UNIVERSITY OF CALIFORNIA, SAN DIEGO

Genetic and Structural Studies of Gross Chromosomal Rearrangements Mediated by Divergent Homology in *S. cerevisiae*

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy

in

Chemistry

by

Sara N. Bell

Committee in charge:

Professor Richard D. Kolodner Chair
Professor Wei Wang, Co-Chair
Professor Edward Dennis
Professor Lorraine Pillus
Professor Mark Thiemens

2017

The dissertation of Sara N. Bell is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Co-Chair

_____

Chair

University of California, San Diego

2017

Dedication

Pax et lux.

Epigraph

Do not go where the path may lead.

Go instead where there is no path, and leave a trail.

Ralph Waldo Emerson

Table of Contents

List of Abbreviations

DSB – Double stranded break

HR – Homologous recombination

NHEJ – Non-homologous end joining

GCR – Gross chromosomal rearrangement

ssDNA – Single stranded DNA

NGS – Next Generation Sequencing

List of Figures

List of Tables

Acknowledgements

I would very much like to acknowledge and thank my advisor, Richard Kolodner, for his support and guidance throughout my training. His relentlessly high standards held me to a standard of accountability that was integral to my development as an independent scientist. From my time working with Richard I learned that with the proper mindset, failure is just a step closer to eventual success. I have no doubt that the lessons that I have learned here will serve me enormously moving forward.

I am extraordinarily grateful for the incessant help and support that I received from Chris Putnam over the years. His intellectualism, patience and passion for teaching have been a tremendous asset to my training and helped me through the toughest failures.

I am beyond thankful to my wonderful parents for their unwavering support for the last 8 years as I moved across the country to pursue my graduate degree. I would like to thank my mother for setting the standard for the strong female role which I have pursued in my professional life, as well as her unconditional love and kindness. I would like to thank my father for always challenging me, pushing me, and always having the tough conversations. Thank you for philosophical conversations under the stars in Florida, and thanks for making me drive your van home that day.

I would like to express my immense gratitude for the friends who have become family, and the co-workers and classmates who have become a part of my soul. I would like to thank Katie Wu for being the closest thing I'll have in this life to a sister, and for being as generous and supportive now as she was when we were fifth graders picking kasa nuts in her uncle's backyard. I would like to thank Mei-Kay, Drew, Ted and Mike for being there from the beginning. I would like to thank my number one adventure buddy Derek for his most nautical enterprises. I would like to thank Christianna for having her mind equally blown by doors and grass. I would like to thank Janny for being my ride-or-die and teaching me so much about myself.

I would like to thank Jessica Karr for being the best partner in crime and fellow Mammoth Lakes groupie. I would literally drive around a wildfire for you. I would like to thank Kristina Hamill for being an unmatchable adventure partner and without whom I might have never been to Spring Valley, Corona or Yuma. I would like to thank both Jessica and Kristina for also not listening to the directions to the Joshua Tree employee campground.

I would like to thank German Gomez for often switching up afternoon coffee for "a cold one", and deeply philosophical conversations about countless topics on miscellaneous patios around campus. I would like to thank Celso Espinoza for being my marathon guru, an endless source of inspiration and support for adventures, and for being my biggest fan. I would like to thank Andrea Local for being a great running buddy, letting me almost kill her father

at her wedding, and always showing up for me. I would like to thank Judy Johns for literally being my second mom and for believing in me to the ends of the earth. I would like to thank Jason for his tremendous friendship, residency on the A-team, and for the legacy of the LICR retreat. I would like to thank Claudio for always telling me how cute my cat is.

I am so grateful to have had the opportunity to work with such incredible colleagues in the Kolodner lab. I would like to thank Nikki for her incredible friendship and conversation; I love dissecting personal growth and what it means to live a meaningful life with you. I would like to thank Anjana for constantly going above and beyond to help me and for the gift of her loyal friendship. I would like to thank Elaine for being a great friend and for sharing my love of Juniper Foxx. I would like to thank Katie Pallis and Cathy Smith for the OG times at Porters and beyond. I would like to thank Betsy for everything she has done for me over the years and for Che. I would like to thank the Black Rock Fish Camp constituent (Matt, Kat, and Dafne) as well as Eva, Binzhong, Hans, Sarah, Sandra and Bill for being wonderful labmates. I am grateful for their time and insight about science and life.

Chapter 2 is in preparation for publication with Richard Kolodner and Christopher Putnam. The dissertation author is the primary author for this paper.

Chapter 3, in full, is a reprint of the material as it appears in A Genetic Network that Suppresses Genome Rearrangements in *Saccharomyces*

cerevisiae and Contains Defects in Cancers. *Nature Communications*. 2016 Apr 13;7:11256. doi: 10.1038/ncomms11256. Putnam CD*, Srivatsan A*, Nene RV*, Martinez SL*, Clotfelter SP, Bell SN, Somach SB, de Souza JE, Fonseca AF, de Souza SJ, Kolodner RD. The dissertation author is third author for this paper (four co-first authors*).

Chapter 4, in part, is in preparation for publication of the material as part of a collaboration between Richard Kolodner, Christopher Putnam and Paul Mischel. The dissertation author is an intermediate author for this paper.

Vita

2009 BS, Biochemistry, Tufts University

2011 MS, Chemistry, University of California, San Diego

2017 PhD, Chemistry, University of California, San Diego

2014 Micro-MBA Certificate, Rady Center for Executive Development

2016 ADMET Process Certificate, UC San Diego Extension

Publications

Putnam CD*, Srivatsan A*, Nene RV*, Martinez SL*, Clotfelter SP, Bell SN, Somach SB, de Souza JE, Fonseca AF, de Souza SJ, Kolodner RD. *A genetic network that suppresses genome rearrangements in Saccharomyces cerevisiae and contains defects in cancers*. Nat Commun. 2016 Apr 13;7:11256. doi: 10.1038/ncomms11256.

ABSTRACT OF THE DISSERTATION

Genetic and Structural Studies of Gross Chromosomal Rearrangements Mediated by Divergent Homology in *S. cerevisiae*

by

Sara N. Bell

Doctor of Philosophy in Chemistry

University of California, San Diego, 2017

Professor Richard D. Kolodner, Chair
Professor Wei Wang, Co-Chair

Maintenance of genome stability is critical for cell viability and survival. Consequently, DNA double strand breaks (DSBs) must be faithfully repaired in order to maintain the integrity of the genome. In S. cerevisiae, the homologous recombination (HR) pathways frequently repair DSBs using a homologous chromosome as a template. However, aberrant recombination with an improper homeologous template may lead to gross chromosomal rearrangements (GCRs). It is not known how divergent homologous sequences are selected as targets for non-allelic HR, or which pathways suppress the formation of such GCRs. To study this, we devised two new GCR assays in which sequences on chromosomes 14 and 15 target imperfect homologies in the rest of the genome. The homologies targeted by these assays differ in length, number of potential rearrangement targets, and percent homology to the original sequence. The formation of GCRs detected by these assays exhibit diverse genetic requirements; surprisingly, we found that many GCRs were not suppressed by mutations in the HR pathways. We characterized the spectrum of GCR structures observed in the assays and sequenced the translocation breakpoint junctions for the most prominent GCRs. We found that mutations in some genes in the heteroduplex rejection pathway, a process that suppresses recombination between imperfect templates, shifted the location of the translocation breakpoints towards regions of lower homology. Another study involving a genome-wide systematic screen of 4 different types of GCR assays involving different types of homology

identified 182 new genes involved in genome instability as well as 483 cooperatively acting genes that suppress the genome instability caused by these genes. Analysis of TCGA data revealed that 93% of ovarian and 66% of colorectal cancers had defects affecting one or more of these genes. Finally, an exploration into the role of the essential TORC2 complex in maintenance of genome stability using the temperature sensitive *tor2-21* allele revealed an increase in GCRs associated with a defect in Tor2 in two different GCR assays. This mutation caused a shift in the spectrum of GCR structures observed in an assay mediated by a single copy sequence but did not distort the spectrum of GCR structures observed in an assay whose GCRs were mediated by divergent homology. Together, these results suggest that the role of divergent homology in formation of GCRs is complex and the properties of the homology as well as its recombination targets influence the mechanisms and genetics of GCR formation

CHAPTER 1

Introduction to the genome instability and genome rearrangements

1.1 Introduction

Gross chromosomal rearrangements (GCRs) underlie a variety of human genetic diseases and cancers. These events include chromosome translocations, deletions, insertions, and inversions. Several pathways are implicated in formation of GCRs, but it is believed that most occur response to DNA double-strand breaks (DSBs). DSBs are a major source of genomic instability and must be properly repaired in order to avoid cell death. Cells utilize a variety of pathways to repair DNA damage sustained through any of a number of sources, ranging from UV light to drug exposure or replication error. Normally, faithful repair occurs through the homologous recombination (HR) pathway, in which an intact copy of a broken chromosome serves as a template for repair [1]. However, if an alternative repair pathway heals the break instead, improper repair may result and a GCR may occur. Examples of this include non-allelic HR utilizing an ectopic sequence, the various end-joining pathways that simply join a DSB to another DSB in the genome, or the *de novo* telomere addition pathway that creates a new telomere at the site of the break [2-4]. Importantly, the characteristics of the DNA sequence at the site of the DSB and any homologous sequences present in the genome influence the types of aberrant repair observed [5].

The human genome is highly repetitive and faithful repair is critical in order to avoid incurring GCRs. Although repetitive DNA accounts for 40-50%

of the human genome, it is replicated with surprising fidelity: the somatic mutation rate is approximately 1.4 E-10 mutations per nucleotide per cell per generation [6-8]. Repetitive elements in the genome vary from each other in terms of both length of the sequence as well as percent identity. These elements range from short, perfectly (or near perfectly) homologous elements to long, divergently homologous sequences. Examples include long terminal repeats (LTRs), the different classes of retrotransposons, SINEs, LINEs, microsatellites, and segmental duplications [9, 10]. Previous work has identified specific pathways that suppress improper recombination between different types of repetitive sequences that can lead to GCRs, but the full range of genetic requirements for suppression of GCRs are still not understood [5]. Furthermore, the properties of divergent homologous sequences that lend themselves to formation of GCRs remain unclear, but have profound implications for human disease.

Despite the prevalence of divergent homologous sequences in the genome and their implications in genetic disease, repetitive elements such as these remain challenging to study because they are often nearly indistinguishable from each other. Genetic studies to identify the genes and pathways involved in the formation of chromosome rearrangements involving repetitive elements can provide insight into the mechanisms by which these events occur. Understanding the structures of the GCRs formed can also provide mechanistic insight as well as implicate the propensity of certain kinds

of sequences to be involved in their formation. Historically, a number of methods have been used to perform structural analysis of chromosome rearrangements. Some of the most straightforward approaches include PCR mapping of the chromosome arm(s) expected to be involved in the rearrangements and PCR amplification across chromosome fusion junctions, Pulsed-Field Gel Electrophoresis (PFGE) to measure overall chromosome size and Southern blotting of the relevant chromosomes, array Comparative Genomic Hybridization (aCGH) to analyze copy number changes genome-wide, Multiplex Ligation-dependent Probe Amplification (MLPA) to assess the presence or absence of chromosome arms to infer rearrangements, and Next Generation Sequencing (NGS) methods to provide comprehensive whole genome analysis [3, 11-15]. Using these tools, our work seeks to understand the pathways by which divergent homologous sequences form GCRs and what properties of these sequences lend themselves to such rearrangements.

1.2 Genome instability and genetic pathways involved in GCR suppression

GCRs have been extensively studied in the yeast *Saccharomyces cerevisiae* and numerous pathways have been implicated in both their formation and suppression. These include the various DSB repair pathways such as HR, NHEJ and its variants, the *de novo* telomere addition machinery, as well as other pathways involved in heteroduplex rejection of improper HR

substrates, the DNA replication checkpoint, and short- and long-range resection. HR repair in particular plays an integral role in the formation of repeat-mediated GCRs and consequently will be discussed more in depth.

1.2.1 Mechanisms of DSB repair and homologous recombination

In the simplest models of HR, as described in Symington's 2014 review, a DNA DSB is resected at its 5' ends to leave two 3' single-stranded DNA (ssDNA) tails. One of these tails invades a separate, homologous double-stranded DNA template to form a D-loop that primes new DNA synthesis. The other tail anneals to the DNA strand that was displaced to form the D-loop and is also extended by new DNA synthesis. A double Holliday junction (dHJ) intermediate is formed and must be removed by endonuclease or topoisomerase activity. Depending on how the dHJ is resolved, crossover/gene conversion (GC) or non-crossover (NCO) products will be formed. Two major variations of this model are the synthesis-dependent strand annealing (SDSA) model which favors a higher ratio of NCO products and break-induced replication (BIR) model which results in the non-reciprocal transfer of genetic material to the recipient chromosome [1, 16]. These processes represent some of the mechanisms through which GCRs may arise.

Several proteins function in the homologous recombination pathways or DSB end processing pathways preceding HR [1]. While some are known to be required for HR, others have been shown to possess redundancy with other factors. When a DSB occurs, the MRX complex binds to both DNA strands at the site of the break. This complex consists of Mre11, Rad50, and Xrs2 and is concurrently recruited with the endonuclease Sae2 [17]. Together, these proteins function in DNA end processing (such as removing covalently bound adducts, secondary structure, or other bound proteins). Further resection is carried out by one of two independent pathways involving either the exonuclease Exo1 or the helicase and endonuclease activities of Sgs1 and Dna2, respectively. Resection may proceed at a rate of 4kb/hour and may reveal up to 50kb of DNA sequence during the search for template homology. During end processing and resection, the ssDNA tails are coated with RPA to prevent the formation of secondary structures, but RPA is removed and the DNA is loaded with Rad51 to form the ssDNA filament for strand invasion in a reaction mediated by Rad52. In addition, Rad52 promotes ssDNA strand annealing necessary for second-end capture and this activity is thought to be aided by Rad59 [1].

Proper cleavage of Holliday junction structures is necessary in order to separate the DNA strands for either GC or NCO products. There are several structure-specific endonucleases that cleave these structures, including

Mus81-Mms4 and Yen1 [18]. Most HR mechanisms require some degree of new DNA synthesis and this is often carried out by DNA polymerase δ [19].

Non-allelic homologous recombination may occur when an improper template is selected as a recombination target. Such ectopic recombination may lead to genome instability or chromosome rearrangements if allowed to progress. The heteroduplex rejection process suppresses improper HR through mispair recognition by the proteins Msh2 and Msh6 and subsequent unwinding of mispaired substrates [20-22]. The Sgs1-Top3-Rmi1 complex has also been implicated in untangling mispaired DNA substrates as well as the helicases Mph1 and Srs2 [23-25]. Rad1 and Rad10 are known to clip the non-homologous tails formed during single strand annealing (SSA) [26]. Recently, these tails were shown to promote heteroduplex rejection by Msh2 [27]. The same study demonstrated the mismatch tolerance of the Rad51 filament that performs strand invasion during HR has been characterized and repair was still observed in substrates containing a mismatch every 6 bases.

1.2.2 Other pathways of DSB repair

Aside from ectopic HR, other repair pathways may heal broken chromosomes and give rise to GCR events. The NHEJ pathways fuse broken chromosomes to other DSBs elsewhere in the genome, often leading to interstitial deletions (if a portion of the same chromosome is captured) or translocations. Alternatively, a terminally deleted chromosome may be capped

with a *de novo* telomere at the site of the break by telomerase and other telomere maintenance proteins [28]. These GCRs are formed by telomerase targeting TG “seed” sequences and extending them into functional telomeres [29].

1.3 GCR assays and genetic methods

1.3.1 Genetic assays in S. cerevisiae

Several genetic assays have been developed in haploid *S. cerevisiae* to study GCRs and the genetic pathways through which they form. These canonical “GCR assays” are based on the insertion of the counter-selectable markers *CAN1* and *URA3* onto the same chromosome arm (generally as a single gene cassette) beyond the most telomeric essential gene, which confer sensitivity to the drugs canavanine and 5-fluoroorotic acid, respectively [3, 30]. Cells containing the *CAN1-URA3* gene cassette cannot survive when grown in the presence of these drugs. However, when cultures of cells are grown under non-selective conditions and subsequently plated on media containing the two drugs, we find that Can^r 5-FOA^r survivors have undergone large-scale chromosome rearrangements involving the portion of the chromosome arm containing the two markers (not two independent point mutations) [3]. Importantly, the properties of the DNA sequence beyond the most telomeric essential gene and preceding the *CAN1-URA3* cassette can influence the types of rearrangements observed if it contains homology to the rest of the

genome because it can be used as a template for repair by HR [5]. This portion of the chromosome arm is called the breakpoint region because rearrangements involving this region of the chromosome arm presumably have undergone some form of DSB leading to the GCR.

The mechanisms by which GCRs form are the subject of various studies including those described here, but one common mechanism of formation is the occurrence of a DSB and subsequent repair by HR. For this reason haploid *S. cerevisiae* is a particularly useful model system to study GCR formation because the cells possess only one copy of each of the 16 chromosomes; that is, no homologous chromosome exists as a template for repair. In the absence of the homologous chromosome pair, DSB repair is necessarily ectopic and all of the survivors that are recovered have undergone GCRs in lieu of faithful repair, allowing us to study GCR formation specifically.

1.3.2 A selection of previously existing GCR assays

The first GCR assay that was characterized extensively contained *CAN1* in its native locus on chromosome 5 while *URA3* was inserted into the distal *HXT13* gene [3, 30]. This assay is notable in that its breakpoint region does not possess homology to any other genomic loci and it is known as the unique GCR or uGCR assay. The primary types of GCR structures observed in this assay are products of NHEJ or *de novo* telomere addition.

By moving the *CAN1* and *URA3* markers on chromosome 5 to a more telomeric location beyond the *DSF1* and *HXT13* genes, the first "duplication-mediated" GCR (or dGCR) assay was developed [5]. The S. cerevisiae genome contains a few nearly identical copies of the *DSF1-HXT13* region, similar to segmental duplications in mammals. This assay tested the influence of sequence homology in the breakpoint region of a GCR assay and the rearrangements observed are primarily ectopic homology driven. Another assay containing a Ty retrotransposon in the breakpoint region was developed and characterized [12, 13]. This assay possessed almost 250 potential targets for non-allelic recombination and all rearrangements observed were mediated by the Ty element.

1.3.3 Two novel *S. cerevisiae* assays

Here we characterize the genetic requirements and structures of GCRs in two novel assays on the right and left arms of chromosomes 14 and 15, respectively. These assays were designed to test the influence of two different regions of homology within their breakpoint regions. The chromosome 14 assays possesses approximately 6kb of very good (over 90%) homology to chromosomes 9 and 10 (identical to each other in this region) and a shorter tract of good (80%) homology to chromosome 3. The homology region of the chromosome 15 assay is about 10kb in length and is nearly identical (almost 100% homologous) to regions of chromosomes 9 and 10 (different regions

than those targeted by the chromosome 14 assay but also identical to each other). The rearrangements observed in these assays are primarily homology-driven.

1.4 Methods of structural analysis of GCRs

Many techniques have been used to elucidate the structure of GCRs. Structural information about chromosome rearrangements is valuable for providing mechanistic insight into they form. Analysis of GCR structures can implicate distinct features of chromosomal context or DNA sequences in GCR formation. Methods that have been used to analyze the structures of GCRs in the past include PCR mapping along the chromosome arms, pulsed-field gels, Southern blots, aCGH, multiplex ligation-dependent probe amplification (MLPA) and next-generation whole genome sequencing [31]. Since analysis of a large number of rearrangements is often required to draw conclusions about the mechanisms by which GCRs form, high-throughput technologies are particularly useful for this work but sometimes bear a higher cost.

1.4.1 PCR mapping

PCR mapping may be used to detect the presence or absence of a region of a chromosome arm and consequently indirect evidence that a GCR has occurred [11]. A series of primers are designed to generate short (~500bp) PCR products tiling along a given chromosome arm between the telomere and the most telomeric essential gene. Once the region containing

the putative breakpoint is narrowed down by PCR mapping, additional primers can be designed to span the breakpoint fusion junction. Breakpoint junctions are often challenging to PCR amplify both in terms of template sequence/ slippage and length, but a variety of specialty kits and enzymes may aid amplification.

1.4.2 Pulsed-field gels (CHEF)

Pulsed-field gels (often clamped homogenous electric field or CHEF gels) may be used to separate individual chromosomes. Reasonably large changes in chromosome size corresponding to GCRs may be detected by ethidium bromide staining. Additionally, Southern blotting with a probe specific to one chromosome may be used to assess changes in size associated with that particular chromosome to identify GCRs [12, 14, 15, 30].

1.4.3 Array comparative genomic hybridization (aCGH)

Array comparative genomic hybridization is a method in which genomic DNA from a parental or wild-type strain is labeled with a fluorophore of one color and genomic DNA from a GCR isolate is labeled with a fluorophore of a second color before the DNAs are competitively hybridized to a chip containing immobilized oligonucleotides that together encode the entire genome. The ratio of the fluorophore signals at each genomic location reveal changes in copy number such as duplications or deletions. [14, 31, 32]. One limitation of aCGH is that it proves no information about the connectivity of

these changes in copy number, and so the inferred GCR structure must be verified through other methods.

1.4.4 Multiplexed ligation-dependent probe amplification (MLPA)

MLPA is a high-throughput, low cost method to quickly detect copy number changes. A series of pairs of probes are designed and hybridized to genomic loci of interest, and adjacent probes are ligated together. Only those probes which have hybridized next to each other will be able to be ligated together to form a ssDNA product, and only these ligated products can be PCR amplified in a multiplexed reaction using distinct sequences incorporated at their ends along with a fluorescent marker. These ligation products are then separated and quantified using capillary electrophoresis and peaks of fluorescence corresponding to the amount of a given ligation product are plotted. Increases or reductions in peak size correspond to changes in copy number (i.e. amplifications or deletions) and this analysis can thus be used to identify the structures of GCRs [12, 13].

1.4.5 Next-generation sequencing (NGS)

Whole-genome sequencing using next-generation sequencing (NGS) technologies is undoubtedly the most complete method for analysis of GCRs (or any other genetic analysis, for that matter). We have used multiplexed paired-end sequencing to identify the structures of GCRs in a relatively inexpensive and high-throughput manner in the past [15]. However, NGS is a

developing technology and a robust bioinformatics analysis pipeline is critical to its success. Challenges include deconvoluting repetitive or nearly identical sequences and extracting precise breakpoint sequences for GCRs from the datasets.

1.5 Summary and aims

This work seeks to understand the complexity of genetic requirements and GCR structures formed in a variety of assays designed to probe different kinds of divergent homologous DNA sequences in the genome. Here, we characterize two novel GCR assays and investigate the genetic pathways that underlie GCR formation in these strains. We also describe the spectrum of chromosome rearrangements formed in a multitude of mutant backgrounds in order to prove mechanistic insight into the role that breakpoint homology plays in recombination target selection. We next describe a large-scale screen to understand the genetic networks that underlie repeat-mediated GCRs as compared to those formed between unique sequences [33]. We follow up with an in-depth analysis of the GCR structures observed in the uGCR and dGCR assays with a mutation in *TOR2*, a homolog of the mammalian mTOR protein, which regulates a cell- growth signaling pathway implicated in many cancers. Genome instability pathways and rearrangement mechanisms play a profound role in many human cancers and genetic diseases and understanding the

properties of divergent homologous sequences that lend themselves to rearrangements has vast implications for human health.

Chapter 2

Genetic requirements and structural analysis of two novel GCR assays to study divergent homologous sequences in *S. cerevisiae*

2.1 Summary

DNA double strand breaks are a major source of genome instability and must be properly repaired in order to maintain cell viability. Normally, DSBs in *S. cerevisiae* are repaired via the homologous recombination (HR) pathway that uses an intact copy of the broken chromosome as a template for repair. Unfaithful repair by HR or other DNA repair pathways may lead to gross chromosomal rearrangements (GCRs). We developed two new genetic assays to test the influence of two different regions of divergent homology on GCR formation. We discovered that the homology properties of the divergent repetitive sequences influenced both the genetic requirements and structures of GCRs observed.

2.2 Introduction

An estimated 50 to 80% of the human genome contains repetitive elements [34, 35] and an estimated 4% is comprised of segmental duplications [36] which are regions greater than 1 kb in size with over 90% sequence identity to at least one other genomic locus. The presence of multiple homologies complicates the role of homologous recombination during the repair of DNA damage. Allelic homologous recombination between sister chromatids or homologs will restore the structure and sequence of the damaged chromosome. In contrast, recombination between non-allelic sites has been implicated as a mechanism leading to both pathogenic and non-

pathogenic changes in copy number [37-39], as well as more complex rearrangements that have been identified in inherited diseases and cancer [40-42]. Challenges in the study of these rearrangements in human systems are that most analyses are by necessity restricted to product analysis and that mechanistic features of the process cannot be easily probed.

Quantitative genetic assays that detect gross chromosomal rearrangements (GCRs) in *Saccharomyces cerevisiae* have provided considerable information about the spontaneous formation of genome rearrangements [31]. In the most studied forms of these assays, two counter-selectable markers, *CAN1* and *URA3*, are placed onto a non-essential end of a chromosome in a haploid *S. cerevisiae* strain [3]. GCRs are selected for using the drugs canavanine (Can) and 5-fluoroorotic acid (5FOA). The rate of gene inactivation of each gene is around 10^{-6} per cell per generation in strains proficient for DNA mismatch repair [30]. Thus, selection for two independent gene inactivation events occurs at a rate of 10^{-12} per cell per generation. In contrast, the rate of accumulating larger chromosomal rearrangements that lead to loss of *CAN1* and *URA3* is several orders of magnitude higher, depending on the GCR assay [3, 5, 12, 33]. Analysis of the interactions between different mutations have provided insight into the pathways that suppress spontaneous GCRs and how these pathways interact [31].

The types of GCRs isolated in these assays are highly dependent upon the nature of the sequences in the GCR breakpoint region [5]. The breakpoint

region is the chromosomal region between the counter selectable markers and the most centromeric essential gene. Chromosomal features studies have included non-repetitive sequence, a retrotransposon homology, and one segmental duplication region with fairly restrictive homology choices [3, 5, 12]. A key limitation of previous studies has been the difficulty of characterizing the structures of the GCRs, which has limited the ability to analyze the effect of different mutants.

Here we have developed and probed additional GCR assays containing segmental duplication-like regions that have a variety of non-allelic homologies that could be utilized as targets during homologous recombination. In addition, we have extensively analyzed the products of these rearrangements using next generation whole-genome sequencing and have developed strategies to perform these analyses in the context of non-uniquely mapping regions of the genome.

2.3 Results

2.3.1 Construction of the Chromosome 14R and Chromosome 15L GCR assays

Most GCR assays to date have been constructed in non-essential chromosomal termini that lie between the telomere and the most telomeric essential gene [3, 5, 12, 28]. To develop GCR assays that would probe genomic regions with homology to multiple genomic loci, we first analyzed the

yeast genome for non-essential chromosome termini containing three features: (i) a relatively large length (>20 kb), (ii) a region of homology to more than one other chromosomal locus, and (iii) the presence of "unique" regions, *e.g.* regions that lacked strong homology to other regions of the genome, flanking the homology. Candidate regions were identified and used to construct GCR assays on the right arm of chromosome 14 (herein called the 14R-GCR assay) and the left arm of chromosome 15 (herein called 15L-GCR assay) (Figure 2.1A, 2.1B).

The ~60kb candidate region for the 14R-GCR assay was bounded by the essential gene *ESF2* (*YNR054C*) and the right telomere (*TEL14R*) (Figure 2.1A). Two homologies were found in this region. The first homology contained Ty-derived sequences (*YNRWsigma4*, *YNRCdelta8*, and *YNRCdelta9*) and a tRNA-encoding gene (*tL(UAA)N*). This first homology was removed by replacing this region with a nourseothricin-resistance marker (Figure 2.1A). The second homology, primarily made up of the genes *YNR065C, YNR066C* and *DSE4*, was retained for the GCR assay. The *YNR065C-DSE4* duplication exhibits ~6 kb of homology with chromosomes 9L and 10L, with ~3 kb being nearly identical, and ~3 kb of homology to chromosome 3 (Figure 2.1C). Homology-mediated rearrangements between the *YNR065C-DSE4* duplication and chromosomes 3, 9 or 10 were predicted to generate monocentric translocation products. In the 14R-GCR assay strains, a *CAN1/URA3* cassette

for selection of strains containing GCR events was inserted telomeric to the *YNR065C-DSE4* duplication in the *BSC5* (*YNR069C*) gene (Figure 2.1A).

The ~45kb candidate region for the 15L-GCR assay was bounded by the essential gene *DCP1* (*YOL149W*) and the left telomere (*TEL15L*) (Figure 2.1B). This region contained one substantial homology consisting of the *IMA2* and *HXT11* genes and 3' end of the *HPF1* gene (Figure 2.1D). The ~8 kb *IMA2-HPF1* duplication has nearly perfect homology to ~8 kb of chromosome 9L and ~5 kb of chromosome 10L. In the 15L-GCR assay strains, a *CAN1/URA3* cassette was inserted telomeric to the *IMA2-HPF1* duplication in the *ENB1* (*YOL158C*) gene (Figure 2.1B).

2.3.2 The 14R-GCR and 15L-GCR assays have increased GCR rates relative to GCR assays containing short homologies or no homologies

Previously constructed GCR assays have shown that the introduction of sequences with homologies to other regions of the genome into the breakpoint region increases the GCR rate more than simply adding additional single copy sequences [3, 5, 12, 33]. Consistent with this, wild-type strains containing either the 14R-GCR or 15L-GCR assay had relatively high rates of accumulating GCRs (Table 2.1). These were 74- and 253-foldhigher than the original GCR assay [3] and were more similar to those of the duplication-containing dGCR and TyGCR assays [5, 12].

2.3.3 Homology-mediated GCRs dominate the rearrangements selected in the 14R-GCR assay

To characterize the Can^R and $5FOA^R$ strains selected using the 14R-GCR assay, we initially characterized the GCR-containing chromosome using PCR mapping. A series of 14 primer pairs were designed to hybridize to unique regions of chromosome 14 and to generate ~500bp PCR products along ~40kb of the breakpoint region of chromosome 14 (Figure 2.2). Which PCR products could be amplified from the genomic DNA of the GCR-containing strain and which PCR products could not be amplified narrowed down the position of the breakpoint. For example, in isolate SNBG836, PCR products could be generated only for primer pairs #1-7, indicating that the breakpoint was adjacent to *YNR065C* within the region that also had homology to chromosomes 9 and 10 (Figure 2.2). To verify that the rearrangement in SNBG836 was due to a t(14;9 or 10) translocation in the *YNR065C-DSE4* duplication, we designed primers to span the predicted t(14;9 or 10) chromosome fusion junction and were able to successfully amplify the t(14;9 or 10) translocation junction. PCR mapping also identified homeology-mediated translocations involving chromosome 3, as evidenced in GCR isolate SNBG842, where PCR mapping indicated a larger portion of chromosome 14 was retained than in SNBG836 (primer pairs #1-11 were able to generate products) and the breakpoint region was narrowed to the vicinity of *YNR066C.*

Subsequent analysis and primer design (SNB183 + SNB209) identified a t(14;3) translocation in this isolate.

Isolates that were difficult to characterize by PCR mapping and/or failed junction amplification were subjected to paired end whole genome next-generation sequencing (NGS) to characterize the structure of the rearranged chromosome 14. These isolates were a combination of those with homology-mediated t(14;9 or 10) and t(14;3) translocations that failed PCR testing for technical reasons as well as other types of unexpected rearrangements. Three types of evidence are available from paired end sequencing to indicate the presence of genomic rearrangements: (1) changes in the copy number (e.g. read depth at each base), (2) the existence of “non-concordant” read pairs that span breakpoint junctions and have aberrant mapping relative to the reference genome, and (3) the existence of read pairs in which one read maps next to the non-concordant read pairs and the other cannot be mapped as it contains the novel sequence of the breakpoint junction.

Junctions mediated by homologies, however, are more complicated to deconvolute from the NGS data than junctions present in single copy regions of the genome. The lack of uniquely mapping reads makes identifying non-concordant reads difficult, and the homologies used to form the junctions means that the novel junctions do map to the reference genome at multiple positions. Previous analyses in the Kolodner laboratory have taken advantage of the paired-read nature of the data; reads mapping uniquely to one side of

the homology have been used to identify reads within the crossover-containing homology. When these crossover-containing reads are aligned, the sequence of the junction can be determined. This strategy, however, also fails when the homology lengths are greater than the length of the fragments in the library used to generate the read pairs, and the 3 to 6 kb homology lengths in the 14R-GCR breakpoint region are far larger than the size of the fragments generated during library production.

To characterize the types of rearrangements isolated in the 14R-GCR assay using the NGS data, we took advantage of being able to analyze copy number information, which was derived by determining the read depth at each base. The relative copy number can be calculated at each position by dividing the read depth at each position by the median read depth for all of the uniquely mapping regions of the nuclear chromosomes. Plots of the uniquely mapping copy number revealed that the majority of the GCRs isolated in the 14R-GCR assay had a loss of terminal region of the right arm of chromosome 14 starting at the *YNR065C-DSE4* repetitive region (Figure). In the parental strain lacking a GCR, the terminal region of chromosome 3R telomeric to the *YCR099C-YCR101C* (coordinates 302-307 kb) possessed a single copy, whereas it was duplicated in t(14;3) translocations (Figure 2.6).

For t(14;9 or 10) translocations, the target sequences lie within nearly identical terminal regions of chromosome 9L and 10L, thus duplications cannot be observed in copy number plots of uniquely mapping regions. We therefore

developed a method to count the read depth for multiply mapped regions; this involved ensuring that the short read mapping program reported all possible mapping positions and that the read depth counts were incremented at each mapped position for each read. For uniquely mapping regions, this multiply mapped read depth is the same as the read depth derived from the uniquely mapping reads. In the parental strain lacking a GCR, the terminal regions of both chromosomes 9L and 10L (coordinates 0-22 kb) have an average of two copies in the multiply mapped read depth plots, whereas this region has three copies in t(14;9 or 10) translocations (Figure 2.5).

The majority of rearrangements in the 14R-GCR assay were either monocentric t(14; 9 or 10) translocations (59 of 98; Figure 2.3A) or monocentric t(14;3) translocations (34 of 98; Figure 2.3A). For the t(14;9 or 10) translocation products, the near perfect identities between chromosomes 9 and 10 homologies precluded identification of the precise target. In addition to the predicted products, some additional rearrangements were observed. One t(14;5) translocation involved recombination between *URA3* in the chromosome 14 *CAN1/URA3* cassette and *ura3-52* on chromosome 5 (Figure 2.4). One recovered translocation was a t(14;5;14) tripartite rearrangement that caused a 100 kb deletion of chromosome 14 and was mediated by a chromosome 14 *URA3*/chromosome 5 *ura3-52* recombination event on the centromeric side and a chromosome 14 *MAN2*/chromosome 5 *DSF1* recombination event on the telomeric side; regions on chromosome 5 between

ura3-52 and *DSF1* were duplicated, indicating that the strain had an intact copy of chromosome 5 in addition to the t(14;5;14) rearrangement (Figure 2.4). Finally, one t(14;15) translocation had the inserted *nat* marker fused to the left arm of chromosome 15; this non-reciprocal translocation copied from the junction position (near the *MHF1* gene) to the left telomere (Figure 2.4).

2.3.4 Homology-mediated GCRs dominate the rearrangements selected in the 15L-GCR assay

As for the 14R-GCR assay, rearrangements in the 15L-GCR assay were identified using a combination of PCR mapping, junction amplification, and/or NGS sequencing. In the NGS data, t(15;9 or 10) rearrangements had a characteristic deletion of the left arm of chromosome 15 telomeric to the *IMA2-HPF1* homology present in unique copy number plots and an increase in the multiple copy number plots of chromosomes 9 and 10 from 2 to 3. The majority of the characterized GCRs isolated in the wild-type 15L-GCR assay (69 of 71) were t(15; 9 or 10) rearrangements (Figure 2.3B; Figure 2.4; Figure 2.5).

In addition to the t(15;9 or 10) rearrangements, we observed two other rearrangements. SNBG912 contained a monocentric t(15;2) rearrangement mediated by a weak homology between the *IMA2* gene on chromosome 15 and the *MAL32* gene on chromosome 2. In contrast, SNBG647 contained a bipartite t(15;9 or 10;15) rearrangement in which the *IMA2-HPF1* homology invaded either chromosome 9 or 10 (Figure 2.1) followed by a reinvasion of

chromosome 15-L using homologies between *PAU20* and *PAU1* (chromosome 10) or *PAU14* (chromosome 9).

2.3.5 Deciphering crossover positions in homology-mediated rearrangements

The homologies involved in the t(14;9 or 10), t(14;3), and t(15;9 or 10) translocations are imperfect and contain some sequence variants and are thus termed “homeologies”. These sequence variations are analogous to Single Nucleotide Polymorphisms (SNPs), but the variants, which we term Homeologous Nucleotide Variants (HNVs), are not at allelic positions and hence are not SNPs. Given the sequence of the junctions, the positions of the crossovers can be localized to lying within regions of identity surrounded by HNVs, assuming that DNA mismatch repair of the joint molecule heteroduplexes is rare.

To identifying the positions of these crossovers, we first PCR amplified the junction-containing regions of the rearranged chromosomes and submitted these PCR products to Sanger sequencing as has been previously performed in a chromosome 5-based GCR assay [5]. Amplifying these junctions in the 14R-GCR and 15L-GCR assays is technically challenging given the repetitive nature of these sequences and the large PCR products being generated.

We also reasoned that the crossover positions, as defined by the HNVs, can be derived from the NGS data. The complication of the NGS data, however, is the need to determine the frequencies of each base at HNV

positions, regardless of which homeology that HNV-associated reads mapped to in the reference genome. Thus, this analysis must modify the traditional analysis strategies whereby reads are only mapped to a reference genome to one where reads are mapped to an alignment of the homeologies (Figure). This new method, as implemented in the program rucola, takes reads mapped to the reference genome by standard read mapping software and then applies them to a prebuilt alignment of the homeologies. For each position within the alignment, the total read depth and the frequency of the observed bases are determined. Differences between the experimental and predicted base frequencies at HNV positions can be converted to a p-value using the chi-squared test. These predicted frequencies can derived from the alignment itself, a reference sample, or modified forms of the alignment. In cases where the predicted frequencies correspond to the parental strain lacking a GCR, HNV frequencies with p-values near 1.0 indicate an unaltered HNV distribution, whereas low p-values indicate positions affected by GCR events.

Analysis of the crossover positions in both the t(14;9 or 10) and the t(14;3) rearrangements (Figure) reveals a number of interesting features. First, the largest stretches of identities between HNV positions do not dominate the crossover positions. Second, the distribution of the crossover positions are distributed across the alignment rather linearly, which is consistent with the notion that the crossover positions are relatively random within the homeologies. Third, there are very few cases in which HNVs switch back and

forth between chromosome 14 and the target chromosome(s), which is consistent with the observations of the wild-type crossovers in the chromosome 5 duplication assay [5].

2.3.6 Mutations affecting homologous recombination have differing effects in the 14R-GCR and 15L-GCR assays

Given that the GCRs formed in the wild-type 14R-GCR and 15L-GCR appeared to be products of homeologous recombination, we tested the effect of deleting genes involved in homologous recombination. In *S. cerevisiae*, most recombination genes primarily act downstream of *RAD52*, which encodes a protein that stimulates the interchange of the single-strand binding protein RPA for the strand exchange protein Rad51 on single-stranded DNA [43]. In previous duplication-mediated GCR assays, loss of *RAD52* suppressed the GCR rate due to the role of Rad52 in the formation of GCRs [5, 12], whereas it increased the GCR rate in single copy-mediated GCR assays presumably due to its role in suppressing the formation of GCRs by recombination with the sister chromatid [3, 5]. Deletion of *RAD52* in the 15L-GCR assay reduced the GCR rate by around 5-fold; however, deletion of *RAD52* in the 14R-GCR assay had a surprisingly, but modestly, increased GCR rate (Table 2.1).

The effect of single mutations that eliminated the two *RAD52*-dependent subpathways, *rad51Δ* and *rad59Δ*, have mixed effects in previous duplication-mediated GCR assays [3, 5, 28] as well as in the GCR assays

studied here (Table 2.1). Deletion of *RAD51* had no effect on the chromosome 5 dGCR assay or the 14R-GCR assay, increased the rate in the chromosome 5 TyGCR assay, and decreased the rate in the 15L-GCR assay. A similarly complicated scenario was observed for deletion of *RAD59*, which increased the rate in the dGCR assay and the 15L-GCR assay, but had little effect in the TyGCR and 14R-GCR assays. The effect of the individual *rad51Δ* and *rad59Δ* mutations likely reflects the relative importance of the individual recombination subpathways in suppressing formation of GCRs. Remarkably, the *rad51Δ rad52Δ rad59Δ* triple mutant, which should eliminate even *RAD52*-independent recombination reactions [44], increased GCR rate in the 14R-GCR assay by 26-fold, whereas it decreased the GCR rate in all other assays (Table 2.1). We then tested the influence of deleting pairs of genes involved in HR by constructing *rad51Δrad59Δ*, *rad51Δrad52Δ*, and *rad52Δrad59Δ* double mutants. Although the *rad51Δrad59Δ* double mutant should eliminate both canonical *RAD52*-dependent subpathways of HR, we still observed a small level of GCR formation independent of these pathways in the various assays. The *rad51Δrad59Δ* mutation suppressed GCR formation in the 14R-GCR assay and the Chr5 dGCR assay, with minimal effect in the 15L-GCR and TyGCR assays. Deleting *RAD52* in conjunction with either *RAD51* or *RAD59* had consistent but disparate effects among the assays. Both the *rad51Δrad52Δ* mutation and the *rad52Δrad59Δ* mutation elevated the GCR

rate substantially in the 15L-GCR and chromosome 5 dGCR assay while possessing almost no effect in the 14R-GCR assay (Table 2.1).

2.3.7 Mutations affecting the DNA Replication Checkpoint and BIR have differing effects in the 14R-GCR and 15L-GCR assays

Since it has been well characterized that translocations between divergent repetitive sequences often proceed through break-induced replication (BIR) [45-47], we then tested the impact of two genes involved in this process, *POL32*, a subunit of the DNA polymerase δ utilized in BIR, as well as the helicase *MPH1*. Similar to what was observed for the chromosome 5 dGCR assay, the *pol32Δ* mutation had little or minimal effect, while *mph1Δ* raised GCR rates across the assays (Table 2.1). Finally, we measured GCR rates in strains with mutations in genes involved in the DNA replication checkpoint. Both *mrc1Δ* and *mec1Δsml1Δ* increased GCR rates across a variety of strain backgrounds, with more prominent effects in the chromosome 5, TyGCR and 14R-GCR assays, and a more subtle effect in the Chr15-L GCR assay [5, 12]. Furthermore, a *POL32* mutation impacting BIR diminished the amount of t(14; 9 or 10) products selected in the 14R-GCR assay but did not exhibit this effect in the 15L-GCR assay (Figure 2.3).

2.3.8 Role of heteroduplex rejection in the control of product formation

Non-allelic recombination between homeologous sites is known to be suppressed by heteroduplex rejection, which acts through the recognition of

mismatches in the nascent heteroduplex by Msh2-Msh6 or Msh2-Msh3 leading to disassembly or degradation of the intermediate by a pathway that includes Sgs1 [25, 48-50]. Deleting *MSH2* or *SGS1* caused increased GCR rates in previous studies examining the Chr5 dGCR assay [5] and TyGCR assay [12, 13] however, the rate increase in *sgs1Δ* strains was much higher than in *msh2Δ* strains likely due to additional roles of Sgs1 outside of heteroduplex rejection in suppressing genome instability or different roles in heteroduplex rejection [23].

In the 14R-GCR assay, the *sgs1Δ* mutation caused a substantial increase in the GCR rate, whereas the *msh2Δ* mutation had little effect (Table 1). Both mutations did, however, shift the distribution of observed GCR products from 59% t(14;9 or 10) in wild-type to 69% and 71% in the *msh2Δ* and *sgs1Δ* strains, respectively (Figure 2.3A). A similar lack of effect for *msh2Δ* was observed in the 15L-GCR assay (Figure 2.3B).

Sequencing the breakpoint junctions and analyzing the homeologous nucleotide variants (HNVs), which are the differing nucleotide positions related by the imperfect homology, identifies the regions containing the breakpoints between homeologous sequences. The HNVs for the breakpoint can be identified either by specific amplification of the junction followed by Sanger sequencing or by analysis of whole-genome next generation sequencing using alignments of the homeologous regions (see Materials and Methods Section 2.5.4 and 2.5.5). Analysis of 20 GCR isolates from the chromosome 5 dGCR

assay background showed a fairly linear distribution of breakpoints located only within the ~4kb *HXT13-DSF1* homology region (Figure 2.10) [5]. For both the t(14;9 or 10) and t(14;3) junctions recovered from the wild-type 14R-GCR assay strain (Figures 2.7A and 2.8A), the cumulative distribution of breakpoints was fairly linear over a ~3 kb and ~1 kb region of the homeologies, respectively (Figure 2.8). This is consistent with a fairly random distribution of the breakpoint position across the chromosome 14 homeology at a macroscopic level. Distributions similar to wild-type were observed for both types of rearrangements in the *msh2Δ* 14R-GCR assay strain (Figures 2.7C and 2.8C). This argues that the effect of heteroduplex rejection in the 14R-GCR assay causes a bias towards the better target, but plays a lesser or no role in controlling which microhomology is used within the larger homeologous region. In contrast the sequenced junctions recovered from the *sgs1Δ* 14R-GCR assay showed a rather different distribution, most strikingly for the t(14;9 or 10) junctions, that are biased towards better local microhomologies within the larger homeologous region (Figures 2.7B and 2.8B), while the junctions from the *sgs1Δ* chr5 dGCR opposite showed the opposite (Figure 2.9).

2.3.9 The presence of a replication origin telomeric to the markers in the 14R-GCR assay does not explain the effect of recombination mutants

The fact that HR defects give rise to increased GCR rates in the 14R-GCR assay indicates that the 14R-GCR assay is different than the other duplication-mediated GCR assays in that HR likely plays a role in avoiding the formation of rearrangements or preventing DNA damage that might cause rearrangements. We therefore investigated potential chromosomal features that might distinguish the 14R-GCR assay. An analysis of the replication timing profiles of the *S. cerevisiae* genome [51] revealed that unlike the other duplication-mediated assays, the 14R-GCR assay had an origin that was telomeric to the *CAN1/URA3* marker genes that fired early enough that replication forks either within or slightly centromeric to the 14R-GCR breakpoint region. We reasoned that the effect of HR defects could be explained if HR played a role in suppressing DNA damage when two DNA replication forks met. We therefore constructed a 14R-GCR strain with a deletion of an unannotated ARS at 730kb and compared the GCR rate to the *rad52Δ* mutant (Table 2.1). We found that deletion of the telomeric origin did not resolve the increased GCR rate caused by loss of HR in the 14R-GCR, suggesting another role for HR in suppressing GCR formation in this assay.

2.3.10 Overexpression of RNase H1 does not diminish genome instability in the absence of HR in the 14R-GCR assay

Enzymes that remove RNA:DNA hybrids such as RNase H1 and H2 are known to suppress genome instability in *S. cerevisiae* [52-56] One explanation for this observation is that these enzymes remove R-loops (consisting of a DNA:RNA duplex and a third displaced DNA strand) that can lead to chromosome instability. Overexpression of RNase H1 has been shown to reverse genome instability caused by accumulation of R-loops [57-60].

We wondered if the GCR rate increases observed in the 14R-GCR HR double and triple mutant backgrounds corresponded to an increase of R-loops in these strains in the absence of functional HR repair pathways. Results from a plasmid-based assay indicated that RNase H1 overexpression did not substantially alter the 14R-GCR rate in the wild-type or *rad51Δrad52Δrad59Δ* background (Figure 2.11).

2.4 Discussion

In this study we found that characteristics of divergent homologous sequences influence the genetic requirements for the formation of chromosome rearrangements, as well as the types of structures of rearrangements observed. We recovered a variety of homology-driven GCRs from the two newly developed 14R- and 15L-GCR assays. Genetically, the 15L-GCR assay behaved more similarly to the previously existing chromosome dGCR and TyGCR assays [5, 12] while the 14R-GCR assay was less consistent. In particular, the behavior of strains with mutations in the

various HR pathways was surprising as the GCR rate increased in both the 14R-GCR *rad52Δ* and *rad51Δrad52Δrad59Δ* mutant strains.

While homology-driven translocations dominated in the 14R-, 15L- and chromosome 5 dGCR assays, mutations in certain pathways shifted the spectrum of GCRs observed substantially. Defects in HR led to the accumulation of products of *de novo* telomere addition in the Chr14-R assay, as observed in other assays. This is unsurprising because in the absence of HR, we do not expect the sister chromatid or homeologous non-allelic loci to be used as a template for repair. More interestingly, we found that mutations in the heteroduplex rejection pathway (*sgs1Δ* and *msh2Δ*) shifted the GCR product spectrum towards the products formed with the most homologous recombination targets, while we would expect the loss of heteroduplex rejection to promote recombination with worse targets. However, subsequent analysis of the breakpoint junction sequences for translocations formed in these mutant backgrounds revealed use of less homologous regions of the "better" recombination targets. This could be explained by a propensity for DSBs to occur more frequently in certain chromosomal regions, or in the case of the *sgs1Δ* mutant, the complicated balance between its roles in heteroduplex rejection and long-range DSB resection.

One intriguing aspect of genome instability revealed by the 14R-GCR results is that different pathways of GCR formation may be masked in certain genetic backgrounds. Although we didn't recover products of *de novo* telomere

addition in the wild-type 14R-GCR assay, these products comprised nearly 75% of the products recovered in the 14R-GCR *rad51Δrad52Δrad59Δ* triple mutant strains and were observed at a 26-fold increase in rate compared to the wild-type. This suggests that in the wild-type background, repair through active HR pathways may often successfully heal broken chromosomes without causing formation of GCRs, although GCRs are still observed (albeit at a lower rate). In the absence of canonical HR, the same amount of damage can only be repaired unfaithfully and thus it is more “visible” in the form of a majority of *de novo* telomere addition products observed at a higher rate. Understanding the propensity of the breakpoint region to sustain DNA damage and the common sources of the damage could shed further light on this phenomenon.

One potential cause of DNA damage that we examined in the 14R-GCR assay was RNA:DNA hybrids or R-loops, which are recombinogenic and known to block incoming replication forks [59, 61-63]. We found that overexpression of RNase H1, which specifically cleaves R-loops [54, 60] was insufficient to decrease the rate of accumulation of GCRs in the 14R-GCR wild-type and 14R-GCR *rad51Δrad52Δrad59Δ* strains. Therefore it seems likely that the major source of damage in this assay is incurred through a different mechanism.

2.5 Materials & Methods

2.5.1 Plasmid construction

The *CAN1/URA3* cassette was digested out of pRDK1378 in RDK5311 with NheI and gel purified. For the chromosome 14 integration site, an integration construct was generated by subcloning the NheI fragment into a plasmid containing the target genes *YNR068C-BSC5* (Chr14 759529-761607), which were amplified by PCR and cloned into pCR2.1-TOPO (Invitrogen) to generate RDK5418. Subcloning the *CAN1/URA3* cassette into the NheI site within *YNR068C* generated RDK5423 and was verified by sequencing. For the chromosome 15 integration site, an integration construct was generated by PCR amplification of the gene and flanking regions of *ENB1* (Chr15 19491-21311) which was then cloned into pCR2.1-TOPO (Invitrogen) to generate RDK5416. Site-directed mutagenesis was used to introduce an NheI cut site into *ENB1* to generate RDK5420. The *CAN1/URA3* cassette was subcloned into the NheI site to generate RDK5429 and verified by sequencing.

The RNase H1 overexpression plasmid pCM184 RNH1 was a gift from Andrés Aguilera (Santos-Pereira et al., 2013). The *TRP1* selection marker was replaced with a *LEU2* selection marker using a standard gap repair protocol to generate RDK5753. The Chr14-R wild-type strain and *rad51Δrad52Δrad59Δ* mutant were transformed with this plasmid and the control plasmid pRS315 and selected for on –LEU plates. For GCR rate determination, transformants were cultured overnight in –LEU liquid media.

2.5.2 Genetic methods

Integration cassettes were amplified from their respective plasmids by PCR and transformed into RDKY5461 (*ura3-52, leu2Δ1, trp1Δ63, his3Δ200, lys2ΔBgl, hom3-10, ade2Δ1, ade8, can1::hisG*) to form the Chromosome 14-R GCR assay (RDKY7706) and the Chromosome 15-L GCR assay (RDKY7734). A nourseothricin-resistance marker (*NAT*) replaced *tL(UAA)N*, *YNRWsigma4* and *YNRCdelta9* on chromosome 14 (Chr14 725830-728145) to remove the possibility of recovering GCRs involving these genetic features. Select mutations were introduced into assay strains with standard PCR based mutagenesis techniques using deletion primers with 50 bp of flanking homology to the target of interest. Transformations were performed with standard lithium acetate protocols and all strains were verified by sequencing. GCR rates for constructed strains were determined for multiple independent isolates using fluctuation tests as previously described [64].

2.5.3 Homology analysis

Percent homology between the chromosome 14 and 15 assay homology regions and their major recombination targets were calculated using a 50bp binning window and plotted along the assay chromosome axis.

2.5.4 Analysis of GCR isolates by PCR

GCR isolates were first characterized by determining the approximate position of deletions in the assay chromosomes by PCR. A series of PCR

reactions were performed to amplify fragments along the breakpoint region, including repetitive portions of the breakpoint region where sequence divergence could be used to design specific primers (Figure 2) [11]). The approximate breakpoint position was then identified by those PCR reactions that did not produce products for the GCR isolate. In almost all cases where this mapping was performed, these approximate breakpoint positions were mapped to the homology regions. In many cases, these novel junctions could by amplified by PCR using primers that spanned the predicted homology-mediated junctions between chromosome 14 and chromosomes 3, 9 or 10 and between chromosome 15 and chromosomes 9 or 10. Breakpoint junctions were amplified using the Roche Expand Long Template PCR system, and unresolved GCR isolates were retained for analysis using next-generation sequencing.

2.5.5 Whole genome sequencing

Multiplexed paired-end libraries were constructed from 2μg of genomic DNA purified using the Puregene kit (Qiagen) and subsequently treated with 0.15mg (45U) RNase A for 1 hour at 37°C. Genomic DNA in Covaris microtube-50 tubes was sheared into 550bp fragments by sonication using a Covaris M220 instrument at peak incident power of 75W, 10% duty factor, 200 cycles per burst, and treatment time of 40s. Samples were then end-repaired using the End-it DNA End-repair kit (Epicentre Technologies) and A-tailed using the Klenow fragment (3'→5' exo-, NEB). Common adaptors from the

Multiplexing Sample Preparation Oligo Kit and TruSeq PCR-Free LT DNA Sample Preparation Kit (Illumina) were then ligated to the genomic DNA fragments using the Quick Ligation Kit (NEB) and then run on an agarose gel to select for 600bp fragments. Samples were then subjected to 18 cycles of amplification using the Library Amplification Readymix (KAPA Biosystems). The amplified products were fractionated on an agarose gel to select 600 bp fragments, which were quantified with the Qubit dsDNA HS assay kit (Thermofisher). Libraries were subsequently sequenced on an Illumina HiSeq 2000 using the Illumina GAII sequencing procedure for paired-end short read sequencing. Reads from each read pair were mapped separately by bowtie version 2.2.1 [65] to a reference sequence that contained revision 64 of the *S. cerevisiae* S288c genome [66], *hisG* from *Salmonella enterica*, and the *kanMX4* marker.

2.5.6 Analysis of GCR structures from sequencing data

Paired-end reads were aligned to the *S. cerevisiae* S288C reference genome release R64.1.1 (February 2011) and R64.2.1 (January 2015) using bowtie-0.12.7. Further analyses to identify copy number changes and identify novel structural variants were performed with version 0.6 of the Pyrus suite (http://www.sourceforge.net/p/pyrus-seq) [15]. Rearrangements relative to the reference S288c genome were identified by analyzing the read depth distributions (Figure), the discordantly mapping read pairs (Figure/ Table), and/or extracting the sequences of the novel junctions (Figure) Associated

junction-sequencing reads, which were reads that did not map to the reference but were in read pairs in which one end was adjacent to discordant reads defining a junction, were used to sequence novel junctions.

While many junctions could be determined using alignments of junction-sequencing reads, for junctions formed by HR between regions of near-perfect homology, the GCR structures often could not be detected because both reads in a given read pair mapped to both chromosomes. In these instances, all reads mapping to a given homology region were analyzed and for every position at which a nucleotide differed between the divergent homologies, the expected ratio of nucleotides mapping to that position was predicted. Changes in the ratio of nucleotides in the homeology region ("homeology nucleotide variants") corresponded to loss or gain of genomic regions involved in a GCR and were thus used to provide indirect evidence for these events that could subsequently be verified through physical methods. Junctions indicated by copy number changes, discordant read pairs, and junction sequencing were identified with a high degree of confidence; however, previous analyses have indicated that even junctions inferred from only copy number changes can be experimentally verified at high frequency [12-15]. Analysis of the sequencing data identified all of the genetic modifications introduced during construction of the starting strains, such as the *his3Δ200* deletion, as well as the molecular features associated with the selected GCRs.

2.6 Acknowledgements

The dissertation author would like to thank Richard D. Kolodner and Christopher D. Putnam for assisting with the conception and design of the research, data analysis, manuscript writing and mentorship. The dissertation author would also like to thank Anjana Srivatsan, Catherine Smith, Eva Goellner, Binzhong Li, Katie Pallis and other members of the Kolodner lab for helpful discussions and comments. This work was also supported by NIH R01 Grant GM26017 to Richard D. Kolodner and Christopher S. Putnam and the Ludwig Institute for Cancer Research. Chapter 2 is in preparation for publication with Richard Kolodner and Christopher Putnam. The dissertation author is the primary author for this paper.

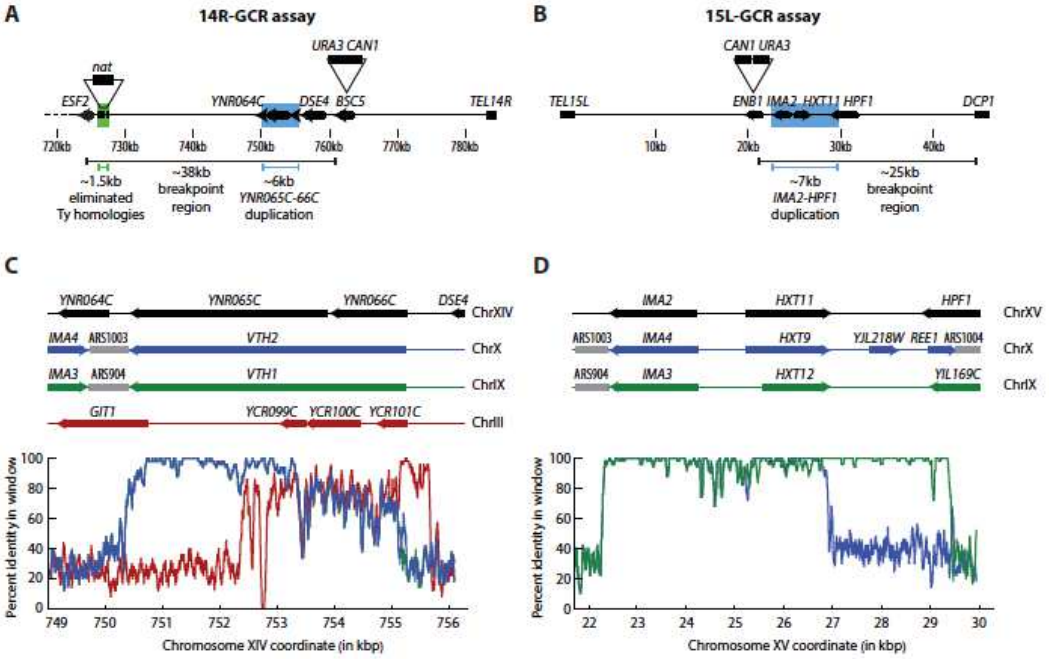


Figure 2.1 Two new assays to study homeology-mediated GCRs

A. The *CAN1/URA3* cassette was inserted into *YNR068C* on the right arm of chromosome 14 and a nearby delta sequence centromeric to the assay cassette was knocked out with a nourseothricin resistance marker (*nat)* in a strain with *ura3-52* and *can1::hisG* mutations. This assay was designed to study GCRs mediated by the ~9kb *YNR065C-YNR066C-DSE4* homology region. B. The *CAN1/URA3* cassette was inserted into *ENB1* on the left arm of chromosome 15 in a strain containing *ura3-52* and *can1::hisG* mutations. This assay was designed to study GCRs involving the ~10kb *IMA2-HPF1* homology region. C. The average percent identity in 50bp windows of the *YNR065C-YNR066C-DSE4* region with regions of chromosomes 9, 10 and 3 is plotted against the chromosome 14 coordinate. D. The average percent identity in 50bp windows of the *IMA2-HPF1* region with regions of chromosomes 9 and 10 is plotted against the chromosome 15 coordinate.

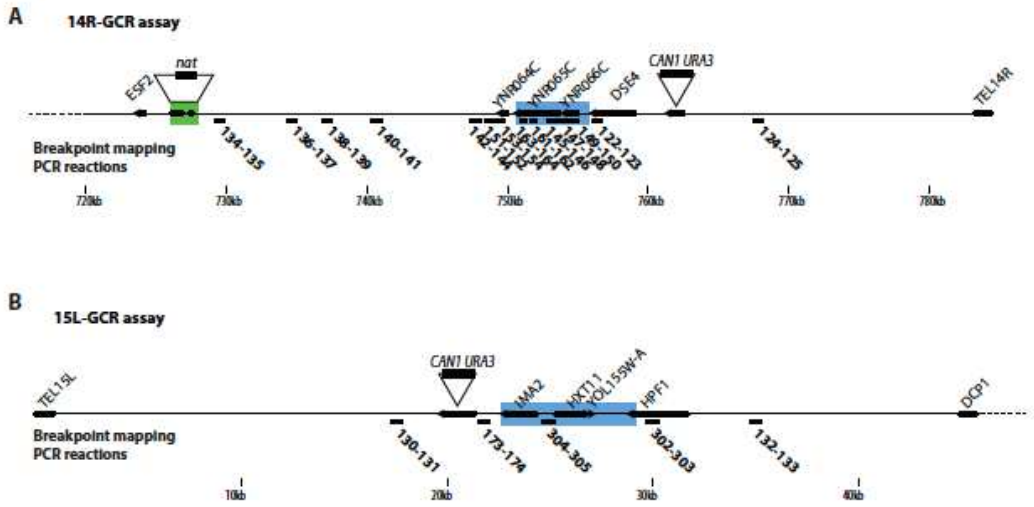


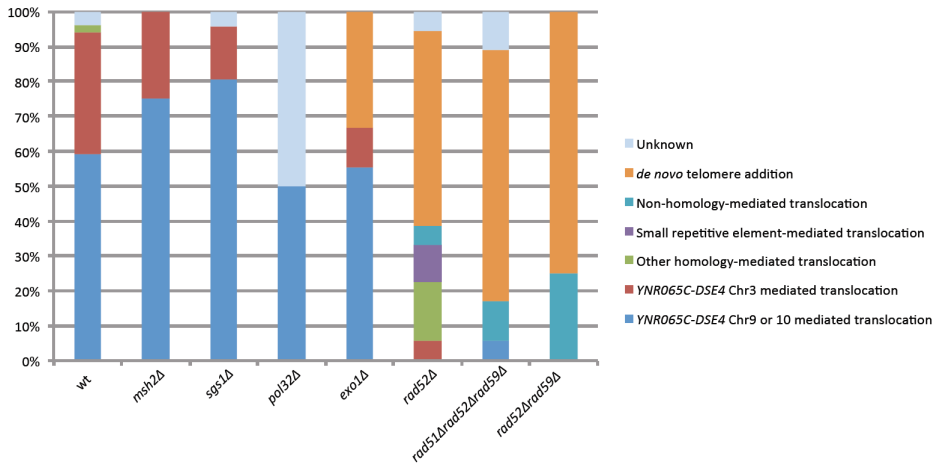
Figure 2.2 PCR mapping of breakpoint regions in GCR assays

A. Primer pairs to generate PCR products (portrayed as short black lines) span the breakpoint region of the Chr14-R GCR assay from ~730-768kb with numeric labels corresponding to SNB primer numbers. Primers were designed to lie in regions of unique sequence. B. Primer pairs to generate PCR products (portrayed as short black lines) span the breakpoint region of the Chr15-L GCR assay from ~18-36kb with numeric labels corresponding to SNB primer numbers. Primers were designed to lie in regions of unique sequence.

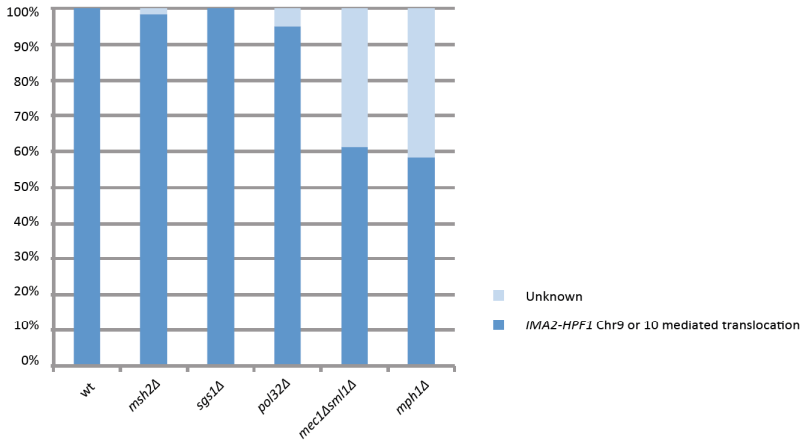
Figure 2.3 Summary of the types of GCRs detected in the Chr14-R and Chr15-L GCR assays

A. Percentage of the different types of GCRs observed in the Chr14-R assay wild-type strain and strains with mutations impacting heteroduplex rejection, BIR, DSB resection and HR. B. Percentage of the different types of GCRs observed in the Chr15-L assay wild-type strain and strains with mutations impacting heteroduplex rejection, BIR, the DNA replication checkpoint and HR.

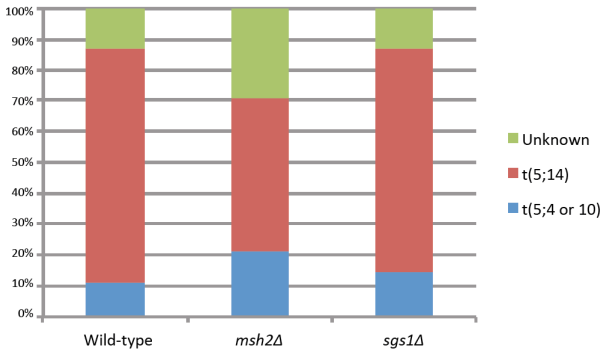
A. **Chr14-R GCR**



B. **Chr15-L GCR**



C. **Chr5 dGCR**



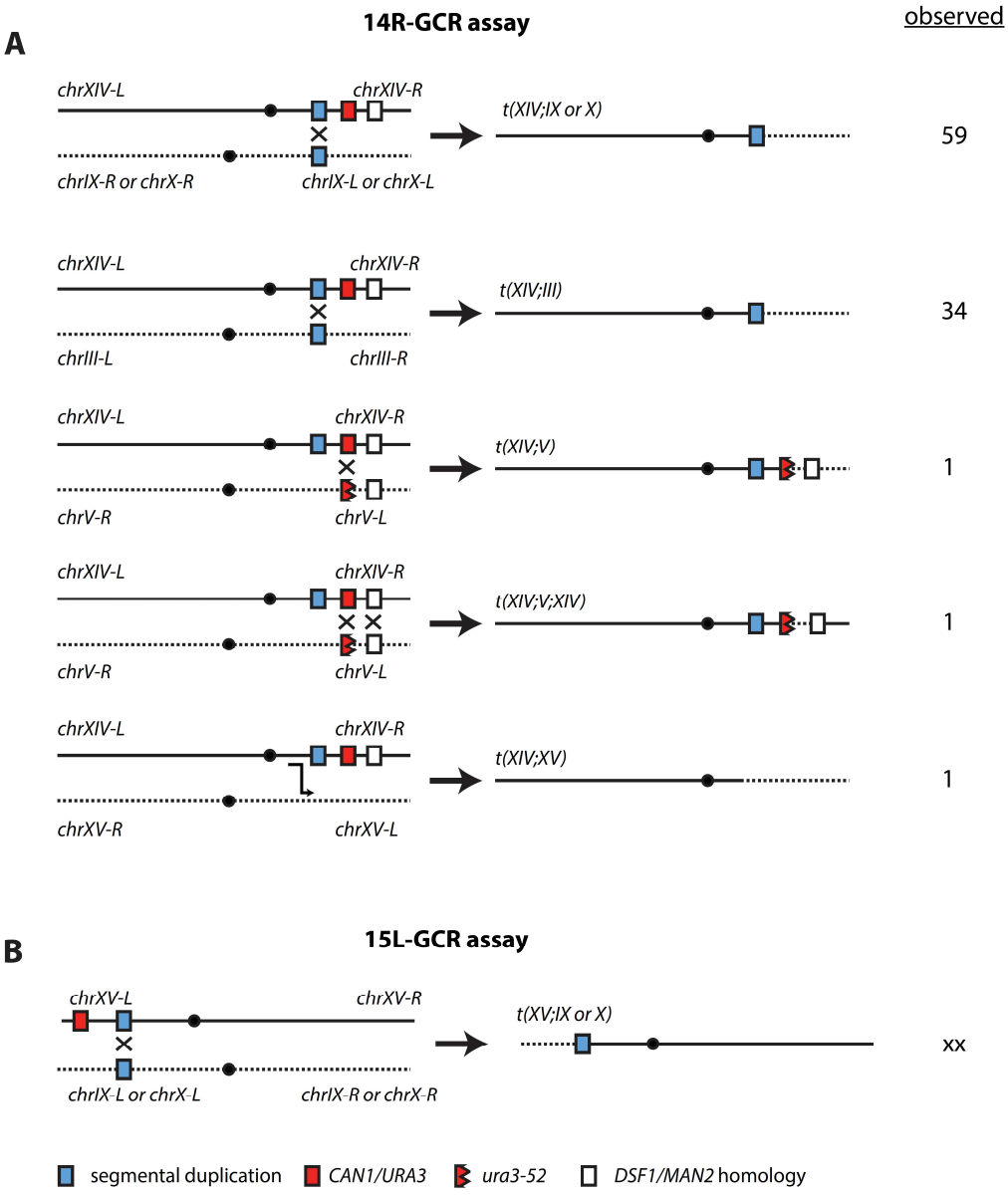


Figure 2.4 Summary of common GCR structures observed in Chr14R-GCR and Chr15L-GCR assays.
A. The Chr14-R assay forms a variety of GCR products including t(14;9/10) translocations (indistinguishable from each other in the region involved in the GCR), t(14;3) translocations, t(14;5) translocations, t(14;5;14) translocations, and t(14;15) translocations in addition to *de novo* telomere addition products. B. The Chr15-L assay forms t(15;9/10) translocations between the left arm of chromosome 15 and the left arms of chromosomes 9 and 10 (indistinguishable from each other in the region involved in the GCR).

Figure 2.5 Representative copy number plots for chromosome regions involved in t(14;9 or 10) GCRs

A. Diagram of the HR event to generate the t(14;9 or 10) translocation product. B. Copy number analysis of the sequenced parental strain and GCR-containing strain shows that GCRs are associated with deletion of the *CAN1/URA3*-containing terminal portion of chromosome 14-R. The duplication homology region is shaded in blue and the *CAN1/URA3* case is shaded in red. C. Copy number analysis of the sequenced parental strain and GCR-containing strain shows that GCRs are associated with duplication of either chromosome 9 or 10 (indistinguishable from each other in the region involved in the event). Because there are 2 copies of this sequence present in the genome in a wild-type strain (2n), we observe an increase to 3 copies of this sequence in a GCR-containing strain (3n) as depicted by the red arrow. D. P-values for predicted base frequencies for parental and GCR-containing isolates calculated using using the chi-squared test. HNV frequencies with p-values near 1.0 indicate an unaltered HNV distribution, while low p-values indicate positions affected by GCR events.

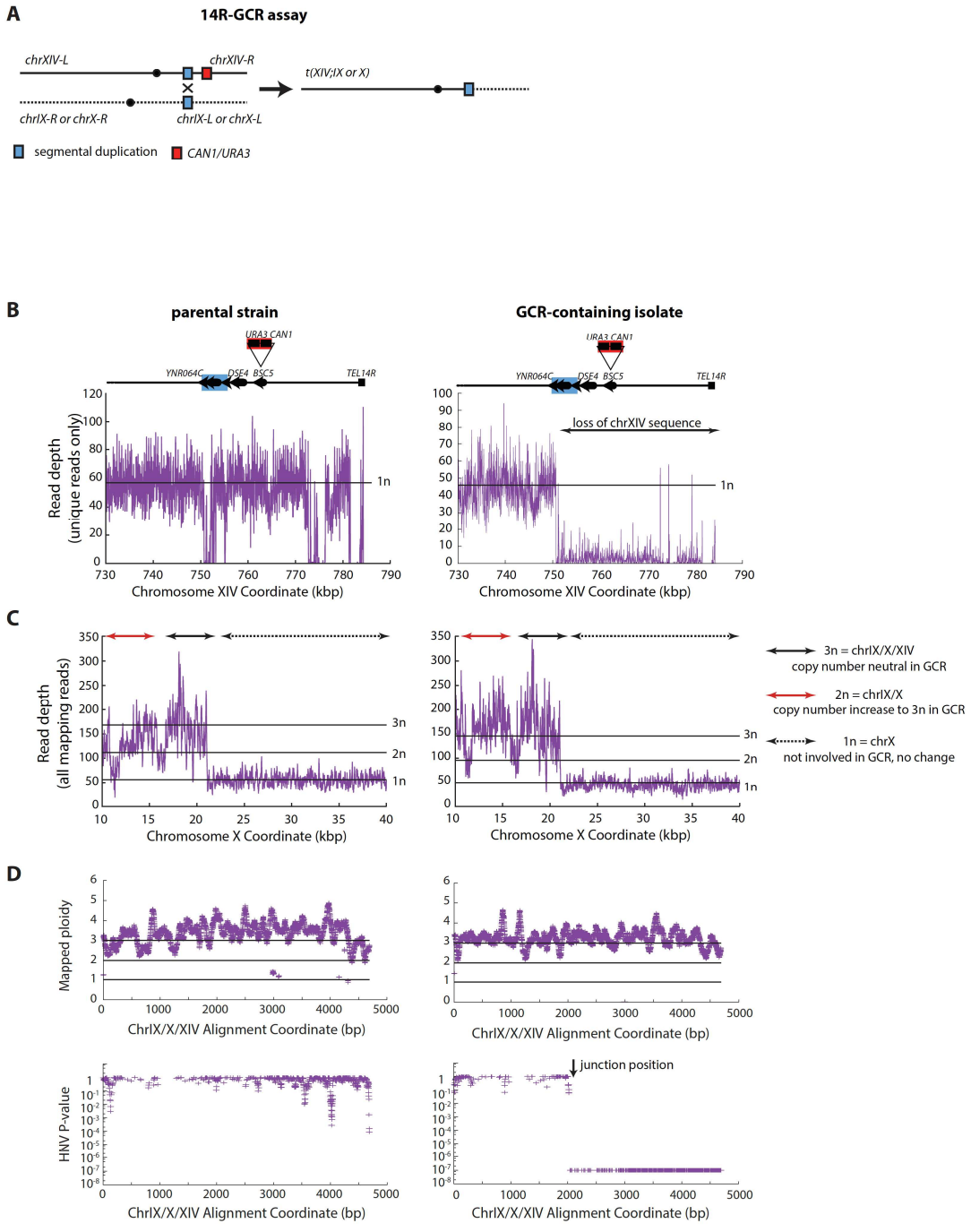
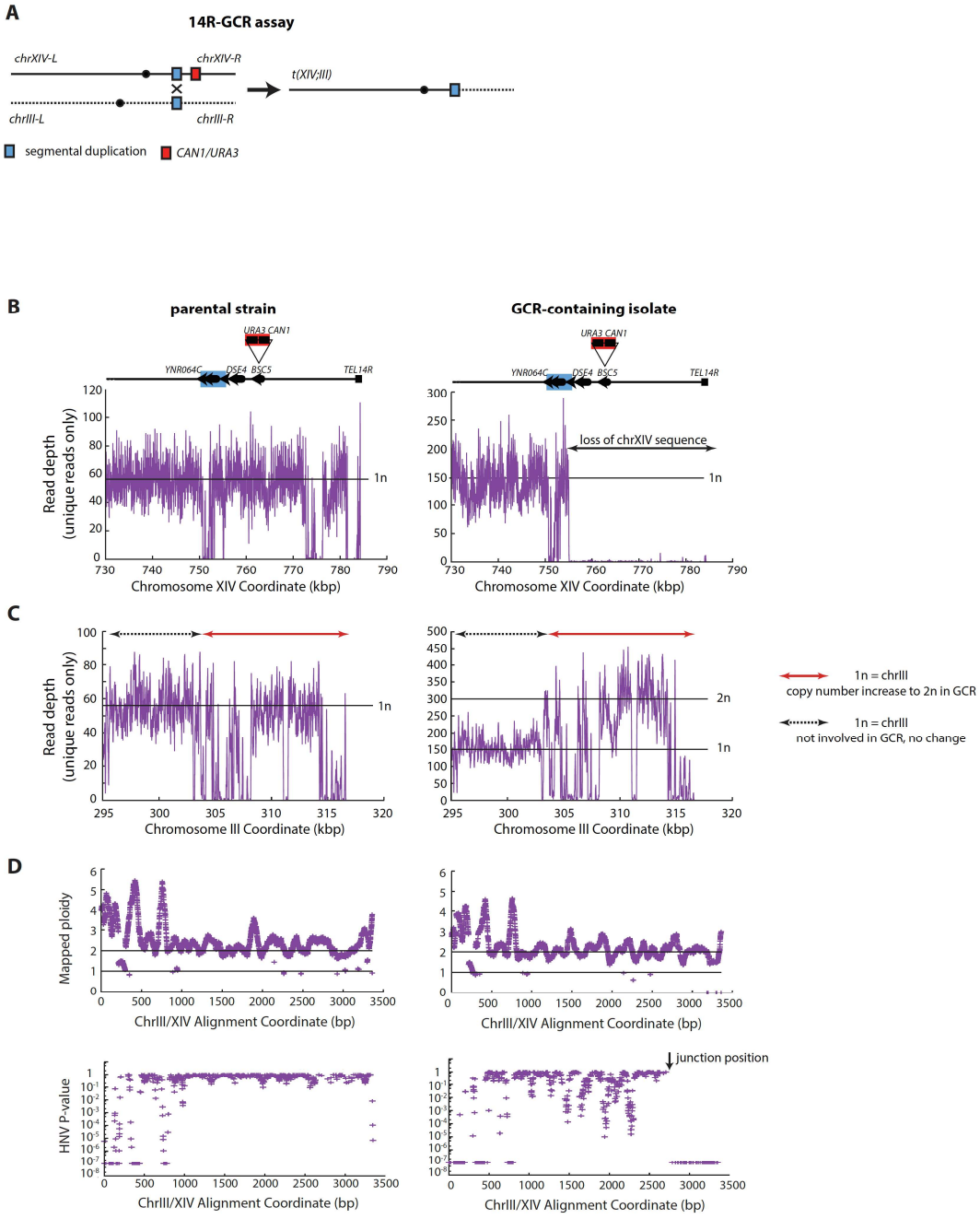

A
14R-GCR assay
chrXIV-L
chrXIV-R
t(XIV;IX or X)
chrIX-R or chrX-R
chrIX-L or chrX-L
segmental duplication
CAN1/URA3
B
parental strain
GCR-containing isolate
URA3 CAN1
YNR064C
DSE4 BSC5
TEL14R
loss of chrXIV sequence
Read depth
(unique reads only)
1n
Chromosome XIV Coordinate (kbp)
C
Read depth
(all mapping reads)
3n
2n
1n
Chromosome X Coordinate (kbp)
3n = chrIX/X/XIV
copy number neutral in GCR
2n = chrIX/X
copy number increase to 3n in GCR
1n = chrX
not involved in GCR, no change
D
Mapped ploidy
ChrIX/X/XIV Alignment Coordinate (bp)
HNV P-value
junction position

Figure 2.6 Representative copy number plots for chromosome regions involved in t(14;3) GCRs

A. Diagram of the HR event to generate the t(14;3) translocation product. B. Copy number analysis of the sequenced parental strain and GCR-containing strain shows that GCRs are associated with deletion of the *CAN1/URA3*-containing terminal portion of chromosome 14-R. The duplication homology region is shaded in blue and the *CAN1/URA3* case is shaded in red. C. Copy number analysis of the sequenced parental strain and GCR-containing strain shows that GCRs are associated with duplication of chromosome 3 (copy number increase to 2n shown by red arrow). D. P-values for predicted base frequencies for parental and GCR-containing isolates calculated using using the chi-squared test. HNV frequencies with p-values near 1.0 indicate an unaltered HNV distribution, while low p-values indicate positions affected by GCR events.


A
14R-GCR assay
chrXIV-L
chrXIV-R
chrIII-L
chrIII-R
t(XIV;III)
segmental duplication
CAN1/URA3
B
parental strain
GCR-containing isolate
URA3 CAN1
YNR064C
DSE4
BSC5
TEL14R
Read depth (unique reads only)
loss of chrXIV sequence
1n
Chromosome XIV Coordinate (kbp)
C
Chromosome III Coordinate (kbp)
2n
1n = chrIII copy number increase to 2n in GCR
1n = chrIII not involved in GCR, no change
D
Mapped ploidy
ChrIII/XIV Alignment Coordinate (bp)
HNV P-value
junction position

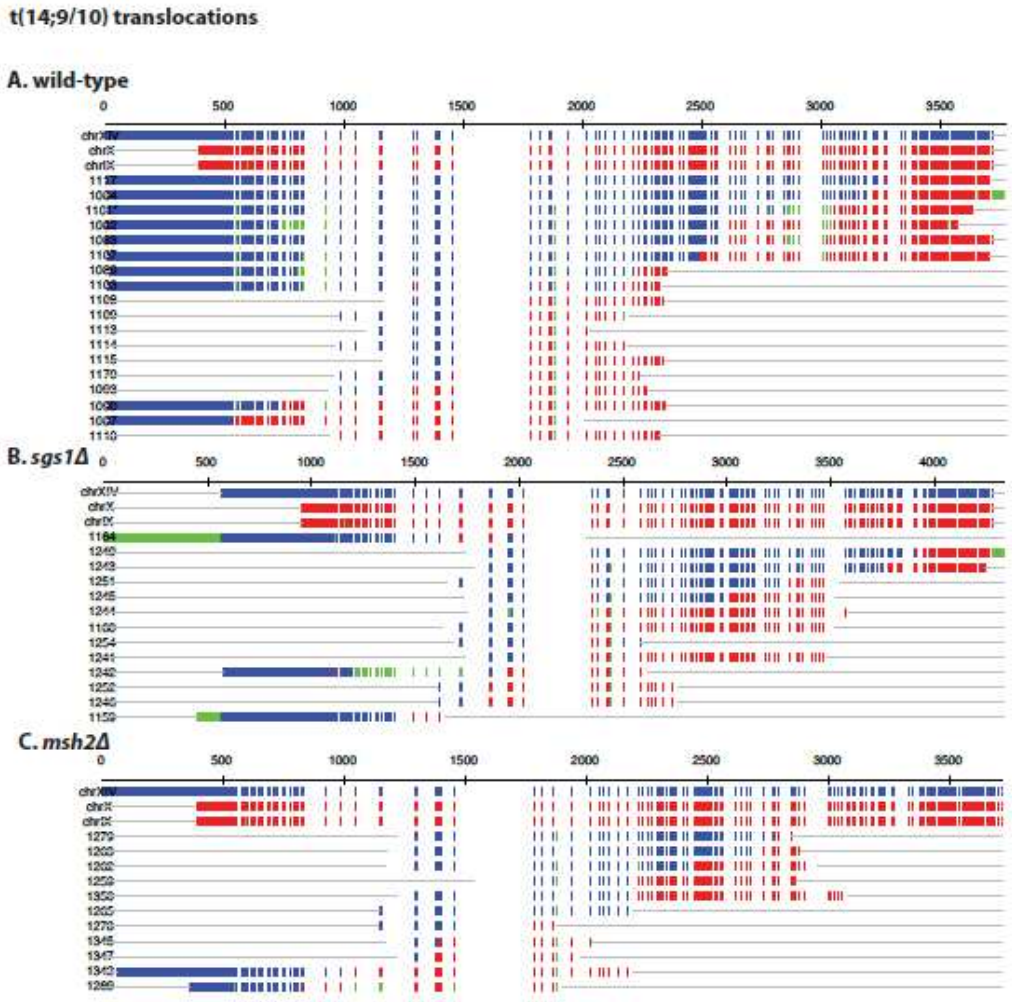


Figure 2.7 Sequence analysis of the breakpoints of t(14;9/10) GCRs mediated by the *YNR065C-YNR066C-DSE4* homology

A. Vertical lines indicate the identity of a single nucleotide polymorphism between Chr14 (blue lines) and Chr9 or Chr10 (red lines; indistinguishable from each other in this region) for 18 sequenced t(14;9/10) fusions from the wild-type Chr14-R GCR assay strain. Three of the sequences demonstrated evidence of template switches resulting in composite rearrangements B. Sequence analysis of 13 t(14;9/10) fusions from a *sgs1Δ* mutant as in (A). One sequence demonstrated evidence of a template switch resulting in a composite rearrangement. C. Sequence analysis of 11 t(14;9/10) fusions from a *msh2Δ* mutant as in (A). Each GCR had only a single junction between Chr14 sequence and Chr9/Chr10 sequence.

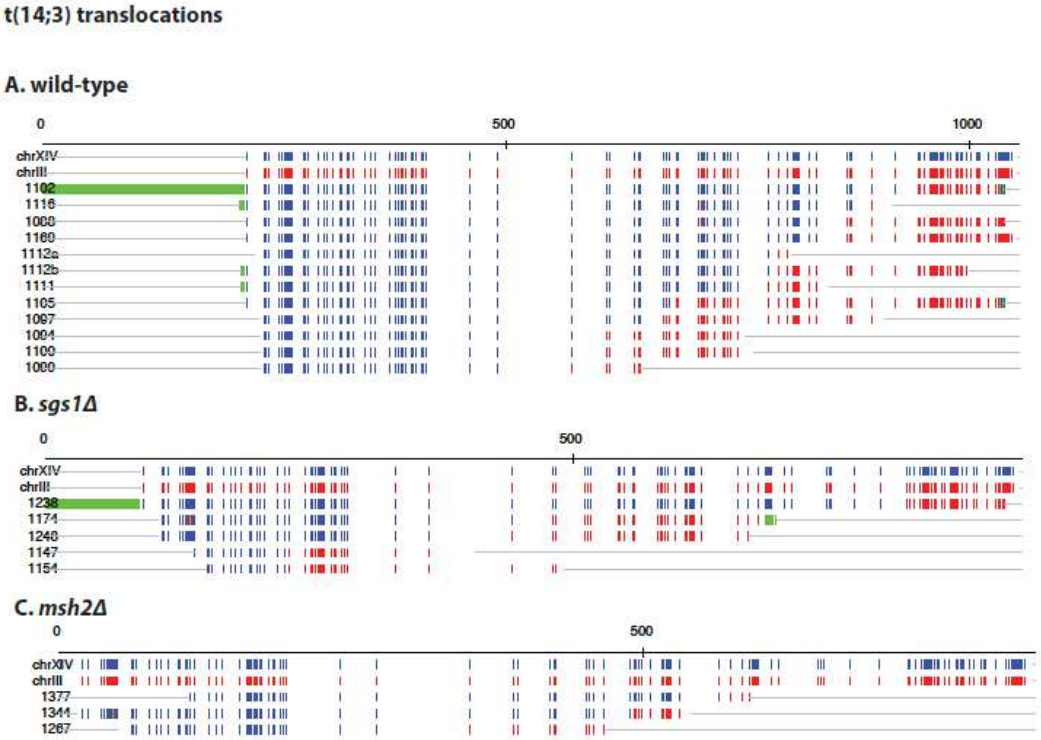


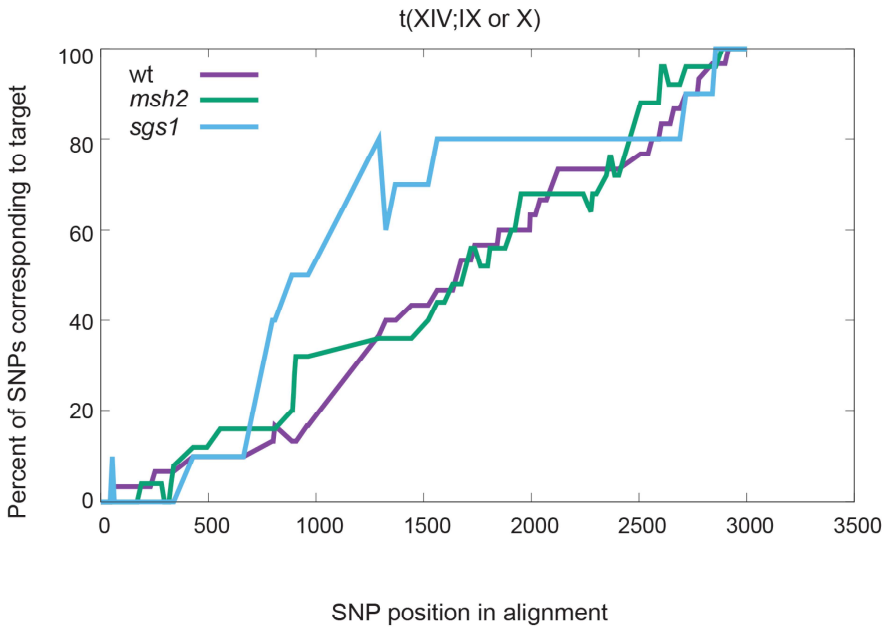
Figure 2.8 Sequence analysis of the breakpoints of t(14;3) GCRs mediated by the *YNR065C-YNR066C-DSE4* homology

A. Vertical lines indicate the identity of a single nucleotide polymorphism between Chr14 (blue lines) and Chr3 (red lines) for 12 sequenced t(14;3) fusions from the wild-type Chr14-R GCR assay strain. Three of the sequences demonstrated evidence of template switches resulting in composite rearrangements (box?). B. Sequence analysis of 5 t(14;3) fusions from a *sgs1Δ* mutant as in (A). One sequence demonstrated evidence of a template switch resulting in a composite rearrangement (box?). C. Sequence analysis of 3 t(14;3) fusions from a *msh2Δ* mutant as in (A). Each GCR had only a single junction between Chr14 sequence and Chr3 sequence.

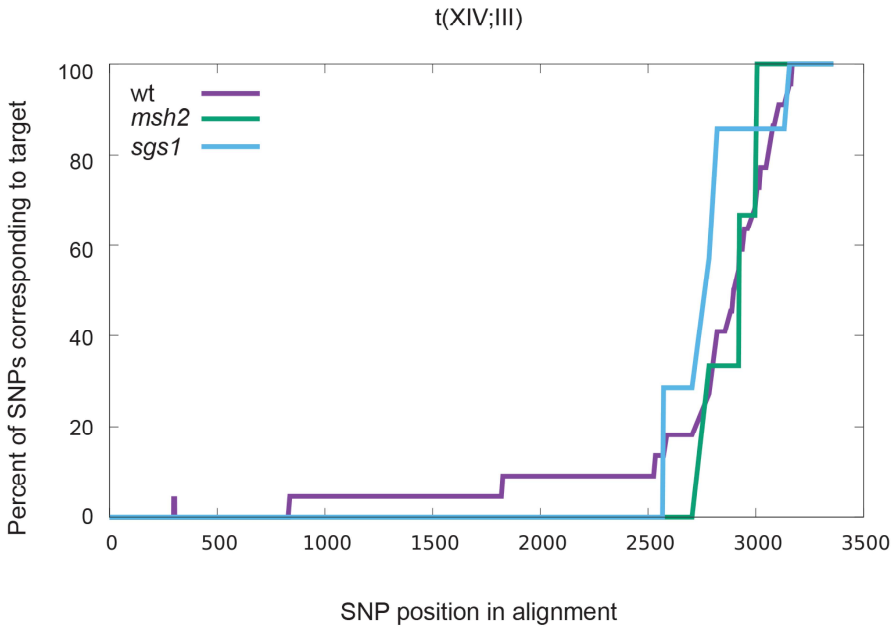
Figure 2.9 Cumulative distribution plots of single nucleotide polymorphisms for t(14;9 or 10) and t(14;3) GCRs in heteroduplex rejection mutants

A. The percentage of GCR-containing isolates (out of X total isolates) with SNPs mapping to Chr9 or Chr10 instead of Chr14 at a given coordinate within the Chr14-R GCR assay homology region plotted along the y-axis (Chr14: 750350-753351; Chr9: 16139-13138; Chr10: 16122-13121). Curves for wild-type, *sgs1Δ* and *msh2Δ* isolates are shown. B. The percentage of GCR-containing isolates (out of X total isolates) with SNPs mapping to Chr3 instead of Chr14 at a given coordinate within the Chr14-R GCR assay homology region plotted along the y-axis (Chr14: 750350-753351; Chr3: 300428-297427). Curves for wild-type, *sgs1Δ* and *msh2Δ* isolates are shown.

A


t(XIV;IX or X)
wt
msh2
sgs1
Percent of SNPs corresponding to target
SNP position in alignment
0
20
40
60
80
100
0
500
1000
1500
2000
2500
3000
3500

B


t(XIV;III)
wt
msh2
sgs1
Percent of SNPs corresponding to target
SNP position in alignment
0
20
40
60
80
100
0
500
1000
1500
2000
2500
3000
3500

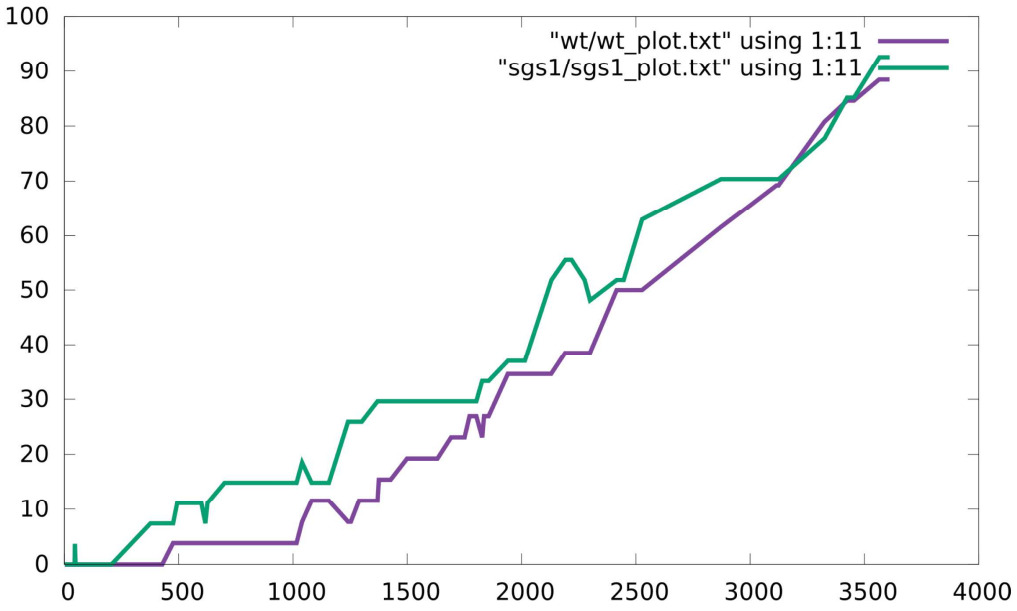


Figure 2.10 Cumulative distribution plot of single nucleotide polymorphisms for t(5; 14) GCRs in heteroduplex rejection mutants.

The percentage of GCR-containing isolates (out of X total isolates) with SNPs mapping to Chr14 instead of Chr5 at a given coordinate within the Chr5 dGCR assay homology region plotted along the y-axis (Chr5: 23321-19589; Chr14: 772657-776300. Curves for wild-type and *sgs1Δ* isolates are shown.

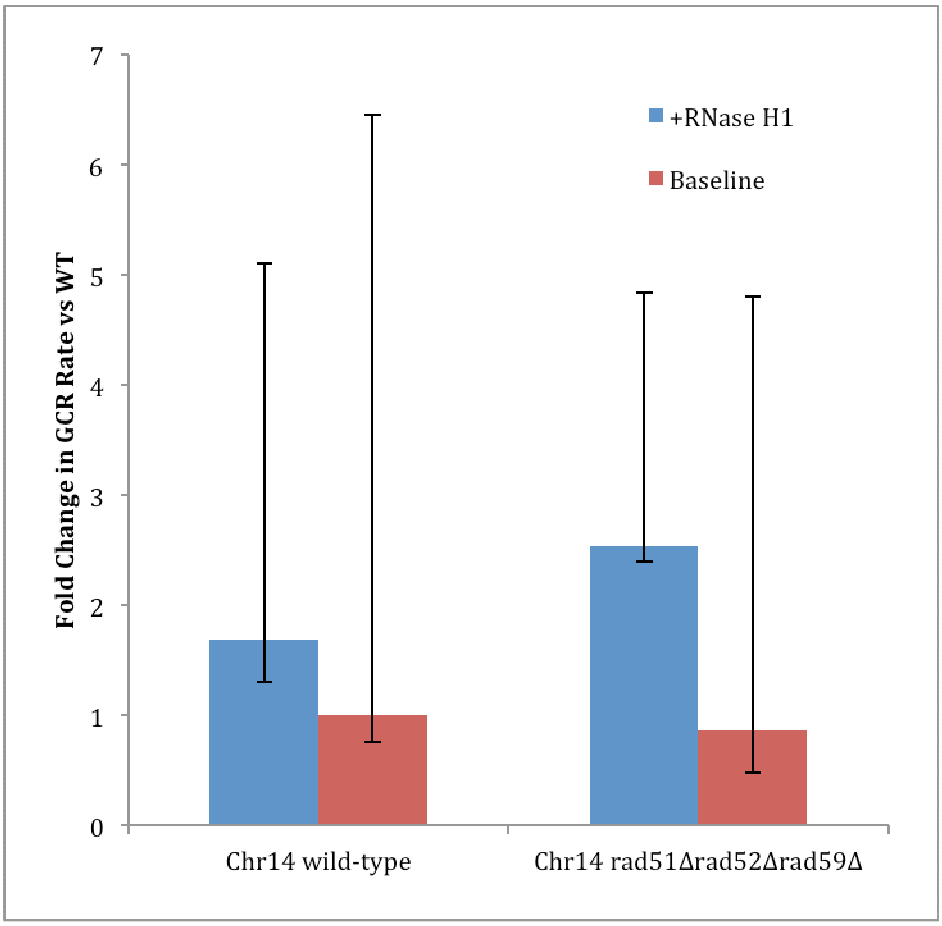


Figure 2.11 Effect of RNase H1 overexpression on Chr14-R GCR rates.

Overexpression of RNase H1 does not affect GCR rates in either the Chr14-R GCR assay wild-type or *rad51Δrad52Δrad59Δ* background. Strains were transformed with RDK5753, which overexpresses RNase H1. GCR rates were measured by fluctuation tests of fourteen independent isolates, with 95% confidence intervals shown.

Table 2.1 Effect of mutations on GCR rates among homeology-mediated assays.

		Chr14R GCR assay		Chr15L GCR assay		
Genotype	**Strain**	**GCR rate [95% CI]**	**Fold increase**	**Strain**	**GCR rate [95% CI]**	**Fold increase**
Wild-type	RDKY7706	2.58[0.28-6.64]E-08		RDKY7734	8.84 [4.7-17.9] E-08	
Homologous recombination						
*rad52*Δ	RDKY7842	6.28 [0.8-12.1]E-09	2.43	RDKY7819	1.76 [0.9-8.31] E-08	0.2
*rad51*Δ	RDKY7840	3.01 [0.54-9.67]E-08	1.16	RDKY7817	2.15 [0.73-8.16] E-08	0.24
*rad59*Δ	RDKY8129	1.73 [0.61-6.85] E-08	0.06	RDKY8121	1.85 [0.9-2.95] E-07	2.39
*rad51*Δ*rad59*Δ	RDKY8131	3.31 [1.80-10.4] E-09	0.16	RDKY8123	1.07E [0.24-4.55] E-07	1.2
*rad51*Δ*rad59*Δ*rad52*Δ	RDKY8344	6.71 [3.9-13.7] E-07	25.98	RDKY8351	1.29 [0.34-5.12] E-08	0.15
*rad51*Δ*rad52*Δ	RDKY9024	2.50 [1.87-8.63]E-08	0.97	RDKY8908	1.85 [1.17-3.27] E-06	20.9
*rad52*Δ*rad59*Δ	RDKY8906	5.11 [3.15-6.75] E-09	0.2	RDKY9026	5.35 [5.24-5.76] E-07	6.1
Heteroduplex rejection and resection						
*msh2*Δ	RDKY7846	6.29 [2.94-25.4] E-08	2.4	RDKY7821	7.84 [0.94-21.6] E-08	0.88
*sgs1*Δ	RDKY7848	7.48 [5.8-17.3]E-08	29	RDKY7777	5.70 [3.2-13.9] E-06	65
*msh2*Δ*sgs1*Δ	RDKY9237	3.87 [1.05-6.27] E-06	150	RDKY9239	1.22 [0.9-1.69] E-05	138
*exo1*Δ	RDKY8910	1.86 [1.10-4.26] E-07	7.2	RDKY8915	3.00 [1.79-9.06] E-06	34
*exo1*Δ*sgs1*Δ	RDKY8912	4.47 [2.69-6.32] E-07	17.3	RDKY8916	4.12 [1.96-6.94] E-06	46.6
*exo1*Δ*msh2*Δ	RDKY9020	3.09 [1.31-3.81] E-07	12	RDKY9022	7.54 [4.05-13.0] E-06	85.3
Break-induced replication						
*pol32*Δ	RDKY9028	2.11 [0.77-5.94] E-08	0.82	RDKY9241	2.53 [1.77-3.29] E-07	2.86
*mph1*Δ	RDKY8415	2.22 [1.26-5.07] E-07	8.6	RDKY9243	1.47 [1.04-2.55] E-06	16.62
Replication checkpoint						
*mrc1*Δ	RDKY7844	2.30 [0.1-3.22] E-07	8.9	RDKY7775	2.09 [1.57-3.49] E-07	2.36
*mec1*Δ*sml1*Δ	RDKY8417	1.43 [1.25-3.18] E-07	5.5	RDKY9245	2.36 (1.55-4.74] E-07	2.67

		Chr5 dGCR assay		Ty GCR assay		
Genotype	**Strain**	**GCR rate [95% CI]**	**Fold increase**	**Strain**	**GCR rate [95% CI]**	**Fold increase**
Wild-type	RDKY6678	1.97 [1.6-4.3]E-08		RDKY6076	8.40 [5.9-9.6]E-08	
Homologous recombination						
*rad52*Δ	RDKY6708	1.09 [0.29-7.08] x 10-8	0.55	RDKY6503	1.3 [0.7-2.5]E-08	0.16
*rad51*Δ	RDKY6709	2.31 [1.02-4.35] x 10-8	1.17	RDKY6555	5.9 [2.5-9.2]E-07	7.02
*rad59*Δ	RDKY6710	6.94 [5.17-9.12] x 10-8	3.5	RDKY6599	6.1[5.0-8.4]E-08	0.73
*rad51*Δ*rad59*Δ	RDKY6711	4.48 [2.49-91.1] x 10-8	0.23	RDKY7083	7.9[5.9-17]E-08	0.94
*rad51*Δ*rad59*Δ*rad52*Δ	RDKY6712	4.53 [2.64-8.80] x 10-9	0.23	RDKY7085	1.7[0.7-4.2]E-07	0.2
*rad51*Δ*rad52*Δ	RDKY9055	1.34 [0.64-3.17]E-06	68.02	RDKY7187	1.9 [0.3-5] E-08	0.2
*rad52*Δ*rad59*Δ	RDKY9058	4.20 [2.54-5.98] E-07	21.32	RDKY7191	7.0 [2.5-17]E-09	0.1
Heteroduplex rejection and resection						
*msh2*Δ	RDKY6713	1.75 [1.333-2.69] x 10-7	8.9	RDKY6607	2.2[0.8-10]E-07	2.62
*sgs1*Δ	RDKY6690	1.93 [1.56-2.49] x 10-6	98	RDKY6501	3.10E-06	36.9
*msh2*Δ*sgs1*Δ	RDKY9018	4.48 [2.20-7.02] E-06	227			
*exo1*Δ	RDKY6746	8.44 [2.08-21.5] x 10-8	4.2			
*exo1*Δ*sgs1*Δ	RDKY8033	5.70 [2.92-10.5] x10-6	289			
*exo1*Δ*msh2*Δ	RDKY9016	4.47 [2.83-5.71] E-07	22.7			
Break-induced replication						
*pol32*Δ	RDKY6720	3.15 [1.76-3.91] x 10-8	1.6			
*mph1*Δ	RDKY6795	1.05 [0.93-1.25] x 10-7	5.34			
Replication checkpoint						
*mrc1*Δ	RDKY6747	3.75 [2.79-5.16] x 10-7	19	RDKY6529	1.20E-06	14.3
*mec1*Δ*sml1*Δ	RDKY6769	1.50 [0.554-2.74] x 10-7	7.61	RDKY6581	1.10E-06	13.1

Rate of accumulating Can^r 5-FOA^r progeny. The number in brackets is the 95% CI and the number in parentheses is the fold increase relative to the relevant wild-type rate for that assay (varies by assay).

Table 2.2 Yeast strains and genotypes

Name	***Genotype****	*The genotype is that of the listed strain with the indicated genetic modifications
RDKY5461	*MATa ura3-52 leu2∆1 trp1∆63 his3∆200 lys2∆Bgl hom3-10 ade2∆1 ade8 can1::hisG*	
RDKY7706	RDKY5461 *ynr068c::CAN1/URA3 ynrcdelta9::NAT*	
RDKY7734	RDKY5461 *enb1::CAN1/URA3*	
RDKY7840	RDKY7706 *rad51::HIS3*	
RDKY7842	RDKY7706 *rad52::HIS3*	
RDKY7848	RDKY7706 *sgs1::HIS3*	
RDKY7844	RDKY7706 *mrc1::TRP1*	
RDKY7846	RDKY7706 *msh2::G418*	
RDKY8129	RDKY7706 *rad59::TRP1*	
RDKY8131	RDKY7706 *rad51::HIS3 rad59::TRP1*	
RDKY8344	RDKY7706 *rad51::HIS3 rad52::hph rad59::TRP1*	
RDKY8415	RDKY7706 *mph1::G418*	
RDKY8417	RDKY7706 *mec1::G418 sml1::hph*	
RDKY8906	RDKY7706 *rad52::hph rad59::TRP1*	
RDKY8910	RDKY7706 *exo1::TRP1*	
RDKY8912	RDKY7706 *exo1::TRP1 sgs1::HIS3*	
RDKY9020	RDKY7706 *exo1::TRP1 msh2::G418*	
RDKY9024	RDKY7706 *rad51::HIS3 rad52::hph*	
RDKY9028	RDKY7706 *pol32::HIS3*	
RDKY7817	RDKY7734 *rad51::HIS3*	
RDKY7819	RDKY7734 *rad52::HIS3*	
RDKY7775	RDKY7734 *mrc1::TRP1*	
RDKY7777	RDKY7734 *sgs1::HIS3*	
RDKY7821	RDKY7734 *msh2::G418*	
RDKY8121	RDKY7734 *rad59::TRP1*	

Table 2.2 cont.

RDKY8123	RDKY7734 *rad51::HIS3 rad59::TRP1*
RDKY8351	RDKY7734 *rad51::HIS3 rad52::hph rad59::TRP1*
RDKY8908	RDKY7734 *rad51::HIS3 rad52::hph*
RDKY8914	RDKY7734 *exo1::TRP1 msh2::G418*
RDKY8915	RDKY7734 *exo1::TRP1*
RDKY8916	RDKY7734 *exo1::TRP1 sgs1::HIS3*
RDKY9022	RDKY7734 *exo1::TRP1 msh2::G418*
RDKY9026	RDKY7734 *rad52:hph rad59::TRP1*
RDKY6678	*MATa ura3-52 leu2∆1 trp1∆63 his3∆200 lys2∆Bgl hom3-10 ade2∆1 ade8 can1::hisG iYEL072W::hph yel072w::CAN1/URA3*
RDKY9016	RDKY6678 *exo1::HIS3 msh2::G418*
RDKY9018	RDKY6678 *sgs1::HIS3 msh2::G418*
RDKY9055	RDKY6678 *rad51::HIS3 rad52::G418*
RDKY9058	RDKY6678 *rad52::G418 rad59::TRP1*

CHAPTER 3

Systematic identification of an extended genetic network that suppresses genome instability

3.1 Summary

Gross chromosomal rearrangements (GCRs) play an important role in human diseases, including cancer. The identity of all Genome Instability Suppressing (GIS) genes is not currently known. Here multiple *Saccharomyces cerevisiae* GCR assays and query mutations were crossed into arrays of mutants to identify progeny with increased GCR rates. One hundred eighty two GIS genes were identified that suppressed GCR formation. Another 438 cooperatively acting GIS genes were identified that were not GIS genes, but suppressed the increased genome instability caused by individual query mutations. Analysis of TCGA data using the human genes predicted to act in GIS pathways revealed that a minimum of 93% of ovarian and 66% of colorectal cancer cases had defects affecting one or more predicted GIS gene. These defects included loss-of-function mutations, copy-number changes associated with reduced expression, and silencing. In contrast, acute myeloid leukaemia cases did not appear to have defects affecting the predicted GIS genes.

3.2 Introduction

Genetic instability is seen in most cancers and is thought to play a critical role in the development and progression of tumors[67]. There are two general types of genetic instability seen in cancer[68]: the accumulation of large numbers of mutations and the accumulation of genome rearrangements

such as translocations, copy number changes and aneuploidy[68, 69]. The study of cancer susceptibility syndromes like Fanconi Anemia and the *BRCA1*- and *BRCA2*-defective breast and ovarian cancer syndromes provided the first evidence for a causal link between defects causing increased genome rearrangements and the development of cancer [70, 71]. However, this understanding is incomplete in part because most relevant studies have focused on a limited number of genes and the lack of genetic screens to identify Genome Instability Suppressing (GIS) genes in mammalian cells.

Genetic studies in *Saccharomyces cerevisiae* have provided considerable information about the spontaneous formation of genome rearrangements due to the development of quantitative genetic assays that can detect gross chromosomal rearrangements (GCRs)[3, 5, 12, 25, 72, 73]. The observed GCRs depend in part on the features of the specific GCR assay, but include 1) terminal deletions healed by *de novo* telomere addition, 2) monocentric translocations, 3) interstitial deletions, and 4) complex GCRs resulting from multiple cycles of rearrangement secondary to the formation of dicentric chromosomes by multiple processes[3, 12-15, 23, 29, 74]. Overall, the GCRs observed parallel those being identified by whole-genome analysis in human diseases including cancer. In addition, GCR assays have been used to identify genes that prevent GCRs from occurring and genes that act in the formation of GCRs[3, 5, 12, 15, 25, 28, 72, 75-83].

Even in *S. cerevisiae*, our knowledge of GIS genes is incomplete. This is in part because most known GIS genes have been identified through limited candidate-gene approaches[3, 5, 12, 15, 25, 28, 75, 84]. Only a small number of additional GIS genes have been identified in systematic screens[72, 76, 77, 82]. Reasons for the limited success of these screens include: 1) the use of assays that were not specific for GCRs, 2) the low GCR rates detected in GCR assays are not well-suited for large scale genetic screens, 3) the use of only a single GCR assay, and 4) the lack of analysis of interacting mutations. Here, we used a 2-stage screen design in which an *in silico* approach was used to develop a highly enriched candidate gene list sorted into candidate pathways[85] followed by an extensive genetic screen utilizing three different GCR assays and 43 query mutations to identify genes and interacting pairs of genes that act to suppress GCRs. Our results have provided a much more detailed picture of the genetic network that acts to prevent GCRs than previously available, and analysis of The Cancer Genome Atlas (TCGA) data[86-89] has suggested that the genes in this network are potentially altered in a large proportion of ovarian and colorectal cancers but not in acute myeloid leukemia.

3.3 Results

3.3.1 Design of the systematic mating screen to identify mutations and genetic interactions causing increased genome instability

Our strategy for identifying new GIS genes was to generate mutant strains using an adaptation of the Synthetic Genetic Array (SGA) method[90] and test them for increased genome instability. We crossed a collection of candidate mutant strains (described below) against strains containing one of three GCR assays (GCR query strains; Fig. 3.1a) and against strains containing a GCR assay and one of 43 mutations (GCR+mutation query strains). The 43 GCR+mutation query strains were included in the crosses because some genes are cooperating Genome Instability Suppressing (cGIS) genes in which mutations only affect genome stability when combined with other mutations[25]. The 43 mutations affected known GIS genes and genes that clustered with known GIS genes[85], and were selected to maximize the number of gene clusters surveyed (Fig. 3.1b-c).

The GCR assays select haploid cells resistant to both canavanine (Can) and 5-fluoroorotic acid (5FOA) due to loss of the *CAN1* and *URA3* genes on the left arm of chromosome V[3]. These GCRs have a breakpoint between the *CAN1* and *URA3* genes and the most telomeric essential gene on the left arm of chromosome V (*PCM1*); genomic features in this breakpoint region influence the types of GCRs that are formed[5, 12]. The short repeated

sequence GCR (sGCR) assay contains single copy sequences in the breakpoint region and ~100 bp of *YCLWdelta5* sequence in the *can1::*P_{LEU2}*-NAT* locus that has homology to the long-terminal repeats from Ty1 and Ty2 retrotransposons (Fig. 3.1a). The segmental duplication GCR (dGCR) assay contains the ~4 kb *DSF1-HXT13* segmental duplication with divergent homology to regions of chromosomes IV, X and XIV[5] in addition to the *YCLWdelta5* fragment (Fig. 3.1a). The Ty912-containing GCR (tyGCR) assay mediates GCRs by HR with the other Ty-related sequences in the genome (Fig. 3.1a)[12]. Single and double-mutant haploid strains generated by the SGA procedure were tested for increased accumulation of GCRs by determining the number of Can^R $5FOA^R$ papillae observed after growing patches from independent spore clones and replica plating the patches onto GCR selection media (Fig. 3.1d-e). A numerical score (0-5) was assigned to each patch by counting the total number of papillae per patch, and a GCR strain score was calculated by averaging the scores for all of the patches analyzed for each mutant (Fig. 3.1d). The GCR strain scores are not the direct equivalent of GCR rates; doubling of the strain score corresponds to an increase in GCR rate of an average of 5-fold.

As determining GCR strain scores is labor-intensive, we implemented a two-stage genetic screening strategy to focus on a subset of non-essential *S. cerevisiae* genes that were enriched in GIS and cGIS genes. The first stage was our previous genome-wide *in silico* screen that identified 1,041 candidate

GIS genes[85]. A preliminary investigation of ~10% of these 1,041 genes identified 34 new GIS genes and 1 new cGIS gene, revealing that this group of genes was enriched for GIS genes but that not all of the 1,041 genes were GIS genes[85]. In the second stage of the screen, which is described in the present study, we eliminated all essential genes from the candidate list of GIS genes and added in all additional non-essential genes known to function in the pathways identified by the 1,041 genes (see Methods; Supplementary Table 1) resulting in 1,055 genes/mutations. Finally, we added two additional *mrc1* and *rad53* alleles and a *leu2Δ* control deletion for a total of 1,058 strains.

The first-generation set of 1,058 mutant strains was crossed to the wild-type dGCR, sGCR and tyGCR query strains and dGCR+query mutation strains containing *dia2Δ*, *exo1Δ*, *rrm3Δ*, and *rtt107Δ* mutations, and the resulting progeny were evaluated for increased GCR rates using patch tests (see Methods). Based on these results, we generated a second-generation set of mutant strains, which reduced the number of strains to 639 (see Methods and Supplementary Table 1). This collection of mutants contained all of the mutations that either increased the GCR rate in at least 1 GCR assay or interacted with at least 1 of the *dia2Δ*, *exo1Δ*, *rrm3Δ*, and *rtt107Δ* mutations in the dGCR assay. The *dia2Δ*, *exo1Δ*, *rrm3Δ*, and *rtt107Δ* mutations were selected for evaluating candidate enhancing mutations because together they interacted with the largest number of bait mutations in a subset of the first-generation set of mutant strains (see Methods 3.5.4). The 419 bait mutations

excluded were identified in the *in silico* screen on the basis of causing increased sensitivity to DNA damaging agents[85], which can reflect processes unrelated to DNA repair like small molecule export and detoxification. Consistent with this, the 419 excluded genes showed little if any genetic similarity to *bona fide* GIS genes[85] and were enriched for roles in the endosome, Golgi complex, the ESCRT complex, the retromer complex and general metabolism but not in DNA or chromosome metabolism. Crossing of the second-generation set of mutants to the remaining 39 dGCR+query mutation strains was then continued, and the resulting double mutant progeny were evaluated for increased GCR rates using patch tests (see Methods).

3.3.2 Identification of GIS genes

Crossing the wild-type dGCR, sGCR, and tyGCR query strains to the first-generation mutant set generated 1,002, 995, and 1,009 single mutant strains, respectively (Supplementary Table 1). The GCR strain scores for the *leu2Δ* control strains were 0.1, 0.94, and 2.67 for the sGCR, dGCR, and tyGCR assays, respectively, and were consistent with quantitative GCR rate measurements (Supplementary Table 2). The fact that the distribution of strain scores for all of the mutations tested in each GCR assay peaked around the strain scores for the *leu2Δ* control strains suggested that most of the mutations tested did not strongly affect genome instability alone (Fig. 3.2a-c).

To determine a cutoff score for identifying mutations causing increased GCR rates, we determined GCR rates for 101 single mutant dGCR strains and all 43 *leu2Δ queryΔ* double mutant dGCR strains from crosses with the dGCR+query mutation strains (Supplementary Table 3). We found a robust correlation between the GCR strain scores and GCR rates (Fig. 3.2d), despite a small but consistent increase in dGCR rates that was observed in strains from the systematic crosses that was potentially due to GCRs mediated by the *YCLWdelta5* fragment at the *can1::*P_{LEU2}*-NAT* locus (Supplementary Table 4; Supplementary Fig. 1). Using the GCR strain scores and rates for the 144 systematically generated dGCR assay-containing strains, we determined that a cutoff score of 1.4 (0.4 above the wild-type score) balanced the false-positive and false-negative errors in identifying mutations in GIS genes (Supplementary Fig. 2; Methods).

We generated a comprehensive list of GIS genes by combining the GIS genes identified here with those previously known. Initially, we selected all single mutations that caused GCR strain scores that were 0.4 or more above the wild-type score in any GCR assay (Supplementary Table 1; Supplementary Fig. 3). We then removed mutations that caused less than a 3-fold increase in GCR rate and included mutations that caused at least a 3-fold increase in rate, regardless of GCR strain scores (Supplementary Tables 3, 5, 6; Supplementary Fig. 4). Finally, we included mutations previously shown to increase the GCR rate by 3-fold or more, including mutations in essential

genes not studied here and mutations in genes identified in studies using GCR assays lacking repetitive sequences in the GCR breakpoint region (single-copy or unique sequence GCR assays; designated here as uGCR assays[3, 5] and previously summarized[85]; Supplementary Table 1). We observed 75 genes that suppressed GCRs in the dGCR assay, 71 genes in the tyGCR assay, 80 genes in the sGCR assay, and 105 genes in the uGCR assays. The higher number of GCR suppressing genes identified in the uGCR assays is primarily the result of candidate gene studies that included alleles of essential genes not tested here and mutations that cause small but significant increases in quantitative GCR assays, which were too small to reliably detect by the semi-quantitative scoring method used here. Together, we identified 182 *S. cerevisiae* GIS genes, 50 of which suppress genome instability in at least 3 of the 4 GCR assays (Fig. 3.3; Supplementary Table 1).

This analysis identified 64 previously unrecognized GIS genes, re-identified 62 known GIS genes including 20 identified in our previous test validation[85], and failed to re-identify 56 previously recognized GIS genes [85]. Of the 56 genes that were not re-identified in this screen, 13 were not discoverable, as these genes were either essential for viability or mating, and 43 caused only a small increase in GCR rate that could not be easily identifiable by patch scores. 14 of these 43 genes were found in our previous test validation[85]. 42 of these 43 were subsequently found as interactors in our cGIS screen (see below), which would be expected to identify weak alleles

as interacting mutations. In total, this study and our previous test validation of the list of candidate GIS genes[85] identified of 98 new GIS genes that were not known when we constructed the candidate list[85]. Examples of previously unrecognized GIS genes included *VID22* and *YDJ1* [84, 91, 92]. *VID22* encodes a partner of Tbf1 involved in transcriptional regulation[93, 94] and DSB repair[95]. *YDJ1* encodes the major cytosolic Hsp40/DnaJ co-chaperone that acts in protein maturation and stabilization[96]. The imperfect overlap of mutations causing increased GCR strain scores in the different assays suggests that some mutations have different effects on GCRs in different genomic contexts[5], which was verified by GCR rates (Supplementary Table 2).

To determine the efficiency of our pre-selection of candidate GIS genes[85], we crossed the dGCR assay strain to five randomly selected 96-well plates of mutant strains from the *S. cerevisiae* deletion collection and determined GCR strain scores for the progeny (Supplementary Table 1). Only 1 of the 463 single mutants scored, *ydl118wΔ*, that was not previously identified caused an increased GCR strain score. This deletion was tested in the initial cross but did not cause an increased GCR strain score, likely because it only causes a small increase in GCR rate. Extrapolating to the entire deletion collection, we estimate that our method potentially missed ~8 GIS genes and that we identified 96% of the GIS genes. However, *ydl118wΔ* was identified in the cGIS gene screen described below; this suggests that at

least some of the ~8 GIS genes that were predicted to not be identified in the single mutant screen were likely identified in the cGIS gene screen.

3.3.3 Identification of cGIS genes

We recovered and tested 25,974 double mutants from the crosses of the 43 dGCR+query mutation strains with the first-generation (*dia2Δ*, *exo1Δ*, *rrm3Δ*, and *rtt107Δ*) and second-generation (the remaining 39 dGCR+query mutation strains) bait strain sets (Methods; Supplementary Table 1). As the 43 query mutations were also present as bait mutations, we obtained 801 pairs of double mutant strains (out of a possible 903) generated as both query × bait or bait × query combinations. The individual pairs of these double mutants had consistent GCR strain scores (Supplementary Fig. 5). The scores of the double mutant strains were distributed about the score of the query mutants as for the single mutant strains, including mutations causing reduced scores (e.g. *rsc30Δ*), scores essentially identical to wild-type (e.g. *lge1Δ*), or increased scores (e.g. *ckb2Δ* and *rad17Δ*) (Fig. 3.4a-h; Supplementary Fig. 6-11).

3,149 (~13%) double mutant strains had GCR strain scores that were at least 0.4 (the single mutant strain differential score) greater than the higher of the two single mutant strain scores, suggestive of a genetic interaction causing a greater than additive increase in GCR rate (Supplementary Table 1). GCR rate determination of 66 selected double mutants predicted to show a genetic interaction revealed that 71% of the double mutation combinations resulted in a synergistic increase in GCR rate compared to that of the respective single

mutants. Thus, increased double mutant GCR strain scores were a good indicator for synergistic interactions (Supplementary Table 7). Raising the strain score differential above 0.4 did not substantially improve the identification of synergistic interactions; this suggests that the selection of false positive double mutants reflects some biological property of the double mutants (*e.g.*, selection of suppressor mutations) affecting the patch scores or rates rather than an inappropriate cutoff score. Despite this, double mutant GCR strain scores re-identified many previously known genetic interactions[25, 75], such as the redundancy between the *REV1-REV3-REV7*- and *MMS2-UBC13*-dependent branches of post-replication repair (PRR), the dependence on *SRS2* of the increases in GCR rates caused by *rad18Δ* and *rad5Δ* single mutations, and the redundancy of *MEC1*- and *TEL1*-mediated suppression of the formation of GCRs as well as many new interacting mutations (Supplementary Fig. 12).

The query mutations interacting with the largest number of mutations were *ckb2Δ*, *exo1Δ*, *rad17Δ*, *yta7Δ*, *mec1Δ*, *mms4Δ*, and *rrm3Δ* (Fig. 3.4i), and the bait mutations interacting with the largest number of query mutations were *est1Δ*, *ckb2Δ*, *mrn1Δ*, *exo1Δ*, *chk1Δ*, *isu1Δ*, *rnh201Δ*, *ckb1Δ*, and *tof1Δ* (Fig. 3.4j). Two mutations illustrating the complexity of these interactions were *ckb2Δ* and *exo1Δ,* which both interacted with checkpoint defects and also interacted with each other (Supplementary Tables 1 and 8), indicating that casein kinase II and Exo1 function in different GCR suppressing pathways

both of which interact with checkpoint pathways. Mutations causing very high (>3) GCR strain scores as single mutations tended to have few interactions, possibility due to difficulties in scoring strains that come close to saturating the assay. In total, 595 mutations interacted with at least one query mutation; 438 of the affected genes were distinct from the 182 GIS genes and hence were cGIS genes (Supplementary Table 1). In total, mutations in 620 genes (182 GIS genes and 438 cGIS genes; 13% of the 4,848 non-essential *S. cerevisiae* ORFs) were identified as causing or enhancing genome instability.

To identify the most robust interactions, we searched for interactions between a query mutation and mutations in multiple genes encoding components of an annotated complex or pathway, which we termed “modules” (Fig. 3.4k; Methods). We found shared interaction for 77 modules (Table 3.1). Mutations affecting an additional 91 modules had interactions that were not shared (Supplementary Table 9); although this included 64 complexes where only a single gene was tested. Mutations affecting only 2 modules, Mre11-Rad50-Xrs2 and Sgs1-Top3-Rmi1, caused significant increases in GCR rates but lacked interactions with other mutations; the lack of interacting mutations in these cases was likely due to the fact that single mutations in the genes encoding these complexes cause high GCR strain scores that saturate the assay (~4.0).

3.3.4 Inactivation of GIS genes in ovarian cancer, colorectal cancer, and acute myeloid leukemia

To determine if defects in GIS genes might occur in cancer, the ovarian cancer, colorectal cancer and acute myeloid leukemia (AML) TCGA data were analyzed[86-88]. The genes analyzed were the human homologs of the 182 *S. cerevisiae* GIS genes plus 13 additional genes that act in pathways and protein complexes defined by the GIS genes (hGIS1, 214 genes; Supplementary Table 10) and an expanded list (hGIS2, 279 genes; Supplementary Table 10) that included human DNA repair genes that function in pathways identified in *S. cerevisiae* but lack an *S. cerevisiae* homolog (e.g., *BRCA1* and *BRCA2*) or have an *S. cerevisiae* homolog that was not initially identified because of a borderline score (e.g., *NHEJ1* and *H2AFX*).

To identify potential cancer genes, we used a scoring system (S-score)[97] that integrates genome-wide data (copy number variation, expression, methylation and mutations) from a set of tumor samples. In the first analysis, human GIS genes were analyzed for signatures consistent with tumor suppressors (S-scores ≤ -2) or proto-oncogenes (S-scores ≥ 2) (Supplementary Table 10). Genes from hGIS1 and hGIS2 with S-scores ≤ -2 were enriched in ovarian cancer cases (hGIS1, 26 genes, $p=0.0008$; hGIS2, 41 genes, $p<0.0001$). In contrast, there was no enrichment in human GIS genes with S-scores ≥ 2 in the ovarian cancer cases (hGIS1, 43 genes, $p=0.31$; hGIS2, 54 genes, $p=0.40$), and these genes were not studied further. The 41 genes from hGIS2 with S-scores ≤ -2 and 4 additional genes with S-scores between -2 and -1.95 in ovarian cancer were analyzed for reduced

copy number (GISTIC scores of -1 or -2) associated with reduced expression (Z scores < -2). Reduced copy number associated with reduced expression of 1 to 19 of these 45 genes was observed in 97% of 527 ovarian cancer cases. A box plot of the data for one such gene, *RAD17*, and the frequency of occurrence for the top 20 such genes in ovarian cancer are shown in Figures 5a & 5b. There were also 3 genes that appeared to be silenced in 12% of 537 ovarian cancer cases (Supplementary Tables 11 & 12). Genes with S-scores ≤ -2 were enriched (hGIS1, 18 genes, p=0.0001; hGIS2, 18 genes. P=0.0015) in colorectal cancer cases; in contrast, human GIS genes with S-scores ≥ 2 were not enriched (hGIS1, 12 genes, p=0.10; hGIS2,16 genes, p=0.058) and were not studied further. The 18 genes with S-scores ≤-2 and 2 genes with S-scores between -2 and -1.95 in colorectal cancer cases were further analyzed. Reduced copy number associated with reduced expression of from 1 to 8 of these 20 genes was observed in 54% of 456 colorectal cancer cases, and 4 genes had apparent silencing in 10% of 463 colorectal cancer cases (Supplementary Tables 11 & 13). In the case of AML (222 samples), there was no enrichment of human GIS genes with S-scores ≤ -2 in hGIS1 (p=0.067) and a marginal enrichment of human GIS genes with S-scores ≤ -2 in hGIS2 (p=0.045). There was no enrichment for human GIS genes with S-scores ≥ 2 in both hGIS1 (p=0.085) and hGIS2 (p=0.194) and no genes with apparent silencing were identified.

In the second analysis, the number of potential loss-of-function (LOF) mutations (nonsense mutations, frameshift insertion/deletions, in-frame insertion/deletions, and splice site mutations) in the hGIS1 and hGIS2 genes was tabulated for 476 ovarian cancer cases and 537 colorectal cancer cases. For ovarian cancer, LOF mutations were not enriched in hGIS1 genes ($p=0.87$) but were enriched in hGIS2 genes ($p<0.0001$); this increase in significance was due to the presence *BRCA1* and *BRCA2* in the hGIS2 gene list, which accounted for 70% of the LOF mutations in hGIS2 genes. Analysis of the enrichment of classes of the LOF mutations for the hGIS2 genes in ovarian cancer revealed that deletions (includes frameshift deletions; $p<0.0001$), insertions (includes frameshift insertions; $p<0.0001$), frameshift deletions ($p<0.0001$), frameshift insertions ($p<0.0001$) and nonsense mutations ($p=0.0015$) were present at significantly increased level; many but not all of these LOF mutations were in *BRCA1* or *BRCA2* (Supplementary Table 10). Overall, 27% of the 476 ovarian cancer samples had LOF mutations in at least 1 of 44 predicted human GIS genes, with 1-3 genes mutated per sample (Supplementary Tables 11 & 14). LOF mutations in both sets of GIS genes were enriched in the colorectal cancer TCGA cases (hGIS1, $p<0.0001$; hGIS2, $p=0.0012$). The frequency of LOF and predicted deleterious missense mutations for the top 20 hGIS2 genes in colorectal cancer is shown in Figure 5c. Deletions (including frameshift deletions; hGIS1, $p=0.0004$; hGIS2, $p=0.0038$), mononucleotide repeat frameshifts (hGIS1, $p<0.0001$;

hGIS2, p<0.0001) and splice site mutations (hGIS1, p=0.0003; hGIS2, p=0.0002) were present at significantly increased levels in both hGIS1 and hGIS2 genes, and nonsense mutations were present at statistically significant increased levels in hGIS1 and at borderline significant levels in hGIS2 (hGIS1, p=0.01; hGIS2, p=0.0613) (Supplementary Table 10). A proportion of colorectal cancer has MMR defects associated with high rates of accumulating mutations [88]. We therefore repeated the analysis using a sample set in which the MMR-defective cases had been excluded and found that deletions (including frameshift deletions; hGIS1, p=0.001; hGIS2, p<0.0001), nonsense mutations (hGIS1, p=0.001; hGIS2, p=0.0046), frameshift deletions (hGIS1, p=0.032; hGIS2, p=0.029), mononucleotide repeat frameshifts (hGIS1, p=0.0016; hGIS2, p=0.0042) and splice site mutations (hGIS1, p=0.0004; hGIS2, p<0.0001) were present at significantly increased levels in both hGIS1 and hGIS2 and frameshift insertions were present at significantly increased levels in only hGIS2 (p=0.022). This indicates that the accumulation of these classes of mutations in the colorectal cancer cases was not due to MMR defects. Overall, 30% of the 537 colorectal cancer samples had LOF mutations in at least 1 of 185 predicted human GIS genes, with 1-36 genes mutated per sample (Supplementary Tables 11 & 15). In the case of AML, there was no enrichment of LOF mutations in the GIS genes (hGIS1, p=1.00; hGIS2, p=0.99), and as a result, individual classes of mutations were not analyzed.

All of the gene inactivation data were merged and the proportion of different classes of gene inactivation was determined (Figure 3.5d; Supplementary Table 16). LOF mutations and LOF mutations plus those missense mutations that scored as “predicted deleterious” in at least 5 of 6 function prediction tests used were considered separately. In ovarian cancer, the gene inactivation signature was dominated by cases with reduced copy number associated with reduced expression. Colorectal cancer showed a different pattern with less overlap between the cases with mutations and the cases with reduced copy number associated with reduced expression. Overall, when only LOF mutations were considered, a minimum of 93% of ovarian cancer cases and 66% of colorectal cancer cases had a signature of inactivation of one or more predicted GIS genes (Fig. 3.5d), although these figures are an underestimate because not all samples were analyzed for all types of alterations. It should be noted that the colorectal cancer cases did include cases with alterations in MMR genes (*MSH2*, *MSH6*, *MSH1*, *PMS2*), including 46 cases with only LOF mutations and 50 cases when LOF + predicted deleterious missense mutations (19 of these cases had silencing of *MLH1*, 1 of which also had LOH of *MLH1*), all but 3 of which had alterations in other GIS genes. In the ovarian cancer cases, there were 23 cases of reduced copy number and reduced expression of *MLH1*, 3 cases with a LOF mutation in an MMR gene and 2 cases with a predicted deleterious missense mutation in an MMR gene; all of these cases had alterations affecting other GIS genes.

This indicates that potential MMR defects account for only a small fraction of the alterations affecting GIS genes; it should be noted that MMR defects cause increased GCR rates in addition to increased mutation rates[5].

The 103 human homologs of the 98 newly identified *S. cerevisiae* GIS genes were analyzed separately [85] (Supplementary Table 17). When only LOF mutations, reduced copy number with reduced expression and silencing were considered 64% of ovarian cancer cases had defects effecting 1 or more of 24 of the 103 human genes and 47% of colorectal cancer cases had defects effecting 1 or more of 67 of the 103 genes. When predicted highly deleterious missense mutations were included 65% of ovarian cancer cases had defects effecting 1 or more of 37 of the 103 genes and 51% of colorectal cancer cases had defects effecting 1 or more of 84 of the 103 genes. This indicate that the newly identified GIS genes likely account for a large number of human GIS genes in which defects can cause increased genome instability in human cancers.

3.4 Discussion

Here we developed methods to screen the *S. cerevisiae* systematic deletion collection to identify new GIS genes, identify genes that interact to suppress the formation of GCRs and identify candidate human genes for the analysis of cancer genomics data to identify potential GIS gene defects in human cancers. This analysis increased the total number of known GIS genes to 182, including 98 new GIS genes identified here and during our targeted

validation of the starting 1,041 candidate genes. We also identified 438 cGIS genes and an extensive catalog of genetic interactions affecting genome stability. Analysis of ovarian and colorectal cancer TCGA data[87, 88] showed that the majority of the cancer cases analyzed (a minimum of 93% of ovarian and 66% of colorectal cancer cases) appeared to have defects affecting one or more genes that were homologs of the *S. cerevisiae* GIS genes or act in the pathways identified by the GIS genes. In contrast, AML, a cancer with little genome instability[89], did not appear to have defects affecting GIS genes. Thus genetic or epigenetic changes causing increased genome instability are likely common in some types of cancer but that due to the large number of GIS genes, the defect signature for any single gene can be weak.

Almost half of the 182 *S. cerevisiae* GIS genes suppress the formation of GCRs detected in multiple GCR assays. The common pathways identified typically include genes involved in DNA metabolism, including DNA replication and repair, and genes involved in checkpoint signaling in response to DNA damage and replication errors. Some of the genes identified, such as *RAD27* and *TSA1*, likely function by suppressing the formation of DNA damage[83, 98]. Other genes, such as those encoding the Mre11-Rad50-Xrs2 and Sgs1-Top3-Rmi1 complexes, likely process DNA damage generated by other mechanisms[99], such as DNA replication errors. A number of genes have roles in suppressing genome instability that are less clear, such as *VID22*, *YDJ1*, *SSZ1*, and *CKB2*. The fact that many GIS genes suppress GCRs

detected in multiple assays that probe different genomic contexts indicates that these genes can suppress the formation of many types of GCRs[5, 12, 14, 15, 25]. A notable exception are *pif1* mutations that cause a defect in suppression of *de novo* telomere additions that appears insensitive to genomic context[5, 28, 100]. In contrast, a number of genes suppress GCRs detected by subsets of GCR assays (Fig. 3.3). In most cases, the mechanisms underlying this specificity is not yet understood; however, in the case of *MSH2* and *MSH6*, the heteroduplexes formed by non-allelic HR during the formation of duplication-mediated GCRs are likely to contain a higher density of mispairs and hence be better recognized by MMR and subjected to heteroduplex rejection[5, 12] than heteroduplexes formed in the tyGCR and sGCR assays.

Mutations that enhance the accumulation of GCRs can in principle act in compensatory or parallel pathways or can have much more complicated relationships involving genes within pathways[101]. In addition, mutations can result in increased levels of DNA damage that can lead to GCRs when repair mechanisms are defective or are saturated by the increased levels of DNA damage. Many mutations showing genetic interactions, such as *exo1Δ*, cause increased GCR rates as single mutations. Mutations like these could affect the response to normal levels of spontaneous DNA damage as well as DNA damage that is either induced in the absence of other pathways or is normally repaired in part by other pathways. In contrast, a number of enhancer mutations, such as *tel1Δ* cause no increase in GCR rates as single

mutations[25]. These mutations may either result in increased DNA damage that is efficiently repaired so long as the relevant repair mechanisms are functional and not overwhelmed by other sources of damage or inactivate a redundant pathway. Defects in the genes encoding complexes can show the same types of interactions, regardless of whether defects in all of the genes encoding a complex behave similarly (such as *RNH201, RNH202,* and *RNH203* as well as *MMS2* and *UBC13*) or whether defects in only a subset of the genes encoding a complex have similar properties (e.g., *SPT3*, *SPT8* encoding part of SAGA).

The systematic identification of *S. cerevisiae* GIS genes [88] has facilitated a pathway-based analysis of human cancer genomics data. We have focused on ovarian and colorectal cancer, two cancers with genome instability that appear to have different relative frequencies of copy number changes and mutation driver alterations[86], as well as AML, a cancer that is associated with little if any genome instability[86]. In the case of the ovarian cancer TCGA data, 23% of the samples with any data had LOF mutations in GIS gene homologs with 65% of the samples with LOF mutations having LOF mutations in *BRCA1* and *BRCA2* as previously reported[87]; no other individual GIS gene homolog had a LOF mutation in more than 0.5 to 1% of the samples. In contrast, there was a high frequency of copy number alterations, including both copy number reductions and homozygous deletions, associated with reduced expression of GIS gene homologs in ovarian cancer.

This included 17% of the samples that had homozygous deletions of 1 to 9 GIS gene homologs per sample, approximating the frequency of samples with *BRCA1* or *BRCA2* LOF mutations. In contrast, the colorectal cancer TCGA data showed a higher proportion of samples and GIS gene homologs with LOF mutations and a lower yet high proportion of samples and GIS gene homologs with copy number alterations associated with reduced expression. A minimum of 93% of ovarian cancer TCGA cases and 66% of colorectal cancer TCGA cases had alterations (not considering predicted deleterious missense mutations) affecting one or more GIS gene homologs with only 5% and 8% of the samples, respectively, having alterations in genes expected to cause a strong mismatch repair defect (*MSH2*, *MSH6*, *MLH1*, *PMS2*) and hence a mutator phenotype. Overall, these results suggest that a high prevalence of alterations in GIS genes can explain how genome stability is compromised in these two cancers. Consistent with this view, there was no evidence for significant alteration of GIS genes in AML, a cancer that is not associated with high levels of genome instability[86]. Defects in some of the human genes identified here have been implicated in cancer (e.g., *BRCA1*, *BRCA2*, *ATM*, *BLM*, *REV3L*, *PBRM1*), and some of the genes have been associated with the suppression of genome instability (e.g., *WRN*, *BLM*, *ATM*, *ATR*, *BRCA1*, *BRCA2*) or with pathways thought to act in the suppression of genome instability (e.g., *RAD17*, *RAD50*, *XRCC6*, *TP53BP1*)[102, 103]. Our functional studies in *S. cerevisiae* provide evidence that many of the human GIS gene

homologs likely act in the suppression of genome instability in human cells and provide a restricted, prioritized list of human genes for genetics and functional validation studies.

Supplementary information available:

https://www.nature.com/articles/ncomms11256#supplementary-information

3.5 Materials & Methods

3.5.1 Plasmid construction

The plasmid pRDK1590, which is a version of pRS315[104] in which the *LEU2* open reading frame was replaced by the nourseothricin-resistance (*NAT*) open reading frame, was constructed by gap repair in the *S. cerevisiae* strain BY4741 as follows. BY4741 was transformed with *Afl*II-digested pRS315, and the *NAT* open reading frame was amplified from plasmid pFA6a-natNT2 (Janke C. et al, Yeast 2004, 21) using the primers CTT TTA CAT TTC AGC AAT ATA TAT ATA TAT TTC AAG GAT ATA CCA TTC TAa tgg gta cca ctc ttg acg a and ATT TCA TTT ATA AAG TTT ATG TAC AAA TAT CAT AAA AAA AGA GAA TCT TTt tag ggg cag ggc atg ctc a, where uppercase letters correspond to pRS315 sequence and the lowercase letters correspond to *NAT* sequence. The plasmid pRDK1593 was generated by sub-cloning a P_{LEU2}-*NAT*-containing *Bsr*GI to *Xba*I fragment from pRDK1590 into pRS305[104]

digested with *Bsr*GI and *Xba*I and was subsequently used as a template for PCR amplification of P_{LEU2}-*NAT* for generating gene disruptions.

3.5.2 Query strain construction

The selectable markers used in the MATα query strains in systematic mating in the original SGA protocol are incompatible with the genetic markers required for GCR assays. Therefore, different selectable markers were introduced into MATα query strains containing GCR assays. The selected markers were as follows. First, because the GCR assay requires *CAN1*, which interferes with use of canavanine in combination with thialysine to kill diploid strains in the SGA protocol[90], we introduced a deletion of *LYP1* and the cycloheximide-resistant *cyh2-Q38K* mutation[105] into our strains, allowing the use of thialysine and cycloheximide to kill diploid strains in our SGA protocol. Second, we introduced a copy of *LEU2* driven by the *MFA1* promoter near the *YFR016C* gene to select for MAT**a** haploid progeny. Third, we replaced the native *CAN1* gene with a selectable nourseothricin-resistance gene driven by the *LEU2* promoter in the dGCR and sGCR assay strains. The MAT**a** and MATα strains with P_{LEU2}-*NAT* were nourseothricin-resistant when grown on complete synthetic media (CSM). However, the MAT**a** strains were not nourseothricin-resistant on YPD (1% Bacto-yeast extract, 2% Bacto-peptone, 2% dextrose) medium, which is potentially due to increased expression of the Leu2 protein in MAT**a** strains, resulting in down-regulation of the *LEU2*

promoter; this did not interfere with the selection scheme because the selections were performed in the appropriate CSM-dropout media.

The required strains were constructed in the following steps. First BY404 (Brachmann CB et al. 1998, Yeast 14(2)) (MAT**a** *ade2::hisG his3Δ200 leu2Δ0 trp1Δ63 ura3Δ0*) was crossed with RDKY3686 (MATα *hom3-10 lys2-10A his3Δ200 leu2Δ1 trp1Δ63 ura3-52*) and sporulated to isolate RDKY7595 (MAT**a** *lys2-10A hom3-10 his3Δ200 leu2Δ0 trp1Δ63 ura3Δ0*) and RDKY7594, a MATα version of RDKY7595. *URA3* was amplified from pRS306 with the primers 5'-GGA GTT TAT GTT TAT ATA CAC CGG TGT AGG CTG TGC GTT GGT GTG AAC ACg agc aga ttg tac tga gag tgc acc-3' and 5'-GGC TGT ATG ACT ACA GTT GCA TGC GGA GAC GGC TTC AAC AGC AAC AGC AAc tcc tta cgc atc tgt gcg gta tttc-3' and inserted 3' to the *YFR016C* gene to generate RDKY7596. The *iYFR016C::URA3* insertion was then replaced with a P_{MFA1}-*LEU2* construct amplified from FYAT258, generously provided by D. Bernard[106], using the primers 5'-GGA GTT TAT GTT TAT ATA CAC CGG TGT AGG CTG TGC GTT GGT GTG AAC ACg taa caa tag atc cac tag-3' and 5'-GGC TGT ATG ACT ACA GTT GCA TGC GGA GGC TTC AAC AGC AAC AGC Aaa ttt aag tat tca ctt tcg-3' to generate RDKY7597 (MAT**a** *lys2-10A hom3-10 his3Δ200 leu2Δ0 trp1Δ63 ura3Δ0 iYFR016C::*P_{MFA1}*-LEU2*). RDKY7594 was crossed to RDKY7597 and sporulated to generate RDKY7598, a MATα version of RDKY7597. *HXT13* was replaced by *URA3* in RDKY7598 to generate RDKY7599. A wild-type copy of *LYS2* was amplified

from BY4741 and used to replace the *lys2-10A* allele in RDKY7599 to generate RDKY6970. The *LYP1* gene in RDKY6970 was then replaced by *TRP1* to generate RDKY6971. A *cyh2* mutation, determined to be *cyh2-Q38K* by sequencing, was selected in RDKY6971 on YPD plates containing 10 µg/mL cycloheximide (Sigma) to generate RDKY6975 (MATα *hom3-10 his3Δ200 leu2Δ0 trp1Δ63 ura3Δ0 lyp1::TRP1 iYFR016C::*P_{MFA1}*-LEU2 cyh2-Q38K hxt13::URA3*). RDKY6975 and RDKY7597 were crossed, and the resulting diploid was sporulated to obtain RDKY7625 (MATα *hom3-10 his3Δ200 leu2Δ0 trp1Δ63 ura3Δ0 lyp1::TRP1 iYFR016C::*P_{MFA1}*-LEU2 cyh2-Q38K*). The P_{LEU2}*-NAT* gene was amplified from pRDK1593 and integrated into the *CAN1* locus in RDKY7625 to generate RDKY7629. The *CAN1/URA3* cassette with flanking targeting sequences was amplified from pRDK1378 and pRDKY1379 and integrated into RDKY7629 to generate the dGCR query strain RDKY7635 (MATα *hom3-10 ura3Δ0 leu2Δ0 trp1Δ63 his3Δ200 lyp1::TRP1 cyh2-Q38K iYFR016C::*P_{MFA1}*-LEU2 can1::*P_{LEU2}*-NAT yel072w::CAN1/URA3*) and the sGCR query strain RDKY7964 (MATα *hom3-10 ura3Δ0 leu2Δ0 trp1Δ63 his3Δ200 lyp1::TRP1 cyh2-Q38K iYFR016C::*P_{MFA1}*-LEU2 can1::*P_{LEU2}*-NAT yel068c::CAN1/URA3*), respectively (Supplementary Table 18). The tyGCR assay strain was constructed by crossing RDKY6975 with RDKY6593[12] and sporulating the resulting diploid to recover RDKY7046 (MATα *hom3-10 ura3Δ0 leu2Δ0 trp1Δ63 his3Δ200 lyp1::TRP1 cyh2-Q38K iYFR016C::*P_{MFA1}*-LEU2 iYEL062W::Ty912-hphNT1*

hxt13::URA3). Disruption of the 43 query genes in RDKY7635 with *HIS3* was performed using standard methods (Supplementary Table 18).

3.5.3 First-generation set of bait strains

The first-generation set of bait strains (Supplementary Table 1) was primarily obtained from strains present in the *S. cerevisiae* deletion collection (Open Biosystems). The mutant strains were chosen based on the 1,041 genes identified in our *in silico* screen for candidate GIS genes[85] (Supplementary Table 1). Among the 1,041 genes, 46 genes were not included; the majority of these 46 genes were either essential for viability or sporulation or encoded *TLC1*, which is a non-protein-coding gene and therefore not present in the available deletion collection (Supplementary Table 1). An additional 12 genes were not included because the 1,041 genes in the *in silico* screen were finalized after the first-generation bait strain set was selected (Supplementary Table 1). Mutations in some of the 1,041 candidate GIS genes were not present in the deletion collection and were subsequently constructed in BY4741, including *mec1::G418 sml1::hph*, *ddc2::G418 sml1::hph*, *rad53::G418 sml1::hph*, and *mrc1-aq.G418*. Additionally, we constructed a control strain by replacing *leu2Δ*0 present in BY4741 with the G418-resistance marker, which allows *leu2::G418*-containing progeny to be selected during systematic mating; these control strains are labeled as *leu2Δ* in the figures. We also added mutations in 60 additional genes associated with pathways implicated by the 1,041 genes identified in the *in silico* screen but

that were not present in that gene set (Supplementary Table 1). We verified all the deletions by PCR amplification using primers that hybridized within the inserted G418-resistance cassette and primers that hybridized to flanking sequences. Deletions that could not be verified were either replaced by crossing a verified BY4742 deletion strain with BY4741 and sporulating the resulting diploid or by constructing new strains by PCR-mediated gene disruption in BY4741 when a verified BY4742 strain was unavailable (Supplementary Table 19). The final first-generation mutation strain set included 1,058 strains (corresponding to deletions of 1,055 genes of interest with two additional *mrc1* and *rad53* alleles and the *leu2Δ* control deletion; Supplementary Table 1) [85].

3.5.4 Second-generation set of bait strains

To facilitate double mutant strain production, we divided the first-generation bait strain collection into two groups, (i) a “high-priority” set (502 strains) and (ii) a “low-priority” set (555 strains; Supplementary Table 1). The high-priority set contained mutations in GIS genes and genes with patterns of genetic interactions that were most similar to those of known GIS genes[85] (Supplementary Fig. 12). During the initial construction and analysis of double mutant strains, we identified four mutations, *dia2Δ*, *exo1Δ*, *rrm3Δ*, and *rtt107Δ*, out of 30 mutations tested at the time, which interacted with the largest number of bait mutations in the high-priority set, resulting in increased GCR strain scores. No other set of the final 43 query mutations interacted with

more than 90% of the mutations that the *dia2Δ*, *exo1Δ*, *rrm3Δ*, and *rtt107Δ* mutations were found to interact with. We crossed these four mutations to the low-priority set of mutants and scored the resulting double mutants. These 4 query mutations showed genetic interactions with a much lower proportion of the mutations in the low-priority mutation set compared to the high-priority mutation set (Supplementary Fig. 13). We then identified mutations in the low-priority set that (i) increased the GCR strain scores in at least one of the dGCR, sGCR, or tyGCR assay-containing strains (22 mutations), (ii) showed interactions with at least one of the *dia2Δ*, *exo1Δ*, *rrm3Δ*, and *rtt107Δ* mutations in the dGCR assay (87 mutations, 9 in common with group (i)), or (iii) could not be evaluated as we did not recover strains when crossing the wild-type query strains or dGCR assay+mutant query strains (39 mutations). We then added strains containing these mutations to the strains containing the high-priority mutations. This resulted in a second-generation bait strain collection containing 639 strains that were then crossed to the remainder of the dGCR+query mutation strains. This resulted in a second-generation bait strain collection containing 639 strains that were then crossed to the remainder of the dGCR+query mutation strains.

3.5.5 Screen for GCR-suppressing genes and interacting genes

Query strains grown on YPD-agar were crossed to arrayed strains containing bait mutations on YPD-agar in quadruplicate by pinning onto a fresh YPD agar plate using a Singer RoToR robot (Singer Instruments, UK)

and grown for 1-2 days at 30˚C. The cells were then subjected to two rounds of pinning onto diploid selection medium (YPD-agar containing 200 µg/mL geneticin (G418; Gibco) and 100 µg/mL nourseothricin; clonNAT (Werner BioAgents)) and grown for 1-2 days at 30˚C. The cells were then pinned onto pre-sporulation medium (containing 15 g Difco nutrient broth, 5 g Bacto-yeast extract (Fisher Scientific), 10 g Bacto-agar (Fisher Scientific), and 62.5 mL 40% glucose per 500 mL) and grown for 3 days at 30˚C. Cells from the pre-sporulation medium were then pinned onto sporulation medium (10 g potassium acetate, 0.05 g zinc acetate, 20 g Bacto-agar per liter, containing a final concentration of 50 µg/mL G418 and 25 µg/mL nourseothricin) and incubated for 7 days at 30˚C. The resulting spore-containing cells were then subjected to two rounds of pinning onto diploid killing medium (1.7 g yeast nitrogen base without amino acids and without ammonium sulfate (Fisher Scientific), 1 g L-glutamic acid monosodium salt (Sigma), 2 g CSM dropout mix (Methods in Molecular Biology, Vol 313, Yeast Protocols: Second Edition, Humana Press 2005) without uracil, lysine, leucine, and, when appropriate, histidine, 20 g Bacto-agar, 50 mL of 40% glucose per liter, containing a final concentration of 50 µg/mL thialysine (S-[2-aminoethyl]-L-cysteine hydrochloride, Sigma), 10 µg/mL cycloheximide, 200 µg/mL G418, and 100 µg/mL nourseothricin) followed by growth for 5 days at 30˚C for the first pinning and 2 days at 30˚C for the second pinning. Cells were then subjected to two rounds of pinning and growth on haploid selection medium (1.7 g yeast

nitrogen base without amino acids and without ammonium sulfate, 1 g L-glutamic acid monosodium salt, 2 g CSM dropout mix without leucine, uracil, and, when appropriate, histidine, 20 g Bacto-agar, 50 mL of 40% glucose per liter, containing a final concentration of 200 µg/mL G418 and 100 µg/mL nourseothricin) and grown for 2 days at 30˚C. Then the cells were pinned and grown on YPD-agar followed by storage at -85˚C.

3.5.6 GCR patch tests

A minimum of 3 independent spore clones were isolated from each mutant progeny pool arising from the SGA protocol and then grown as patches on a YPD-agar plate at 30°C for two days and replica-plated onto CSM –Arg media containing 60 mg/L canavanine (Sigma) and 1 g/L 5-fluoroorotic acid (US Biological). The number of papillae growing on the GCR medium was scored using a semi-quantitative scoring system as follows: no papillae, 1: 1-5 papillae (this was on average the number of papillae observed with the *leu2Δ* control strain for the dGCR assay), 2: 6-15 papillae, 3: 16- a countable number of papillae (~150-200), 4: papillae that were too many or too close together to count, 5: a lawn of papillae covering the entire patch) (Fig. 1d). Then the scores for all independent patches analyzed for each mutant were averaged to generate a GCR strain score (Supplementary Table 1). Negative scores were assigned to strains that did not grow and so that these strains could be ignored during the analysis.

3.5.7 Determination of GCR rates

The media and protocol for strain propagation and measuring GCR rates were as described previously[11].

3.5.8 Determination of an optimal cutoff score

Using 101 paired GCR rates and average GCR patch scores for single mutants in the dGCR assay and 43 strains resulting from the crosses of mutant dGCR query strains with the *leu2Δ* control strain, we determined an optimal cutoff score as described[107]. Briefly, for any given cutoff value, c_i, we calculated the sensitivity, which is the fraction of mutations causing increased GCR rates that we include as $SENS_i = TP_i / (TP_i + FN_i)$, where TP_i is the number of true positives (mutants with a GCR rate at least 3-fold higher than wild-type with a score >= c_i) and FN_i is the number of false negatives (mutants with a GCR rate at least 3-fold higher than wild-type with a score < c_i). For each cutoff value, we also calculated the specificity, which is the fraction of mutants that do not have increased GCR rates that we reject: $SPEC_i = TN_i / (TN_i + FP_i)$, where TN_i is the number of true negatives (mutants with a GCR rate less than 3-fold higher than wild-type with a score < c_i) and FP_i is the number of false positives (mutants with a GCR rate less than 3-fold higher than wild-type with a score > c_i). An optimal cutoff for balancing sensitivity and specificity can be determined by optimizing the cost function $w_1 SENS_i + w_2 SPEC_i$ as a function of c_i. Here, we weighted sensitivity slightly higher than

specificity (w_1=2, w_2=1) with the rationale that false negatives were more problematic because false positives could be identified by quantitative rate testing. We found that the optimal cutoff c_i was 1.38 for the set of 101 rate/score pairs solely from the wild-type dGCR cross and for the set of 144 rate/score pairs from the wild-type dGCR cross and the *leu2* double mutants from the mutant dGCR crosses (Supplementary Fig. 2). With equal weights, the optimal cutoff was slightly higher, ~1.69. As expected, analysis of Receiver-Operator Characteristic (ROC) curves[108] showed that mutations causing higher GCR rates were clearly better detected by these patch-based GCR strain scores than mutations that only weakly increased the GCR rates (Supplementary Fig. 2). We also used the Kolmogorov-Smirnov test as extended for discrete null distributions as implemented in R[109] to calculate p-values for differences between the distribution of patches from the *leu2Δ* control strain and each single mutant. Unlike calculations based on the average GCR patch score, this test included the number and distribution of all observed patches. We found that the list of mutant strains with significantly different patch scores that were higher than the *leu2Δ* control strain (p<0.01) was essentially the same as the list of strains identified by minimizing the false-positive and false-negative errors as described above.

3.5.9 Analysis of *S. cerevisiae* modules

Protein complex and pathway (module) definitions were extracted from a variety of studies[110-117] as well as manually curated complexes such as

CYC2008v2 and YHTP2008[118], the *Saccharomyces* Genome Database GO complex and pathway definitions[119], *S. cerevisiae* KEGG pathways[120], and *S. cerevisiae* MetaCyc pathways[121]. Modules containing genes that showed increased GCR scores alone or enhanced the GCR scores of query mutations were identified. Hits were manually curated to identify well-supported modules, and these modules were divided into two groups. The first group contained modules with more than one genes that when mutated shared at least one query mutation that caused increased GCR scores. The second group contained modules for which only a single gene caused increased GCR interactions when mutated or for which multiple genes caused increased GCR interactions but lack shared interacting partners.

3.5.10 Analysis of cancer genomics data

TCGA data [87, 88], including expression z-scores, methylation and GISTIC CNV (copy number variation) data were obtained from the cBIO portal (http://www.cbioportal.org) through the CGDS-R package. Somatic mutation data were obtained from a local compilation[97] that includes data from the TCGA and COSMIC as well as a compilation of data from the literature. As previously described [97], all mutations for a given tumor were used in the S-score calculation. For all other analyses, only TCGA mutation data were used. As defined by TCGA, putative copy-number calls on samples were determined using the GISTIC algorithm [122]. Boxplots were generated using ggplot2, a

graphics tool for the R statistical package (http://ggplot2.org). For expression data, the Z-score metrics adopted by TCGA were used.

3.5.11 Computational prediction of the functional impact of missense mutations

To identify putative deleterious missense mutations in our gene set, we used 5 different computational algorithms resulting in 6 different tests per mutation: SIFT[123], PolyPhen-2[124], MutationTester[125], Fathmm[126] and LTR[127]. Two versions of PolyPhen-2 were used, each one trained by a different dataset (HDIV and HVAR). Each missense mutation was assigned a score called the "Ndamage score" that was the number of prediction tests in which the mutation scored as deleterious. To be considered “predicted deleterious”, a given missense mutation had to have an Ndamage score of 5 or 6.

3.5.12 Simulations to determine statistical significance in cancer genomics analyses

Two types of simulations were used. First, a gene-set enrichment analysis was performed to evaluate whether the set of GIS genes were enriched with genes with extreme S-scores (≤-2 or ≥2). Ten thousand random sets of the same size (number of genes) were selected from the pool of all human genes, and for each set the number of genes with extreme S-scores was defined. A p-value for the enrichment analysis was determined by ranking

the real set in the random set distribution. Second, we evaluated whether a given set of genes was enriched for different types of mutations (or combinations of different types). To avoid any bias due to different gene lengths, we normalized the analysis for the total length of the corresponding gene set (in amino acids of the longest coding region for each gene). The total number of amino acids for the real set was randomly selected from the total pool of human genes (ten thousand random sets). The number of mutations in the real set was then compared to all random sets, and a p-value for the enrichment analysis was determined by ranking the real set within the distribution of the random sets.

3.6 Acknowledgements

We thank Vincent Pennaneach and Jorrit Enserink for assistance in construction of early versions of the query strains. We thank Renan Valieris for assistance in the construction of the MySQL database containing TCGA data. This work was supported by NIH grant R01-GM26017 to R.D.K, NIH grant F30-CA177240 to R.N., CAPES (Brazil) grant (23038.004629/2014-19) to S.J.S., and support from the Ludwig Institute for Cancer Research to R.D.K. and C.D.P. Chapter 3, in full, is a reprint of the material as it appears in A Genetic Network that Suppresses Genome Rearrangements in *Saccharomyces cerevisiae* and Contains Defects in Cancers. *Nature Communications*. 2016 Apr 13;7:11256. doi: 10.1038/ncomms11256. Putnam

CD*, Srivatsan A*, Nene RV*, Martinez SL*, Clotfelter SP, Bell SN, Somach SB, de Souza JE, Fonseca AF, de Souza SJ, Kolodner RD. The dissertation author is third author for this paper (four co-first authors*).

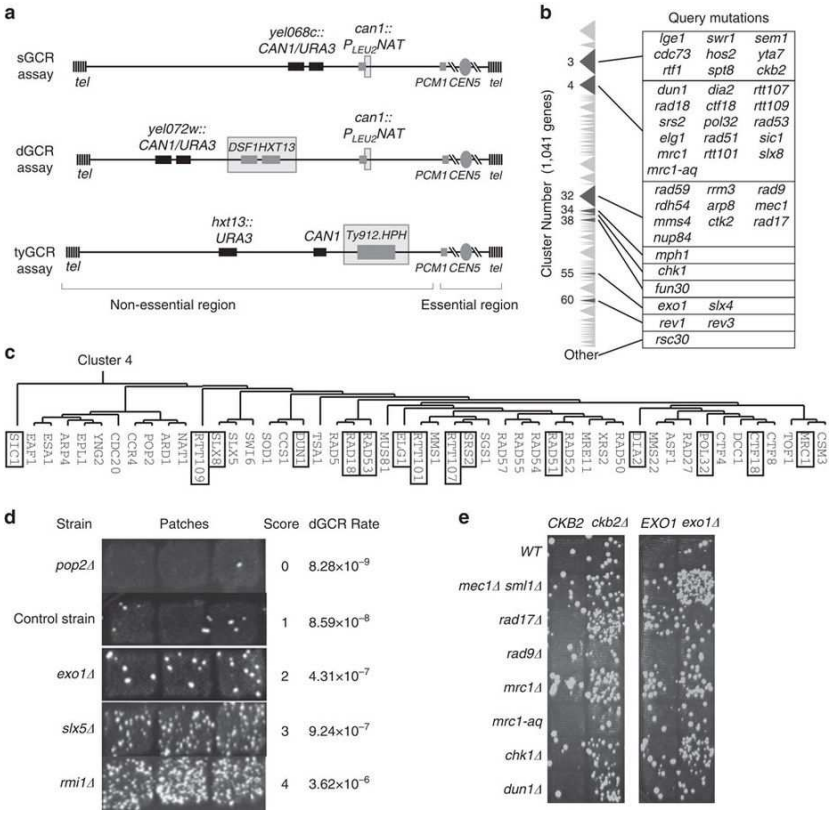


Figure 3.1 Details of the systematic screen

a. The sGCR, dGCR, and tyGCR assays involve simultaneous selection against the *CAN1* and *URA3* genes inserted into the terminal non-essential region of the left arm of chromosome V. The GCR breakpoint region is the region between the telomeric *CAN1* and *URA3* genes and the first centromeric essential gene, *PCM1*. Homologies within the GCR breakpoint regions, including the ~100 bp fragment of *YCLWdelta5* sequence introduced by *can1::*P_{LEU2}*-NAT*, the *DSF1/HXT13* segmental duplication, and the inserted Ty912 element, are indicated with grey boxes. **b.** The query mutations were primarily selected from the previously described gene clusters 3, 4, 32, 55, and 60 generated by clustering the candidate GCR-suppression genes by genetic interactions [85]. Clusters 3, 4, and 32 had the greatest number of GCR-suppressing genes. Triangles indicate the relative size of the cluster in terms of the number of genes, and the darker triangles are the clusters from which query mutations were selected. **c.** Query mutations (indicated by the boxes) in non-essential genes in cluster 4 were selected to provide the greatest genetic diversity by picking 1 or 2 mutations from most sub-clusters. Query mutations were similarly selected from clusters 3 and 32. **d.** The semi-quantitative scoring strategy assigns a number between 0 and 5 to each patch depending on the number of papillae (0: no papillae, 1: 1-5 papillae, 2: 6-15 papillae, 3: 16- a countable number of papillae (~150-200), 4: papillae that were too many or too close together to count, 5 [not shown]: a lawn of papillae covering the entire patch). For each strain, a minimum of 3 individual GCR patch scores were averaged to calculate the GCR strain score. Increases in the GCR strain score were paralleled by increases in GCR rates measured by the fluctuation method. **e.** Patch tests documenting genetic interactions involving mutations in either *CKB2* or *EXO1*. The status of *CKB2* or *EXO1* is indicated across the top of each set of patches, and the bait mutations tested are indicated along the left side of each set of patches.

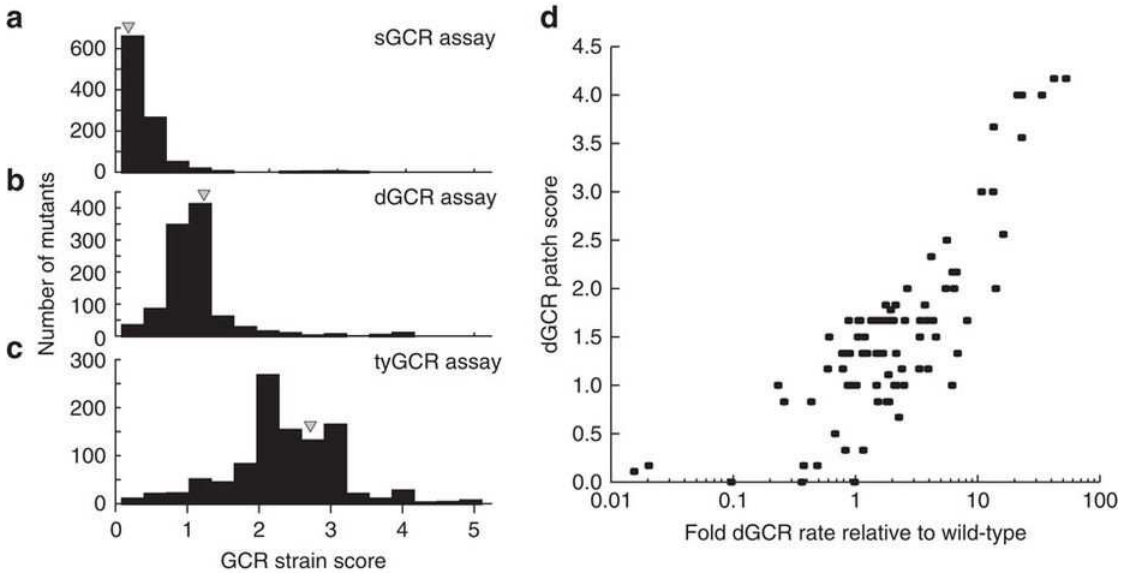


Figure 3.2 Assaying single mutant strains using GCR strain scores

a.-c. Histograms of the distribution of GCR strain scores for single mutant strains from the sGCR (panel a), dGCR (panel b), and tyGCR (panel c) assays reveal that the average GCR strain score increases with the GCR rate for each GCR assay and that the score of the *leu2Δ* control strain (grey triangle) generally lies at the peak of each histogram, suggesting that many of the mutations tested do not substantially affect the GCR strain score as single mutations. **d.** The fold increase in the GCR rate is correlated with the GCR strain score for systematically generated strains containing the dGCR assay.

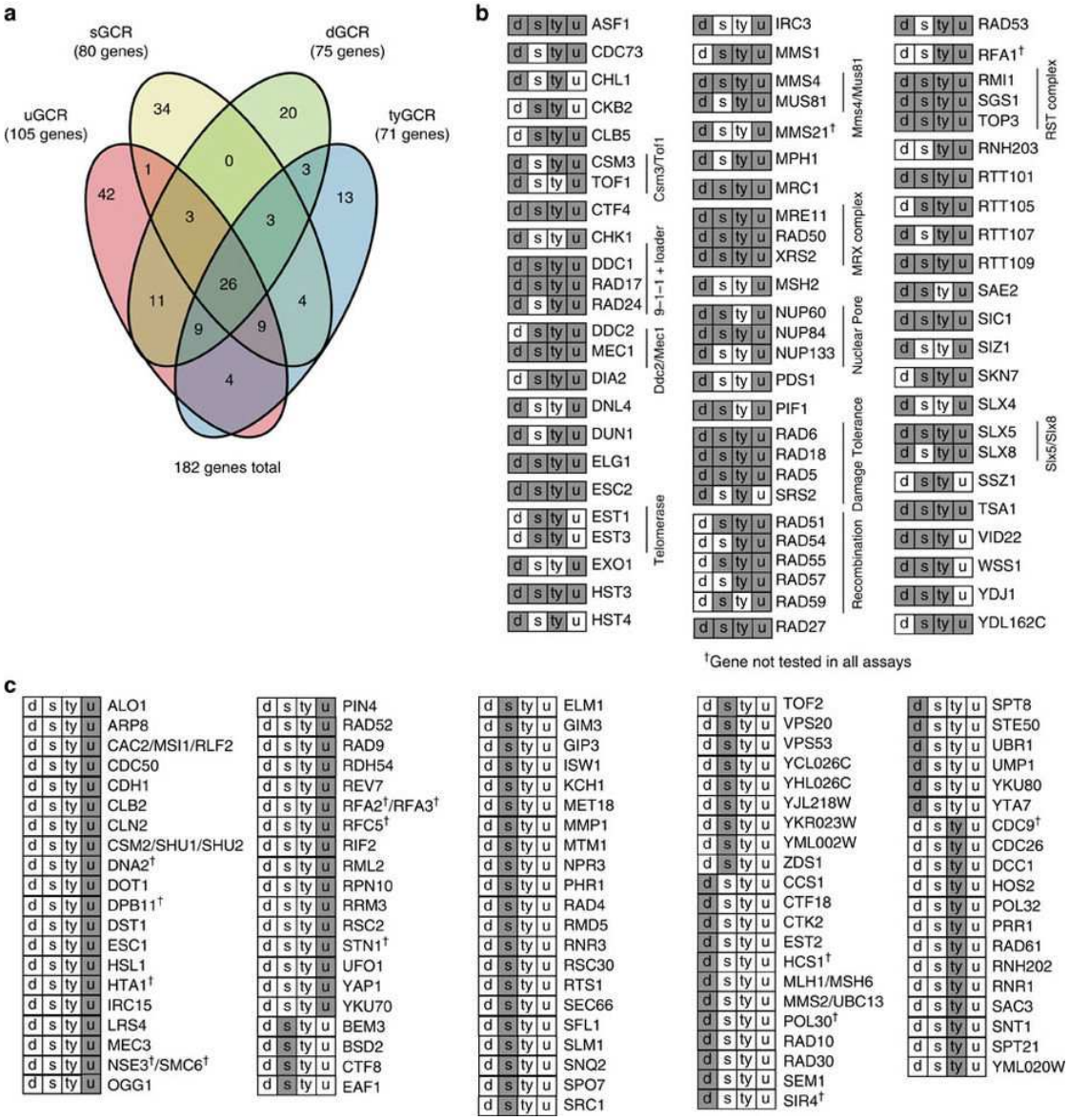


Figure 3.3 Summary of the increased GCR rate of single mutant strains identified using patch tests

a. Venn diagram indicating the number of genes that suppress GCRs in each of the GCR assays used. **b.** Genes implicated in suppressing GCRs in more than one GCR assay. The boxes indicate the assays (d=dGCR, s=sGCR, ty=tyGCR, u=uGCR) in which the listed gene suppresses (grey) or does not suppress (white) GCRs. Note that uGCR assays are GCR assays lacking repetitive sequences in the GCR breakpoint region that have been utilized in previous studies[3, 5]. Many genes unique to the uGCR assay are primarily genes in which mutations cause small but significant increases in GCR rates, which were identified using fluctuation assays but are difficult to identify by the semi-quantitative patch score method used here. **c.** Genes implicated in suppressing GCRs in only one GCR assay, annotated as in panel b.

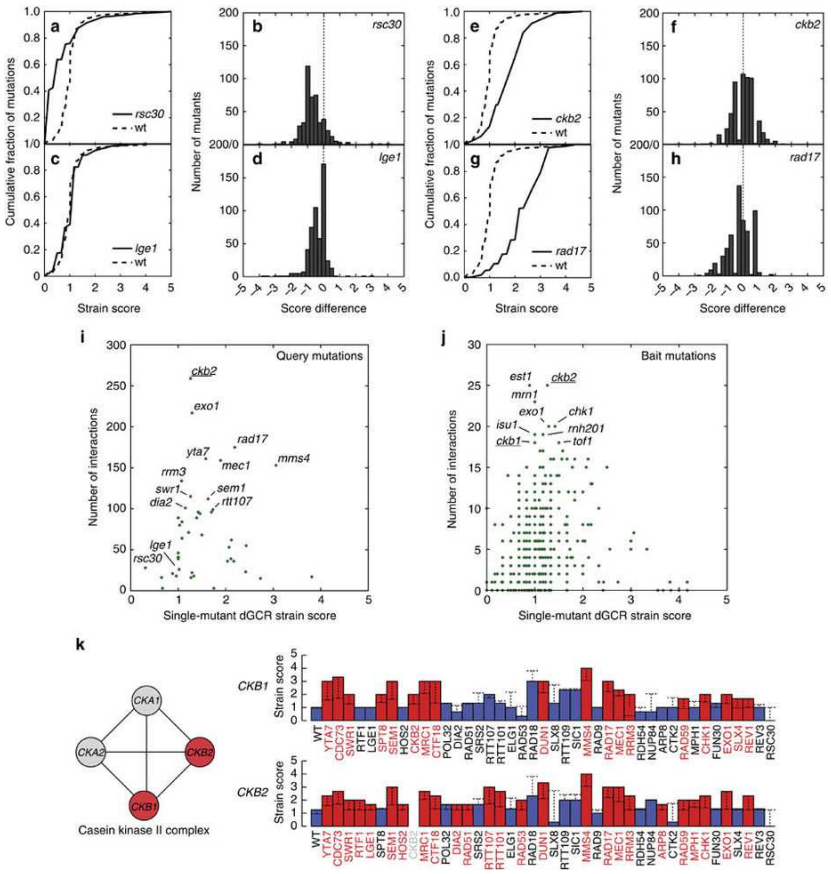


Figure 3.4 Identification of genetic interactions involved in suppressing genome instability

a, c, e, g. Plots of the cumulative fraction of mutations below specific GCR strain scores for crosses for strains containing bait mutations and in addition one of the *rsc30Δ*, *lge1Δ*, *ckb2Δ*, or *rad17Δ* query mutations (solid line) compared with the distribution from the crosses of the bait mutations to the wild-type strain (dashed line). **b, d, f, h.** Histograms of the number of mutations in combination with *rsc30Δ*, *lge1Δ*, *ckb2Δ*, or *rad17Δ* as a function of the GCR strain score difference, which is the GCR strain score of the double mutant strain (*aΔ bΔ*) minus the GCR strain score of the higher of the two single mutant strains (*aΔ* or *bΔ*). **i.** Plot of the number of GCR-based interactions as a function of the single mutant GCR strain score for the 43 mutant query strains. Query mutations with large numbers of interactions or those displayed in panels a-h are indicated. **j.** Plot of the number of GCR-based interactions as a function of the single mutant GCR strain score for bait mutations. Bait mutations with large numbers of interactions are indicated. **k.** Analysis of physical interaction data for the casein kinase II complex is shown (left) with reported physical interactions in BioGrid (lines) between complex components (circles). Components with known GCR interactions are in red; untested components (*CKA2*) or those tested with only 4 query mutations (*CKA1*) are in grey. Display of the genetic interactions between the *ckb1Δ* and *ckb2Δ* bait mutations and the 43 query mutations (right). Bar heights indicate the strain score for the double mutant, and bar colors correspond to the presence (red) or absence (blue) of an increased level of genome instability in the double mutant as observed in patch tests relative to the respective single mutant with the highest level of increased genome instability; the horizontal dashed line corresponds to the GCR strain score of the higher of the two single mutations. Missing bars and query names in grey correspond to double mutant strains that were not generated in the crosses performed.

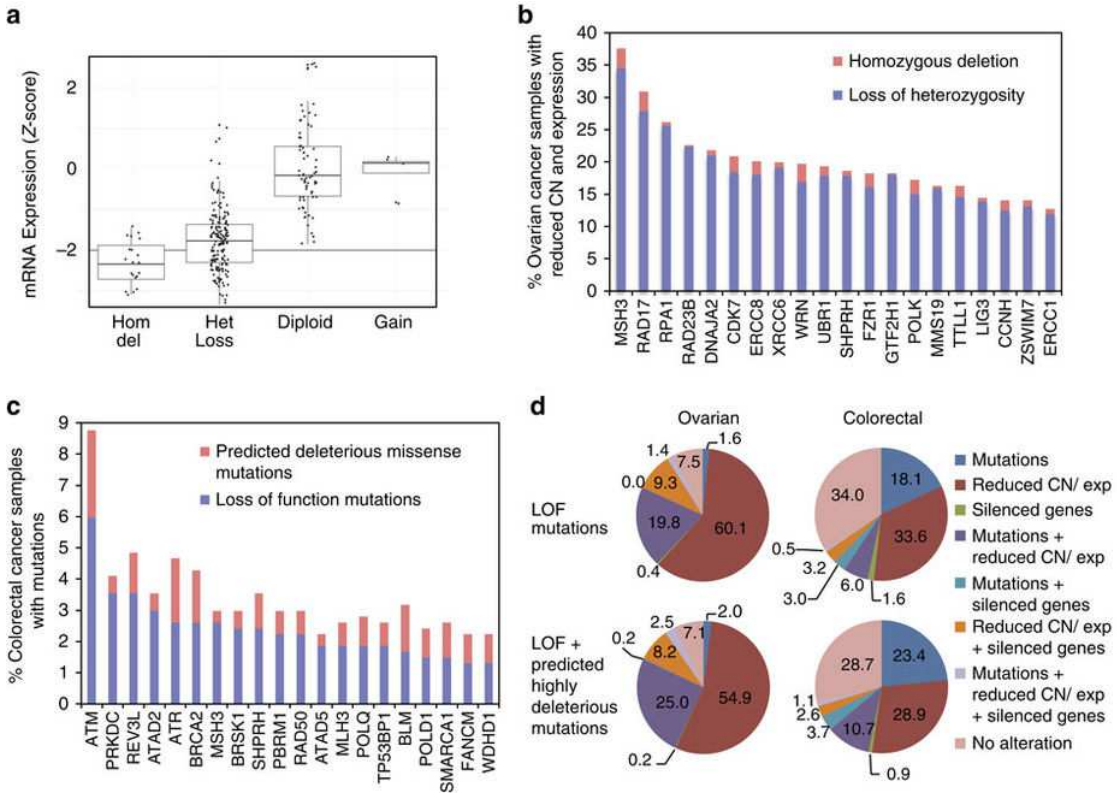


Figure 3.5 Analysis of the ovarian and colorectal cancer TCGA data for alterations in GIS genes

This figure summarizes the data analysis presented in Supplementary Tables 11 - 16. **a.** Box plot of the RNA Seq data for the copy number (GISTIC -2, Homozygous Deletion; GISTIC -1, Heterozygous Loss; GISTIC 0, Diploid; GISTIC 1, Gain) vs. the Z-score for mRNA expression of *RAD17* in ovarian cancer. **b.** Histogram of the frequency of reduced copy number with reduced mRNA expression for the top 20 most-altered GIS genes in ovarian cancer. **c.** Histogram of the frequency of mutations in the top 20 most-altered GIS genes in colorectal cancer. Data for *MSH2*, *MSH6* and *MLH1* were excluded as defects in these genes predominantly cause increased rates of accumulation of point mutations. Predicted deleterious missense mutations are those that scored as deleterious in 5 or 6 out of 6 functional prediction tests. **d.** Pie charts showing the % of ovarian (left) and colorectal (right) cancer samples with different combinations of mutations, reduced copy number with reduced expression and silencing among all samples for which any type of genomics data were available. Analysis of LOF mutations alone (Top) and LOF + predicted deleterious missense mutations (Bottom) are presented separately. Note that 19% of the ovarian and 25% of the colorectal cancer cases were not analyzed for all types of potential alterations, and consequently the values presented are an underestimate.

Table 3.1 Modules with shared interactions in the dGCR enhancer screen

Process	***Module****
DNA repair	*core mitotic homologous recombination* *base excision repair* *Cul8-RING ubiquitin ligase complex* *DNA ligase IV complex* *DNA polymerase zeta – Rev1 complex* *Ku complex* *Mlh1-Mlh2 complex* *Mlh1-Mlh3 complex* *Mlh1-Pms1 complex* *Mms2-Ubc13 complex* *Mms4-Mus81 complex* *Msh2-Msh6 complex* *nucleotide-excision repair factor 1 (NEF1) complex* *nucleotide-excision repair factor 4 complex* *Rad1-Rad10-Saw1 complex* *ribonuclease H2 complex* *Shu complex* *Slx1-Slx4 complex*
DNA replication	*DNA polymerase epsilon complex* *ribonucleoside-diphosphate reductase complex* *telomerase*
Chromosome cohesion and segregation	*Ctf18 RFC-like complex* *Ctf19 complex (includes COMA complex)* *dynactin complex* *monopolin* *Msh4-Msh5 complex* *prefoldin complex*
Cell cycle checkpoints	*anaphase-promoting complex (APC/C)* *Cdc28 cyclin-dependent kinase complexes* *Mec1-Ddc2 complex* *protein phosphatase (PP4) complex* *Rad17-Ddc1-Mec3 complex + Rad24-Rfc2-5 clamploader* *spindle checkpoint* *Tof1-Csm3 complex*
Chromatin/ Transcription/ mRNA processing	*carboxy-terminal domain protein kinase complex* *CCR4-NOT core complex* *Cdc73/Paf1 complex* *chromatin assembly complex* *Chz1-Htz1-Htb1 complex* *COMPASS complex* *cytoplasmic mRNA processing body* *cytoplasmic Sm-like complex* *Elogin-Cullin-Socs (ECS) ligase complex* *HIR complex* *Ino80 complex* *ISW1a chromatin remodeling complex* *mediator complex*

Table 3.1 cont.

Chromatin/Transcription/ mRNA processing	*Rpd3S complex* *RNA polymerase I complex* *RSC complex* *SAGA complex* *Set3C complex* *SLIK (SAGA-like) complex* *Spt3-Spt8 SAGA subunit of SAGA complex* *Swr1 complex* *U6 snRNP*
Nuclear pore	*nuclear pore nuclear basket* *nuclear pore outer ring*
Proteasome/Protein degradation	*Doa10 ubiquitin ligase complex* *Hrd1p ubiquitin ligase ERAD-L complex* *proteasome 19/22S regulator* *proteasome 20S complex + Ump1 chaperone,* *Rad6-Ubr1 complex* *Ula-Uba3 complex*
Other	*AP-3 adaptor complex* *casein kinase II complex* *Chs5p/Arf-1 binding proteins (ChAPs)* *ESCRT III complex* *Golgi transport complex* *HMC complex* *Kel1-Kel2 complex* *NatA complex* *Sod1-Ccs1 complex* *Ssk1-Ssk2 complex*

CHAPTER 4

A role for *TOR2* in maintenance of genome stability in *S. cerevisiae*

4.1 Summary

The TORC2 complex is known to function in a signaling network that controls a multitude of cell growth processes. It is known that defects in this complex are associated with sensitivity to DNA damaging agents. Using the temperature sensitive *tor2-21* allele we found that deletion of Tor2 led to an increase in accumulation of gross chromosomal rearrangements (GCRs) in two different GCR assays containing either a short, highly repetitive sequence (sGCR assay) or a longer, lower-copy repetitive sequence (dGCR assay). We also observed a change in the structure of GCRs observed for the sGCR assay corresponding to the TORC2 defect.

4.2 Introduction

The *S. cerevisiae* genome contains two Tor genes, *TOR1* and *TOR2*. *TOR1* and *TOR2* encode phosphatidylinositol-related protein kinases that are paralogs of the mammalian TOR (target of rapamycin) kinase, the center of a signaling network that controls cell growth [128]. Tor signaling, which is conserved in eukaryotes, regulates gene transcription, translation, ribosome biogenesis, autophagy and actin polarization among other processes [129]. Both Tor genes act in promoting cell cycle progression and protein synthesis in the presence of nutrients [130] loss of both genes causes arrest in early G1 with characteristics similar to starved cells entering a quiescent stationary phase (G0) [131]. The Tor kinase is found in two multi-protein complexes

comprising different combinations of various subunits. In yeast, these complexes are called TORC1 and TORC2, and in humans they are known as mTORC1 and mTORC2. While TORC1 promotes ribosome biogenesis as it pertains to cell growth [132], the TORC2 complex has an essential and non-redundant role in promoting organization of the actin cytoskeleton [130, 133] via activation of *RHO1* and *RHO2*, which are Rho-like GTPases [134, 135]. Inhibition of the TORC2 causes cells to become sensitive to DNA damaging agents and suggests a role in maintenance of genome stability [136-138]. The *tor2-21* allele, with a permissive temperature of 24ºC and a restrictive temperature of 37ºC, is defective for both the nutrient sensing and cytoskeletal organization functions of *TOR2* [130]. This temperature-sensitive mutation allows for examination of the role of TORC2 deficiency in genome instability.

Although previous work has associated defects in TORC2 with sensitivity to DNA damaging agents [136], this phenotype has not been well characterized. We examined the effect of a *tor2-21* mutation in two different GCR assays (sGCR and dGCR) to understand if the TORC2 complex suppresses chromosome rearrangements mediated by shorter or longer repetitive DNA sequences specifically. In the short sequence homology (sGCR) assay, GCRs in wild-type strains form equally readily by rearrangements mediated short homologies (a ~100 bp fragment of *YCLWdelta5* or a 114 bp repetitive *SUP53* tRNA gene) or by single- copy sequence rearrangements (Nene et al. submitted). In the duplication GCR

(dGCR) assay, GCRs predominantly form via homology-mediated rearrangements between the *HXT13-DSF1* region and other homologies elsewhere in the genome [5]. In the Chr5 dGCR assay used here that contains the *can1::P LEU2 -NAT* locus, short homology-mediated rearrangements and single copy sequence-mediated rearrangements can also be observed.

We sought insight into the rate at which genome rearrangements arise in the absence of Tor2 in these assays as well as the structures of the rearrangements observed. Elucidating the structures of GCRs found in these strains can provide mechanistic insight into whether certain regions of the genome have a propensity to form chromosome rearrangements in the absence of TORC2 as well as the pathways responsible for their formation.

4.3 Results

4.3.1 The *tor2-21* mutation suppresses GCR formation in the sGCR assay

To determine if defects in the Tor pathway can give rise to increased genome instability as measured by the accumulation of gross chromosomal rearrangements (GCRs), we introduced the *tor2-21* allele and other temperature-sensitive mutations affecting the Tor complexes TORC1 and TORC2 into strains containing genetic assays that allow for the detection of GCRs (Figure 4.1A). We found that *tor2-21* caused a 3.5-fold increase in the dGCR rate at room temperature and a 14.8-fold increase at 30ºC and caused a 7.4-fold increase in sGCR rate at room temperature (Table 4.1). Thus, the

strongest effect of a *TOR2* defect was observed in the sGCR assay at the non-permissive temperature. In contrast to the increased GCR accumulation in *tor2-21* strains, we did not observe increased GCR accumulation in temperature sensitive alleles of *LST8* or *TSC11*, which encode some of the subunits of the TORC1 and TORC2 complexes respectively, along with several other proteins [128, 139].

4.3.2 The *tor2-21* mutation distorts the spectrum of GCRs observed in the sGCR assay

We examined the spectrum of GCR structures observed in the sGCR and dGCR assays to understand how the TORC2 complex interacts with the different pathways through which GCRs may form. We characterized the structure of the GCRs selected in these assays either by PCR amplification of the breakpoints or by whole genome sequencing of GCR-containing strains (Figures 4.2-4.4). The GCR spectrum of the *can1::P LEU2 -NAT* dGCR assay (Figure 4.1B) was dominated by the *HXT13-DSF1*-mediated rearrangements and was essentially unchanged from the dGCR lacking the *can1::P LEU2 -NAT* locus (87% vs. 92%) [32]. GCRs selected in the *tor2-21 can1::P LEU2 -NAT* dGCR assay strain were also dominated by *HXT13-DSF1*-mediated rearrangements (75% of isolates); although these rearrangements were slightly reduced relative to the wild-type strain. The remaining isolates were dominated by rearrangements involving the *YCLWdelta5* and *SUP53* tRNA homologies. In contrast, analysis of the *tor2-21* sGCR spectrum showed more

substantial differences from the wild-type sGCR spectrum, and the proportion of *de novo* telomere addition GCRs and *SUP53* homology-mediated GCRs were reduced whereas the number of *YCLWdelta5*-mediated inversions increased (Figure 4.1C).

4.4 Discussion

The *tor2-21* mutation caused increased GCRs in both assays at the temperatures that could be lethal. The genetic data suggest that the TORC2 complex prevents genome rearrangements preferentially between specific kinds of sequences. The *tor2-21* mutation caused increased GCRs in the sGCR assay at the non-permissive temperature, implicating the TORC2 complex in the formation of GCRs in this background. Meanwhile the *tor2-21* mutation promotes the accumulation of GCRs in the dGCR background.

The influence of the *tor2-21* mutation on the spectrum of GCRs observed was different for the two assays. In the dGCR assay, the *tor2-21* mutation decreases the GCR rate and does not affect the spectrum of rearrangements observed. In the sGCR assay, the *tor2-21* mutation increases the GCR rate and distorts the spectrum of GCRs observed to include structures generated by more complex intermediates such as foldback inversions and inverted duplications.

One explanation for this observation is that the mechanisms by which the GCRs arise in the two assays are different, and consequently the

interaction between the TORC2 complex and the relevant pathways leading to GCR formation is different. GCRs in the sGCR background may depend on the nutrient sensing or cytoskeletal organization functions of the TORC2 complex. Analysis of the GCR structures formed in strains with the sGCR *tor2-21* and dGCR *tor2-21* backgrounds provides insight into the relationship between TORC2 activity and sequence homology characteristics recombination targets.

4.5 Materials & Methods

4.5.1 Strain construction

The *tor2-21-G418* allele was integrated into the pre-existing dGCR assay RDKY7635 background to form RKDY8492 (*MATa ura3-delta0 leu2-delta0* (*trp1-delta63* or *TRP1*) *(his3-delta200* or *his3-delta1) met15-delta0 lyp1::TRP1 cyh2-Q38K iYFR016::PMFA1-LEU2 can1::PLEU2-NAT yel072w::CAN1-URA3 tor2-21.G418*). The *tor2-21-G418* allele was integrated into the pre-existing sGCR assay RDKY7964 background to form RKDY8768 (*MATa ura3-delta0 leu2-delta0 (trp1-delta63* or *TRP1) (his3-delta200* or *his3-delta1) met15-delta0 lyp1::TRP1 cyh2-Q38K iYFR016::PMFA1-LEU2 can1::PLEU2-NAT yel068c::CAN1-URA3 tor2-21.G418*). GCRs isolates were collected using standard methods.

4.5.2 Measurement of GCR rates

General methods, including use of YPD and synthetic dropout medias, have been described previously [64]. For each strain, we used 14 or more independent cultures from 2 biological isolates in our fluctuation analyses [140] to calculate the median rates [141].

4.5.3 Analysis of GCR structures by PCR

The Gentra Puregene Yeast/Bacteria Kit (Qiagen) was used to isolate DNA from *S. cerevisiae* strains for further analysis. PCR reactions were performed using the Roche Expand Long-Template PCR System and 500ng of genomic DNA along with previously described primer pairs to identify t(5;14) and t(5;4 or 10) breakpoint junctions [5].

4.5.4 Whole genome sequencing

Isolates whose GCR structure could not be determined by PCR screening were sequenced using next-generation sequencing. Multiplexed paired-end libraries were constructed from 2µg of genomic DNA purified using the Puregene kit (Qiagen) and subsequently treated with 0.15mg (45U) RNase A for 1 hour at 37°C. Genomic DNA in Covaris microtube-50 tubes was sheared into 550bp fragments by sonication using a Covaris M220 instrument at peak incident power of 75W, 10% duty factor, 200 cycles per burst, and treatment time of 40s. Samples were then end-repaired using the End-it DNA End-repair kit (Epicentre Technologies) and A-tailed using the Klenow

fragment (3'→5' exo-, NEB). Common adaptors from the Multiplexing Sample Preparation Oligo Kit and TruSeq PCR-Free LT DNA Sample Preparation Kit (Illumina) were then ligated to the genomic DNA fragments using the Quick Ligation Kit (NEB) and then run on an agarose gel to select for 600bp fragments. Samples were then subjected to 18 cycles of amplification using the Library Amplification Readymix (KAPA Biosystems). The amplified products were fractionated on an agarose gel to select 600 bp fragments, which were quantified with the Qubit dsDNA HS assay kit (Thermofisher). Libraries were subsequently sequenced on an Illumina HiSeq 2000 using the Illumina GAII sequencing procedure for paired-end short read sequencing. Reads from each read pair were mapped separately by bowtie version 2.2.1 [65] to a reference sequence that contained revision 64 of the *S. cerevisiae* S288c genome [66], *hisG* from *Salmonella enterica*, and the *kanMX4* marker. Reads are available from National Center for Biotechnology Information Sequence Read Archive under accession number: SRP107803.

4.5.5 Analysis of GCR structures from sequencing data

Paired-end reads were aligned to the *S. cerevisiae* S288C reference genome release R64.1.1 (February 2011) and R64.2.1 (January 2015) using bowtie-0.12.7. Further analyses to identify copy number changes and identify novel structural variants were performed with version 0.6 of the Pyrus suite (http://www.sourceforge.net/p/pyrus-seq) [15]. Rearrangements relative to the

reference S288c genome were identified by analyzing the read depth distributions, the discordantly mapping read pairs, and/or extracting the sequences of the novel junctions. Associated junction-sequencing reads, which were reads that did not map to the reference but were in read pairs in which one end was adjacent to discordant reads defining a junction, were used to sequence novel junctions. Analysis of the sequencing data identified all of the genetic modifications introduced during construction of the starting strains, such as the *his3Δ200* deletion.

4.6 Acknowledgements

The dissertation author would like to thank Richard D. Kolodner, Paul S. Mischel and Christopher D. Putnam for assisting with the conception and design of the research, data analysis, manuscript writing and mentorship. The dissertation author would also like to thank Anjana Srivatsan and Binzhong Li for strain construction, measurement of GCR rates and NGS library construction as well as members of the Kolodner lab for helpful discussions and comments. This work was also supported by NIH R01 Grant GM26017 to Richard D. Kolodner and Christopher S. Putnam and the Ludwig Institute for Cancer Research. Chapter 4, in part, is in preparation for publication of the material as part of a collaboration between Richard Kolodner, Christopher Putnam and Paul Mischel. The dissertation author is an intermediate author for this paper.

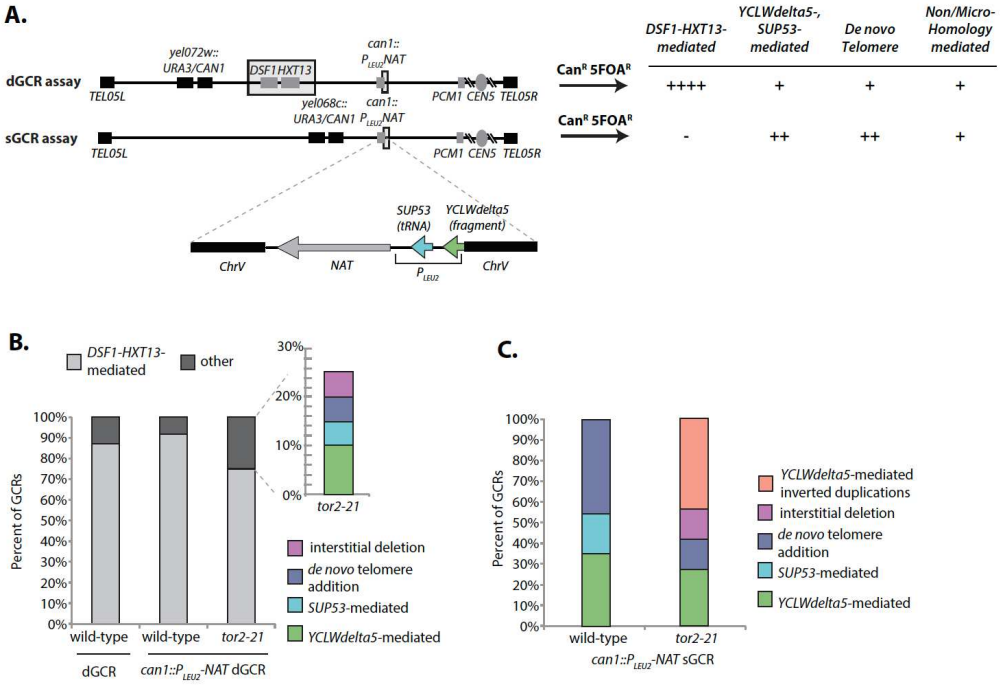
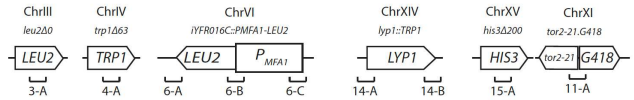


Figure 4.1 The effect of a *tor2-21* mutation on GCRs structures observed in the sGCR and dGCR assays

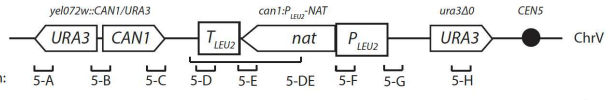
A. The dGCR and sGCR assays involve selection against the *CAN1* and *URA3* genes placed on the terminal non-essential region of the left arm of chromosome 5. Breakpoints must occur between the most telomeric essential gene, *PCM1*, and the *CAN1* and *URA3* genes. The dGCR assay primarily selects GCRs mediated by non-allelic HR between the *DSF1/HXT13* segmental duplication (grey outline) and regions of divergent homology on chromosomes 4, 10 and 14. The sGCR assay contains a portion of chromosome 3 containing the *SUP53* tRNA and ~100 bp fragment of *YCLWdelta5* at the *can1::P LEU2 -NAT* insertion (also found in the version of the dGCR assay used here) and allows selection of HR-mediated rearrangements that target many tRNA and Ty-related sequences in the S. cerevisiae genome as well as nonhomology- and microhomology-mediated translocations, interstitial deletions, and de novo telomere addition-mediated GCRs. The number of "+" symbols indicate the relative importance of different types of GCRs in each GCR assay observed in wild-type strains. B. Distribution of GCRs identified in the dGCR assay. Note that isolate SNBG1758 was a diploid containing both a *de novo* telomere chromosome 5 GCR and a *DSF1-HXT13* homology-mediated chromosome 5 GCR; this isolate was counted twice in the distribution. C. Distribution of GCRs identified in the sGCR assay.

Engineered junctions in *tor2-21* strains



	Junction:	3-A	4-A	6-A	6-B	6-C	14-A	14-B	15-A	11-A
dGCR	isolate 1758	46/29	108/51	73/75	101/56	81/123	73/73	37/135	35/4	95/46
	isolate 1762	51/21	100/40	92/69	127/45	98/136	69/72	30/167	53/23	104/39
	isolate 1763	55/39	106/34	96/84	157/51	123/156	38/134	50/69	60/33	100/41
	isolate 1764	12/6	36/12	32/19	49/13	36/53	23/20	25/59	21/13	32/12
	isolate 1765	86/33	168/51	182/97	238/75	171/207	127/92	48/220	38/47	146/45
	isolate 1767	110/36	226/69	191/117	279/83	189/222	146/128	53/232	59/45	175/50
sGCR	isolate 1769	137/33	566/150	263/157	368/97	260/290	202/135	86/263	69/56	230/50
	isolate 1777	12/7	4/0	20/15	43/16	24/35	19/15	15/41	4/0	18/10
	isolate 1780	1/0	1/0	5/0	6/0	2/0	4/0	4/0	-/-	6/0
	isolate 1782	43/12	73/25	69/32	83/25	60/72	55/36	19/77	34/16	76/19
	isolate 1783	28/37	33/18	54/70	72/43	53/75	34/30	22/105	38/26	56/28
	isolate 1784	20/8	43/8	48/30	62/23	39/47	35/31	15/65	29/11	57/17
	isolate 1785	13/6	16/8	29/17	37/8	28/33	25/10	17/35	12/4	26/11
	isolate 1786	14/3	10/16	23/11	31/10	32/28	21/12	12/29	3/0	28/7
	isolate 1787	5/7	22/11	13/7	19/3	10/19	11/7	7/7	6/2	9/4
	isolate 1788	71/19	2/0	182/97	238/75	171/207	127/92	48/220	38/47	118/70
	isolate 1791	8/0	-/-	28/11	35/16	17/31	31/31	18/37	6/6	19/12
	isolate 1792	10/8	-/-	34/12	43/12	34/41	13/9	18/33	6/4	25/9
	isolate 1794	9/0	-/-	19/10	38/8	25/25	19/11	4/0	13/4	21/9

Engineered ChrV junctions in *tor2-21* strains



	Junction:	5-A	5-B	5-C	5-D	5-E	5-DE	5-F	5-G	5-H
dGCR	isolate 1758	-/-	-/-	-/-	16/22	9/0	56/45	30/5	49/35	55/37
	isolate 1762	-/-	-/-	-/-	26/45	15/35	58/76	60/34	50/39	76/38
	isolate 1763	-/-	-/-	-/-	-/-	-/-	-/-	-/-	46/27	67/29
	isolate 1764	-/-	-/-	-/-	14/15	5/23	37/19	24/12	18/14	26/24
	isolate 1765	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	128/49
	isolate 1767	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	141/69
sGCR	isolate 1769	-/-	-/-	-/-	40/100	16/0	219/170	179/98	140/70	157/75
	isolate 1777	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	19/10
	isolate 1780	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	3/0
	isolate 1782	-/-	-/-	-/-	16/22	3/0	31/20	30/12	37/20	59/22
	isolate 1783	-/-	-/-	-/-	-/-	-/-	-/-	-/-	21/10	41/32
	isolate 1784	-/-	-/-	-/-	2/0	21/7	15/20	5/0	13/10	39/12
	isolate 1785	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	33/9
	isolate 1786	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	13/8
	isolate 1787	-/-	-/-	-/-	1/0	2/0	11/11	7/5	6/5	8/7
	isolate 1788	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	144/51
	isolate 1791	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	30/9
	isolate 1792	-/-	-/-	-/-	-/-	-/-	-/-	-/-	-/-	57/20
	isolate 1794	-/-	-/-	-/-	3/0	16/0	17/8	16/0	14/7	17/6

Figure 4.2 Identification of the starting chromosomal features by whole genome sequencing.

A. For each junction along chromosome 5 (junctions 5-A to 5-H), the evidence for each junction in the paired-end sequencing data is reported. The number preceding the slash is the number of junction-defining read pairs (those for which one read maps to one side of the junction and the other read maps to the other side of the junction). The number following the slash is the number of junction-sequencing reads (those that can be aligned to derive the sequence of the junction). "-/- " indicates a junction that could have been observed but was not observed, which is typically due to a GCR-related deletion. Note that some sequences are short enough that some read pairs span multiple junctions, e.g. junction 5-DE contains read pairs that span both junctions 5-D and 5-E. B. Junctions on chromosomes other than chromosome 5 depicted as in panel A.

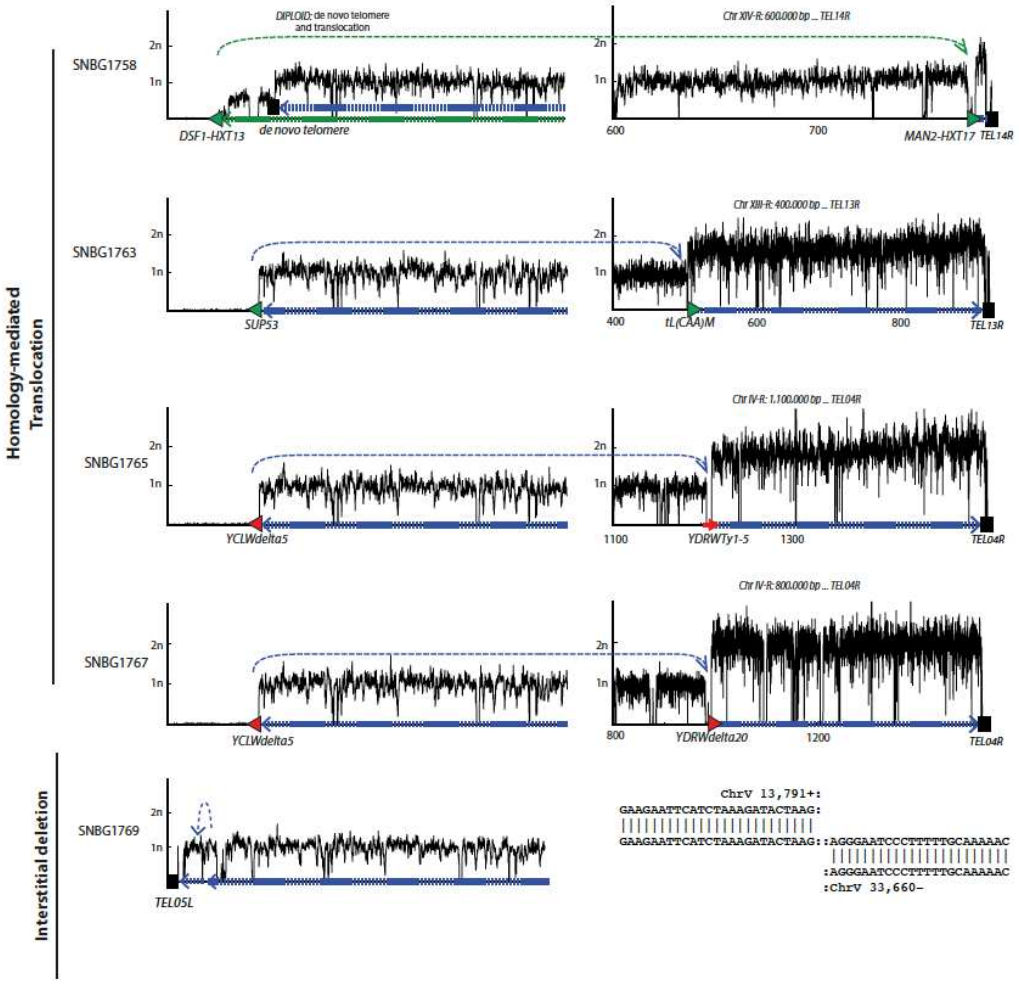


Figure 4.3 Analysis of GCRs selected in the sGCR assay in a *tor2-21* strain by whole genome sequencing

Copy number analysis of the sequenced parental strain and GCR-containing strains shows that GCRs are associated with deletion of the *CAN1/URA3*-containing terminal portion of chromosome 5 L (left) and either duplication of a terminal region of a target chromosome or the junction sequence associated with de novo telomere (right). The thick- hashed blue arrow indicates sequences within the GCR; the thin dashed blue arrow indicates connectivity between portions of the GCR that map to different regions of the reference chromosome. When more than one rearrangement is observed, thick-hashed green arrows and thin dashed green arrows are used for the second rearranged chromosome. Duplicated sequence involved in GCR-related HR events are shown as triangles; red triangles are Ty-related homologies and green triangles are other homologies.

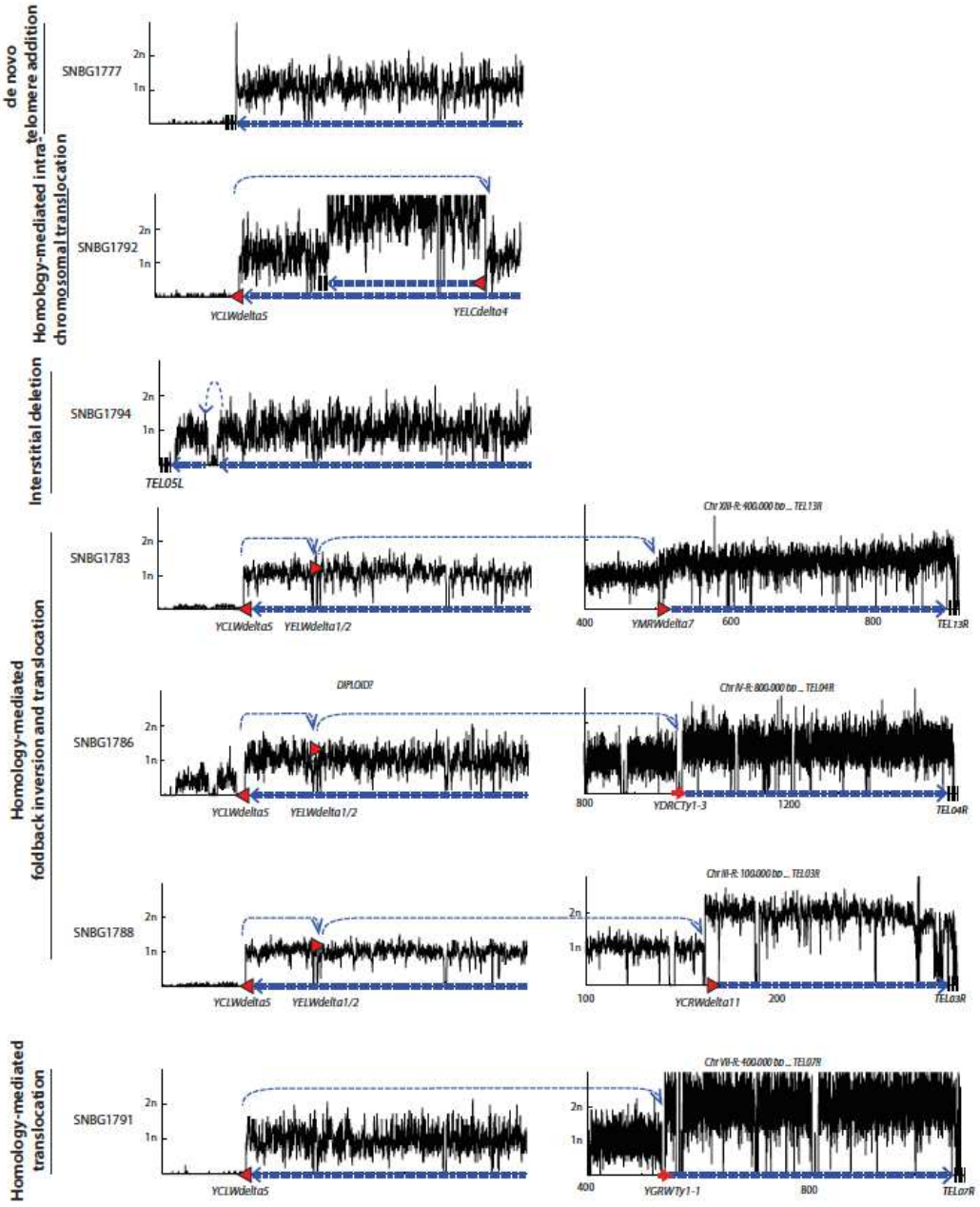


Figure 4.4 Analysis of GCRs selected in the sGCR assay in a *tor2-21* strain by whole genome sequencing

Copy number analysis and breakpoint junction sequences of the sequenced parental strain and GCR-containing strains displayed as for Figure 3.

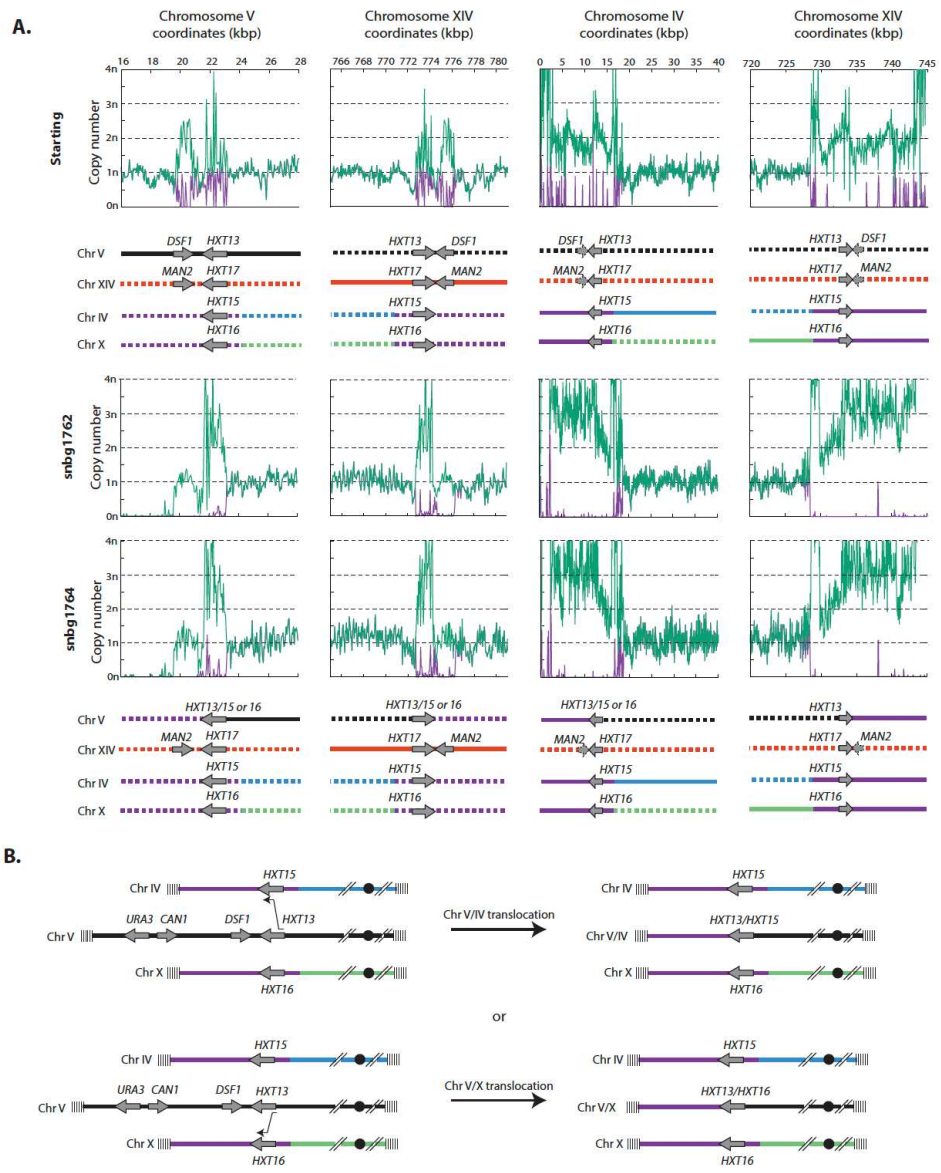


Figure 4.5 Evidence of copy number changes corresponding to GCR events in dGCR

A. Copy number plots for regions of chromosomes 5, 14, 4 and 10 involved in either t(5; 4 or 10) or t(5;14) non-reciprocal translocations. Plots for the starting strain (RDKY8492) and GCR isolates snbg1762 and snb1764 are included. B. Mechanisms through which t(5;4) or t(5;10) translocations arise through recombination between *HXT13* on chromosome 5 and either *HXT15* (chromosome 4) or *HXT16* (chromosome 10).

Table 4.1. Rate of accumulation of GCRs in *tor2-21* sGCR and dGCR strains

Temp + Genotype	GCR Rate	95% CI		Fold change vs wt leu2 (respective assay)
		Low	High	
RT dGCR leu2	1.89E-07	1.29E-07	3.01E-07	1.0
RT dGCR tor2-21	6.54E-07	3.87E-07	1.07E-06	3.47
30C dGCR leu2	8.14E-08	3.67E-08	1.79E-07	1.0
30C dGCR tor2-21	1.20E-06	7.89E-07	2.18E-06	14.75
RT sGCR leu2	4.79E-09	3.01E-09	8.73E-09	1.0
RT sGCR tor2-21	3.56E-08	6.90E-09	6.06E-08	7.43
30C sGCR leu2	6.17E-09	4.53E-09	9.51E-09	1.0 (sick)
30C sGCR tor2-21	<4.56E-10			0.07 (sick)

Table 4.2. Statistics for Next-Generation Sequencing results for dGCR isolates

Sample (relevant gentoype)	No. Read Pairs*	% Read 1 Mapped	% Read 2 Mapped	No. Uniquely Mapping Read Pairs	Median Intra-Read Pair Distance (bp)	Median Read Depth**
Snbg1758 (dgcr *tor2-21*)	52,058,216 (56,768,752)	26.08%	22.38%	7,381,316	123	70
Snbg1762 (dgcr *tor2-21*)	32,418,535 (32,873,464)	24.48%	19.70%	5,341,492	364	51
Snbg1763 (dgcr *tor2-21*)	70,710,236 (74,222,748)	21.75%	18.12%	9,225,469	149	88
Snbg1764 (dgcr *tor2-21*)	8,177,383 (8,396,622)	35.23%	28.81%	1,947,841	374	17
Snbg1765 (dgcr *tor2-21*)	21,987,267 (22,947,519)	60.06%	49.80%	9,526,449	344	87
Snbg1767 (dgcr *tor2-21*)	25,290,458 (26,340,687)	62.27%	48.97%	11,465,760	346	106
Snbg1769 (dgcr *tor2-21*)	36,795,463 (38,216,068)	58.12%	47.37%	15,572,872	340	135
Snbg1782 (dgcr *tor2-21*)	31,083,900 (34,202,618)	24.91%	21.27%	3,911,487	358	37
Snbg1784 (dgcr *tor2-21*)	15,777,744 (17,600,689)	30.70%	25.98%	2,503,208	377	23

Table 4.3. Statistics for Next-Generation Sequencing results for dGCR isolates

Sample (Relevant Genotype)	GCR junction description	GCR junction evidence
snbg1758 (dGCR *tor2-21*)	Modified chromosome 1: Homology-mediated translocation between chr5 L *DSF1-HXT13* and chr14 R *MAN2-HXT17*	Copy number (Figure 4)
	Modified chromosome 2: *De novo* telomere addition	Junction sequence
snbg1762 (dGCR *tor2-21*)	Homology-mediated translocation between chr5 L *HXT13* and chr4 L *HXT15* or chr10 R *HXT16*	Copy number
snbg1763 (dGCR *tor2-21*)	Homology-mediated translocation between chr5 L *SUP53* in *can1::P_{LEU2}-NAT* locus and chr13 R *tL(CAA)M*	Copy number (Figure 4)
snbg1764 (dGCR *tor2-21*)	Homology-mediated translocation between chr5 L *HXT13* and chr4 L *HXT15* or chr10 R *HXT16*	Copy number
snbg1765 (dGCR *tor2-21*)	Homology-mediated translocation between chr5 L *YCLWdelta5* in *can1::P_{LEU2}-NAT* and chr4 R *YDRWTy1-5*	Copy number (Figure 4)
snbg1767 (dGCR *tor2-21*)	Homology-mediated translocation between chr5 L *YCLWdelta5* in *can1::P_{LEU2}-NAT* and chr4 R *YDRWdelta20*	Copy number (Figure 4)
snbg1769 (dGCR *tor2-21*)	Interstitial deletion between chr5 L *CAN1* and chr5 L *YEL072W*	Copy number (Figure 4)

CHAPTER 5

Discussion

5.1 Conclusion and future directions

These studies demonstrate the complexity of the pathways that maintain genomic integrity in *S. cerevisiae* and the diversity and intricacy of the types of genome rearrangements that arise when they fail. Together, these findings will help us better understand the properties of divergent homologous sequences that lend themselves to genome rearrangements as well as the pathways that prevent these processes from occurring.

In Chapter 2 we describe the genetic requirements and GCR structures for two novel duplication-mediated GCR assays. While the Chr15-L assay behaves similarly to the dGCR assay, the Chr14-R assay exhibits surprising behavior, including a 26-fold increase in GCR rate in the absence of all canonical HR. This could be due to an underlying propensity to form a certain kind of GCR that is normally "masked" by repair involving the sister chromatid. Additionally, the interplay between DSB resection and homologous recombination may influence both the GCR rate and type of GCRs formed. The accessibility of DNA sequences for recombination may depend on how extensively a DSB is resected as well as the rate at which this occurs. Furthermore, it would be interesting to examine the exact chromosomal context of the assays and whether replication timing of a nearby ARS or other chromosomal features impact GCR formation in these strains.

In Chapter 3, we used a modified synthetic genetic array (SGA) approach to cross three GCR assays possessing different types of

homologous sequences in their breakpoint regions against a subset of the yeast deletion collection, with the goal of understanding how genetic interactions vary between the assays. We report on 182 genes that suppress genome instability in the various assays, including 64 genes that had not been previously identified. An additional 43 query mutations were also crossed against the deletion library and subsequent screening of double mutants for synergistic increases in genome instability identified 438 cooperatively interacting genes- that is, genes that suppress genome instability in the context of another gene deletion. Analysis of ovarian and colorectal cancers using TCGA data indicated that many tumors possess defects in one or more genes that are homologs of the GIS genes identified in *S. cerevisiae* in this screen. These findings have broad implications for cancers that exhibit characteristics of genome instability, and vast amounts of data merit future investigation including the 64 newly identified GIS genes from the screen.

The results in Chapter 4 identify a role for the TORC2 complex in the suppression of genome rearrangements involving shorter repetitive sequences. Incorporation of the temperature-sensitive allele *tor2-21* of the essential gene *TOR2* in *S. cerevisiae* resulted in greater suppression of GCR formation in the sGCR assay compared to the dGCR assay. The spectrum of GCR structures observed in the dGCR assay remained essentially unchanged in the *tor2-21* mutant and was dominated by homology-mediated rearrangements involving the *HXT13-DSF1* region. However, the spectrum of

structures observed in the sGCR *tor2-21* mutant shifted away from a spectrum consisting primarily of translocation events involving very small homology regions (i.e. delta sequences) and *de novo* telomere addition products to a more varied spectrum including many delta sequence-mediated inverted duplications. These results may imply that the mechanisms by which GCRs arise in the two different assays also differ; the TORC2 complex possesses numerous activities in the cell and future work may elucidate the role that these different functions play in GCR formation.

While we characterized many different aspects of homology-mediated GCRs through a variety of assays and genetic experiments, there is still much be discovered about how cells prevent rearrangements between different kinds of divergent homologous sequences. The precise roles of chromosomal context as well as chromosomal localization within the nucleus (particularly in mammalian systems) have not been thoroughly elucidated. The 64 new GIS genes obtained from the screen implicate a variety of new pathways in genome instability. The central role of the TORC2 complex has been expanded even further but its precise functions in genome maintenance are unclear. Future work into the details of how homologous recombination functions during DSB repair and to combat DNA damage sustained by the cell will require further studies.

References

1. Symington, L.S., R. Rothstein, and M. Lisby, *Mechanisms and regulation of mitotic recombination in Saccharomyces cerevisiae.* Genetics, 2014. **198**(3): p. 795-835.

2. Pennaneach, V. and R.D. Kolodner, *Recombination and the Tel1 and Mec1 checkpoints differentially effect genome rearrangements driven by telomere dysfunction in yeast.* Nat Genet, 2004. **36**(6): p. 612-7.

3. Chen, C. and R.D. Kolodner, *Gross chromosomal rearrangements in Saccharomyces cerevisiae replication and recombination defective mutants.* Nat Genet, 1999. **23**(1): p. 81-5.

4. Symington, L.S. and J. Gautier, *Double-strand break end resection and repair pathway choice.* Annu Rev Genet, 2011. **45**: p. 247-71.

5. Putnam, C.D., T.K. Hayes, and R.D. Kolodner, *Specific pathways prevent duplication-mediated genome rearrangements.* Nature, 2009. **460**(7258): p. 984-9.

6. Monnat, R.J., Jr., *Molecular analysis of spontaneous hypoxanthine phosphoribosyltransferase mutations in thioguanine-resistant HL-60 human leukemia cells.* Cancer Res, 1989. **49**(1): p. 81-7.

7. Loeb, L.A., *Mutator phenotype may be required for multistage carcinogenesis.* Cancer Res, 1991. **51**(12): p. 3075-9.

8. Oller, A.R., et al., *A statistical model to estimate variance in long term-low dose mutation assays: testing of the model in a human lymphoblastoid mutation assay.* Mutat Res, 1989. **216**(3): p. 149-61.

9. Deininger, P.L. and M.A. Batzer, *Alu repeats and human disease.* Mol Genet Metab, 1999. **67**(3): p. 183-93.

10. Gordenin, D.A. and M.A. Resnick, *Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability.* Mutat Res, 1998. **400**(1-2): p. 45-58.

11. Schmidt, K.H., et al., *Analysis of gross-chromosomal rearrangements in Saccharomyces cerevisiae.* Methods Enzymol, 2006. **409**: p. 462-76.

12. Chan, J.E. and R.D. Kolodner, *A genetic and structural study of genome rearrangements mediated by high copy repeat Ty1 elements.* PLoS Genet, 2011. **7**(5): p. e1002089.

13. Chan, J.E. and R.D. Kolodner, *Rapid analysis of Saccharomyces cerevisiae genome rearrangements by multiplex ligation-dependent probe amplification.* PLoS Genet, 2012. **8**(3): p. e1002539.

14. Pennaneach, V. and R.D. Kolodner, *Stabilization of dicentric translocations through secondary rearrangements mediated by multiple mechanisms in S. cerevisiae.* PLoS One, 2009. **4**(7): p. e6389.

15. Putnam, C.D., et al., *DNA repair pathway selection caused by defects in TEL1, SAE2, and de novo telomere addition generates specific chromosomal rearrangement signatures.* PLoS Genet, 2014. **10**(4): p. e1004277.

16. Anand, R.P., S.T. Lovett, and J.E. Haber, *Break-induced DNA replication.* Cold Spring Harb Perspect Biol, 2013. **5**(12): p. a010397.

17. Daley, J.M., et al., *Biochemical mechanism of DSB end resection and its regulation.* DNA Repair (Amst), 2015. **32**: p. 66-74.

18. Wyatt, H.D. and S.C. West, *Holliday junction resolvases.* Cold Spring Harb Perspect Biol, 2014. **6**(9): p. a023192.

19. Li, X., et al., *PCNA is required for initiation of recombination-associated DNA synthesis by DNA polymerase delta.* Mol Cell, 2009. **36**(4): p. 704-13.

20. White, J.H., K. Lusnak, and S. Fogel, *Mismatch-specific post-meiotic segregation frequency in yeast suggests a heteroduplex recombination intermediate.* Nature, 1985. **315**(6017): p. 350-2.

21. Alani, E., R.A. Reenan, and R.D. Kolodner, *Interaction between mismatch repair and genetic recombination in Saccharomyces cerevisiae.* Genetics, 1994. **137**(1): p. 19-39.

22. Tham, K.C., R. Kanaar, and J.H. Lebbink, *Mismatch repair and homeologous recombination.* DNA Repair (Amst), 2016. **38**: p. 75-83.

23. Schmidt, K.H., J. Wu, and R.D. Kolodner, *Control of translocations between highly diverged genes by Sgs1, the Saccharomyces cerevisiae homolog of the Bloom's syndrome protein.* Mol Cell Biol, 2006. **26**(14): p. 5406-20.

24. Spell, R.M. and S. Jinks-Robertson, *Examination of the roles of Sgs1 and Srs2 helicases in the enforcement of recombination fidelity in Saccharomyces cerevisiae.* Genetics, 2004. **168**(4): p. 1855-65.

25. Myung, K., A. Datta, and R.D. Kolodner, *Suppression of spontaneous chromosomal rearrangements by S phase checkpoint functions in Saccharomyces cerevisiae.* Cell, 2001. **104**(3): p. 397-408.

26. Lyndaker, A.M. and E. Alani, *A tale of tails: insights into the coordination of 3' end processing during homologous recombination.* Bioessays, 2009. **31**(3): p. 315-21.

27. Anand, R., et al., *Rad51-mediated double-strand break repair and mismatch correction of divergent substrates.* Nature, 2017. **544**(7650): p. 377-380.

28. Myung, K., C. Chen, and R.D. Kolodner, *Multiple pathways cooperate in the suppression of genome instability in Saccharomyces cerevisiae.* Nature, 2001. **411**(6841): p. 1073-6.

29. Putnam, C.D., V. Pennaneach, and R.D. Kolodner, *Chromosome healing through terminal deletions generated by de novo telomere*

additions in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A, 2004. **101**(36): p. 13262-7.

30. Chen, C., K. Umezu, and R.D. Kolodner, *Chromosomal rearrangements occur in S. cerevisiae rfa1 mutator mutants due to mutagenic lesions processed by double-strand-break repair.* Mol Cell, 1998. **2**(1): p. 9-22.

31. Putnam, C.D. and R.D. Kolodner, *Pathways and Mechanisms that Prevent Genome Instability in Saccharomyces cerevisiae.* Genetics, 2017. **206**(3): p. 1187-1225.

32. Putnam, C.D., E.J. Jaehnig, and R.D. Kolodner, *Perspectives on the DNA damage and replication checkpoint responses in Saccharomyces cerevisiae.* DNA Repair (Amst), 2009. **8**(9): p. 974-82.

33. Putnam, C.D., et al., *A genetic network that suppresses genome rearrangements in Saccharomyces cerevisiae and contains defects in cancers.* Nat Commun, 2016. **7**: p. 11256.

34. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.

35. Kuhn, R.M., et al., *The UCSC genome browser database: update 2007.* Nucleic Acids Res, 2007. **35**(Database issue): p. D668-73.

36. Zhang, L., et al., *Patterns of segmental duplication in the human genome.* Mol Biol Evol, 2005. **22**(1): p. 135-41.

37. Hastings, P.J., et al., *Mechanisms of change in gene copy number.* Nat Rev Genet, 2009. **10**(8): p. 551-64.

38. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome.* Science, 2004. **305**(5683): p. 525-8.

39. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome.* Nat Genet, 2004. **36**(9): p. 949-51.

40. Zhang, F., C.M. Carvalho, and J.R. Lupski, *Complex human chromosomal and genomic rearrangements.* Trends Genet, 2009. **25**(7): p. 298-307.

41. Weckselblatt, B. and M.K. Rudd, *Human Structural Variation: Mechanisms of Chromosome Rearrangements.* Trends Genet, 2015. **31**(10): p. 587-99.

42. Chen, J.M., et al., *Genomic rearrangements in inherited disease and cancer.* Semin Cancer Biol, 2010. **20**(4): p. 222-33.

43. Sung, P., *Catalysis of ATP-dependent homologous DNA pairing and strand exchange by yeast RAD51 protein.* Science, 1994. **265**(5176): p. 1241-3.

44. Signon, L., et al., *Genetic requirements for RAD51- and RAD54-independent break-induced replication repair of a chromosomal double-strand break.* Mol Cell Biol, 2001. **21**(6): p. 2048-56.

45. Davis, A.P. and L.S. Symington, *RAD51-dependent break-induced replication in yeast.* Mol Cell Biol, 2004. **24**(6): p. 2344-51.

46. Lydeard, J.R., et al., *Break-induced replication and telomerase-independent telomere maintenance require Pol32.* Nature, 2007. **448**(7155): p. 820-3.

47. Malkova, A., et al., *RAD51-dependent break-induced replication differs in kinetics and checkpoint responses from RAD51-mediated gene conversion.* Mol Cell Biol, 2005. **25**(3): p. 933-44.

48. Chakraborty, U., et al., *A Delicate Balance Between Repair and Replication Factors Regulates Recombination Between Divergent DNA Sequences in Saccharomyces cerevisiae.* Genetics, 2016. **202**(2): p. 525-40.

49. Goldfarb, T. and E. Alani, *Distinct roles for the Saccharomyces cerevisiae mismatch repair proteins in heteroduplex rejection, mismatch*

repair and nonhomologous tail removal. Genetics, 2005. **169**(2): p. 563-74.

50. Sugawara, N., et al., *Heteroduplex rejection during single-strand annealing requires Sgs1 helicase and mismatch repair proteins Msh2 and Msh6 but not Pms1.* Proc Natl Acad Sci U S A, 2004. **101**(25): p. 9315-20.

51. Raghuraman, M.K. and B.J. Brewer, *Molecular analysis of the replication program in unicellular model organisms.* Chromosome Res, 2010. **18**(1): p. 19-34.

52. Lazzaro, F., et al., *RNase H and postreplication repair protect cells from ribonucleotides incorporated in DNA.* Mol Cell, 2012. **45**(1): p. 99-110.

53. O'Connell, K., S. Jinks-Robertson, and T.D. Petes, *Elevated Genome-Wide Instability in Yeast Mutants Lacking RNase H Activity.* Genetics, 2015. **201**(3): p. 963-75.

54. Wahba, L., et al., *RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability.* Mol Cell, 2011. **44**(6): p. 978-88.

55. Zimmer, A.D. and D. Koshland, *Differential roles of the RNases H in preventing chromosome instability.* Proc Natl Acad Sci U S A, 2016. **113**(43): p. 12220-12225.

56. Allen-Soltero, S., et al., *A saccharomyces cerevisiae RNase H2 interaction network functions to suppress genome instability.* Mol Cell Biol, 2014. **34**(8): p. 1521-34.

57. Aguilera, A. and T. Garcia-Muse, *R loops: from transcription byproducts to threats to genome stability.* Mol Cell, 2012. **46**(2): p. 115-24.

58. Aguilera, A. and T. Garcia-Muse, *Causes of genome instability.* Annu Rev Genet, 2013. **47**: p. 1-32.

59. Chavez, S. and A. Aguilera, *The yeast HPR1 gene has a functional role in transcriptional elongation that uncovers a novel source of genome instability.* Genes Dev, 1997. **11**(24): p. 3459-70.

60. Santos-Pereira, J.M., et al., *The Npl3 hnRNP prevents R-loop-mediated transcription-replication conflicts and genome instability.* Genes Dev, 2013. **27**(22): p. 2445-58.

61. Bermejo, R., M.S. Lai, and M. Foiani, *Preventing replication stress to maintain genome stability: resolving conflicts between replication and transcription.* Mol Cell, 2012. **45**(6): p. 710-8.

62. Chan, Y.A., P. Hieter, and P.C. Stirling, *Mechanisms of genome instability induced by RNA-processing defects.* Trends Genet, 2014. **30**(6): p. 245-53.

63. Gomez-Gonzalez, B., et al., *Genome-wide function of THO/TREX in active genes prevents R-loop-dependent replication obstacles.* Embo j, 2011. **30**(15): p. 3106-19.

64. Putnam, C.D. and R.D. Kolodner, *Determination of gross chromosomal rearrangement rates.* Cold Spring Harb Protoc, 2010. **2010**(9): p. pdb prot5492.

65. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

66. Engel, S.R., et al., *The reference genome sequence of Saccharomyces cerevisiae: then and now.* G3 (Bethesda), 2014. **4**(3): p. 389-98.

67. Loeb, L.A., *A mutator phenotype in cancer.* Cancer Res, 2001. **61**(8): p. 3230-9.

68. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.

69. Inaki, K. and E.T. Liu, *Structural mutations in cancer: mechanistic and functional insights.* Trends Genet, 2012. **28**(11): p. 550-9.

70. D'Andrea, A.D., *Susceptibility pathways in Fanconi's anemia and breast cancer.* N Engl J Med, 2010. **362**(20): p. 1909-19.

71. Kobayashi, H., et al., *Hereditary breast and ovarian cancer susceptibility genes (review).* Oncol Rep, 2013. **30**(3): p. 1019-29.

72. Kanellis, P., et al., *A screen for suppressors of gross chromosomal rearrangements identifies a conserved role for PLP in preventing DNA lesions.* PLoS Genet, 2007. **3**(8): p. e134.

73. Hackett, J.A., D.M. Feldser, and C.W. Greider, *Telomere dysfunction increases mutation rate and genomic instability.* Cell, 2001. **106**(3): p. 275-86.

74. Putnam, C.D., V. Pennaneach, and R.D. Kolodner, *Saccharomyces cerevisiae as a model system to define the chromosomal instability phenotype.* Mol Cell Biol, 2005. **25**(16): p. 7226-38.

75. Putnam, C.D., T.K. Hayes, and R.D. Kolodner, *Post-replication repair suppresses duplication-mediated genome instability.* PLoS Genet, 2010. **6**(5): p. e1000933.

76. Smith, S., et al., *Mutator genes for suppression of gross chromosomal rearrangements identified by a genome-wide screening in Saccharomyces cerevisiae.* Proc Natl Acad Sci U S A, 2004. **101**(24): p. 9039-44.

77. Stirling, P.C., et al., *The complete spectrum of yeast chromosome instability genes identifies candidate CIN cancer genes and functional roles for ASTRA complex components.* PLoS Genet, 2011. **7**(4): p. e1002057.

78. De Piccoli, G., et al., *Smc5-Smc6 mediate DNA double-strand-break repair by promoting sister-chromatid recombination.* Nat Cell Biol, 2006. **8**(9): p. 1032-4.

79. Banerjee, S., et al., *Mph1p promotes gross chromosomal rearrangement through partial inhibition of homologous recombination.* J Cell Biol, 2008. **181**(7): p. 1083-93.

80. Motegi, A., et al., *Regulation of gross chromosomal rearrangements by ubiquitin and SUMO ligases in Saccharomyces cerevisiae.* Mol Cell Biol, 2006. **26**(4): p. 1424-33.

81. Schmidt, K.H. and R.D. Kolodner, *Suppression of spontaneous genome rearrangements in yeast DNA helicase mutants.* Proc Natl Acad Sci U S A, 2006. **103**(48): p. 18196-201.

82. Huang, M.E., et al., *A genomewide screen in Saccharomyces cerevisiae for genes that suppress the accumulation of mutations.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11529-34.

83. Huang, M.E. and R.D. Kolodner, *A biological network in Saccharomyces cerevisiae prevents the deleterious effects of endogenous oxidative DNA damage.* Mol Cell, 2005. **17**(5): p. 709-20.

84. Albuquerque, C.P., et al., *Distinct SUMO ligases cooperate with Esc2 and Slx5 to suppress duplication-mediated genome rearrangements.* PLoS Genet, 2013. **9**(8): p. e1003670.

85. Putnam, C.D., et al., *Bioinformatic identification of genes suppressing genome instability.* Proc Natl Acad Sci U S A, 2012. **109**(47): p. E3251-9.

86. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers.* Nat Genet, 2013. **45**(10): p. 1127-1133.

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

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

89. Cancer Genome Atlas Research, N., *Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia.* N Engl J Med, 2013. **368**(22): p. 2059-74.

90. Tong, A.H. and C. Boone, *Synthetic genetic array analysis in Saccharomyces cerevisiae.* Methods Mol Biol, 2006. **313**: p. 171-92.

91. Badin-Larcon, A.C., et al., *Suppression of nuclear oscillations in Saccharomyces cerevisiae expressing Glu tubulin.* Proc Natl Acad Sci U S A, 2004. **101**(15): p. 5577-82.

92. Mullen, J.R., C.F. Chen, and S.J. Brill, *Wss1 is a SUMO-dependent isopeptidase that interacts genetically with the Slx5-Slx8 SUMO-targeted ubiquitin ligase.* Mol Cell Biol, 2010. **30**(15): p. 3737-48.

93. Fourel, G., et al., *Cohabitation of insulators and silencing elements in yeast subtelomeric regions.* EMBO J, 1999. **18**(9): p. 2522-37.

94. Preti, M., et al., *The telomere-binding protein Tbf1 demarcates snoRNA gene promoters in Saccharomyces cerevisiae.* Mol Cell, 2010. **38**(4): p. 614-20.

95. Bonetti, D., et al., *Tbf1 and Vid22 promote resection and non-homologous end joining of DNA double-strand break ends.* EMBO J, 2013. **32**(2): p. 275-89.

96. Sahi, C. and E.A. Craig, *Network of general and specialty J protein chaperones of the yeast cytosol.* Proc Natl Acad Sci U S A, 2007. **104**(17): p. 7163-8.

97. de Souza, J.E., et al., *S-score: a scoring system for the identification and prioritization of predicted cancer genes.* PLoS One, 2014. **9**(4): p. e94147.

98. Debrauwere, H., et al., *Links between replication and recombination in Saccharomyces cerevisiae: a hypersensitive requirement for homologous recombination in the absence of Rad27 activity.* Proc Natl Acad Sci U S A, 2001. **98**(15): p. 8263-9.

99. Mimitou, E.P. and L.S. Symington, *DNA end resection: many nucleases make light work.* DNA Repair (Amst), 2009. **8**(9): p. 983-95.

100. Schulz, V.P. and V.A. Zakian, *The saccharomyces PIF1 DNA helicase inhibits telomere elongation and de novo telomere formation.* Cell, 1994. **76**(1): p. 145-55.

101. Zinovyev, A., et al., *Synthetic lethality between gene defects affecting a single non-essential molecular pathway with reversible steps.* PLoS Comput Biol, 2013. **9**(4): p. e1003016.

102. Ciccia, A. and S.J. Elledge, *The DNA damage response: making it safe to play with knives.* Mol Cell, 2010. **40**(2): p. 179-204.

103. Friedberg, E.C., et al., *DNA Repair and Mutagenesis*. 2006, Washington, D.C.: ASM Press.

104. Sikorski, R.S. and P. Hieter, *A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae.* Genetics, 1989. **122**(1): p. 19-27.

105. Kaufer, N.F., et al., *Cycloheximide resistance in yeast: the gene and its protein.* Nucleic Acids Res, 1983. **11**(10): p. 3123-35.

106. Loeillet, S., et al., *Genetic network interactions among replication, repair and nuclear pore deficiencies in yeast.* DNA Repair (Amst), 2005. **4**(4): p. 459-68.

107. De Smet, F., et al., *Balancing false positives and false negatives for the detection of differential expression in malignancies.* Br J Cancer, 2004. **91**(6): p. 1160-5.

108. Fawcett, T., *An introduction to ROC analysis.* Pattern Recognition Letters, 2006. **27**: p. 861-874.

109. Arnold, T.B. and J.W. Emerson, *Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions.* The R Journal, 2011. **3**(2): p. 34-39.

110. Babu, M., et al., *Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae.* Nature, 2012. **489**(7417): p. 585-9.

111. Bandyopadhyay, S., et al., *Functional maps of protein complexes from quantitative genetic interaction data.* PLoS Comput Biol, 2008. **4**(4): p. e1000065.

112. Benschop, J.J., et al., *A consensus of core protein complex compositions for Saccharomyces cerevisiae.* Mol Cell, 2010. **38**(6): p. 916-28.

113. Hart, G.T., I. Lee, and E.R. Marcotte, *A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality.* BMC Bioinformatics, 2007. **8**: p. 236.

114. Pu, S., et al., *Identifying functional modules in the physical interactome of Saccharomyces cerevisiae.* Proteomics, 2007. **7**(6): p. 944-60.

115. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.

116. Ulitsky, I., et al., *From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions.* Mol Syst Biol, 2008. **4**: p. 209.

117. Hong, E.L., et al., *Gene Ontology annotations at SGD: new data sources and annotation methods.* Nucleic Acids Res, 2008. **36**(Database issue): p. D577-81.

118. Pu, S., et al., *Up-to-date catalogues of yeast protein complexes.* Nucleic Acids Res, 2009. **37**(3): p. 825-31.

119. Cherry, J.M., et al., *Saccharomyces Genome Database: the genomics resource of budding yeast.* Nucleic Acids Res, 2012. **40**(Database issue): p. D700-5.

120. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.

121. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res, 2014. **42**(Database issue): p. D459-71.

122. Beroukhim, R., et al., *Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.* Proc Natl Acad Sci U S A, 2007. **104**(50): p. 20007-12.

123. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat Protoc, 2009. **4**(7): p. 1073-81.

124. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.

125. Schwarz, J.M., et al., *MutationTaster evaluates disease-causing potential of sequence alterations.* Nat Methods, 2010. **7**(8): p. 575-6.

126. Shihab, H.A., et al., *Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models.* Hum Mutat, 2013. **34**(1): p. 57-65.

127. Chun, S. and J.C. Fay, *Identification of deleterious mutations within three human genomes.* Genome Res, 2009. **19**(9): p. 1553-61.

128. Pracheil, T., J. Thornton, and Z. Liu, *TORC2 signaling is antagonized by protein phosphatase 2A and the Far complex in Saccharomyces cerevisiae.* Genetics, 2012. **190**(4): p. 1325-39.

129. Loewith, R. and M.N. Hall, *Target of rapamycin (TOR) in nutrient signaling and growth control.* Genetics, 2011. **189**(4): p. 1177-201.

130. Helliwell, S.B., et al., *TOR2 is part of two related signaling pathways coordinating cell growth in Saccharomyces cerevisiae.* Genetics, 1998. **148**(1): p. 99-112.

131. Barbet, N.C., et al., *TOR controls translation initiation and early G1 progression in yeast.* Mol Biol Cell, 1996. **7**(1): p. 25-42.

132. Yerlikaya, S., et al., *TORC1 and TORC2 work together to regulate ribosomal protein S6 phosphorylation in Saccharomyces cerevisiae.* Mol Biol Cell, 2016. **27**(2): p. 397-409.

133. Schmidt, A., J. Kunz, and M.N. Hall, *TOR2 is required for organization of the actin cytoskeleton in yeast.* Proc Natl Acad Sci U S A, 1996. **93**(24): p. 13780-5.

134. Schmidt, A., et al., *The yeast phosphatidylinositol kinase homolog TOR2 activates RHO1 and RHO2 via the exchange factor ROM2.* Cell, 1997. **88**(4): p. 531-42.

135. Helliwell, S.B., et al., *The Rho1 effector Pkc1, but not Bni1, mediates signalling from Tor2 to the actin cytoskeleton.* Curr Biol, 1998. **8**(22): p. 1211-4.

136. Shimada, K., et al., *TORC2 signaling pathway guarantees genome stability in the face of DNA strand breaks.* Mol Cell, 2013. **51**(6): p. 829-39.

137. Ungar, L., et al., *Tor complex 1 controls telomere length by affecting the level of Ku.* Curr Biol, 2011. **21**(24): p. 2115-20.

138. Weisman, R., et al., *Opposite effects of tor1 and tor2 on nitrogen starvation responses in fission yeast.* Genetics, 2007. **175**(3): p. 1153-62.

139. Liao, H.C. and M.Y. Chen, *Target of rapamycin complex 2 signals to downstream effector yeast protein kinase 2 (Ypk2) through adheres-voraciously-to-target-of-rapamycin-2 protein 1 (Avo1) in Saccharomyces cerevisiae.* J Biol Chem, 2012. **287**(9): p. 6089-99.

140. Luria, S.E. and M. Delbruck, *Mutations of Bacteria from Virus Sensitivity to Virus Resistance.* Genetics, 1943. **28**(6): p. 491-511.

141. Lea, D.E. and C.A. Coulson, *The distribution of the numbers of mutants in bacterial populations.* J Genet, 1949. **49**(3): p. 264-85.