L. The art and design of genetic screens: Yeast. *Nature Reviews Genetics*, 2 (September):659–668, September 2001. Cited on pages 26 and 29.

Foury F. Human genetic diseases: a cross-talk between man and yeast. *Gene*, 195:1–10, August 1997. Cited on pages 20 and 25.

Frazer K. a., Murray S. S., Schork N. J., and Topol E. J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4):241–51, April 2009. Cited on pages 1 and 10.

Fujiwara T., Bandi M., Nitta M., Ivanova E. V., Bronson R. T., and Pellman D. Cytokinesis failure generating tetraploids promotes tumorigenesis in p53-null cells. *Nature*, 437(7061): 1043–7, October 2005. Cited on page 71.

Ghaemmaghami S., Huh W., Bower K., Howson R. W., Belle A., Dephoure N., O'Shea E. K., and Weissman J. S. Global analysis of protein expression in yeast. *Nature*, 425:737–741, October 2003. Cited on page 22.

Giaever G. and Nislow C. The Yeast Deletion Collection: A decade of functional genomics. *Genetics*, 197(2):451–465, June 2014. Cited on pages 21 and 22.

Giaever G., Chu A. M., Connelly C., Riles L., Véronneau S., Dow S., Lucau-Danila A., Anderson K., Arkin A. P., Astromoff A. et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418:387–391, July 2002. Cited on pages 20, 21, 22, 23, 46, and 108.

Giaever G., Flaherty P., Kumm J., Proctor M., Nislow C., Jaramillo D. F., Chu A. M., Jordan M. I., Arkin A. P., and Davis R. W. Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 101(3):793–8, January 2004. Cited on page 23.

Gietz R. D. and Schiestl R. H. Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nature protocols*, 2(1):38–41, January 2007. Cited on pages 94 and 100.

Gkikopoulos T., Schofield P., Singh V., Pinskaya M., Mellor J., Smolle M., Workman J. L., Barton G. J., and Owen-Hughes T. A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. *Science*, 333(6050):1758–60, September 2011. Cited on page 82.

Goffeau A., Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., Galibert F., Hoheisel J. D., Jacq C., Johnston M. et al. Life with 6000 genes. *Science*, 274:546–567, October 1996. Cited on pages 20, 33, and 46.

Goffeau A., Aert R., Agostini-Carbone M. L., Ahmed A., Aigle M., Alberghina L., Albermann K., Albers M., Aldea M., Alexandraki D. et al. The Yeast Genome Directory. *Nature*, 387Suppl:1–105, May 1997. Cited on page 20.

Greenwell P. W., Kronmal S. L., Porter S. E., Gassenhuber J., Obermaier B., and Petes T. D. TEL1 , a gene involved in controlling telomere length in S . cerevisiae , is homologous to the human ataxia telangiectasia gene. *Cell*, 82:823–829, September 1995. Cited on page 20.

Gu W., Zhang F., and Lupski J. R. Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1):4, January 2008. Cited on page 8.

Haag-Liautard C., Dorris M., Maside X., Macaskill S., Halligan D. L., Houle D., Charlesworth B., and Keightley P. D. Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature*, 445(7123):82–5, January 2007. Cited on page 27.

Haber J. E. The Many Interfaces of Mre11 Minireview. *Cell*, 95:583–586, November 1998. Cited on page 82.

Handsaker R. E., Korn J. M., Nemesh J., and McCarroll S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature genetics*, 43 (3):269–76, March 2011. Cited on pages 17 and 88.

Hartman J. L., Garvik B., and Hartwell L. Principles for the buffering of genetic variation. *Science*, 291:1001–1004, February 2001. Cited on page 24.

Hastings P. J., Ira G., and Lupski J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genetics*, 5(1):e1000327, January 2009a. Cited on pages 7, 9, and 30.

Hastings P. J., Lupski J. R., Rosenberg S. M., and Ira G. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–64, August 2009b. Cited on pages 6, 7, and 8.

Hawk J. D., Stefanovic L., Boyer J. C., Petes T. D., and Farber R. A. Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America*, 102(24):8639–43, June 2005. Cited on page 26.

Heinicke S., Livstone M. S., Lu C., Oughtred R., Kang F., Angiuoli S. V., White O., Botstein D., and Dolinski K. The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. *PLoS One*, 2(8):e766, January 2007. Cited on page 20.

Her C., Wu X., Griswold M. D., and Zhou F. Human MutS Homologue MSH4 Physically Interacts with von Hippel-Lindau Tumor Suppressor-binding Protein 1. *Cancer research*, 63: 865–872, February 2003. Cited on pages 71 and 82.

Hirschberg J. and Simchen G. Commitment to the Mitotic Cell Cycle in Yeast in Relation to Meiosis. *Experimental Cell Research*, 105:245–252, March 1977. Cited on page 70.

Ho A. S., Kannan K., Roy D. M., Morris L. G. T., Ganly I., Katabi N., Ramaswami D., Walsh L. A., Eng S., Huse J. T. et al. The mutational landscape of adenoid cystic carcinoma. *Nature genetics*, 45(7):791–8, July 2013. Cited on page 11.

Hoeijmakers J. H. J. Genome maintenance mechanisms for preventing cancer. *Nature*, 411: 366–374, May 2001. Cited on pages 6 and 26.

Hohmann A. F. and Vakoc C. R. A rationale to target the SWI/SNF complex for cancer therapy. *Trends in genetics*, 30(8):356–63, August 2014. Cited on page 69.

Honigberg S. M. and Purnapatre K. Signal pathway integration in the switch from the mitotic cell cycle to meiosis in yeast. *Journal of cell science*, 116(11):2137–47, June 2003. Cited on page 70.

Horigome C., Oma Y., Konishi T., Schmid R., Marcomini I., Hauer M. H., Dion V., Harata M., and Gasser S. M. SWR1 and INO80 chromatin remodelers contribute to DNA double-strand break perinuclear anchorage site choice. *Molecular cell*, 55(4):626–39, August 2014. Cited on page 45.

Huang M.-E., Rio A.-G., Nicolas A., and Kolodner R. D. A genomewide screen in Saccharomyces cerevisiae for genes that suppress the accumulation of mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11529–11534, September 2003. Cited on pages 27, 45, and 83.

Huddleston J., Ranade S., Malig M., Antonacci F., Chaisson M., Hon L., Sudmant P. H., Graves T. A., Alkan C., Dennis M. Y. et al. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome research*, 24(4):688–96, April 2014. Cited on page 12.

Hughes T. R., Roberts C. J., Dai H., Jones A. R., Meyer M. R., Slade D., Burchard J., Dow S., Ward T. R., Kidd M. J. et al. Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature genetics*, 25(3):333–337, July 2000. Cited on page 65.

Huxley C., Green E. D., and Dunham I. Rapid assessment of S. cerevisiae mating type by PCR. *Trends in genetics*, 6(8):236, August 1990. Cited on pages 100, 108, and 124.

Iafrate A. J., Feuk L., Rivera M. N., Listewnik M. L., Donahoe P. K., Qi Y., Scherer S. W., and Lee C. Detection of large-scale variation in the human genome. *Nature genetics*, 36(9): 949–51, September 2004. Cited on page 13.

Indjeian V. B., Stern B. M., and Murray A. W. The Centromeric Protein Sgo1 Is Required to Sense Lack of Tension on Mitotic Chromosomes. *Science*, 307(January):130–134, January 2005. Cited on pages 34 and 46.

Iraqui I., Chekkal Y., Jmari N., Pietrobon V., Fréon K., Costes A., and Lambert S. a. E. Recovery of arrested replication forks by homologous recombination is error-prone. *PLoS genetics*, 8(10):e1002976, January 2012. Cited on page 70.

Jackson S. P. and Bartek J. The DNA-damage response in human biology and disease. *Nature*, 461(7267):1071–8, October 2009. Cited on page 1.

Jambhekar A. and Amon A. Control of Meiosis by Respiration. *Current biology*, 18(13):969–975, July 2009. Cited on page 70.

Jin Y. H., Clark A. B., Slebos R. J. C., Al-Refai H., Taylor J. A., Kunkel T. A., Resnick M. A., and Gordenin D. A. Cadmium is a mutagen that acts by inhibiting mismatch repair. *Nature genetics*, 34(3):326–9, July 2003. Cited on page 26.

Jones D. T. W., Jäger N., Kool M., Zichner T., Hutter B., Sultan M., Cho Y.-J., Pugh T. J., Hovestadt V., Stütz A. M. et al. Dissecting the genomic complexity underlying medulloblastoma. *Nature*, 488(7409):100–5, August 2012. Cited on page 75.

Jorgensen P., Nishikawa J. L., Breitkreutz B.-J., and Tyers M. Systematic identification of pathways that couple cell growth and division in yeast. *Science*, 297(5580):395–400, July 2002. Cited on page 23.

Jung D., Giallourakis C., Mostoslavsky R., and Alt F. W. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annual review of immunology*, 24: 541–70, January 2006. Cited on page 5.

Kanellis P., Gagliardi M., Banath J. P., Szilard R. K., Nakada S., Galicia S., Sweeney F. D., Cabelof D. C., Olive P. L., and Durocher D. A screen for suppressors of gross chromosomal rearrangements identifies a conserved role for PLP in preventing DNA lesions. *PLoS Genetics*, 3(8):e134, August 2007. Cited on page 49.

Kao K. C. and Sherlock G. Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. *Nature genetics*, 40(12):1499–504, December 2008. Cited on page 24.

Keeney S. and Neale M. J. Initiation of meiotic recombination by formation of DNA double-strand breaks: mechanism and regulation. *Biochemical Society Transactions*, 34(Pt 4):523–5, August 2006. Cited on page 5.

Kleinman C. L., Gerges N., Papillon-Cavanagh S., Sin-Chan P., Pramatarova A., Quang D.-A. K., Adoue V., Busche S., Caron M., Djambazian H. et al. Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR_Supp. *Nature genetics*, 46(1):39–44, January 2014. Cited on page 77.

Klement K., Luijsterburg M. S., Pinder J. B., Cena C. S., Del Nero V., Wintersinger C. M., Dellaire G., van Attikum H., and Goodarzi A. A. Opposing ISWI- and CHD-class chromatin remodeling activities orchestrate heterochromatic DNA repair. *The Journal of cell biology*, 207(6):717–733, December 2014. Cited on page 70.

Klinz F. and Gallwitz D. Size and position of intervening sequences are critical for the splicing efficiency of pre-mRNA in the yeast Saccharomyces cerevisiae. *Nucleic acids research*, 13 (11):3791–3804, June 1985. Cited on page 52.

Klochendler-Yeivin A., Picarsky E., and Yaniv M. Increased DNA damage sensitivity and apoptosis in cells lacking the Snf5/Ini1 subunit of the SWI/SNF chromatin remodeling complex. *Molecular and cellular biology*, 26(7):2661–74, April 2006. Cited on page 69.

Kloosterman W. P., Guryev V., van Roosmalen M., Duran K. J., de Bruijn E., Bakker S. C. M., Letteboer T., van Nesselrooij B., Hochstenbach R., Poot M. et al. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Human molecular genetics*, 20(10):1916–24, May 2011. Cited on pages 9 and 31.

Knop M., Siegers K., Pereira G., Zachariae W., Winsor B., Nasmyth K., and Schiebel E. Epitope tagging of yeast genes using a PCR-based strategy: more tags and improved practical routines. *Yeast*, 15(10B):963–72, July 1999. Cited on page 94.

Kool M., Koster J., Bunt J., Hasselt N. E., Lakeman A., van Sluis P., Troost D., Meeteren N. S., Caron H. N., Cloos J. et al. Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *PloS one*, 3(8):e3088, January 2008. Cited on page 74.

Korbel J. O. and Campbell P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell*, 152(6):1226–36, March 2013. Cited on page 12.

Korbel J. O., Urban A. E., Affourtit J. P., Godwin B., Grubert F., Simons J. F., Kim P. M., Palejev D., Carriero N. J., Du L. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–6, October 2007. Cited on pages 17 and 32.

Korbel J. O., Abyzov A., Mu X. J., Carriero N., Cayting P., Zhang Z., Snyder M., and Gerstein M. B. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome biology*, 10 (2):R23, January 2009. Cited on page 17.

Korshunov A., Ryzhova M., Jones D. T. W., Northcott P. A., van Sluis P., Volckmann R., Koster J., Versteeg R., Cowdrey C., Perry A. et al. LIN28A immunoreactivity is a potent diagnostic marker of embryonal tumor with multilayered rosettes (ETMR). *Acta neuropathologica*, 124 (6):875–81, December 2012. Cited on page 77.

Korshunov A., Sturm D., Ryzhova M., Hovestadt V., Gessi M., Jones D. T. W., Remke M., Northcott P., Perry A., Picard D. et al. Embryonal tumor with abundant neuropil and true rosettes (ETANTR), ependymoblastoma, and medulloepithelioma share molecular similarity and comprise a single clinicopathological entity. *Acta neuropathologica*, 128(2):279–89, August 2014. Cited on page 77.

Krzywinski M., Schein J., Birol I., Connors J., Gascoyne R., Horsman D., Jones S. J., and Marra M. a. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–45, September 2009. Cited on page 77.

Kupfer D. M., Drabenstot S. D., Buchanan K. L., Lai H., Zhu H., Dyer D. W., Roe B. A., and Murphy J. W. Introns and splicing elements of five diverse fungi. *Eukaryotic cell*, 3(5): 1088–100, October 2004. Cited on page 52.

Lada A. G., Stepchenkova E. I., Waisertreiger I. S. R., Noskov V. N., Dhar A., Eudy J. D., Boissy R. J., Hirano M., Rogozin I. B., and Pavlov Y. I. Genome-wide mutation avalanches induced in diploid yeast cells by a base analog or an APOBEC deaminase. *PLoS genetics*, 9 (9):e1003736, January 2013. Cited on page 71.

Lambert S. and Carr A. M. Replication stress and genome rearrangements: lessons from yeast models. *Current opinion in genetics & development*, 23(2):132–9, April 2013. Cited on page 71.

Lambert S., Mizuno K., Blaisonneau J., Martineau S., Chanet R., Fréon K., Murray J. M., Carr A. M., and Baldacci G. Homologous recombination restarts blocked replication forks at the expense of genome rearrangements by template exchange. *Molecular cell*, 39(3):346–59, August 2010. Cited on page 71.

Lang G. I. and Murray A. W. Estimating the per-base-pair mutation rate in the yeast Saccharomyces cerevisiae. *Genetics*, 178(1):67–82, January 2008. Cited on pages 27 and 38.

Lang G. I. and Murray A. W. Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome biology and evolution*, 3:799–811, January 2011. Cited on page 27.

Lang G. I., Parsons L., and Gammie A. E. Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast. *G3*, 3(9):1453–65, September 2013. Cited on page 24.

Ledbetter D. H., Riccardi V. M., Airhart S. D., Strobel R. J., Keenan B. S., and Crawford J. D. Deletions of chromosome 15 as a cause of the Prader-Willi syndrome. *New England Journal of Medicine*, 304(6):325–329, February 1981. Cited on page 68.

Lee W., Jiang Z., Liu J., Haverty P. M., Guan Y., Stinson J., Yue P., Zhang Y., Pant K. P., Bhatt D. et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, 465(7297):473–7, May 2010. Cited on page 46.

Lehner K. R., Stone M. M., Farber R. a., and Petes T. D. Ninety-six haploid yeast strains with individual disruptions of open reading frames between YOR097C and YOR192C, constructed for the Saccharomyces genome deletion project, have an additional mutation in the mismatch repair gene MSH3. *Genetics*, 177(3):1951–3, November 2007. Cited on page 65.

Lettier G., Feng Q., de Mayolo A. A., Erdeniz N., Reid R. J. D., Lisby M., Mortensen U. H., and Rothstein R. The role of DNA double-strand breaks in spontaneous homologous recombination in S. cerevisiae. *PLoS Genetics*, 2(11):e194, November 2006. Cited on page 69.

Leung G. P., Lee L., Schmidt T. I., Shirahige K., and Kobor M. S. Rtt107 is required for recruitment of the SMC5/6 complex to DNA double strand breaks. *The Journal of biological chemistry*, 286(29):26250–7, July 2011. Cited on page 47.

Li F. P. and Fraumeni J. F. Soft-tissue sarcomas, breast cancer, and other neoplasms. *Annals of Internal Medicine*, 71(4):747–752, October 1969. Cited on page 10.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., and Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16): 2078–9, August 2009. Cited on pages 14, 87, and 88.

Lieber M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual review of biochemistry*, 79:181–211, January 2010. Cited on page 8.

Lindahl T. and Barnes D. Repair of endogenous DNA damage. *Cold Spring Harbor Symposia on Quantitative Biology*, 65:127–133, 2000. Cited on page 1.

Lisby M., Rothstein R., and Mortensen U. H. Rad52 forms DNA repair and recombination centers during S phase. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15):8276–82, July 2001. Cited on page 66.

Liskay R. M., Letsou A., and Stachelek J. L. Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. *Genetics*, 115:161–167, January 1987. Cited on page 6.

Liu L. F., Desai S. D., Li T.-K., Mao Y., Sun M., and Sim S.-P. Mechanism of action of camptothecin. *Annals of the New York Academy of Sciences*, 922(1):1–10, January 2006. Cited on pages 56 and 70.

López Castel A., Cleary J. D., and Pearson C. E. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nature Reviews Molecular Cell Biology*, 11(3): 165–70, March 2010. Cited on page 80.

Love M. I., Huber W., and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014. Cited on page 98.

Lum P. Y., Armour C. D., Stepaniants S. B., Cavet G., Wolf M. K., Butler J. S., Hinshaw J. C., Garnier P., Prestwich G. D., Leonardson A. et al. Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell*, 116:121–137, January 2004. Cited on page 23.

Lynch M., Sung W., Morris K., Coffey N., Landry C. R., Dopman E. B., Dickinson W. J., Okamoto K., Kulkarni S., Hartl D. L. et al. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 105(27):9272–7, July 2008. Cited on pages 26, 27, 28, 38, 44, 46, and 49.

Mable B. and Otto S. Masking and purging mutations following EMS treatment in haploid, diploid and tetraploid yeast (Saccharomyces cerevisiae). *Genetic Research Cambridge*, 77: 9–26, February 2001. Cited on page 71.

Maere S., Heymans K., and Kuiper M. BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, August 2005. Cited on page 99.

Maki H. Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annual review of genetics*, 36:279–303, January 2002. Cited on page 46.

Maltby V. E., Martin B. J. E., Schulze J. M., Johnson I., Hentrich T., Sharma A., Kobor M. S., and Howe L. Histone H3 lysine 36 methylation targets the Isw1b remodeling complex to chromatin. *Molecular and cellular biology*, 32(17):3479–85, September 2012. Cited on page 69.

Mancera E., Bourgon R., Brozzi A., Huber W., and Steinmetz L. M. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–85, July 2008. Cited on pages 36 and 89.

Mardis E. R. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, February 2011. Cited on page 13.

Mardis E. R. Next-generation sequencing platforms. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 6:287–303, January 2013. Cited on pages 13 and 14.

Marotta C. A., Wilson J. T., Forget B. G., and Weissman S. M. Human beta-globin messenger RNA. III. Nucleotide sequences derived from complementary DNA. *Journal of Biological Chemistry*, 252:5040–5053, July 1977. Cited on page 10.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, May 2011. Cited on page 97.

McCarroll S. A., Kuruvilla F. G., Korn J. M., Cawley S., Nemesh J., Wysoker A., Shapero M. H., de Bakker P. I. W., Maller J. B., Kirby A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics*, 40(10):1166–74, October 2008. Cited on page 13.

McClintock B. The fusion of broken ends of chromosomes following nuclear fusion. *Proceedings of the National Academy of Sciences of the United States of America*, 28:458–463, September 1942. Cited on page 9.

McEachern M. J. and Haber J. E. Break-induced replication and recombinational telomere elongation in yeast. *Annual review of biochemistry*, 75:111–35, January 2006. Cited on page 8.

McVey M. and Lee S. E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends in genetics*, 24(11):529–38, November 2008. Cited on page 8.

Metzker M. L. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1): 31–46, January 2010. Cited on page 14.

Mézard C., Pompon D., and Nicolas A. Recombination between similar but not identical DNA sequences during yeast transformation occurs within short stretches of identity. *Cell*, 70: 659–670, August 1992. Cited on pages 68 and 69.

Mi H., Muruganujan A., Casagrande J. T., and Thomas P. D. Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, 8(8):1551–66, August 2013. Cited on page 99.

Mills R. E., Luttig C. T., Larkins C. E., Beauchamp A., Tsui C., Pittard W. S., and Devine S. E. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research*, 16(9):1182–90, September 2006. Cited on pages 4, 32, and 83.

Mills R. E., Walter K., Stewart C., Handsaker R. E., Chen K., Alkan C., Abyzov A., Yoon S. C., Ye K., Cheetham R. K. et al. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, February 2011. Cited on pages 5 and 80.

Mizuguchi G., Shen X., Landry J., Wu W.-H., Sen S., and Wu C. ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science*, 303: 343–349, January 2004. Cited on pages 45 and 81.

Mogilyansky E. and Rigoutsos I. The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell death and differentiation*, 20(12):1603–14, December 2013. Cited on page 78.

Mohammad D. H. and Yaffe M. B. 14-3-3 proteins, FHA domains and BRCT domains in the DNA damage response. *DNA repair*, 8(9):1009–17, September 2009. Cited on page 47.

Moncunill V., Gonzalez S., Beà S., Andrieux L. O., Salaverria I., Royo C., Martinez L., Puiggròs M., Segura-Wang M., Stütz A. M. et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature Biotechnology*, 32(11), October 2014. Cited on page 74.

Montgomery S. B., Goode D. L., Kvikstad E., Albers C. A., Zhang Z. D., Mu X. J., Ananda G., Howie B., Karczewski K. J., Smith K. S. et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome research*, 23 (5):749–61, May 2013. Cited on pages 4 and 5.

Morillo-Huesca M., Clemente-Ruiz M., Andújar E., and Prado F. The SWR1 histone replacement complex causes genetic instability and genome-wide transcription misregulation in the absence of H2A.Z. *PloS one*, 5(8):e12143, January 2010. Cited on page 81.

Motegi A. and Myung K. Measuring the rate of gross chromosomal rearrangements in Saccharomyces cerevisiae: A practical approach to study genomic rearrangements observed in cancer. *Methods*, 41(2):168–76, February 2007. Cited on page 49.

Mullaney J. M., Mills R. E., Pittard W. S., and Devine S. E. Small insertions and deletions (INDELs) in human genomes. *Human molecular genetics*, 19(R2):R131–6, October 2010. Cited on page 4.

Mullighan C. G., Zhang J., Kasper L. H., Lerach S., Payne-Turner D., Phillips L. A., Heatley S. L., Holmfeldt L., Collins-Underwood J. R., Ma J. et al. CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature*, 471(7337):235–9, March 2011. Cited on page 76.

Myerowitz R. Tay-Sachs disease-causing mutations and neutral polymorphisms in the Hex A gene. *Human mutation*, 9:195–208, January 1997. Cited on page 10.

Myung K., Chen C., and Kolodner R. D. Multiple pathways cooperate in the suppression of genome instability in Saccharomyces cerevisiae. *Nature*, 411(6841):1073–6, June 2001a. Cited on page 49.

Myung K., Datta A., and Kolodner R. Suppression of spontaneous chromosomal rearrangements by S phase checkpoint functions in Saccharomyces cerevisiae. *Cell*, 104:397–408, February 2001b. Cited on pages 49 and 52.

Nachman M. W. Single nucleotide polymorphisms and recombination rate in humans. *Trends in Genetics*, 17(9):481–485, September 2001. Cited on pages 3 and 83.

Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D., Gerstein M., and Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320 (5881):1344–9, June 2008. Cited on pages 36 and 89.

Nagarajan N. and Pop M. Sequence assembly demystified. *Nature Reviews Genetics*, 14(3): 157–67, March 2013. Cited on page 19.

Negritto M. C., Qiu J., Ratay D. O., Shen B., and Bailis A. M. Novel function of Rad27 (FEN-1) in restricting short-sequence recombination. *Molecular and cellular biology*, 21(7): 2349–58, April 2001. Cited on page 45.

Neuvéglise C., Feldmann H., Bon E., Gaillardin C., and Casaregola S. Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts. *Genome research*, 12 (6):930–43, June 2002. Cited on page 68.

Ng P. and Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–4, July 2003. Cited on page 10.

Ng R. and Abelson J. Isolation and sequence of the gene for actin in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 77(7): 3912–3916, July 1980. Cited on page 52.

Nik-Zainal S., Alexandrov L. B., Wedge D. C., Van Loo P., Greenman C. D., Raine K., Jones D., Hinton J., Marshall J., Stebbings L. a. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–93, May 2012a. Cited on page 47.

Nik-Zainal S., Van Loo P., Wedge D. C., Alexandrov L. B., Greenman C. D., Lau K. W., Raine K., Jones D., Marshall J., Ramakrishna M. et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, May 2012b. Cited on page 24.

Nishant K. T., Wei W., Mancera E., Argueso J. L., Schlattl A., Delhomme N., Ma X., Bustamante C. D., Korbel J. O., Gu Z. et al. The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS Genetics*, 6(9), September 2010. Cited on pages 4, 26, 27, 31, 44, and 46.

Northcott P. A., Korshunov A., Witt H., Hielscher T., Eberhart C. G., Mack S., Bouffet E., Clifford S. C., Hawkins C. E., French P. et al. Medulloblastoma comprises four distinct molecular variants. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 29(11):1408–14, May 2011. Cited on page 74.

Northcott P. A., Jones D. T. W., Kool M., Robinson G. W., Gilbertson R. J., Cho Y.-J., Pomeroy S. L., Korshunov A., Lichter P., Taylor M. D. et al. Medulloblastomics: the end of the beginning. *Nature Reviews Cancer*, 12(12):818–34, December 2012a. Cited on pages 11, 74, and 75.

Northcott P. A., Shih D. J. H., Peacock J., Garzia L., Morrissy A. S., Zichner T., Stütz A. M., Korshunov A., Reimand J., Schumacher S. E. et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*, 488(7409):49–56, August 2012b. Cited on pages 12 and 75.

Northcott P. A., Lee C., Zichner T., Stütz A. M., Erkek S., Kawauchi D., Shih D. J. H., Hovestadt V., Zapatka M., Sturm D. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*, 511(7510):428–34, June 2014. Cited on pages 12 and 75.

Novak J. E., Ross-Macdonald P. B., and Roeder G. S. The Budding Yeast Msh4 Protein Functions in Chromosome Synapsis and the Regulation of Crossover Distribution. *Genetics*, 158:1013–1025, July 2001. Cited on page 64.

Odagiri N., Seki M., Onoda F., Yoshimura A., Watanabe S., and Enomoto T. Budding yeast mms4 is epistatic with rad52 and the function of Mms4 can be replaced by a bacterial Holliday junction resolvase. *DNA Repair*, 2:347–358, March 2003. Cited on page 36.

Ojesina A. I., Lichtenstein L., Freeman S. S., Pedamallu C. S., Imaz-Rosshandler I., Pugh T. J., Cherniack A. D., Ambrogio L., Cibulskis K., Bertelsen B. et al. Landscape of genomic alterations in cervical carcinomas. *Nature*, 506(7488):371–5, February 2014. Cited on page 11.

Ooi S. L., Shoemaker D. D., and Boeke J. D. A DNA microarray-based genetic screen for nonhomologous end-joining mutants in Saccharomyces cerevisiae. *Science*, 294(5551):2552–6, December 2001. Cited on page 25.

Ossowski S., Schneeberger K., Lucas-Lledó J. I., Warthmann N., Clark R. M., Shaw R. G., Weigel D., and Lynch M. The Rate and Molecular Spectrum in Arabidopsis thaliana. *Science*, 327:92–95, January 2010. Cited on page 27.

Pan J., Sasaki M., Kniewel R., Murakami H., Blitzblau H. G., Tischfield S. E., Zhu X., Neale M. J., Jasin M., Socci N. D. et al. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, 144(5):719–31, March 2011. Cited on pages 36 and 89.

Pang A. W., MacDonald J. R., Pinto D., Wei J., Rafiq M. A., Conrad D. F., Park H., Hurles M. E., Lee C., Venter J. C. et al. Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, 11(5):R52, January 2010. Cited on page 5.

Papamichos-Chronakis M., Watanabe S., Rando O. J., and Peterson C. L. Global regulation of H2A.Z localization by the INO80 chromatin-remodeling enzyme is essential for genome integrity. *Cell*, 144(2):200–213, January 2011. Cited on page 81.

Papavasiliou F. N. and Schatz D. G. Somatic Hypermutation of Immunoglobulin Genes: Merging Mechanisms for Genetic Diversity. *Cell*, 109:S35–S44, April 2002. Cited on page 26.

Pâques F. and Haber J. E. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiology and molecular biology reviews*, 63(2):349–404, June 1999. Cited on pages 6, 8, and 68.

Patel S., Sprung A. U., Keller B. A., Heaton V. J., and Fisher L. M. Identification of yeast DNA topoisomerase II mutants resistant to the antitumor drug doxorubicin: implications for the mechanisms of doxorubicin action and cytotoxicity. *Molecular pharmacology*, 52:658–666, October 1997. Cited on page 56.

Pavelka N., Rancati G., Zhu J., Bradford W. D., Saraf A., Florens L., Sanderson B. W., Hattem G. L., and Li R. Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature*, 468(7321):321–5, November 2010. Cited on pages 34, 41, 91, and 109.

Paz-Elizur T., Krupsky M., Blumenstein S., Elinger D., Schechtman E., and Livneh Z. DNA repair activity for oxidative damage and risk of lung cancer. *Journal of the National Cancer Institute*, 95(17):1312–1319, September 2003. Cited on page 46.

Pelechano V., Wei W., and Steinmetz L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, 497(7447):127–131, April 2013. Cited on page 19.

Perry G. H., Dominy N. J., Claw K. G., Lee A. S., Fiegler H., Redon R., Werner J., Villanea F. a., Mountain J. L., Misra R. et al. Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, 39(10):1256–60, October 2007. Cited on pages 3 and 11.

Peter M. E. Let-7 and miR-200 microRNAs: Guardians against pluripotency and cancer progression. *Cell Cycle*, 8(6):843–852, October 2009. Cited on page 78.

Petermann E., Orta M. L., Issaeva N., Schultz N., and Helleday T. Hydroxyurea-stalled replication forks become progressively inactivated and require two different RAD51-mediated pathways for restart and repair. *Molecular cell*, 37(4):492–502, February 2010. Cited on pages 56 and 70.

Pfeifer G. P. DNA Methylation: Basic Mechanisms. *Current Topics in Microbiology and Immunology*, 301:259–281, 2006. Cited on page 4.

Pfeifer G. P. and Hainaut P. On the origin of G>T transversions in lung cancer. *Mutation Research*, 526(1-2):39–43, May 2003. Cited on page 46.

Pickrell J. K., Marioni J. C., Pai A. A., Degner J. F., Engelhardt B. E., Nkadori E., Veyrieras J.-B., Stephens M., Gilad Y., and Pritchard J. K. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–72, April 2010. Cited on page 3.

Piekarska I., Rytka J., and Rempola B. Regulation of sporulation in the yeast Saccharomyces cerevisiae. *Acta Biochimica Polonica*, 57(3):241–250, September 2010. Cited on page 70.

Pierce S. E., Davis R. W., Nislow C., and Giaever G. Genome-wide analysis of barcoded Saccharomyces cerevisiae gene-deletion mutants in pooled cultures. *Nature protocols*, 2(11): 2958–2974, January 2007. Cited on page 21.

Pinkel D., Segraves R., Sudar D., Clark S., Poole I., Kowbel D., Collins C., Kuo W.-l., Chen C., Zhai Y. et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature*, 20(october):207–211, October 1998. Cited on page 13.

Pinskaya M., Nair A., Clynes D., Morillon A., and Mellor J. Nucleosome remodeling and transcriptional repression are distinct functions of Isw1 in Saccharomyces cerevisiae. *Molecular and cellular biology*, 29(9):2419–30, May 2009. Cited on page 60.

Pleasance E. D., Cheetham R. K., Stephens P. J., McBride D. J., Humphray S. J., Greenman C. D., Varela I., Lin M.-L., Ordóñez G. R., Bignell G. R. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–6, January 2010a. Cited on page 11.

Pleasance E. D., Stephens P. J., O'Meara S., McBride D. J., Meynert A., Jones D., Lin M.-L., Beare D., Lau K. W., Greenman C. et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184–90, January 2010b. Cited on page 46.

Pritchard C. C., Morrissey C., Kumar A., Zhang X., Smith C., Coleman I., Salipante S. J., Milbank J., Yu M., Grady W. M. et al. Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. *Nature communications*, 5:4988, January 2014. Cited on page 45.

Qian W., Ma D., Xiao C., Wang Z., and Zhang J. The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. *Cell reports*, 2(5):1399–410, November 2012. Cited on page 24.

Quail M. A., Smith M., Coupland P., Otto T. D., Harris S. R., Connor T. R., Bertoni A., Swerdlow H. P., and Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1): 341, January 2012. Cited on page 16.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. Cited on pages 90 and 99.

Rahner N., Friedrichs N., Wehner M., Steinke V., Aretz S., Friedl W., Buettner R., Mangold E., Propping P., and Walldorf C. Nine novel pathogenic germline mutations in MLH1, MSH2, MSH6 and PMS2 in families with Lynch syndrome. *Acta oncologica*, 46(6):763–9, January 2007. Cited on page 45.

Raphael B. J. Chapter 6: Structural variation and medical genomics. *PLoS computational biology*, 8(12):e1002821, January 2012. Cited on page 16.

Rausch T., Jones D. T. W., Zapatka M., Stütz A. M., Zichner T., Weischenfeldt J., Jäger N., Remke M., Shih D., Northcott P. A. et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, 148(1-2):59–71, January 2012a. Cited on pages 9, 11, 12, 30, 74, and 75.

Rausch T., Zichner T., Schlattl A., Stütz A. M., Benes V., and Korbel J. O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18): i333–i339, September 2012b. Cited on pages 16, 17, 75, and 87.

Reagan M. S., Pittenger C., Siede W., and Friedberg E. C. Characterization of a mutant strain of Saccharomyces cerevisiae with a deletion of the RAD27 gene, a structural homolog of the RAD2 nucleotide excision repair gene. *Journal of Bacteriology*, 177(2):364–371, January 1995. Cited on page 45.

Redon C., Pilch D. R., Rogakou E. P., Orr A. H., Lowndes N. F., and Bonner W. M. Yeast histone 2A serine 129 is essential for the efficient repair of checkpoint-blind DNA damage. *EMBO reports*, 4(7):678–84, July 2003. Cited on page 81.

Regairaz M., Zhang Y.-W., Fu H., Agama K. K., Tata N., Agrawal S., Aladjem M. I., and Pommier Y. Mus81-mediated DNA cleavage resolves replication forks stalled by topoisomerase I-DNA complexes. *The Journal of cell biology*, 195(5):739–49, November 2011. Cited on page 70.

Rhee H. S. and Pugh B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389):295–301, March 2012. Cited on pages 36 and 89.

Rowley J. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, 243:290–293, June 1973. Cited on page 11.

Saintigny Y. and Lopez B. S. Homologous recombination induced by replication inhibition, is stimulated by expression of mutant p53. *Oncogene*, 21(3):488–92, January 2002. Cited on page 70.

Saleh A. D., Savage J. E., Cao L., Soule B. P., Ly D., DeGraff W., Harris C. C., Mitchell J. B., and Simone N. L. Cellular stress induced alterations in microRNA let-7a and let-7b expression are dependent on p53. *PloS one*, 6(10):e24429, January 2011. Cited on page 78.

Saleh-Gohari N., Bryant H. E., Schultz N., Parker K. M., Cassel T. N., and Helleday T. Spontaneous homologous recombination is induced by collapsed replication forks that are caused by endogenous DNA single-strand breaks. *Molecular and cellular biology*, 25(16): 7158–69, August 2005. Cited on page 70.

San Filippo J., Sung P., and Klein H. Mechanism of eukaryotic homologous recombination. *Annual review of biochemistry*, 77:229–57, January 2008. Cited on page 6.

Scheifele L. Z., Cost G. J., Zupancic M. L., Caputo E. M., and Boeke J. D. Retrotransposon overdose and genome integrity. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33):13927–32, August 2009. Cited on page 29.

Scherens B. and Goffeau A. The uses of genome-wide yeast mutant collections. *Genome biology*, 5(7):229, January 2004. Cited on pages 21 and 25.

Selva E., Maderazo A., and Lahue R. Differential effects of the mismatch repair genes MSH2 and MSH3 on homeologous recombination in Saccharomyces cerevisiae. *Molecular gene genetics*, 257:71–82, December 1997. Cited on page 45.

Serero A., Jubin C., Loeillet S., Legoix-Né P., and Nicolas A. G. Mutational landscape of yeast mutator strains. *Proceedings of the National Academy of Sciences of the United States of America*, 111(5):1897–902, February 2014. Cited on pages 27, 29, and 30.

Shain A. H. and Pollack J. R. The spectrum of SWI/SNF mutations, ubiquitous in human cancers. *PloS one*, 8(1):e55119, January 2013. Cited on page 69.

Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–504, November 2003. Cited on pages 64 and 99.

Shaw C. J. and Lupski J. R. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Human molecular genetics*, 13 Spec No(1): R57–64, April 2004. Cited on page 5.

Sheltzer J. M., Blank H. M., Pfau S. J., Tange Y., George B. M., Humpton T. J., Brito I. L., Hiraoka Y., Niwa O., and Amon A. Aneuploidy drives genomic instability in yeast. *Science*, 333(6045):1026–30, August 2011. Cited on page 47.

Sherman F. Getting started with yeast. *Methods in enzymology*, 350:3–41, August 2002. Cited on page 19.

Sinclair D. A., Mills K., and Guarente L. Accelerated Aging and Nucleolar Fragmentation in Yeast sgs1 Mutants. *Science*, 277(5330):1313–1316, August 1997. Cited on page 20.

Smith A. M., Ammar R., Nislow C., and Giaever G. A survey of yeast genomic assays for drug and target discovery. *Pharmacology & therapeutics*, 127(2):156–64, August 2010a. Cited on page 23.

Smith A. M., Heisler L. E., St Onge R. P., Farias-Hesson E., Wallace I. M., Bodeau J., Harris A. N., Perry K. M., Giaever G., Pourmand N. et al. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic acids research*, 38(13): e142, July 2010b. Cited on page 25.

Smith S., Hwang J.-Y., Banerjee S., Majeed A., Gupta A., and Myung K. Mutator genes for suppression of gross chromosomal rearrangements identified by a genome-wide screening in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 101(24):9039–44, June 2004. Cited on pages 49 and 69.

Smolle M., Venkatesh S., Gogol M. M., Li H., Zhang Y., Florens L., Washburn M. P., and Workman J. L. Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange. *Nature structural & molecular biology*, 19(9): 884–92, September 2012. Cited on pages 34, 45, 60, 69, and 81.

SnapGene. Software (from GSL Biotech; available at snapgene.com). Cited on page 92.

Snow R. Mutants of yeast sensitive to ultraviolet light. *Journal of Bacteriology*, 94(3):571–575, September 1967. Cited on page 71.

Solomon D. A., Kim T., Diaz-Martinez L. A., Fair J., Elkahloun A. G., Harris B. T., Toretsky J. A., Rosenberg S. A., Shukla N., Ladanyi M. et al. Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science*, 333(6045):1039–43, August 2011. Cited on page 47.

Song W., Dominska M., Greenwell P. W., and Petes T. D. Genome-wide high-resolution mapping of chromosome fragile sites in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 111(21):E2210–8, May 2014. Cited on page 68.

Spingola M., Grate L., Haussler D., and Ares M. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *RNA*, 5:221–234, February 1999. Cited on page 52.

St. Charles J. and Petes T. D. High-resolution mapping of spontaneous mitotic recombination hotspots on the 1.1 Mb arm of yeast chromosome IV. *PLoS genetics*, 9(4):e1003434, April 2013. Cited on pages 68 and 69.

Stankiewicz P. and Lupski J. R. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, 18(2):74–82, February 2002. Cited on page 5.

Stankiewicz P. and Lupski J. R. Structural variation in the human genome and its role in disease. *Annual review of medicine*, 61:437–55, January 2010. Cited on page 5.

Steinmetz L. M., Scharfe C., Deutschbauer A. M., Mokranjac D., Herman Z. S., Jones T., Chu A. M., Giaever G., Prokisch H., Oefner P. J. et al. Systematic screen for human disease genes in yeast. *Nature genetics*, 31(4):400–4, August 2002. Cited on pages 25, 64, and 98.

Stephens P. J., McBride D. J., Lin M.-L., Varela I., Pleasance E. D., Simpson J. T., Stebbings L. A., Leroy C., Edkins S., Mudie L. J. et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005–10, December 2009. Cited on page 11.

Stephens P. J., Greenman C. D., Fu B., Yang F., Bignell G. R., Mudie L. J., Pleasance E. D., Lau K. W., Beare D., Stebbings L. a. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, January 2011. Cited on pages 9, 12, 30, and 75.

Stephens P. J., Tarpey P. S., Davies H., Van Loo P., Greenman C., Wedge D. C., Nik-Zainal S., Martin S., Varela I., Bignell G. R. et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400–4, June 2012. Cited on page 11.

Strand M., Prolla T. A., Liskay R. M., and Petes T. D. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, 365:274–276, September 1993. Cited on page 20.

Stratton M. R., Campbell P. J., and Futreal P. A. The cancer genome. *Nature*, 458(7239): 719–24, April 2009. Cited on pages 11 and 12.

Sudmant P. H., Kitzman J. O., Antonacci F., Alkan C., Malig M., Tsalenko A., Sampas N., Bruhn L., Shendure J., and Eichler E. E. Diversity of human copy number variation and multicopy genes. *Science*, 330(6004):641–6, October 2010. Cited on page 17.

Sundararajan A., Lee B., and Garfinkel D. J. The Rad27 (Fen-1) nuclease inhibits Ty1 mobility in Saccharomyces cerevisiae. *Genetics*, 163:55–67, January 2003. Cited on page 45.

Sung P. and Klein H. Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nature Reviews Molecular Cell Biology*, 7(10):739–50, October 2006. Cited on pages 6 and 8.

Szamecz B., Boross G., Kalapis D., Kovács K., Fekete G., Farkas Z., Lázár V., Hrtyan M., Kemmeren P., Groot Koerkamp M. J. a. et al. The Genomic Landscape of Compensatory Evolution. *PLoS Biology*, 12(8):e1001935, August 2014. Cited on page 24.

Szostak J. W., Orr-Weaver T. L., Rothstein R. J., and Stahl F. W. The double-strand-break repair model for recombination. *Cell*, 33(1):25–35, May 1983. Cited on page 8.

Taylor M. D., Northcott P. A., Korshunov A., Remke M., Cho Y.-J., Clifford S. C., Eberhart C. G., Parsons D. W., Rutkowski S., Gajjar A. et al. Molecular subgroups of medulloblastoma: the current consensus. *Acta neuropathologica*, 123(4):465–72, April 2012. Cited on page 74.

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, October 2010. Cited on page 3.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, November 2012. Cited on pages 3, 4, and 10.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, June 2007. Cited on page 3.

Thompson-Jager S. and Domdey H. Yeast pre-mRNA splicing requires a minimum distance between the 5 ' splice site and the internal branch acceptor site. *Molecular and cellular biology*, 7(11):4010–4016, November 1987. Cited on page 50.

Tishkoff D. X., Filosi N., Gaida G. M., and Kolodner R. D. A novel mutation avoidance mechanism dependent on S . cerevisiae RAD27 is distinct from DNA mismatch repair. *Cell*, 88:253–263, January 1997. Cited on page 45.

Tkach J. M., Yimit A., Lee A. Y., Riffle M., Costanzo M., Jaschob D., Hendry J. A., Ou J., Moffat J., Boone C. et al. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nature cell biology*, 14 (9):966–976, September 2012. Cited on page 71.

Toiber D., Erdel F., Bouazoune K., Silberman D. M., Zhong L., Mulligan P., Sebastian C., Cosentino C., Martinez-Pastor B., Giacosa S. et al. SIRT6 recruits SNF2H to DNA break sites, preventing genomic instability through chromatin remodeling. *Molecular cell*, 51(4): 454–68, August 2013. Cited on page 69.

Tomlins S. A., Rhodes D. R., Perner S., Dhanasekaran S. M., Mehra R., Sun X. W., Varambally S., Cao X., Tchinda J., Kuefer R. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310(5748):644–8, October 2005. Cited on page 12.

Tong A. H., Evangelista M., Parsons A. B., Xu H., Bader G. D., Pagé N., Robinson M., Raghibizadeh S., Hogue C. W., Bussey H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–8, December 2001. Cited on page 23.

Treangen T. J. and Salzberg S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, January 2012. Cited on page 82.

Ullal P., Vilella-Mitjana F., Jarmuz A., and Aragón L. Rtt107 phosphorylation promotes localisation to DNA double-stranded breaks (DSBs) and recombinational repair between sister chromatids. *PloS one*, 6(5):e20152, January 2011. Cited on page 47.

Untergasser A., Nijveen H., Rao X., Bisseling T., Geurts R., and Leunissen J. A. M. Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids research*, 35:W71–4, July 2007. Cited on page 91.

Vary J. C., Gangaraju V. K., Qin J., Landel C. C., Kooperberg C., Bartholomew B., and Tsukiyama T. Yeast Isw1p Forms Two Separable Complexes In Vivo. *Molecular and Cellular Biology*, 23(1):80–91, January 2003. Cited on pages 60 and 69.

Verstrepen K. J., Jansen A., Lewitter F., and Fink G. R. Intragenic tandem repeats generate functional variability. *Nature genetics*, 37(9):986–90, September 2005. Cited on page 26.

Vidi P.-A., Liu J., Salles D., Jayaraman S., Dorfman G., Gray M., Abad P., Moghe P. V., Irudayaraj J. M., Wiesmüller L. et al. NuMA promotes homologous recombination repair by regulating the accumulation of the ISWI ATPase SNF2h at DNA breaks. *Nucleic acids research*, 42(10):6365–79, June 2014. Cited on page 69.

Viswanathan S. R., Daley G. Q., and Gregory R. I. Selective blockade of microRNA processing by Lin28. *Science*, 320:97–100, April 2008. Cited on page 78.

Voineagu I., Narayanan V., Lobachev K. S., and Mirkin S. M. Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 105(29):9936–41, July 2008. Cited on page 52.

Voineagu I., Freudenreich C. H., and Mirkin S. M. Checkpoint responses to unusual structures formed by DNA repeats. *Molecular carcinogenesis*, 48(4):309–318, April 2009. Cited on page 80.

Wach A., Brachat A., Pöhlmann R., and Philippsen P. New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. *Yeast*, 10(13):1793–1808, December 1994. Cited on pages 21 and 99.

Wang K., Li M., and Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164, September 2010. Cited on pages 39, 41, and 89.

Warren S., Zhang F., Licameli G., and Peters J. The fragile X sites in somatic cell hybrids: an approach for molecular cloning of fragile sites. *Science*, 237:420–423, July 1987. Cited on page 10.

Weigand J. E. and Suess B. Tetracycline aptamer-controlled regulation of pre-mRNA splicing in yeast. *Nucleic acids research*, 35(12):4179–85, January 2007. Cited on page 52.

Weischenfeldt J., Symmons O., Spitz F., and Korbel J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2): 125–138, February 2013. Cited on pages 12 and 18.

Wellinger R. J. and Zakian V. A. Everything you ever wanted to know about Saccharomyces cerevisiae telomeres: beginning to end. *Genetics*, 191(4):1073–105, August 2012. Cited on page 46.

Wicks S. R., Yeh R. T., Gish W. R., Waterston R. H., and Plasterk R. H. A. Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. *Nature*, 28: 160–164, June 2001. Cited on page 4.

Wilkening S., Tekkedil M. M., Lin G., Fritsch E. S., Wei W., Gagneur J., Lazinski D. W., Camilli A., and Steinmetz L. M. Genotyping 1000 yeast strains by next-generation sequencing. *BMC genomics*, 14(1):90, January 2013. Cited on pages 87 and 111.

Willingham S., Outeiro T. F., DeVit M. J., Lindquist S. L., and Muchowski P. J. Yeast genes that enhance the toxicity of a mutant huntingtin fragment or alpha-synuclein. *Science*, 302 (5651):1769–72, December 2003. Cited on page 25.

Wilson B. G. and Roberts C. W. M. SWI/SNF nucleosome remodellers and cancer. *Nature Reviews Cancer*, 11(7):481–92, July 2011. Cited on pages 69 and 81.

Wilson J. F., Weale M. E., Smith A. C., Gratrix F., Fletcher B., Thomas M. G., Bradman N., and Goldstein D. B. Population genetic structure of variable drug response. *Nature genetics*, 29(3):265–9, November 2001. Cited on page 3.

Winzeler E. A. Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis. *Science*, 285(5429):901–906, August 1999. Cited on pages 21, 22, and 100.

Xi R., Luquette J., Hadjipanayis A., Kim T.-m., and Park P. J. BIC-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data. *Genome Biology*, 11(Suppl 1):O10, October 2010. Cited on pages 17 and 87.

Xu Y., Ayrapetov M. K., Xu C., Gursoy-Yuzugullu O., Hu Y., and Price B. D. Histone H2A.Z controls a critical chromatin remodeling step required for DNA double-strand break repair. *Molecular cell*, 48(5):723–33, December 2012. Cited on page 81.

Yang L., Luquette L. J., Gehlenborg N., Xi R., Haseley P. S., Hsieh C.-H., Zhang C., Ren X., Protopopov A., Chin L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153(4):919–29, May 2013. Cited on page 11.

Ye K., Schulz M. H., Long Q., Apweiler R., and Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–71, November 2009. Cited on pages 16 and 87.

Yen P. H., Li X.-M., Tsai S.-P., Johnson C., Mohandas T., and Shapiro L. J. Frequent deletions of the human X chromosome distal short arm result from recombination between low copy repetitive elements. *Cell*, 61(4):603–610, May 1990. Cited on page 68.

Yu X. and Gabriel A. Patching broken chromosomes with extranuclear cellular DNA. *Molecular cell*, 4:873–881, November 1999. Cited on page 52.

Yuen K. W. Y., Warren C. D., Chen O., Kwok T., Hieter P., and Spencer F. A. Systematic genome instability screens in yeast and their potential relevance to cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 104(10):3925–30, March 2007. Cited on page 29.

Zack T. I., Schumacher S. E., Carter S. L., Cherniack A. D., Saksena G., Tabak B., Lawrence M. S., Zhang C.-Z., Wala J., Mermel C. H. et al. Pan-cancer patterns of somatic copy number alteration. *Nature genetics*, 45(10):1134–1140, September 2013. Cited on page 11.

Zhang F., Gu W., Hurles M. E., and Lupski J. R. Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, 10:451–81, January 2009a. Cited on pages 5 and 11.

Zhang F., Khajavi M., Connolly A. M., Towne C. F., Batish S. D., and Lupski J. R. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature genetics*, 41(7):849–53, July 2009b. Cited on page 9.

Zhang Z. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic acids research*, 31(18):5338–5348, September 2003. Cited on page 4.

Zhang Z. and Dietrich F. S. Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. *Nucleic acids research*, 33(9):2838–51, January 2005. Cited on pages 36 and 89.

Zheng L., Dai H., Zhou M., Li M., Singh P., Qiu J., Tsark W., Huang Q., Kernstine K., Zhang X. et al. Fen1 mutations result in autoimmunity, chronic inflammation and cancers. *Nature medicine*, 13(7):812–9, July 2007. Cited on page 45.

Zhu J., Pavelka N., Bradford W. D., Rancati G., and Li R. Karyotypic determinants of chromosome instability in aneuploid budding yeast. *PLoS Genetics*, 8(5):e1002719, January 2012. Cited on page 47.

Zhu Y. O., Siegal M. L., Hall D. W., and Petrov D. A. Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 111(22):E2310–8, June 2014. Cited on pages 4, 27, 29, 40, 46, and 49.