Chapter 2

Analysis of cancer-associated polymerase mutations

Overarching hypothesis

DNA polymerase mutations identified in cancer samples can be constructed in the model organism *S. cerevisiae* to examine their relevance to tumuor progression and whole-genome sequencing of budding yeast samples can yield relevant biological insights.

Aims:

- To compile a list of relevant mutations in DNA polymerases identified in cancer samples
- To prioritise mutations and determine their *S. cerevisiae* equivalents
- To conduct mutation accumulation experiments to identify the consequences of mutations in DNA polymerase on a genome-wide scale
- To establish sequence analysis protocols for budding yeast whole-genome sequencing data
- To show that these sequence analysis protocols are functional and can be applied beyond this project

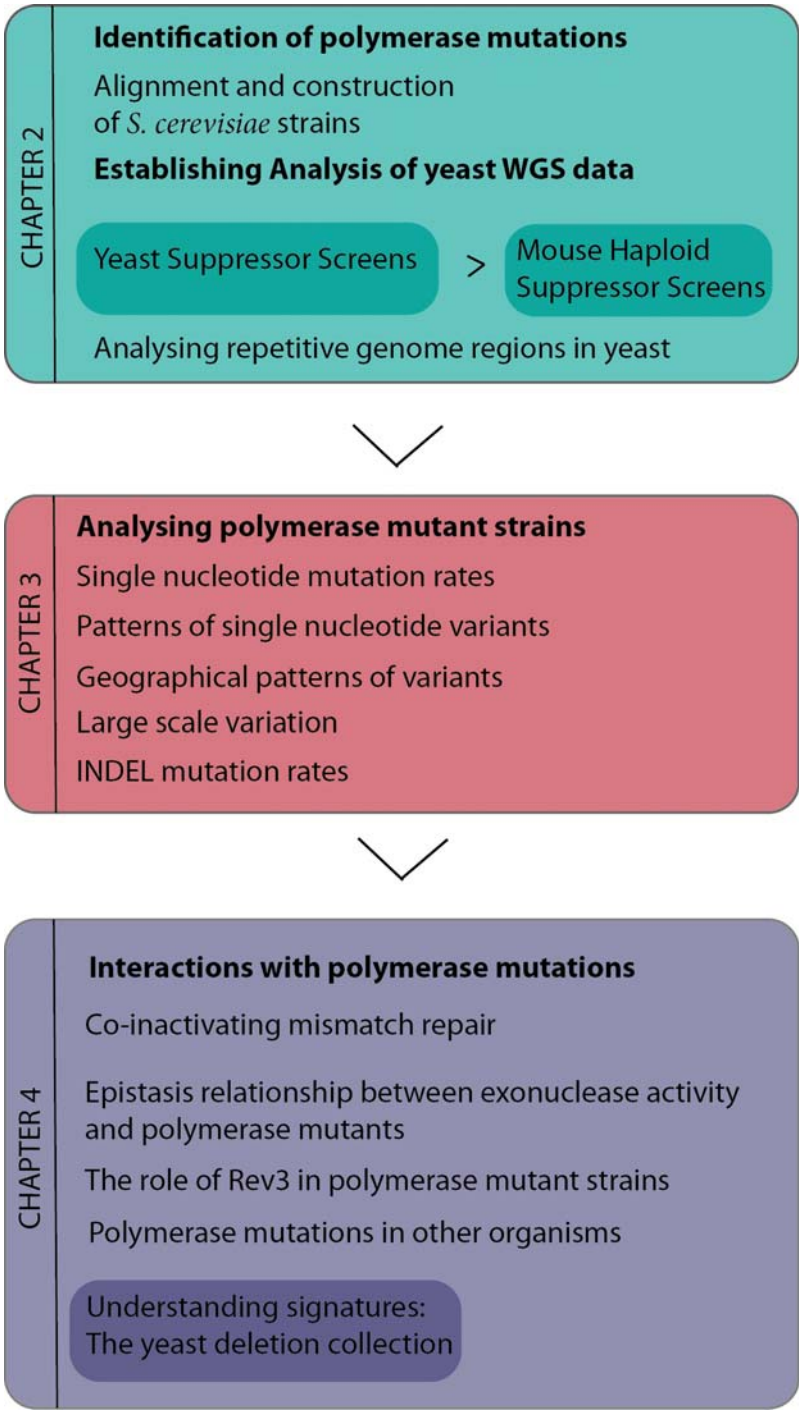


Figure 2.1: Methodology of the work carried out during my PhD
The chapter in which each step is covered is indicated at the left. Projects loosely associated with the my principal DNA polymerase mutation project are highlighted in darker boxes.

2.1 Introduction

Methods and results detailed in sub-sections 2.4.3 and 2.4.4 have been published or are accepted for publication (see [801] and [802]). Figures and Figure legends have partially been reproduced from this work in accordance with the copyright provisions of the publisher.

Cancer is a disease of mutations and defects in the mechanisms that maintain replication fidelity are likely underlying mutations in genes involved in tumourigenesis. It has been described that germline mutations in the DNA mismatch repair (MMR) machinery predispose to hereditary colorectal cancer[803, 804], but the case has been less clear-cut for polymerases. Due to the relatively recent advent of tumour genome sequencing, we now have the tools to actually get information on which polymerase genes are commonly mutated, the frequency of such mutations, which tumour types are affected and the characteristics of such tumours. So far, sequencing of a number of cancers has revealed somatic mutations in *POLE* coding for the catalytic subunit of Polε[259]. At the same time, pedigree-sequencing of families with a history of colorectal cancer identified two predisposing germline variants *POLE* L424V and *POLD1* S478N[768]. The condition was termed polymerase proofreading-associated polyposis (PPAP)[768, 805] though there is currently no genome-wide association studies (GWAS) evidence for associated risk between polymerase SNPs and colorectal cancer[806].

It is known that mutations in the exonuclease domain (EDM) of Polε and Polδ in yeast cause a base substitution phenotype of varying severity. Mutations affecting the catalytic residues of the proofreading domain of *POL3* (*pol3-01*) cause a mutator phenotype with increased base substitution and frameshift mutations[338]. Similar mutations in Polε (*pol2-4*) reduce proofreading activity about 100-fold *in vitro*, while leaving polymerase activity at wild-type levels[312]. *In vivo*, these mutations cause a mutator phenotype and using different reporter assays the increase in mutation rate was found to be between 5- and 43-fold(Table 2.1), highlighting the significance of proofreading for genome maintenance as well as the limitations of classical reporter assays to accurately describe mutator phenotypes.

Mice carrying mutations in the proofreading domain of polymerase ε (Pol$\varepsilon^{\text{exo-/exo-}}$; the mouse equivalent of the yeast *pol2-4* mutation) showed a predisposition to cancer, while Pol$\varepsilon^{\text{exo-/+}}$were virtually indistinguishable from wild-type in this respect[769]. Spontaneous mutations were more frequent in Pol$\varepsilon^{\text{exo-}}$ mice than in Pol$\delta^{\text{exo-}}$ mice, in contrast to the budding yeast, where the *pol3-01* mutation causes a higher mutational frequency than *pol2-4*. This either reflects a true discrepancy between yeast and mice or results from the fact that mutation frequency is estimated usually at single genetic loci (e.g. *Atp1a1* and *Hprt* in mice versus *URA3*, *CAN1* and *SUP4-o* in yeast), further confirming the need for improved methods to assess mutation rate increases. Mice deficient for both Polε and Polδ proofreading activity

Mutation	Assay gene	Fold change to wt	Publication
pol3-01	*his7-2*	240	[338]
pol3-01	*URA3* [a]	130	[338]
pol3-01	*URA3* [a]	52	[282]
pol3-01	*lys2::InsLD*	0.6	[308]
pol3-01	*his7-2*	74	[308]
pol3-01	*his7-2*	630	[282]
pol3-01	*CAN1*	110	[308]
pol3-01	*SUP4-o* [b]	32-106	[303]
pol3-01	*trp1–289*	100	[282]
pol3-01	*lys2::InsE* [c]	26 - 188	[807]
pol2-4	*CAN1*	5	[312]
pol2-4	*ade5-1*	43	[312]
pol2-4	*URA3* [a]	15	[282]
pol2-4	*his7-2*	24	[312]
pol2-4	*his7-2*	63	[282]
pol2-4	*leu2-1*	18	[312]
pol2-4	*hom3-10*	9	[312]
pol2-4	*his1-7*	31	[312]
pol2-4	*SUP4-o* [b]	2.9	[303]
pol2-4	*trp1–289*	3.9	[282]
pol2-4	*lys2::InsE* [c]	1.2 - 6	[807]
pol2-16	*URA3* [a]	1.6	[282]
pol2-16	*his7–2*	1.4	[282]
pol2-16	*trp1–289*	1.9	[282]

Table 2.1: Polymerase exonuclease domain mutations in *S. cerevisiae*
Figures were taken from publications as indicated. Fold change shows the ratio between mutant value and wild-type. All strain mutations are haploid unless otherwise indicated. As a comparison mutation rates for the strain pol2-16 are shown, in which all of *POL2* except the non-catalytic C-terminus is deleted[285]. [a]Forward mutation of *URA3*. [b]SUP4-o orientation was altered to be both on leading and lagging strand, which gave vastly different mutation rates in the case of *pol3-01*. [c]lys2::InsE alleles contain various sizes of dA homonucleotide runs. For similar experiments, see [338, 339, 808].

were viable, but died earlier of thymic lymphoma.

Not much is known about whether these mutations are passenger mutations or promote tumour progression. Additionally, it is unclear whether these mutations affect polymerase fidelity and to what degree. In my thesis, I will explore these questions, first, by assembling a list of mutations in DNA polymerases, then, using the budding yeast *S. cerevisiae* to test the effects of altered DNA polymerases on genomes, I will identify the most striking candidates to explore further in yeast, mouse and human (Fig. 2.1).

2.2 Identification of polymerase mutations

2.2.1 Literature search for DNA polymerase mutations in cancer

Whole-exome and whole-genome sequencing of cancer samples has identified mutations in DNA polymerases and the list is growing with little follow-up work on the nature of these variants. The Cancer Genome Altlas (TCGA), a project to catalogue genetic mutations responsible for cancer, has identified DNA polymerase mutations in 3% of colorectal cancers (CRC)[718] and 7% of endometrial cancers they sequenced[809]. While recurrent mutations in *POLE* could be identified, none were found for *POLD1*. A different CRC project identified another recurrent change p.Pro286Arg[751]. Only a minority of tumours show LOH or inactivating mutations for *POLE* or *POLD1*[806].

For this project, the mutations described in the work from Palles and co-workers[768], Church and co-workers[810] and the TGCA endometrial sequencing project[809] were assembled into a list of mutations and, in order to properly locate these mutations in whole-genome datasets, amino acid changes were converted to their genomic coordinates(Table 2.2). The mutations are all found within the N-terminal exonuclease domains of the polymerases (Fig. 2.2), which may reflect a real increased prevalence of mutations in this part of the protein, but is more likely due to the identification of several mutations by specifically sequencing the exonuclease domain of Polε and Polδ[810].

2.2.2 Query of COSMIC database, discarding single nucleotide polymorphisms and unconserved residues

The availability of vast amounts of cancer sequencing data allows the assessment of the recurrence of individual mutations as a base for further prioritisation as well as their distribution among different types of cancer.The Catalogue of Somatic Mutations in Cancer (COSMIC)

Gene	AA change	Chr	Pos(37)	Pos(38)	REF	ALT	
POLD1	p.Arg311Cys	19	50905959	50402702	C	T	[810]
POLD1	p.Gly426Ser	19	50909472	50406215	G	A	[768]
POLD1	p.Pro327Leu	19	50906319	50403062	C	T	[768]
POLD1	p.Ser370Arg	19	50906449	50403192	C	A	[768]
POLD1	p.Ser478Asn	19	50909713	50406456	G	A	[768]
POLD1	p.Val392Met	19	50906786	50403529	G	A	[810]
POLE	p.Met444Lys	12	133250189	132673603	A	T	[809]
POLE	p.Ala456Pro	12	133249857	132673271	C	G	[810]
POLE	p.Ala465Val	12	133249829	132673243	G	A	[809]
POLE	p.Arg446Gln	12	133250183	132673597	C	T	[810]
POLE	p.Asp275Val	12	133253217	132676631	T	A	[810]
POLE	p.Gln453Arg	12	133250162	132673576	T	C	[809]
POLE	p.Leu424Val	12	133250250	132673664	G	C	[768]
POLE	p.Pro286Arg	12	133253184	132676598	G	C	[810]
POLE	p.Pro436Arg	12	133250213	132673627	G	C	[809]
POLE	p.Ser297Phe	12	133253151	132676565	G	A	[810]
POLE	p.Val411Leu	12	133250289	132673703	C	A	[810]

Table 2.2: Genomic locations of mutations in DNA polymerases in different human genome assemblies

Genomic locations and nucleotide changes for the DNA polymerase mutations were identified using the human reference genome assemblies GRCh37 and CRCh38. Re-mapping between assemblies was done using the NCBI Genome Remapping Service[811]. Locations and nucleotide changes were computed using the reference genomes, their annotations and the codon table (see Fig. 1.27). **AA Change** stands for amino acid change, **Chr** for chromosome, **Pos(37)** for the position along the chromosome in genome assembly GRCh37, **Pos(38)** reflects the position in assembly GRCh38, **REF** is the base found in the reference genome and **ALT** is the base identified in the cancer samples. The source for the mutation can be found in the last column.

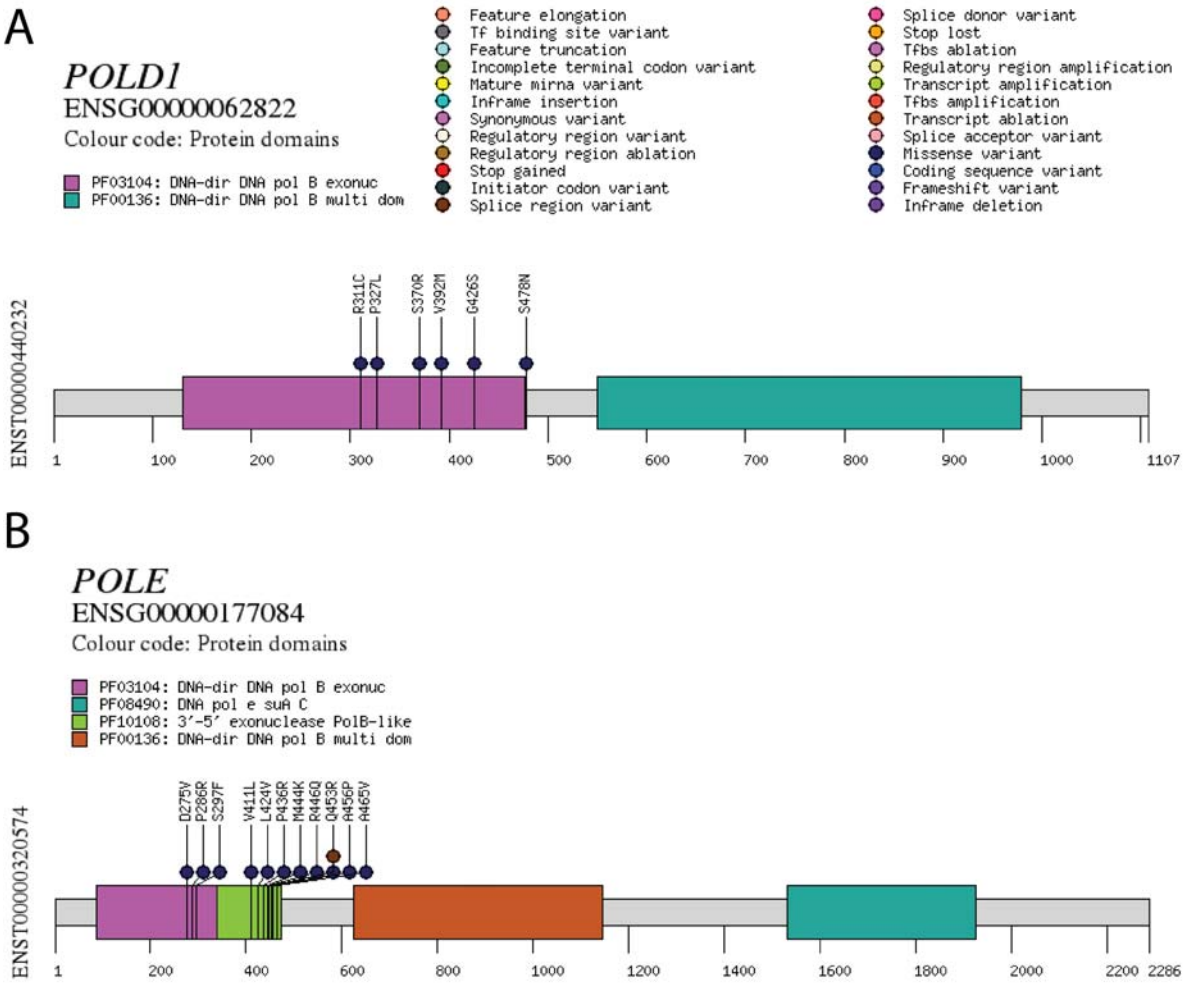


Figure 2.2: Locations of DNA polymerase mutations within the proteins
The locations of the mutations within the protein with reference to the domain structure is given. Plot was generated by Dr. Carla Daniela Robles Espinoza using a custom written script.

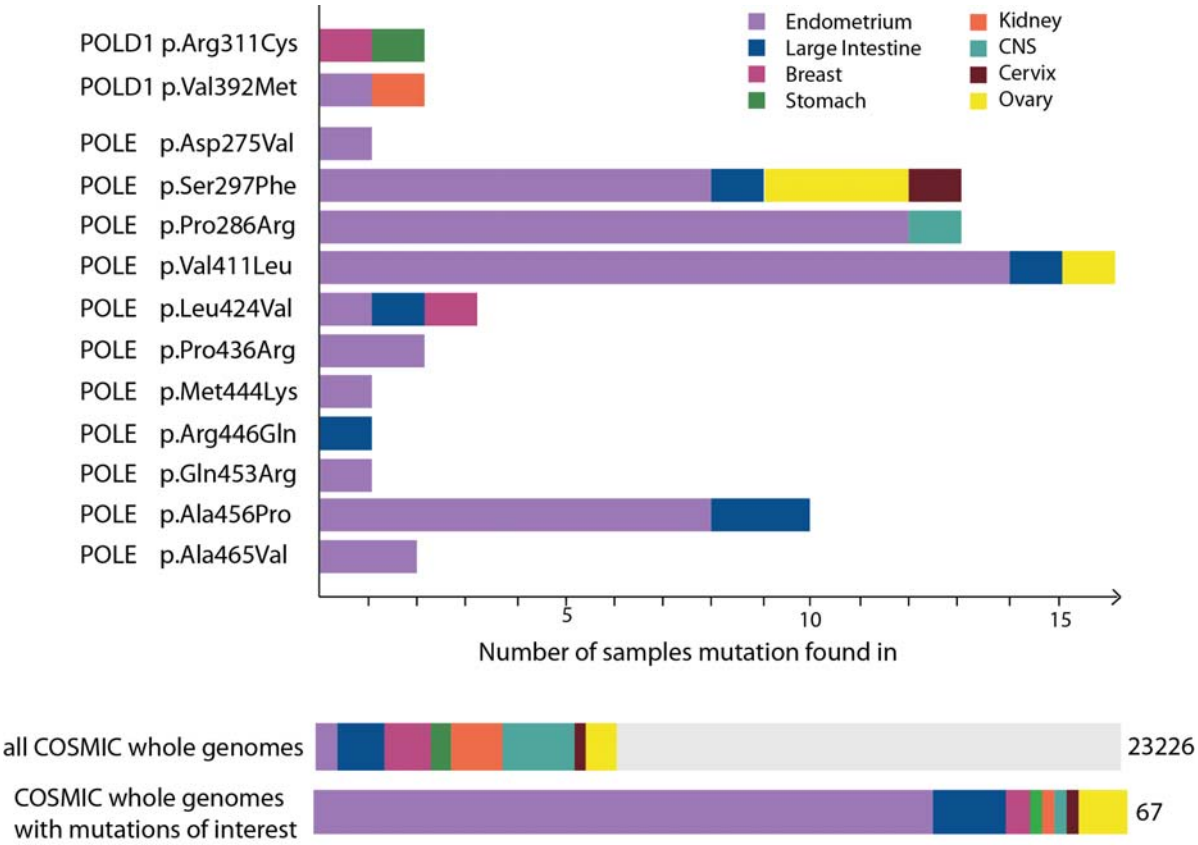


Figure 2.3: Prevalence of polymerase mutations of interest in COSMIC
The list of DNA polymerase mutations were cross-referenced with the COSMIC whole-genome data (v74)[812]. Recurrence of mutations in the whole dataset is displayed with information about the tissue of origin. For comparison, the composition of tumour origins across the whole database for the relevant tissue types is featured.

is a vast database of somatic changes observed in human cancer samples[812]. To assess the prevalence of these mutations, I accessed their curation of 22,690 whole cancer genomes and analysed mutation recurrence and tumour origin(Fig. 2.3). Recurrence indicates that DNA polymerase mutations to prioritise for testing include *POLE* S297F, *POLE* P286R, *POLE* V411L and *POLE* A456P. Indeed, DNA polymerase mutations are enriched in endometrial cancers and to a lesser extent colorectal cancers, which is not due to an overrepresentation of those cancer types in the dataset as a whole (endometrial cancers are 2.7% of all samples, colorectal cancers are 5.8%).

None of these variants were excluded from the list of candidates on the basis of occurrence in sequencing projects aiming to capture common variation in the human population (Table 2.3) considering the most common variant was found in 0.03% of the population. To get preliminary information on the severity of these mutations I ran bioinformatic predictions soft-

Gene	AA change	dbSNP	1000Genomes	500Exomes & CGP
POLD1	p.Arg311Cys	rs201010746 T=0.00001 (ExAC)	T=0.0002/1	rs201010746
POLD1	p.Gly426Ser	-	-	lowQual
POLD1	p.Pro327Leu	rs397514633 (OMIM)	-	-
POLD1	p.Ser370Arg	-	-	-
POLD1	p.Ser478Asn	rs397514632 (OMIM)	-	-
POLD1	p.VAL392Met	rs778843530 A=0.000008 (ExAC)	-	-
POLE	p.Met444Lys	-	-	-
POLE	p.Ala456Pro	-	-	-
POLE	p.Ala465Val	-	-	-
POLE	p.Arg446Gln	rs151273553 T=0.0003 (ExAC)	-	-
POLE	p.Asp275Val	-	-	-
POLE	p.Gln453Arg	-	-	-
POLE	p.Leu424Val	rs483352909 A=0.000008 (ExAC)	-	-
POLE	p.Pro286Arg	-	-	-
POLE	p.Pro436Arg	-	-	-
POLE	p.Ser297Phe	-	-	-
POLE	p.Val411Leu	-	-	-

Table 2.3: Checking DNA polymerase mutations for common variants
DNA polymerase mutations were cross-referenced with dbSNP, build 139 [813], 1000 Genomes, release May 2013[814] and in-house common variation sequencing projects (500 Exome Project and 300 control exomes of the cancer genome project). The submitter to dbSNP is denoted in parentheses. The minor allele frequency (MAF) is denoted in the Table with the minor allele. MAF refers to the frequency of the least common allele in a given population.

Gene	Amino acid change	Provean	Ppopen	PolyPhen	SIFT
POLD1	P327L	-9.824	31	0.999	Affect Protein Function
POLD1	S370R	-4.259	-5	0.407	Tolerated (score of 0.08)
POLD1	G426S	-0.579	-53	0.042	Tolerated (score of 0.37)
POLD1	R311C	-7.837	53	0.999	Affect Protein Function
POLD1	S478N	-2.82	21	0.998	Affect Protein Function
POLD1	V392M	-1.935	-29	0.946	Affect Protein Function
POLE	L424V	-2.78	85	1	Affect Protein Function
POLE	R446Q	-2.881	-41	0.994	Affect Protein Function
POLE	D275V	-8.139	92	1	Affect Protein Function
POLE	P286R	-8.139	94	1	Affect Protein Function
POLE	S297F	-5.426	84	1	Affect Protein Function
POLE	V411L	-2.763	88	1	Affect Protein Function
POLE	A456P	-3.963	67	1	Affect Protein Function
POLE	A428T	-2.234	-52	0.041	Tolerated (score of 0.23)
POLE	M444K	-5.388	83	1	Affect Protein Function
POLE	Q483R	-2.9	-30	1	Affect Protein Function
POLE	A465V	-3.751	39	1	Affect Protein Function

Table 2.4: Polymerase mutations identified from the literature with predicted consequences Polymerase mutations were identified from the literature [768, 809, 810] and their potential effects on protein structure and function was predicted using bioinformatic mutation prediction software. Scores are judged as follows: PROVEAN | If the score is <= -2.5 (predefined threshold), the protein variant is predicted "deleterious". SIFT | Score ranges from 0-1 and any score <0.05 is considered “deleterious”. Poly-Phen2 | The score is the probability of the substitution being deleterious. PredictProtein(PPopen) | Scores range from -100 to 100 and score > 50 indicated a “strong signal for effect”, a score between 50 and -50 indicates a “weak effect” and scores below -50 signify “no effect”.

ware that employ strategies from evolutionary sequence comparisons to structure-based predictions: PROVEAN/SIFT [815–819], Poly-phen2 [820–823], PredictProtein(PPopen) [824], Mechismo [825] and Mutation Taster [826]. When considering all the scores for one mutation combined, *POLE* S297F, *POLE* P286R, *POLE* V411L and *POLE* A456P score as highly damaging to protein function across different software tools.

To overcome the limitations of single-gene reporter assays, a strategy employing mutation accumulation followed by whole-genome sequencing was developed. Rather than testing mutations in human cells, mutations were to be tested in budding yeast. The evolutionary conservation of Polε and Polδ makes this approach possible, as the routine methods for strain construction, short doubling time, expertly curated reference genome and low sequencing costs makes it advantageous. Alignment of human *POLD1* and *POLE* with *S. cerevisiae POL2* and *POL3*, respectively (Fig. 2.4), shows that most candidates can be constructed in yeast as the residues in question are conserved. Four variants from the list of DNA polymerase mutations

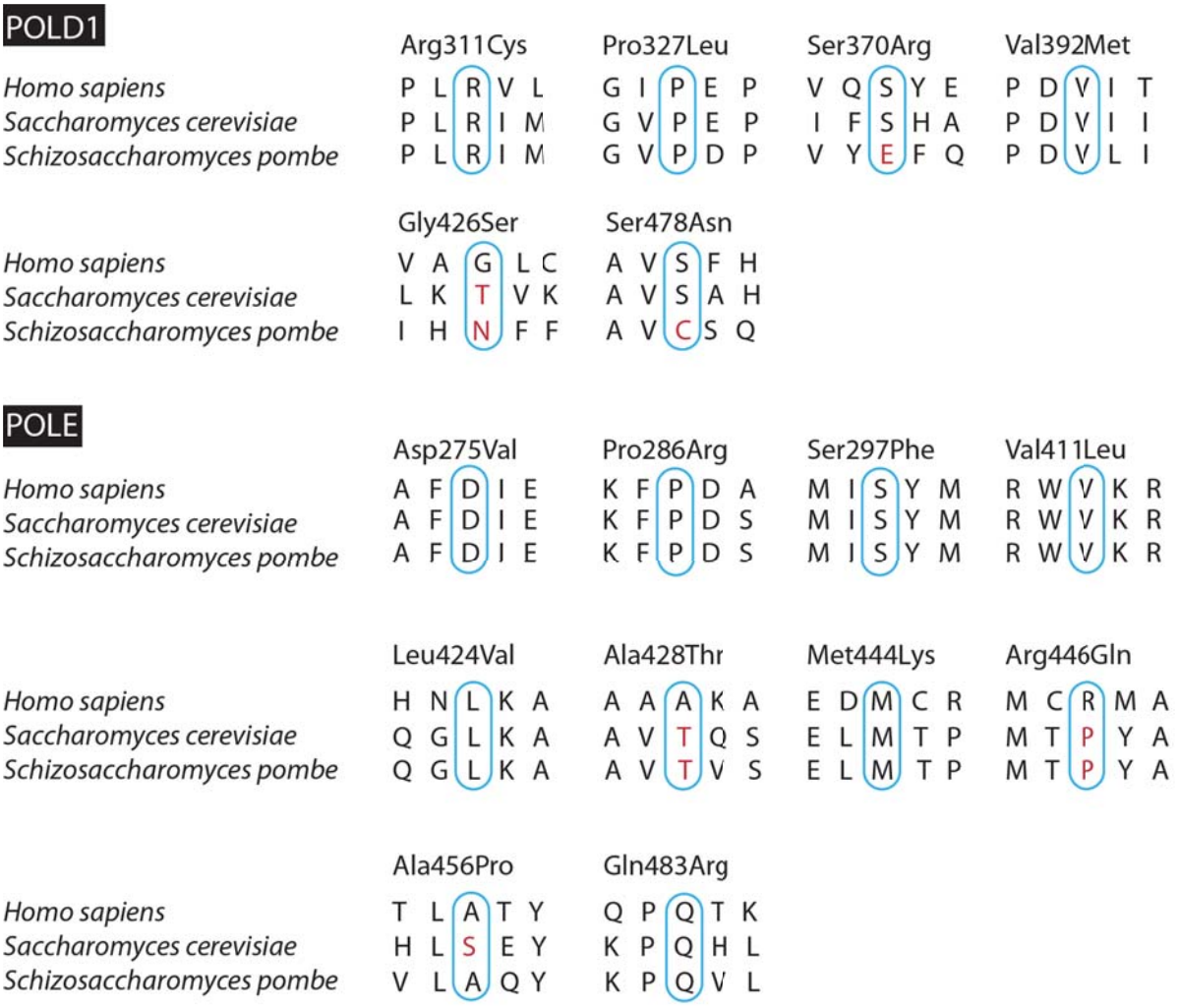


Figure 2.4: Alignment of polymerase residues of interest to the yeast proteins
Sequences were aligned using Clustal Omega version 1.2.1[827–829]. Sequences used for alignment (uniprot ID in parenthesis): *Homo sapiens* POLE (Q07864), *Saccharomyces cerevisiae POL2* (P21951), *Homo sapiens POLD1* (P28340), *Saccharomyces cerevisiae POL3* (P15436), *Schizosaccharomyces pombe POL2* (P87154) and *Schizosaccharomyces pombe POL3* (P30316). The residue identified as mutated in [768],[810] and [809] is encircled and unconserved residues are marked red. The amino acid change identified in the human samples is given at the top of each alignment.

Human variant	Conserved	*S. cerevisiae* variant
POLD1 p.Arg311Cys	Yes	*pol3 p.Arg3116Cys*
POLD1 p.Gly426Ser	No, (T)	-
POLD1 p.Pro327Leu	Yes	*pol3 p.Pro322Leu*
POLD1 p.Ser370Arg	Yes	*pol3 p.Ser375Arg*
POLD1 p.Ser478Asn	Yes	*pol3 p.Ser483Asn*
POLE p.Met444Lys	Yes	*pol2 p.Met459Lys*
POLE p.Ala456Pro	No, (S)	-
POLE p.Ala428Thr	No, (T)	-
POLE p.Ala465Val	Yes	*pol2 p.Ala480Val*
POLE p.Arg446Gln	No, (P)	-
POLE p.Asp275Val	Yes	*pol2 p.Asp290Val*
POLE p.Gln453Arg	Yes	*pol2 p.Gln468Arg*
POLE p.Leu424Val	Yes	*pol2 p.Leu439Val*
POLE p.Pro286Arg	Yes	*pol2 p.Pro301Arg*
POLE p.Ser297Phe	Yes	*pol2 p.Ser312Phe*
POLE p.Val411Leu	Yes	*pol2 p.Val426Leu*

Table 2.5: Budding yeast equivalents of human DNA polymerase mutations of interest
Using protein alignments equivalents of human DNA polymerase mutations were determined when possible. In cases where the affected amino acid is not conserved, the amino acid found in the budding yeast protein at that position is given in brackets.

to test, including the *POLE* A456P variant, were removed due to lack of conservation (Table 2.5).

2.3 Generation and propagation of polymerase mutants in *S. cerevisiae*

2.3.1 Constructing single mutant polymerase strains

All polymerase mutations were introduced into a W303 MAT a haploid *S. cerevisiae* strain generating twelve single mutants and mating them to the isogenic Mat α strain generating heterozygous diploid strains. Point mutations were introduced by plasmid integration: two different plasmid constructs were made for *POL2* and *POL3*(Fig. 2.5-A). Integration of each plasmid results in a functional copy of the gene carrying the mutation and a truncated, non-functional fragment(Fig. 2.5-B), C-terminal for *POL3* and N-terminal for *POL2*.

To allow wild-type expression of the ensuing mutated *POL2* gene, we also included 1kb of the upstream region containing the promoter. This does, however, lead to an N-terminal

truncation which is likely transcribed, but also targeted by nonsense-mediated decay (NMD). See YMH8-YMH41 in 6.3.2 for genotypes of all strains generated.

As reference, strains deficient for the proofreading activity of *POL2* and *POL3* were generated by introducing mutations in the exonuclease domain. As discussed earlier, the exonuclease domain is crucial for the preferential hydrolysis of non-complementary nucleotides at the 3´-terminus of a nascent DNA strand. Elimination of the exonuclease activity of yeast polδ or ε is known to result in a mutator phenotype and can thus act as a positive-control[339]. Three conserved amino acid motifs (called Exo I, II and III) in the N-terminal regions of the proteins form the active site of the exonuclease domain and are conserved in polymerases[313]. The alleles *pol3-01* and *pol2-4* (see Table 2.1) contain mutations of two acidic amino acids (one aspartic acid and one glutamic acid), thought to be involved in metal ion coordination, to alanines, which are known to affect proofreading, but not polymerase activity of these proteins (see red triangles in Fig. 2.6). I introduced these two point mutations using my plasmid constructs to generate haploid *pol2-4* (YMH28) and *pol3-01* (YMH32) equivalents.

2.3.2 Mutation accumulation experiment: Propagation of single mutant polymerase strains

There are several classical reporter gene assays to measure mutagenic activities in yeast. Assays measuring resistance to thialysine (Thia^r) or canavanine (Can^r) measure different types of mutation events inactivating the lysine permease (*LYP1*) or arginine permease (*CAN1*) genes, respectively[830, 831]. Beyond that other constructs have been used to study frameshifts (reversion of *hom3–10* or *lys2ΔBgl*)[832]. Proxies for gross chromosomal rearrangements and aneuploidy events are also available[833, 834]. These assays have been instrumental in identifying mutator phenotypes (Table 2.1), but they do have considerable limitations. For instance, counting resistant colonies provides no measure of phenotypically silent, synonymous mutations. Furthermore, usually only a specific type of mutation in a single gene in a single locus of the genome is used as a proxy for the whole-genome, neglecting factors such as sequence composition and context, variable DNA damage and repair frequencies across the genome, as well as chromatin states and physical conformation of the DNA. Additionally, if one wanted to study the whole mutational spectrum, one would have to combine a vast array of assays to cover the entire catalogue of mutation types. Additionally, forward mutation assays do not allow the experimenter to distinguish between frameshifts and single base changes unless reporter genes are sequenced, which is labour intensive and relatively expensive. Recent work indicates that when compared to whole-genome sequencing measurements of particular muta-

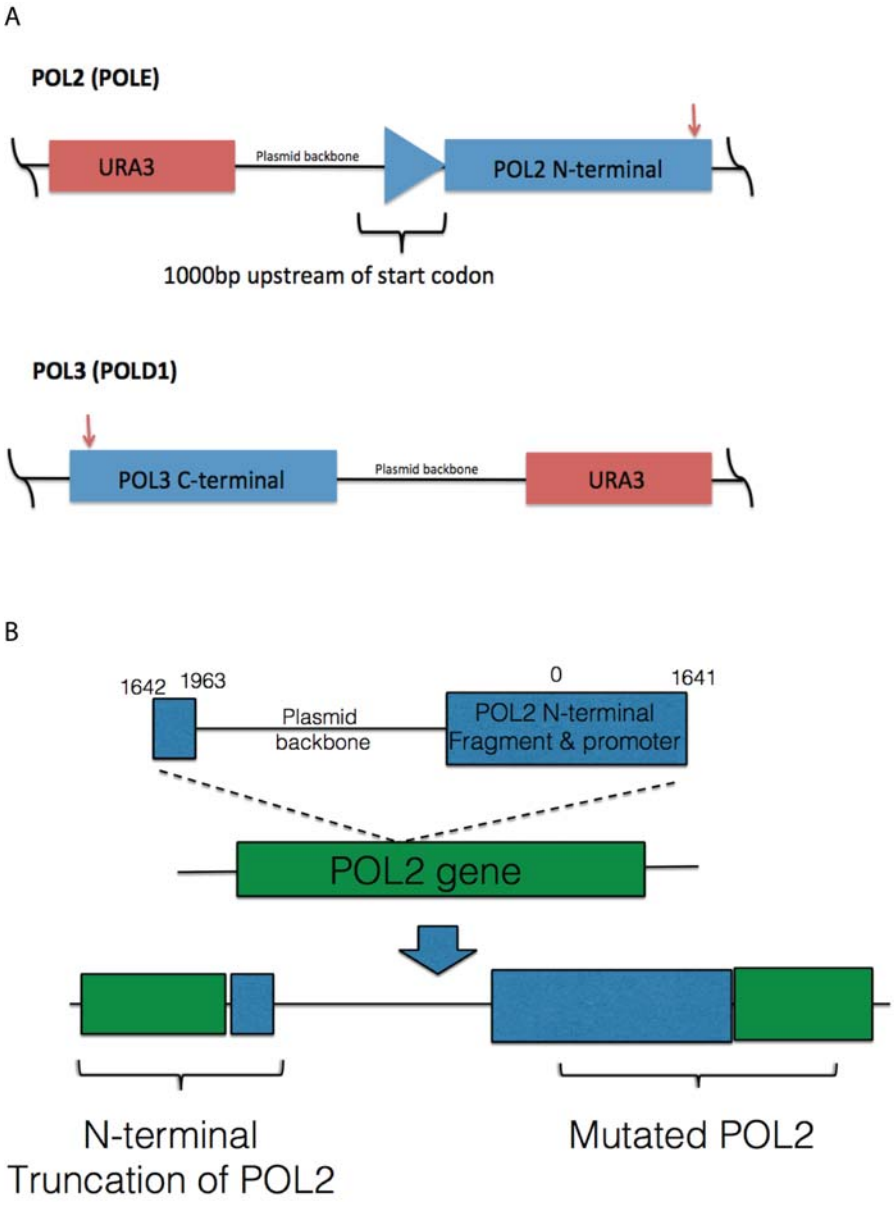


Figure 2.5: Rationale for plasmid construction

A| Two different types of vectors were designed - one for *POL2* and on for *POL3* mutations - which contain a selectable marker and a fragment of the gene. The vector pRS306 was modified to generate appropriate integrating plasmids. This vector contains an ampicillin resistance for selection in *E. coli* and *URA3* for selection and counter-selection in *S. cerevisiae* and no centromere allowing integration after linerarisation. The red arrows denote approximate sites for linerarisation, the black vertical lines at either side of the vectors symbolise that they are circular. **B|** Linearised vectors (here the N-terminal example is shown) will insert into the gene by HR creating a truncated gene as well as a functional gene fusion carrying the mutation introduced in the plasmid.

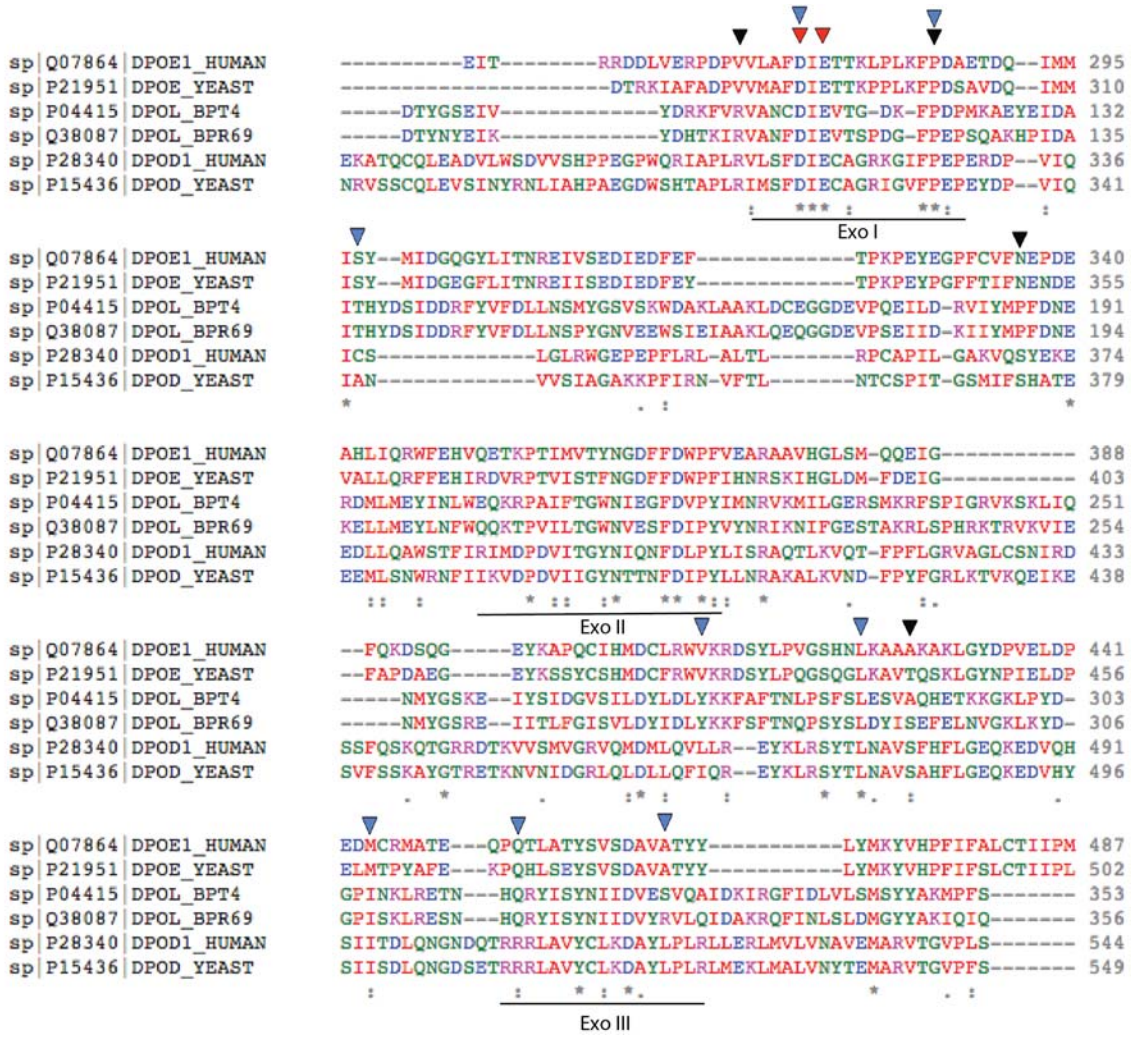


Figure 2.6: Exonuclease domains conserved in B family polymerases
Sequences were aligned using Clustal Omega version 1.2.1[827–829]. Sequences used for alignment (uniprot ID in parenthesis): *Homo sapiens POLE* (Q07864), *Saccharomyces cerevisiae* POL2 (P21951), *Homo sapiens POLD1* (P28340), *Saccharomyces cerevisiae* POL3 (P15436), *Enterobacteria phage T4* 43 (P04415), *Enterobacteria phage RB69* 43 (Q38087), *Bacillus phage phi29* 2 (P03680). The three exonuclease motifs (ExoI, ExoII and ExoIII) are underlined. The residues mutated to generate exonuclease deficient strains are highlighted by red triangles. The polymerase mutations in *POLD1* are highlighted by black triangles, those in *POLE* by blue triangles.

tions, some reporter assays provided reasonably accurate results, while others were not optimal proxies for the whole-genome[835]. With this in mind, I have decided to test the effects of the polymerase mutations by propagating the strains carrying mutated DNA polymerases and detecting mutations acquired during the process by whole-genome sequencing.

2.3.2.1 Single-colony bottleneck propagation of mutant polymerase strains

To obtain a significant number of mutations per strain, mutations were allowed to accumulate in parallel over 26 passages through single colony bottlenecks while cells were grown on non-selective rich medium for a total of three months. As illustrated in Fig. 2.7, in each case the starting strain was sequenced as well as each parallel line that was propagated. To determine the number of parallel lines needed to obtain sufficient mutations, I considered the fact that in a similar experiment, wild-type yeast cells accumulated on average 10.25 mutations after 100 passages[835]. Considering that for examination of mutational spectra, a significantly higher number of mutations is needed, the wild-type YMH9 strain was propagated in 72 parallel lines (projected to result in ~180 mutations in total), the YMH29 strain (carrying the *pol2-4* variant) in 54 parallel lines and all others in 18 parallel lines (see Table B.1.2.1). The shorter time span (25 instead of 100 passages) is aimed to reduce any contributions from secondary arising mutations. However, even in the case of 100 passages (using the Can^r assay) no change in mutation rate between starting and final strains was detected[835], suggesting that alterations in mutation frequencies are most likely due to the query mutation rather than secondary mutations.

2.3.2.2 Population bottleneck propagation of mutant polymerase strains

The main drawback of using single-colony bottlenecks, is that, if sequencing reveals an insufficient number of mutations, one cannot simply sequence more strains. Instead, the experiment would have to be repeated. As an alternative, the same strains as well as the haploid precursors (Table B.1.2.2) were propagated using population (10^4 cells) bottlenecks. This avoids the need for extensive parallel lines and more than one sample from the final population can be sequenced. Final populations can also be stored frozen and more colonies sequenced later. However, since these samples are not independent (as they are in the case of parallel lines with single colony bottlenecks) the actual number of independent mutations depends on the complexity of the final population.

In this experiment, the strains were propagated automatically by a serial-propagation platform in conjunction with our collaborators Ville Mustonen (WTSI) and Jonas Warringer (Uni-

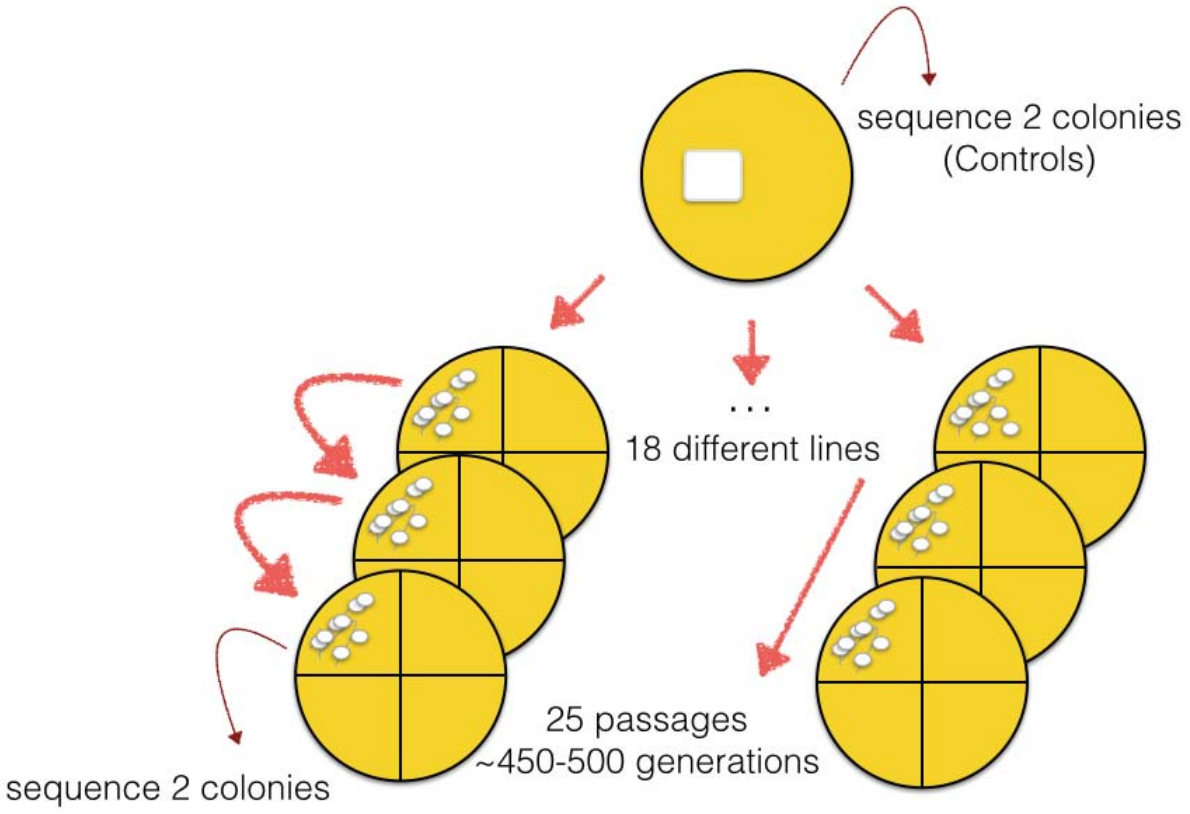


Figure 2.7: Mutation accumulation experiment: manual propagation of mutated *S. cerevisiae* strains
Experimental strategy: the heterozygous diploid polymerase mutant strains (all derived from the same wild-type W303 strain) were patched onto YPAD. From each patch 18 different parallel mutation accumulation lines were derived, by streaking small amounts of cells for single colonies on fresh YPAD plates. The remainder of the patch was frozen for later DNA extraction and serves as a starting points. The cells were grown to single colonies at 25°C (~20-25 generations) and cells were moved to a fresh plate using single-colony bottlenecks for 25 passages. Starting colonies and 2 colonies from each parallel line were whole-genome sequenced.

versity of Gothenburg, Sweden). This involved using a robot to transfer populations of cells onto new agar plates every two-three days for three months (see 6.7 and [836]). Twenty-eight colonies each for YMH8 and YMH9 (wild-type background) and YMH28 and YMH29 (*pol2-4* mutation) and eighteen each for all other strains were sequenced.

2.4 Establishing sequence analysis practices

The majority of the work in this thesis uses the budding yeast *Saccharomyces cerevisiae* and next-generation sequencing. This chapter also describes the establishment of DNA sequencing analysis protocols in budding yeast and their application to other projects as a validation of the analysis strategy.

2.4.1 Automating genomic DNA extraction and whole-genome sequencing of *Saccharomyces cerevisiae* strains

Extracting high quality genomic DNA (gDNA) from yeast cultures by standard protocols is a low throughput method for extracting DNA for sequencing (see 6.6 for protocol). For the scale of this and other work a more high-throughput protocol for extracting gDNA was needed. Dr. Fabio Puddu, with the assistance of Nicola Geisler, developed a protocol to extract gDNA from 96 samples at a time using a robot, which I tested for sequencing by comparing the sequencing data I generated from samples extracted by phenol-chloroform extraction and those that were extracted using the robot (see 6.6 for protocols).

To assess whether the sequencing data obtained from DNA extracted using this high-throughput protocol was of similar high quality as the data acquired from DNA obtained by conventional phenol–chloroform extraction, samples subjected to either of these methods were compared for quality using key quality control measurements.

The Sequencing Facility at the Wellcome Trust Sanger Institute assesses all DNA for concentration, volume and total amount. From over 1000 samples prepared with the high throughput method 96% passed their quality control thresholds to proceed to library preparation and sequencing. For whole-genome deep sequencing, a mean genome-wide coverage of at least 30× is ideal and so far all samples that were sequenced after DNA extraction using this protocol have a coverage of at least that (Fig. 2.8-A). DNA sequencing of samples extracted using the high-throughput protocol is of comparable quality to sequencing of DNA extracted using phenol-chloroform in metrics regarding read alignment (Fig. 2.8-B), coverage of the entire genome (Fig. 2.8-C) and insert size distribution (Fig. 2.8-D) as well as GC content. Thus,

DNA extraction using this high-throughput extraction protocol allows us to obtain DNA of sufficient quantity and concentration for sequencing and the data obtained after sequencing compares favourably to previously sequenced samples in key quality measures. DNA extraction using this protocol was used for the remainder of this work.

2.4.2 Establishing sequencing analysis protocols for the identification of SNVs and INDELs

One of the main issues with identifying mutations from sequencing data is that one has to make decisions about which variants to retain as true variants and which to filter out as likely artifacts or errors, all the while usually not knowing what the true answer is. To tackle this problem, I developed variant calling and filtering strategies while continuously monitoring the approximate false negative and false positive rate under the supervision of Dr. Thomas .

Comparing to a capillary sequence reference The yeast reference genome was generated from a strain of the S288c background, whereas most of the strains featured in this work are of the W303 background, a strain generated in the 1970s. The genome of W303 is 85.4% identical to the S288c background and divergent sequences resemble those of Σ1278b. 799 proteins differ between the W303 and S288c strains, but most of the time only one or two residues differ[837]. Running variant calling and filtering on previously generated sequencing data from the Jackson lab of 22 strains from the W303 background, showed that, on average, MATa W303 lab strains carry 9,534 variants before filtering and 9192 after default filtering when compared to the S288c reference genome(Fig. 2.4.2). It also confirmed that the *rad5-535* allele (a G535R missense mutation in *RAD5* carried by the original W303 strain) has been corrected in our K699 and K700 strains.

The *Saccharomyces* Genome Resequencing Project completed ABI sequencing on a haploid W303 strain to a depth of between 1x and 3x which is freely available to download[838]. Compared to Illumina HiSeq data, ABI or capillary sequencing produces high quality long reads with a high degree of accuracy[774]. Comparing my W303 background data to the capillary sequencing data can provide some insight into the accuracy of my variant calling and filtering strategy. Due to the fact that they are not the exact same strain, discrepancies are expected, but I will be able to get an estimate for the false negative rate. False-positive rate estimates are much more problematic, due to the low coverage of the ABI sequencing data (2.3X), meaning that there will be regions of zero coverage, and won't be calculated here.

Dr. Thomas Keane performed long-read alignment on the capillary sequencing data (see

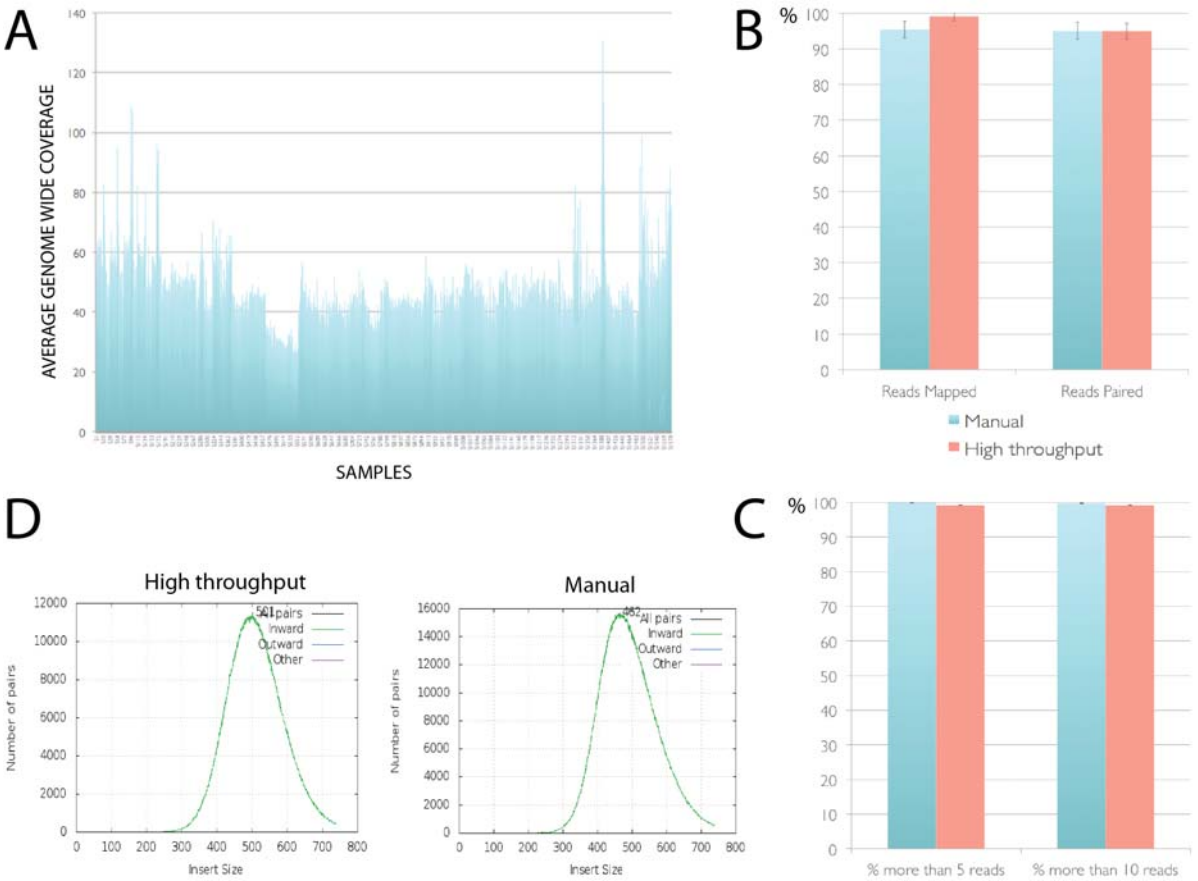


Figure 2.8: DNA extracted using a high-throughput protocol produces high quality sequencing data

A | Mean genome wide coverage of 1577 samples sequenced after DNA was extracted using the high-throughput extraction protocol. **B |** Comparison of the percentage of reads that could be mapped to the reference genome and the percentage of reads that were paired between the 1577 samples whose DNA was extracted using the high-throughput extraction protocol and 168 samples extracted manually using phenol-chloroform. **C |** The same samples were compared for which fraction of the reference genome was sequenced to more than a depth of 5 and more than a depth of 10, respectively. **D |** Representative examples of insert size distributions for a high-throughput extraction (see 6.6) and a manual extraction (see 6.6) are shown.

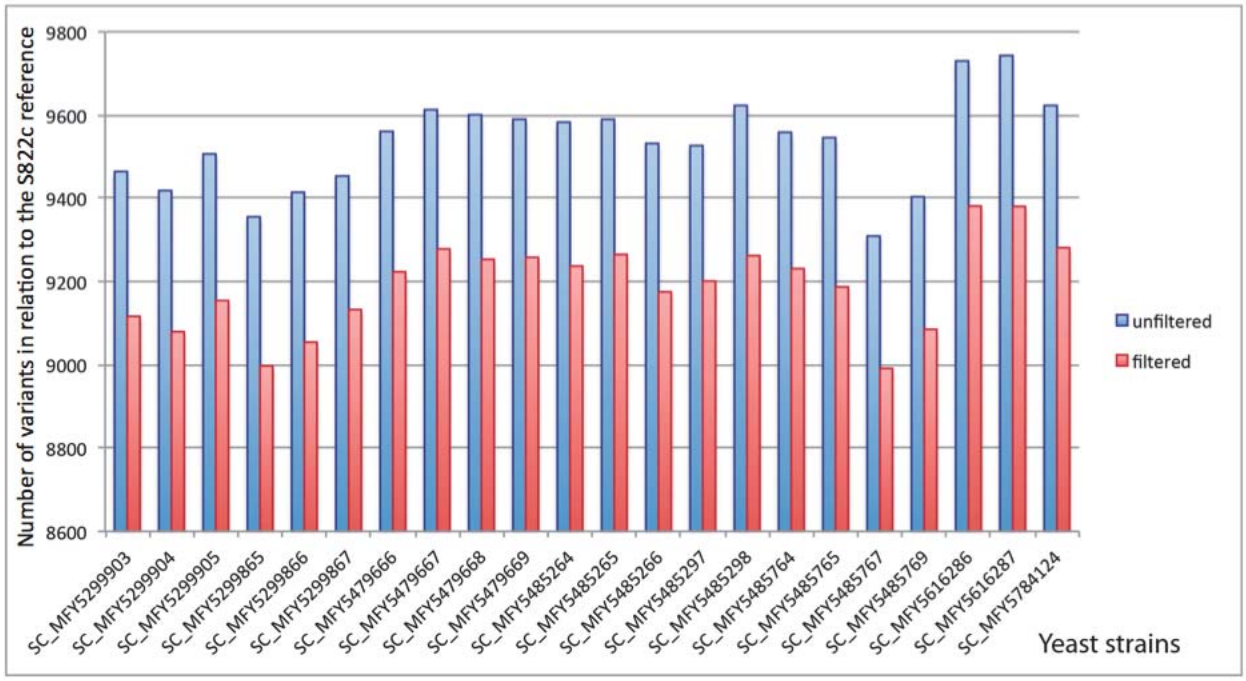


Figure 2.9: The number of variants in W303 strains compared to the S288c reference genome Aligned sequencing data from 22 *S. cerevisiae* strains of the W303 background strains was used to identify the number of background mutations to be expected when sequencing W303 *S. cerevisiae* strains. The samples are all control samples taken from other sequencing projects performed in the lab (see Table 6.3.2). Variant calling was carried out with samtools mpileup using parameters as specified in Table B.1.1.2 and filtering was done by vcf-annotate using its default filtering parameters. Total numbers of mutations per sample before and after filtering were counted and plotted.

Table B.1.1.1) and I performed variant calling as well as filtering on the ABI sequenced sample as well as 10 Illumina sequenced samples. Initially, when intersecting the variants called from the ABI sequencing with different samples of the Illumina sequenced set, we found 44.5%-50.8% of INDELs and 77.7%-78.5% of SNPs from the W303 Capillary data in the Illumina calls. Using the Integrative Genomics Viewer (IGV)[839] to look at the alignments in regions where the variant calling called a variant for the capillary sequencing data, but not the Illumina sequencing, suggested sensible ways to "tweak" the filtering step of the analysis. The alignments revealed that many of those variants were not captured due to mapping quality and depth filter thresholds (as well as many variants mapping to mitochondrial DNA) and adjustments of those reduced the approximate false negative rate to 2.3% meaning we can capture >97% of variants identified in capillary sequencing in the Illumina sequencing data. Running the GATK indel realignment tool to account for misalignment around an INDEL did not improve the calling sensitivity.

Simulated genome data Another, albeit imperfect, approach is to include simulated sample data in every analysis. Simulated data effectively avoids the issue of unknown results: the mutations in the samples are known and analysis should find them with minimal false negatives and false positive rates. The major shortcoming of the technique is, clearly, that it is simulated and can only approximate the realities of next-generation sequencing. Most of my project's analysis will involve experimental samples and controls. Both sets of samples will have their variants called in relation to the reference genome and in order to identify the mutations experimental samples acquired during the experiment, mutations identified in control samples should be discarded from the experimental data (Fig. 2.10-A). This set-up is also reflected in the simulated data set we generated. Using pIRS (profile-based Illumina pair-end reads simulator)[840], several simulated samples were generated: control samples and experimental samples (containing all control sample mutations and additional ones). The control dataset had 8000 mutations inserted. This dataset was further mutated computationally to simulate experimental settings. The number of mutations to add was chosen considering the wild-type mutation rate (base-substitutional mutation rate: 0.33×10^{-9} per site per cell division, [841]) and the suggested fold increase for a polymerase exonuclease deficient strain[768]. At the chosen parameter, around 200-300 SNPs were introduced. An INDEL dataset with 800 INDELs was also generated. After alignment and variant calling a false-negative and a false-positive frequency were determined. The false-negative frequency for SNV calls was 4-5.5% and 39.2% for INDELs. When the same adjustments for mapping quality, low depth and mitochondrial mutations as before were made, this number drops to less than 1% for SNVs

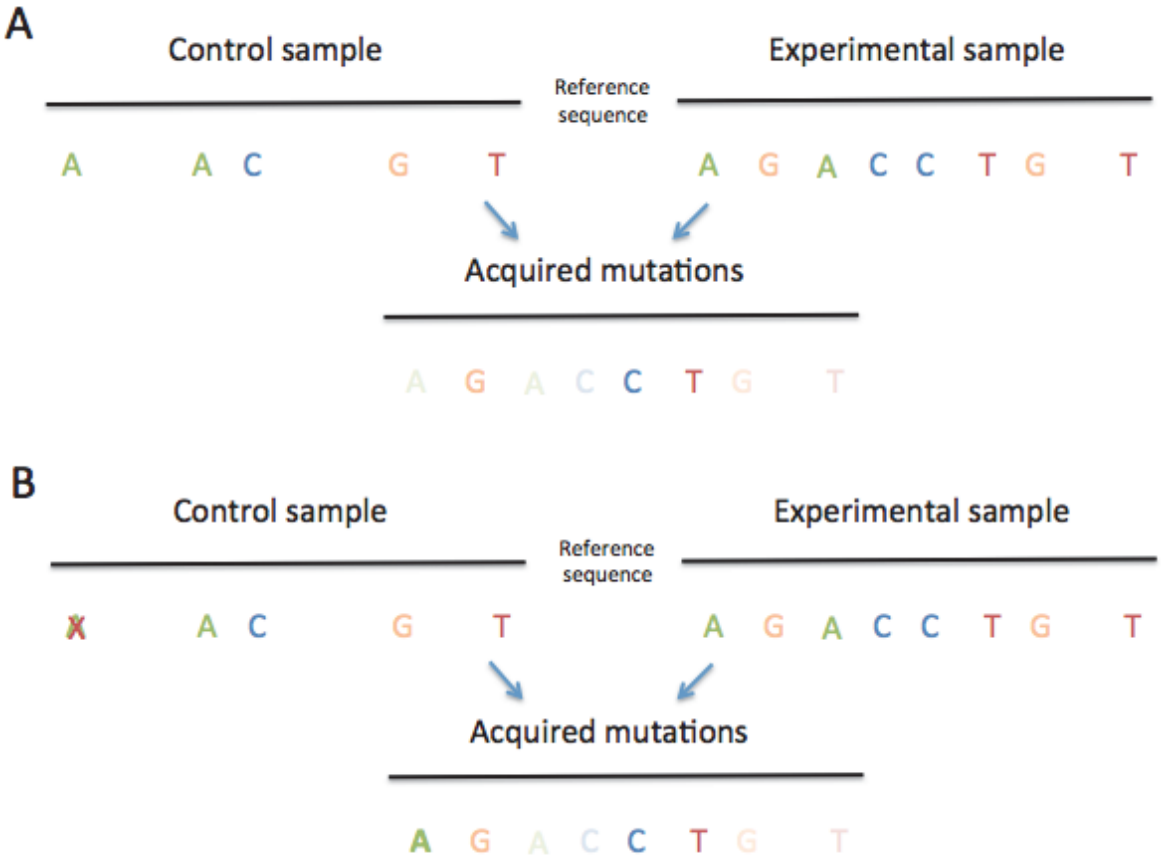


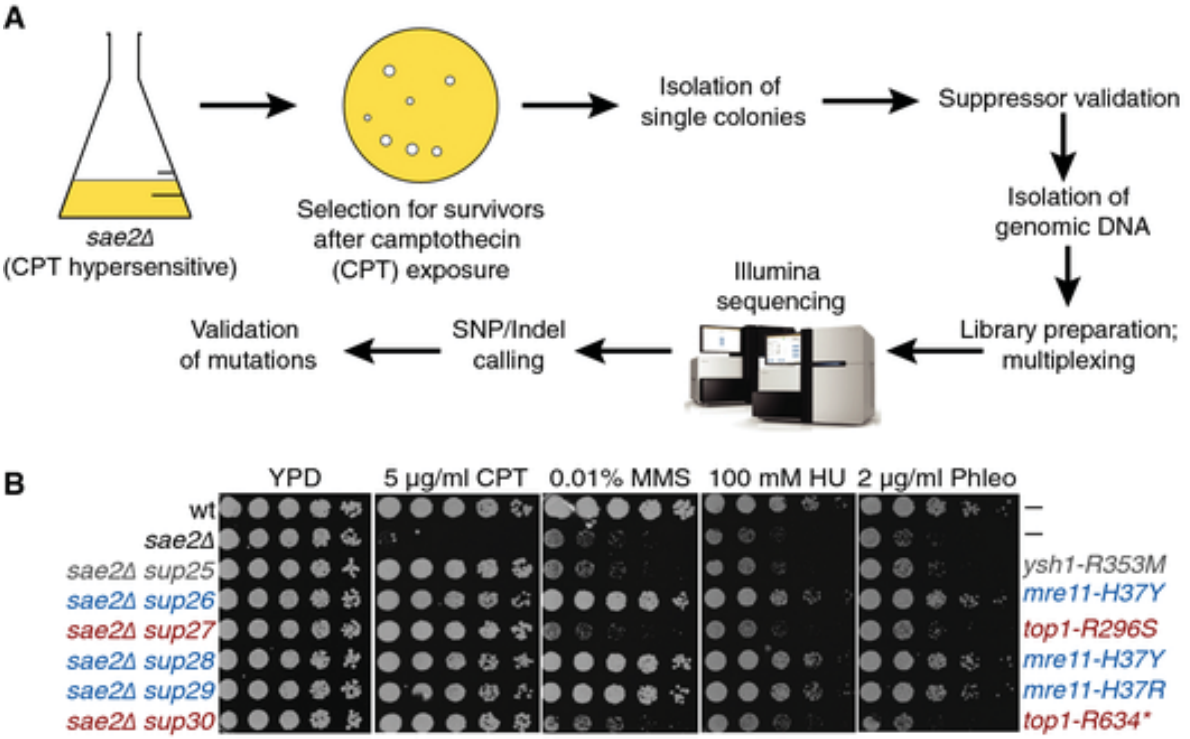
Figure 2.10: Experimental strategy to identify acquired mutations
A| In most sequencing experiments performed in this work, single nucleotide variants and small INDELs have been called with respect to a reference genome for both a control (pre-treatment) and an experimental (post-treatment) sample. The list of identified mutations will be intersected to identify mutations only present in the latter, giving us a list of acquired mutations. **B|** In some cases a mutation may not be detected due to sequencing errors or filtering of low quality even though it is present in the DNA (see stricken out A mutation). This will lead to an apparent false positive in the list of acquired mutations (see bold A mutation).

and 12.6% for INDELs. The false-positive frequency for INDELs was found to be ~25%, whereas for SNVs the false-negative frequency was less than 1%. However, interestingly, not a single case of a true false positive was found (a variant call where no variant was present). Instead, variants that were mistakenly not called or filtered out from the control sample (false negative), could then not be removed from experimental samples creating effectively a false positive (Fig. 2.10-B). This highlights the case for more lax filtering to be applied to control samples and/or using more than one control sample to minimise the number of "false positives" generated this way. Sequencing was also simulated at different coverages (20X, 30X, 40X and 50X) and no difference in variant calling accuracy was found at these coverage levels.

2.4.3 Testing analysis protocol on *Saccharomyces cerevisiae* genetic screens

Screens in budding yeast have been used extensively and successfully to identify gene interactions. One example is synthetic lethality where two mutations result in lethality when co-occuring in one cell while cells carrying only one of the two are viable. Possibly more interesting are suppressor mutations (synthetic viability), where a mutation results in a phenotype which is reversed by a second mutation. While synthetic lethality can occur due to the inactivation of two parallel important pathways and not reflect true genetic interaction, suppressor mutations are often more informative about underlying molecular processes. Until recently, identifying a suppressor mutation involved laborious cloning of the suppressor loci. However, with the advances in sequencing technology and the associated reduction in costs, high-throughput synthetic viability genomic screening has become more and more feasible. To address a long-standing question in yeast DNA repair biology - the DNA damage sensitivities of *sae2*Δ cells - Dr. Tobias Oelschlägel performed a synthetic viability genomic screening identifying *sae2*Δ cells spontaneously resistant to camptothecin (CPT). 48 suppressor were sent for sequencing at the Wellcome Trust Sanger Institute as detailed in [801] and Chapter 6.8.

Since CPT is an inhibitor of DNA enzyme topoisomerase I (*TOP1*), stabilising the *TOP1*-DNA complex and resulting in replication-dependent DSBs, we expected that inactivating mutations of *TOP1* would likely be among the suppressor mutations. Such expectations, together with the fact that this project would likely involve confirming identified suppressor mutation with an orthogonal sequencing technology, this screen was ideal to test our analysis strategy. Together with Dr. Thomas Keane, I analysed the bwa-aligned bam files using the filtering strategy we developed in Chapter 2.4.2 (see Chapter 6.9 for more details). Similar to the set-up of my mutation accumulation experiments, this work involved sequencing a sensitive



C

ORF containing a putative driving mutation	Type				Times found	Number of clones
	Nonsense	Missense	Frameshift	Total		
TOP1	4	4	5	13	28	28
MRE11	0	2	0	2	10	10
YSH1	0	1	0	1	14	3
DNF3	0	1	0	1	12	2
DYS1	0	1	0	1	9	1
IMD2	0	1	0	1	8	1
Not determined						2
Not available						1
Total analysed						48

Figure 2.11: Sequencing analysis identifies mutations capable of suppressing *sae2*Δ DNA damage hypersensitivity
A| Outline of the screening approach that was used to identify suppressors of sae2Δ camptothecin (CPT) hypersensitivity. **B**| Validation of the suppression phenotypes; a subset (sup25–sup30) of the suppressors recovered from the screening is shown along with mutations identified in each clone. **C**| Summary of the results of the synthetic viability genomic screening (SVGS) for sae2Δ camptothecin (CPT) hypersensitivity. The ORF and the type of mutation are reported together with the number of times each ORF was found mutated and the number of clones in which each ORF was putatively driving the resistance. Figure and text reproduced from [801] in accordance with the terms of the Creative Commons Attribution License.

starting strain and multiple suppressors. Retaining only mutations found in the suppressors and not in the starting strain will ideally reveal the suppressor mutations. We found that 24 of the clones possessed *TOP1* mutations and, interestingly, 10 contained either *mre11-H37R* or *mre11-H37Y* mutations (Fig. 2.11). Further strengthening our hypothesis, that these were real suppressors, was the fact that *MRE11* and *TOP1* mutation never occurred in the same samples and, intriguingly, the 10 colonies with *MRE11* mutations were not just resistant to CPT, but also other DNA damaging agents: phleomycin, which generates DSBs, the replication inhibitor hydroxyurea (HU), DNA-alkylating compound methyl methanesulphonate (MMS) and ultraviolet light (UV). Follow-up work to characterize the *mre11-H37R* mutant and elucidate its role as a suppressor of *sae2*Δ-dependent CPT hypersensitivity was largely carried out by Dr. Fabio Puddu and the work has been published[801].

We have extended this method to other questions in yeast DNA replication biology. For instance, the absence of the Tof1/Csm3 complex causes hypersensitivity of cells to CPT. To identify mutations that can alleviate this hypersensitivity, Dr. Fabio Puddu carried out a suppressor screen as above for sae2Δ cells and sequenced 16 suppressors of tof1Δ cells' hypersensitivity to CPT (Fig. 2.12). I performed the analysis as described above and in [801](see Chapter 6.9 for more details). Two of the strongest suppressors were found to have TOP1 mutations. Two different inactivating nonsense mutations in the *SIR3* gene were found in three clones, while eight other suppressor clones carried a nonsense mutation in the *SIR4* gene. Further work by Dr. Puddu confirmed that inactivating members of the Sir complex mediated suppression of camptothecin hypersensitivity and this is likely due to disruption of sir-dependent heterochromatin. We suggest a model that Topoisomerase 1 inhibition in proximity of sir-dependent heterochromatin causes intense topological stress that leads to DNA hypercatenation, especially in the absence of the Tof1/Csm3 complex.

We have also applied this approach to phenotypes outside of replication in collaboration with other researchers and are pursuing the molecular mechanism behind the suppression of other replication stress associated phenotypes. This demonstrates that, not only does the bioinformatical analysis I carried out retrieve relevant mutations that we can confirm by other techniques in the lab, but, while designed for mutation accumulation experiments, it can also be applied to a wide variety of genetic experiments and will be used to generate biological insights beyond the realm of its initial conception.

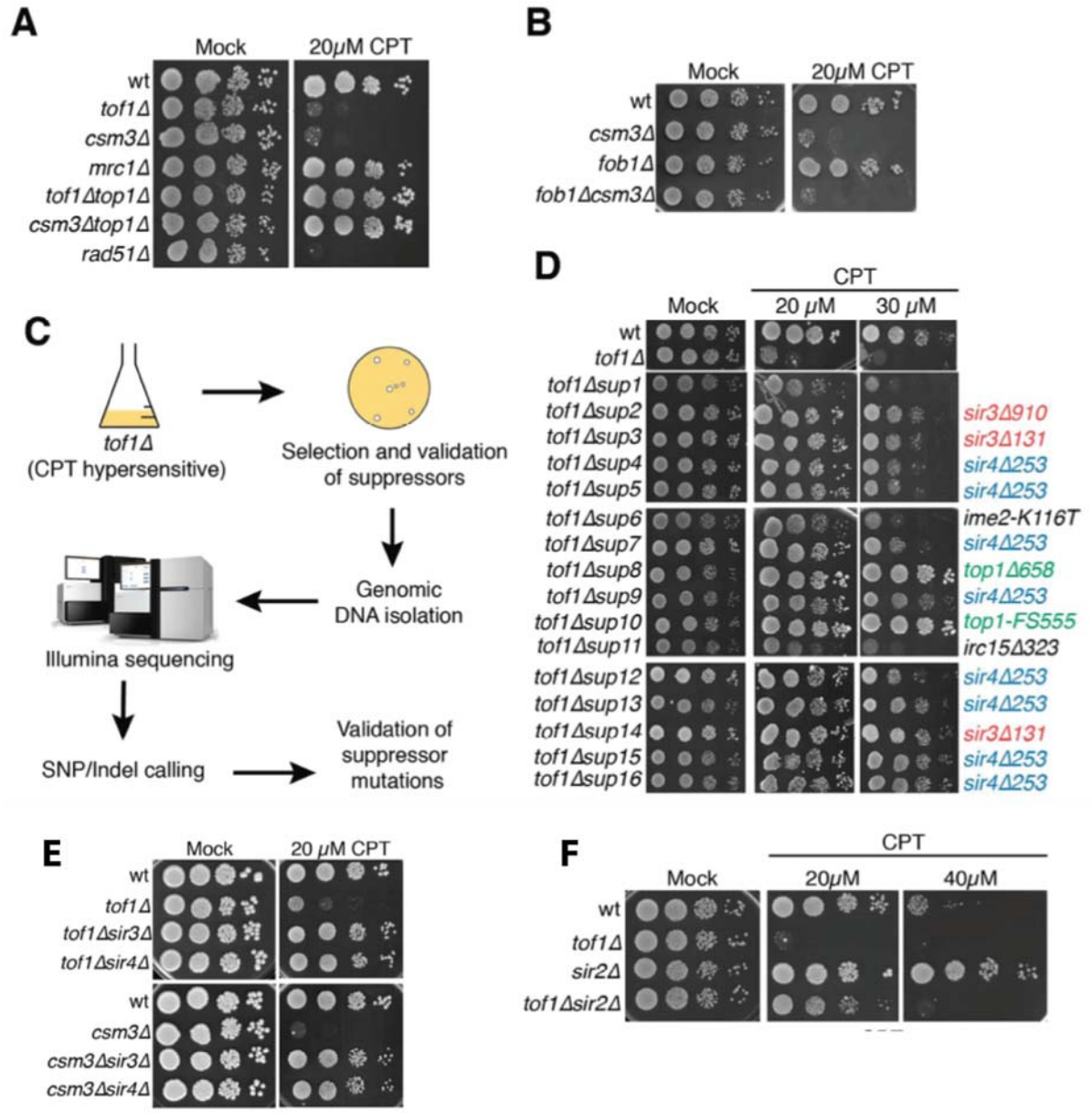


Figure 2.12: Mutations in *SIR3* and *SIR4* identified as the cause for the hypersensitivity of *tof1*Δ cells to camptothecin

(A) Loss of Tof1 and Csm3 but not Mrc1 causes hypersensitivity to camptothecin in a Top1-dependent manner. **(B)** Loss of pausing at the replication fork barrier on rDNA does not cause camptothecin hypersensitivity. **(C)** Outline of the procedure for a synthetic viability screen. **(D)** Synthetic viability screening identifies *sir3* and *sir4* alleles as suppressors of the camptothecin hypersensitivity of *tof1*Δ strains. **(E)** sir3 and sir4 deletions suppress the hypersensitivity of *tof1*Δ cells. **(F)** Deletion of SIR2 (encoding the third member of complexes containing Sir3p and Sir4p) also suppresses the hypersensitivity of *tof1*Δ cells and reduces the sensitivity of a wild-type strain. Drop tests were performed by Dr. Fabio Puddu, the assignments of mutations depicted in **D** were added by me.

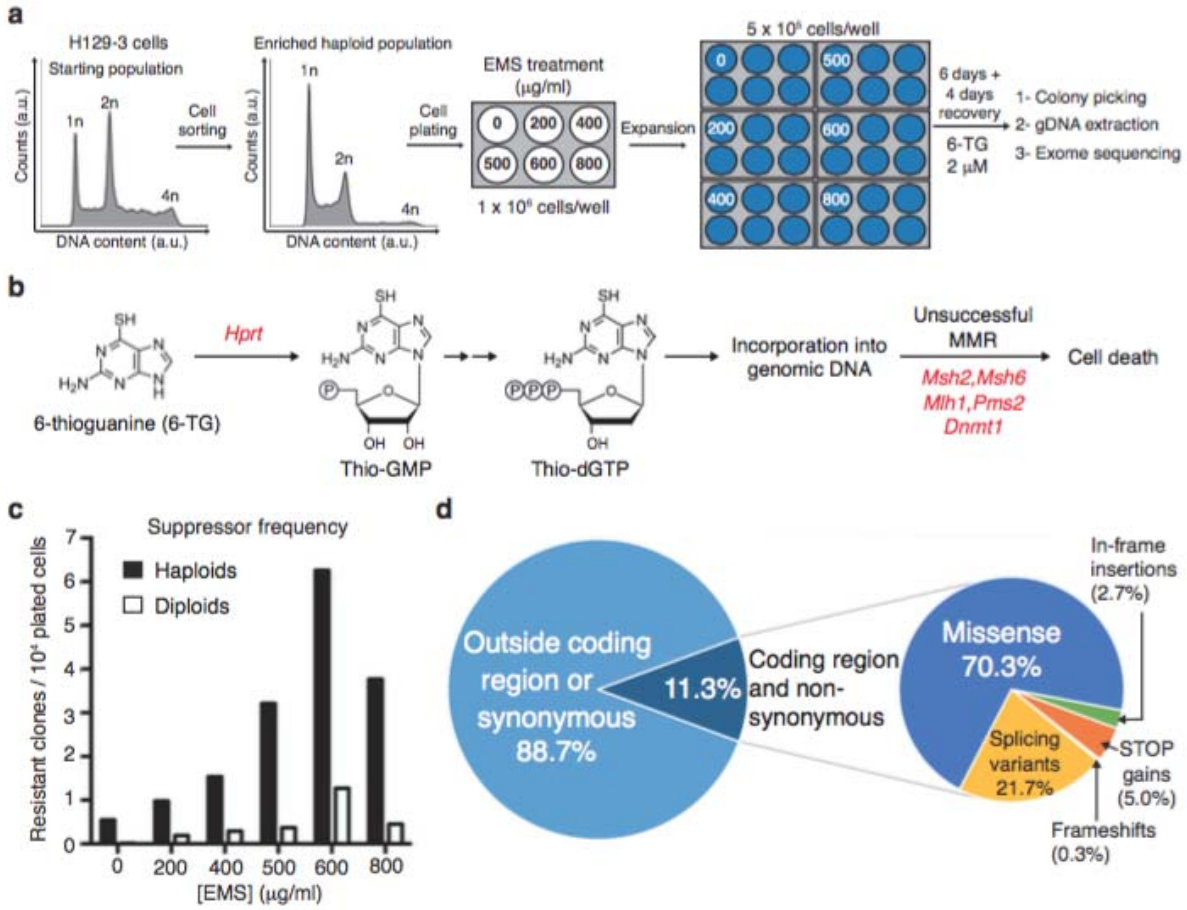


Figure 2.13: Generation of mutagenized libraries

(a) Experimental workflow. (b) Schematic of 6-TG metabolism and genotoxicity. Inactivating mutations in the genes highlighted in red have been shown to confer resistance to 6-TG. (c) Number of suppressors recovered at increasing concentrations of ethyl methanesulfonate (EMS) treatment. (d) Mutation consequences identified by whole-exome sequencing of 7 suppressor clones.

2.4.4 Applying analysis protocols to mouse genetic screens

The main advantage haploid yeast cells have for suppressor screens is that their haploid genome makes phenotypes, that would be recessive in a diploid, visible and selectable. Carrying out suppressor screens in diploid cells requires the appearance of dominant mutations, or mutations in both alleles of the same gene for a phenotype to be visible, making identification of suppressors more difficult. The success with next-generation sequencing of suppressors in haploid yeasts induced us to explore options in mammalian systems. Forward genetic screening in human cell lines has been feasible with the discovery of RNA interference (RNAi)[842], and more recently with insertional mutagenesis[843] and CRISPR/Cas9 libraries in near-haploid human cell lines[844–846]. And while loss-of-function (LOF) approaches like these are powerful, they have their limitations. Suppressor phenotypes caused by separation-of-function, gain-of-function or by mutations in essential genes[801, 847] are unlikely identifiable in these types of screens. The development of H129-3 haploid mouse embryonic stem cells (mESCs)[848] allowed us to circumvent the problems posed by diploid genomes. In collaboration with Dr. Josep Forment, haploid cells were treated with varying doses of the DNA-alkylating agent ethylmethanesulfonate (EMS) and 196 suppressors to the toxic nucleotide precursor 6-TG were isolated(Fig. 2.13-a). HPRT is known to initiate the cytotoxic mechanism of 6-tioguanine (6-TG) conversion to 2'deoxy-6-thioguanosine triphosphate (a cytotoxic nucleotide) in cells(Fig. 2.13-b)[849]. HPRT is thus a prime candidate for suppressor mutations since the loss of HPRT abolishes the cytotoxic effects of 6-TG. To test whether we could identify suppressors in the mouse genome, which is much larger than that of the budding yeast (2,716Mbp as opposed to 12Mbp in the reference genome), DNA from seven of these resistant clones and from a control mESC sample not treated with EMS was subjected to whole-exome sequencing.

Similar to the suppressor screen analysis detailed in Chapter 2.4.3, I performed variant calling (see Chapter 6.9 for details and Table B.1.1.2 for all parameters) on sequencing data aligned to the GRCm38 mouse reference genome by the Sanger Institute. I used my own scripts to remove any variants detected outside the bait regions and heterozygous variants where appropriate (see Chapter 6.9.5 for a list, description and location of Scripts). Low quality variants were filtered using standard and custom filters, variants present in the control sample were discarded and the remaining variants annotated for their functional consequences. Due to the much larger size of the genome, this alone proved not enough to remove all background mutations from the samples. This is likely due to the phenomenon described in Chapter 2.4.2, where false negatives in the control sample lead to an accumulation of apparent false positives in the data. To further filter the variants, Dr. Thomas Keane from the Vertebrate

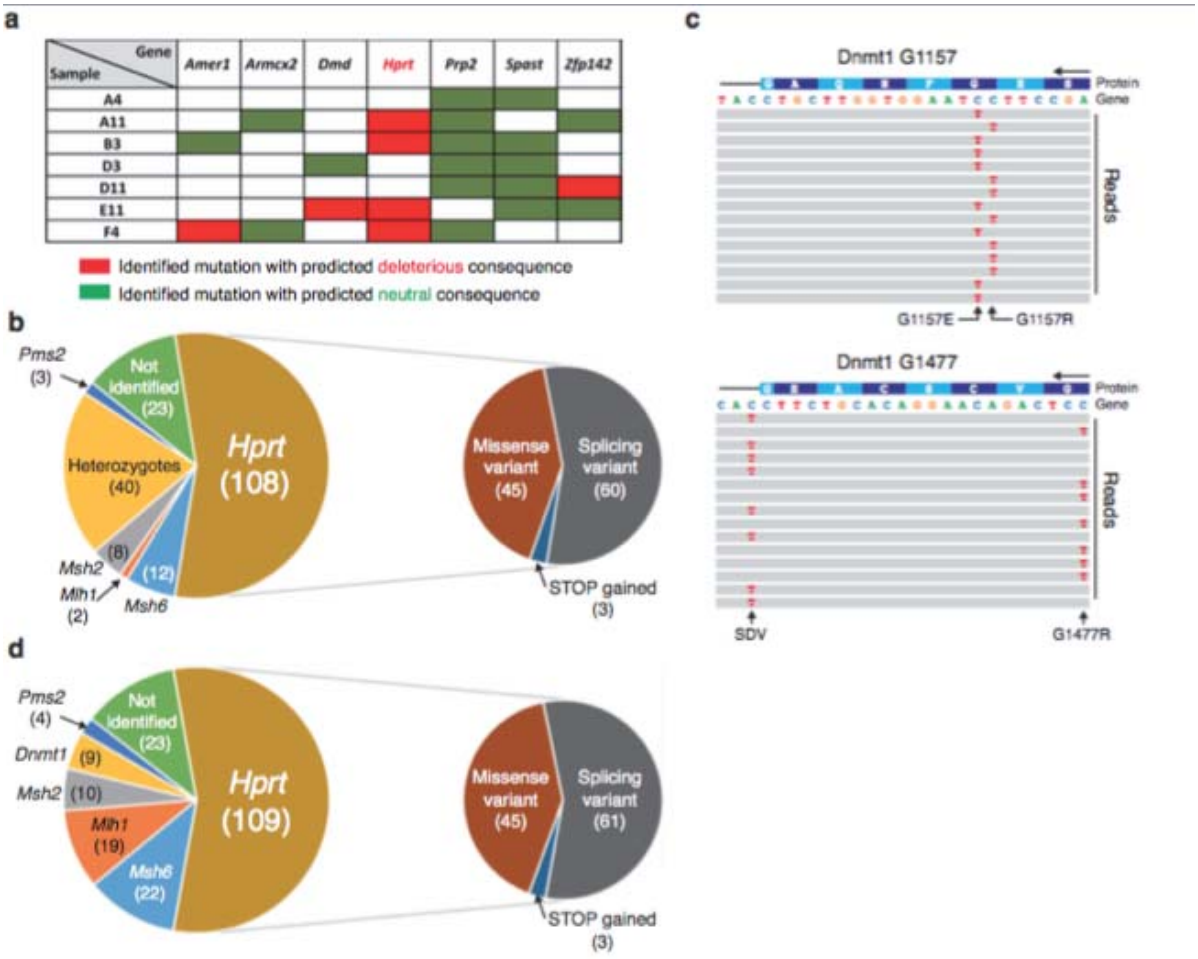


Figure 2.14: Identification of suppressor mutations
(a) Genes harboring independent mutations in different clones. Mutations were assigned as deleterious or neutral according to PROVEAN and SIFT software. (b) Distribution of homozygous mutations identified in suppressor gene candidates; numbers of independent clones are in brackets and types of Hprt mutations are shown in detail. (c) Examples of sequencing reads obtained for heterozygous mutations affecting the Dnmt1 gene. SNVs causing missense mutations G1157E or G1157R (top panel) and G1477R or affecting the splicing donor sequence on intron 36 (bottom panel; see also Supp. Fig. 2), were never detected in the same sequencing read, indicating that they locate to different alleles. (d) Distribution of suppressor gene mutations identified, including heterozygous deleterious mutations.

Resequencing Team at the Sanger provided data from sequencing of a strain from the 129S5 background[850]. While this helped to dramatically reduce the number of likely incorrect single nucleotide variants (SNVs), the number of small INDELs remained unreasonably high, especially since EMS is a DNA-alkylating agent mainly producing SNVs. While SNV detection can generally be very reliable, INDEL detection has been less accurate[851, 852]. In order to retain only high-confidence INDEL variants I supplemented the alignment-based variant calling, with Scalpel, an INDEL caller that uses micro-assembly to identify INDELs and supports "somatic" mutation detection, whereby the algorithm will only report variants found in the sample, but not the control[853]. INDELs that were not identified by both callers were discarded from the dataset. This allowed a drastic reduction of the number of likely incorrect variants in our dataset (Fig. 2.15, for a more detailed description of the workflow see Chapter 6.9.6). Analysis of the 7 suppressors identified 189 different mutations that were either missense mutations, nonsense mutations, frameshift variants, inframe insertions or mutations affecting splice sites (Fig. 2.13-d).

To evaluate candidates for suppressor mutations, genes that were mutated in more than one sample, ideally carrying different mutations, were identified. To further aid in the determination of causative suppressor mutations, PROVEAN and SIFT[815–819] mutation prediction tools were used to evaluate mutations. Taking all these methods into account, the most striking candidate for a suppressor gene was, interestingly, Hprt (Fig. 2.14-a). In four of the samples three different missense mutations and one nonsense mutation were identified, and a fifth sample (D3) carried a mutation affecting a splice donor site, which can also have severe consequences at the protein level. While Hprt is a known suppressor gene, this clearly shows that even without prior knowledge of the 6-TG mechanism of action we would have identified Hprt as a candidate gene for suppression and we would have been able to assign causative mutations in 5 out of seven cases. In addition to Hprt, inactivating mutations of genes encoding for mismatch repair (MMR) proteins Msh2, Msh6, Mlh1 and Pms2 are also known to confer resistance to 6-TG[854], as well as mutations in DNA methyltransferase Dnmt1[855], and in fact the two remaining clones from our initial analysis of 7 carried nonsense mutations in Msh6 and Pms2.

To analyze the frequency of these mutations in suppressors, the remaining 189 suppressor clones were subjected to targeted sequencing of known suppressor mutations (see Table B.1.3). Deleterious mutations in most of these genes were identified (Fig. 2.14-b), confirming that if we had carried out whole-exome sequencing, as for the first 7 clones, we would have identified Hprt, Msh2, Msh6, Mlh1 and Pms2 as strong suppressor candidates, confirming that this approach is feasible for other screens with little or no prior knowledge of suppres-

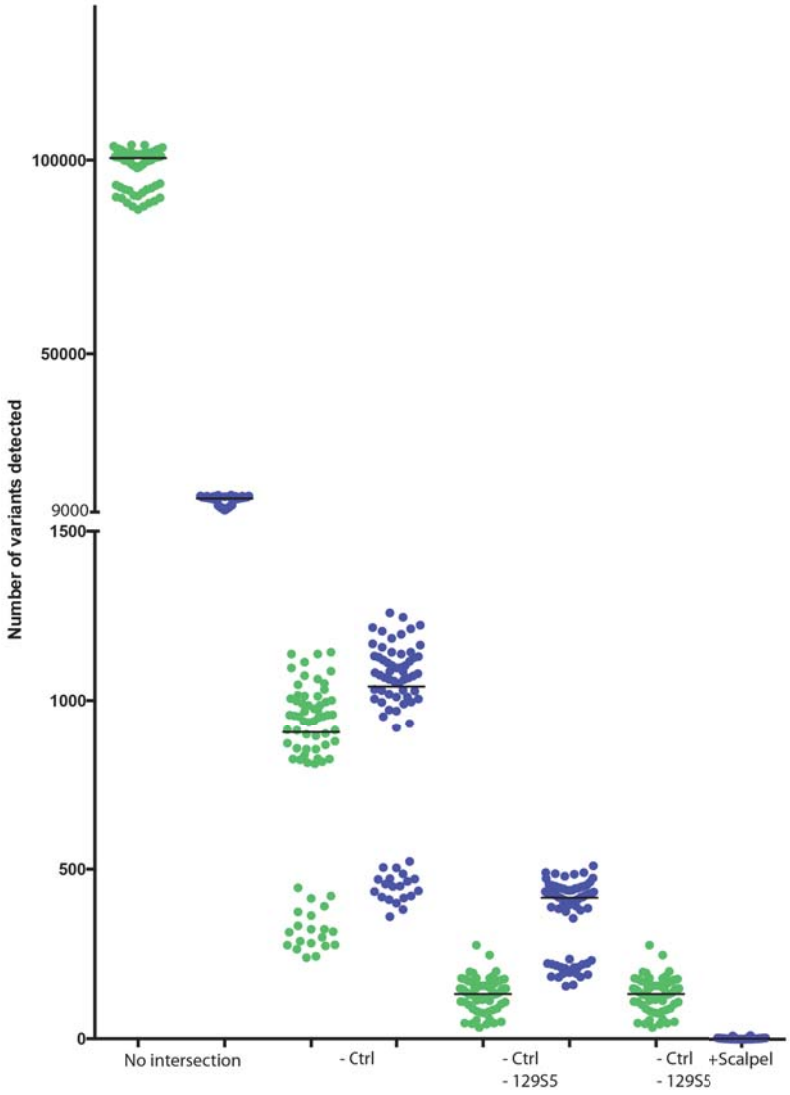
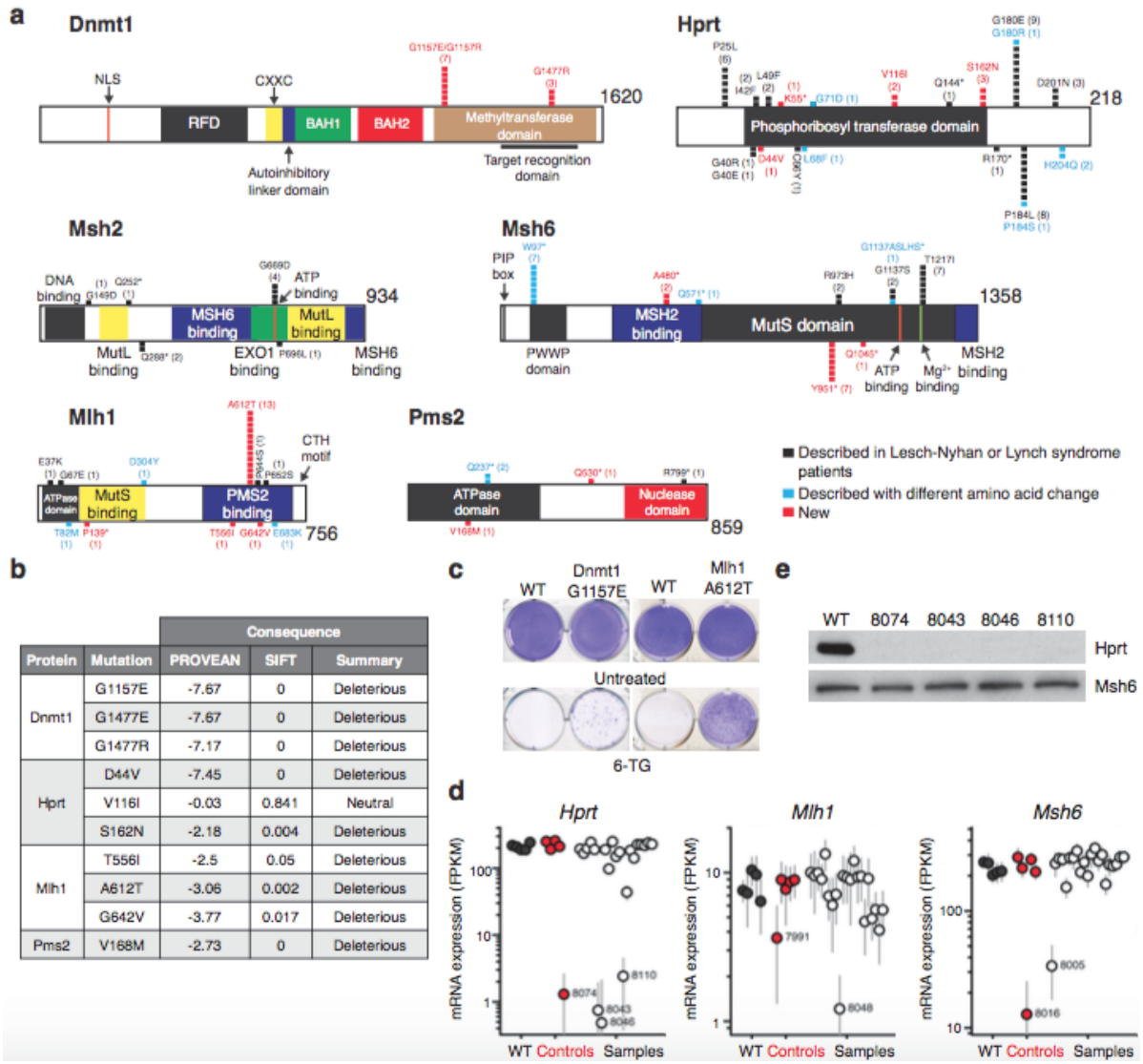


Figure 2.15: Using multiple controls and multiple variant callers to enrich for high confidence variants

Effects of using more than a sequenced control sample to clear samples of background mutations and using more than one INDEL caller to enrich for high confidence INDEL calls using 74 WES mouse samples. SNVs are labelled green, INDELs are labelled blue and median values are represented as horizontal lines. From left to right the data shows successive intersection steps. "No intersection": Number of variants after variant calling and filtering to remove low quality variants are shown. These variants are mostly differences between the 129S5 and the reference background. "- Ctrl": All variants also identified in an untreated mESC sample were removed from the samples. "- Ctrl - 129S5": Additionally, any variants identified in a 129S5 background strain sequenced at the Sanger Institute were removed [850] "- Ctrl - 129S5 + Scalpel": Since INDEL calling tends to be more error prone than SNV calling, we only included variants that were called by a second variant caller "Scalpel"[853].



Protein	Mutation	Consequence		
		PROVEAN	SIFT	Summary
Dnmt1	G1157E	-7.67	0	Deleterious
	G1477E	-7.67	0	Deleterious
	G1477R	-7.17	0	Deleterious
Hprt	D44V	-7.45	0	Deleterious
	V116I	-0.03	0.841	Neutral
	S162N	-2.18	0.004	Deleterious
Mlh1	T556I	-2.5	0.05	Deleterious
	A612T	-3.06	0.002	Deleterious
	G642V	-3.77	0.017	Deleterious
Pms2	V168M	-2.73	0	Deleterious

Figure 2.16: Clinically-relevant and newly-identified suppressor mutations
(a) Distribution of point mutations on Dnmt1, Hprt and MMR proteins; each square represents an independent clone. Asterisks (*) denote STOP-codon gains. (b) Predicted consequences of potential new suppressor mutations. Consequences were predicted as in Fig. 1e. (c) *De novo* introduction of new mutations Dnmt1 G1157E and Mlh1 A612T confers cellular resistance to 6-TG. (d) Hprt, Mlh1 and Msh6 mRNA expression levels (fragments per kilobase per million reads). Black dots indicate wild-type (WT) samples, red dots represent clones with already identified mutations (controls), and white dots represent samples for which no causative mutations were identified. Error bars represent uncertainties on expression estimates. (e) Reduced Hprt mRNA levels correspond to reduced protein production as detected by western blot.

sors. Intriguingly, a subset of clones presented heterozygous deleterious mutations in known suppressor genes. While these cells are sorted for haploid clones on a regular basis, diploid cells do remain and these particular cases could have arisen in the small diploid population or spontaneously after EMS treatment in a diploidized cell. Regardless, in order to be true suppressors these clones would each have to carry heterozygous mutations affecting both alleles of the gene, resulting in homozygous loss of the protein function. While our sequencing data is not phased, we have identified examples, where mutations occurred in such a way that they could be covered by a read (they are less than 150bp apart) or by the different members of a pair. Examples are shown in Fig. 2.14-c which demonstrate that these heterozygous mutations do not co-occur in the same reads indicating that, indeed, these cases are compound heterozygotes. Their scores in PROVEAN and SIFT predictions indicated that they are likely causing the 6-TG sensitivity suppression. When the clones carrying heterozygous mutations were also taken into account, we could also include Dnmt1 in the list of identified suppressor genes (Fig. 2.14-d).

When searching the literature, Dr. Josep Forment was able to assign many of the missense and nonsense variants to clinically-relevant mutations in Hprt (causing Lesch-Nyhan syndrome and its variants[856]) and DNA MMR (linked to Lynch Syndrome[803, 804])(Fig. 2.16-a), as well as previously not identified variants that are predicted deleterious(Fig. 2.16-b), highlighting the ability of this method to identify critical regions of a protein. Mutations affecting splicing donor and acceptor residues were also identified and confirmed by Dr. Josep Forment to reduce total protein level. To test whether some of the newly identified mutations are as deleterious as predicted, Dr. Josep Forment introduced the A612T and G1157E mutations in Mlh1 and Dnmt1 (which I identified as heterozygous mutations), respectively, into wild-type mESCs as homozygous mutations by CRISPR/Cas9 gene editing and showed that cells carrying these mutations were resistant to 6-TG treatment (Fig. 2.16-c).

For a small group of clones, no mutation in the targeted genes could be identified (Fig. 2.14-a,c) and we subjected the clones to whole-exome DNA sequencing and RNA sequencing (and included some in which we were able to identify potential causative mutations as controls). This allowed the production of an unprecedented description of EMS mutagenic preferences on the whole exome level, confirming its preference for producing SNVs, especially C:G>T:A transitions (Fig. S2.17-a,b,c), which could explain the high number of mutants affecting splice sites we recovered. While I was able to successfully retrieve previously identified mutations in the control samples, the DNA sequencing data identified no other obvious gene candidate. However, the RNA sequencing analysis carried out by Dr. Tomasz Konopka revealed significant reductions in expression levels of Hprt, Msh6 or Mlh1 in several clones

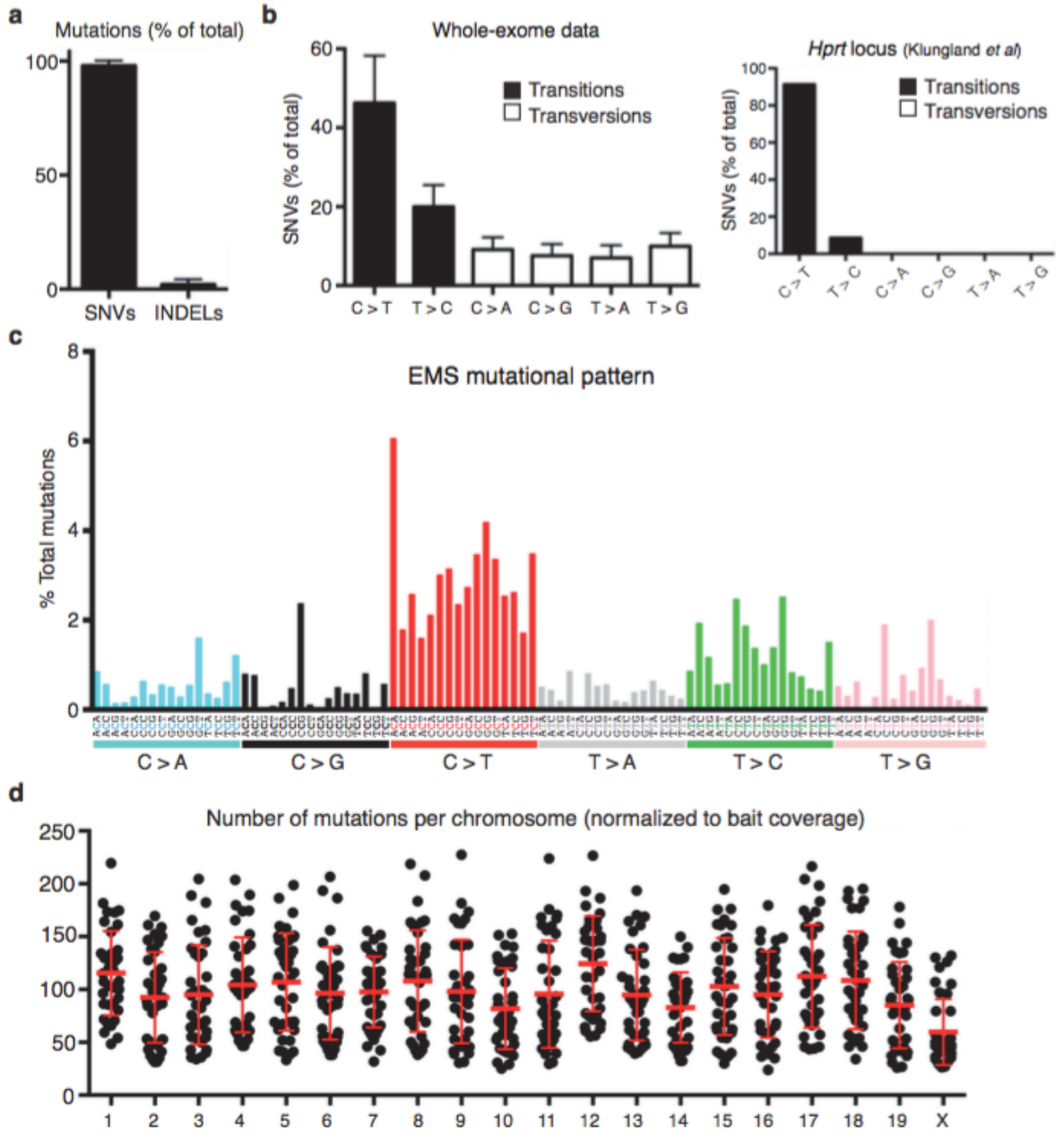


Figure 2.17: EMS mutagenic action

(a) Distribution of mutation types identified by whole-exome sequencing of 66 suppressor clones. SNV, single-nucleotide variant. INDEL, insertion or deletion. Only homozygous mutations were considered. (b) Distribution of identified SNVs. (c) EMS mutational pattern. (d) Number of mutations per chromosome in sequenced clones. Mutation numbers (both homozygous and heterozygous) were normalized to exon bait coverage.

(Fig. 2.16-d), which could explain the 6-TG resistance of these samples. Further work may help to elucidate whether in such clones epigenetic alterations or mutations in regulatory regions not covered by exome-sequencing could explain the suppression mechanism in these clones.

Taken together, my work with Dr. Fabio Puddu and Dr. Josep Forment has shown, not only that we can exploit next-generation sequencing to unravel complex genetic interactions in haploid *S. cerevisiae* and mouse cells, with the potential to extend to human cells and essential gene biology, but also that our bioinformatical analysis is robust and recovers SNVs with high fidelity. Moreover, by using more than one variant caller strategy we can efficiently reduce INDEL false positive levels.

2.4.5 Establishing a sequencing analysis protocols for large genomic changes

We have established that this analysis can identify SNVs and small INDELs with a satisfactory sensitivity and accuracy. While polymerases with a low fidelity are not known for causing large-scale genomic rearrangements, a comprehensive genome analysis will address such changes. A structural variant(SV) is any form of rearrangement in chromosome structure and includes any or a combination of translocations, inversions, copy number variation (CNVs) as well as large insertions and deletions. These changes are critical as contributors to genetic diversity and evolution, but are also frequently involved in disease (see 1.2.1). Several methods exist to detect SVs such as microscopy-based chromosome banding and fluorescence *in situ* hybridisation (FISH), pulse-field gel electrophoresis, microarrays and sequencing-based mate-pair sequencing (sequencing the ends of large, kilobase-long DNA fragments) and whole-genome sequencing(WGS). Next-generation sequencing can detect many SVs by analysis of the mate pairs: for instance, in the event of a translocations the two mates of a pair (which by definition originated from the same DNA fragment) will align to different chromosomes of the reference genome and in the case of insertions or deletions the mate pairs will be much closer or further apart, respectively, than the average insert size dictates (Fig. 2.18). Since read pairs are a key source of evidence for SV detection, the quality of the underlying sequencing library is key and routine quality control (QC) of measures such as insert size is required. A second line of evidence for SVs can be split reads, reads that span a breakpoint and thus only align to parts of the reference in a continuous manner, and this depends highly on the alignment program and its ability to process split reads. A third source of evidence for SVs, especially CNVs is the read depth, following the assumption that an increase in copy number will be accompanied by a roughly proportional increase in coverage. There is a plethora of available

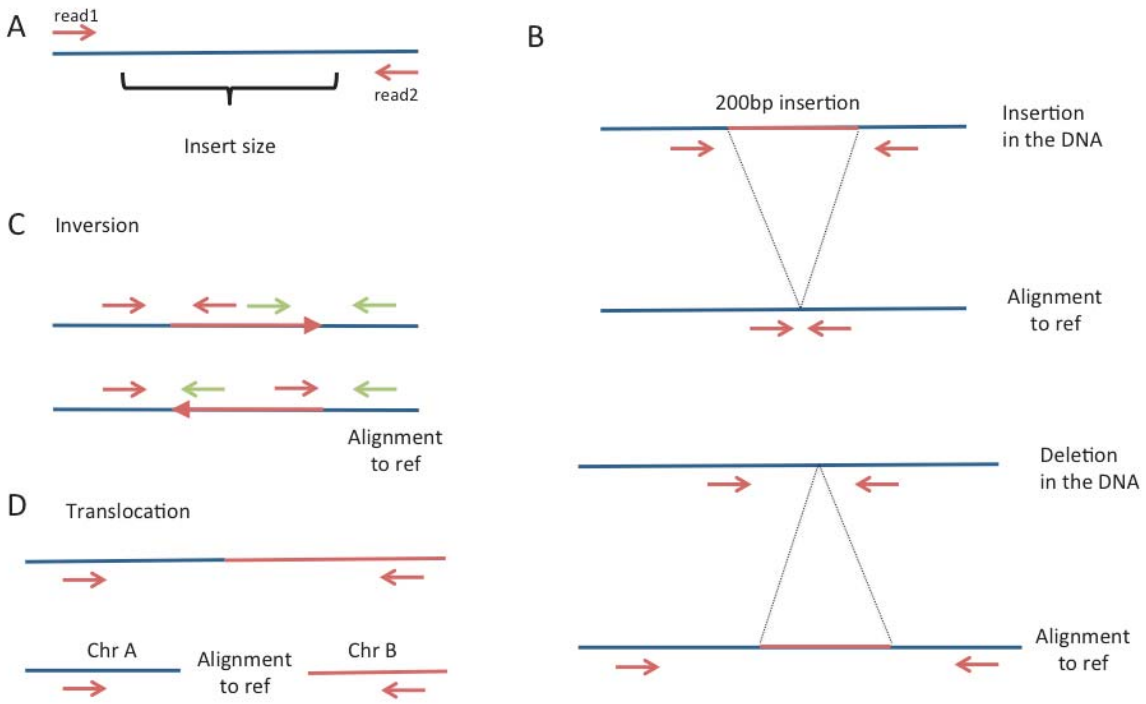


Figure 2.18: Relationship between read pairs and structural variants
A | Schematic of Illumina paired-end sequencing: a fragment of DNA is sequenced from both sides inwards for 150bp (may vary depending on sequencing machine) leaving a fragment in the middle unsequenced. Its size depends on the library prep, but should be similar for all DNA fragments in the library. **B |** Large insertions and deletions: large insertions and deletions will be visible in the sequencing by alterations in the distance between the paired reads. **C |** An inversions most striking effect on a pair of reads is that they will now both be aligning to the forward or the reverse strand. **D |** In a translocation event members of a read pair may now align to different chromosomes.

SV callers that use one of those evidence sources (e.g. BreakDancer[857] uses read-pair information) or a combination (e.g. Lumpy[858] uses read pair, depth and split read information). Since SV callers can usually not detect the full spectrum of SVs and each one has advantages and limitations, Dr. Kim Wong in David Adams' lab developed SVMerge a meta SV calling pipeline[859], which uses a variety of callers to make SV predictions (Fig. 2.19).

To complement the use of a program like this, we wanted to be able to visualise aneuploidy and large copy number changes in budding yeast WGS data. To this end, Dr. Puddu and I wrote compact scripts, that extract positional genome coverage data from bam files. The coverage values are normalised to the whole-genome median and ploidy information given by user input. As a control, this tool was used to visualise aneuploidy in a haploid strain that is diploid for Chromosome IX (Fig. 2.20).

2.4.6 Analysing repetitive DNA regions in the yeast genome

One of the biggest technical challenges facing NGS analysis are repetitive DNA sequences, sequences that are similar or often identical to other regions of the genome. That is especially problematic, because most genomes are abundant in repetitive sequences: about half of the human genome and >80% of the maize genome are covered by repeats[861]. From a computational point of view, repeats create uncertainty in alignments (as well as *de novo* assembly, which will not be further discussed), which can lead to errors when analysing sequences for genome variation. The main computational challenges are due to repeats that are >97% identical across more than one copy and that are longer than typical NGS read length (typically longer than 100-200bp).

After alignment of deep sequencing data, one major challenge remains: how to deal with reads that align to more than one location (multi-reads). In the human genome, the number of short reads (25bp or longer) that can be uniquely mapped tends to be around 70-80% even though the repeat content of the human genome is about 50%[860]. This level of accuracy can be achieved due to the fact that repeats are often non-identical and many reads will have a unique "best match" (Fig. 2.21). "Best match" alignments are a simple way to resolve a significant portion of reads but this is not always correct[860]. Structural and copy number variant detection in unique regions has become relatively reliable, but the short read length of NGS sequencing data prevents similarly accurate detection in repetitive regions[860]. The most reliable sources for SV, coverage and read-pairs, pose more of a challenge in repetitive regions. Suppose an example of two transposable elements (TE), one on chromosome II another on chromosome V. Reads from either will relative equally be distributed between both

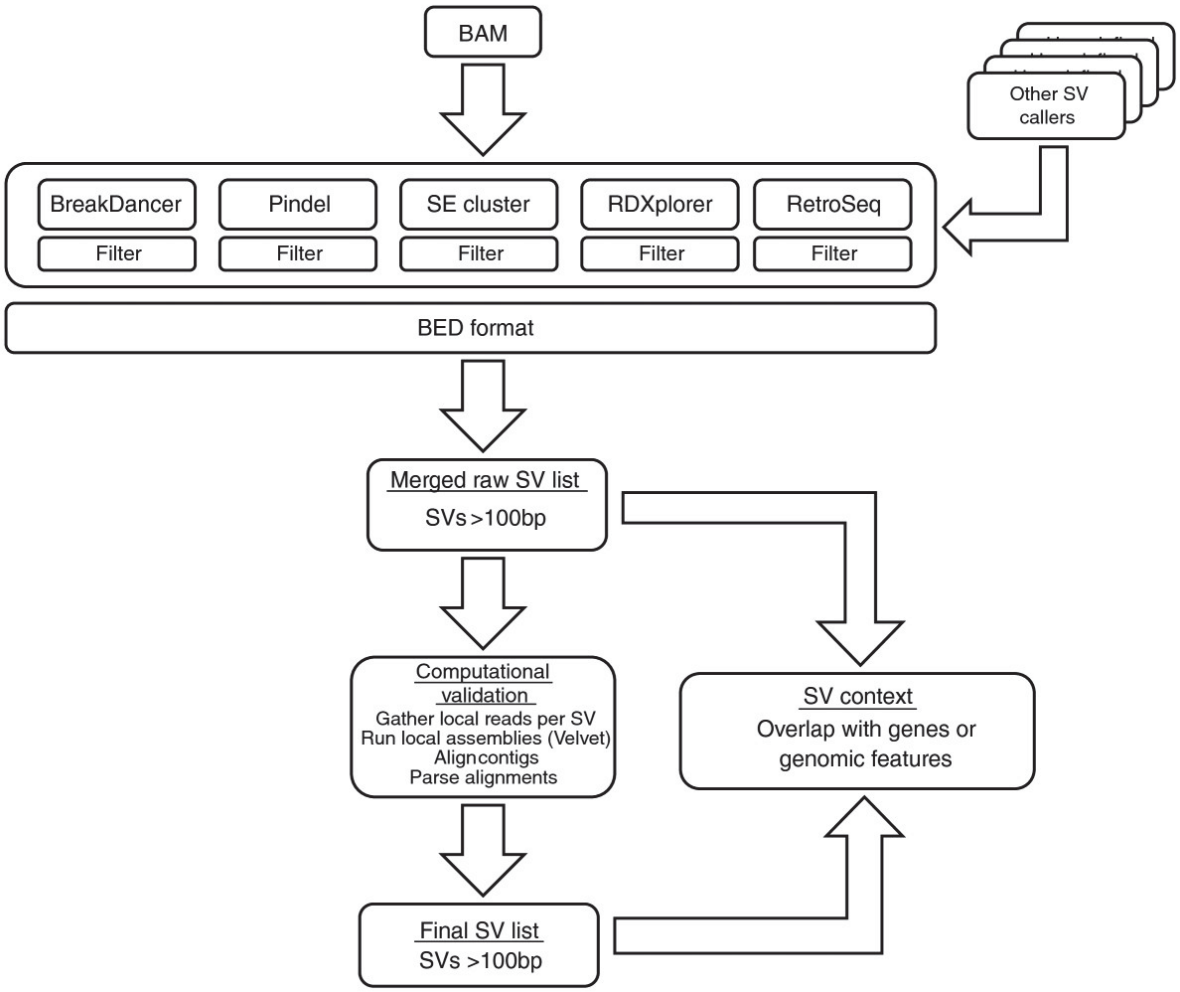


Figure 2.19: An overview of the SVMerge pipeline

"SVMerge uses a suite of software tools to detect structural variants (SVs) from mapped reads. The calls are filtered, merged and then validated computationally by local *de novo* assembly. The output is in BED format, allowing for easy downstream analysis or viewing in a genome browser. The SVMerge pipeline is extendable so that calls made by other software can be included in the downstream analysis. BAM, Binary Alignment/Map format." Figure and text reproduced from [859] in accordance with the publisher's terms of use.

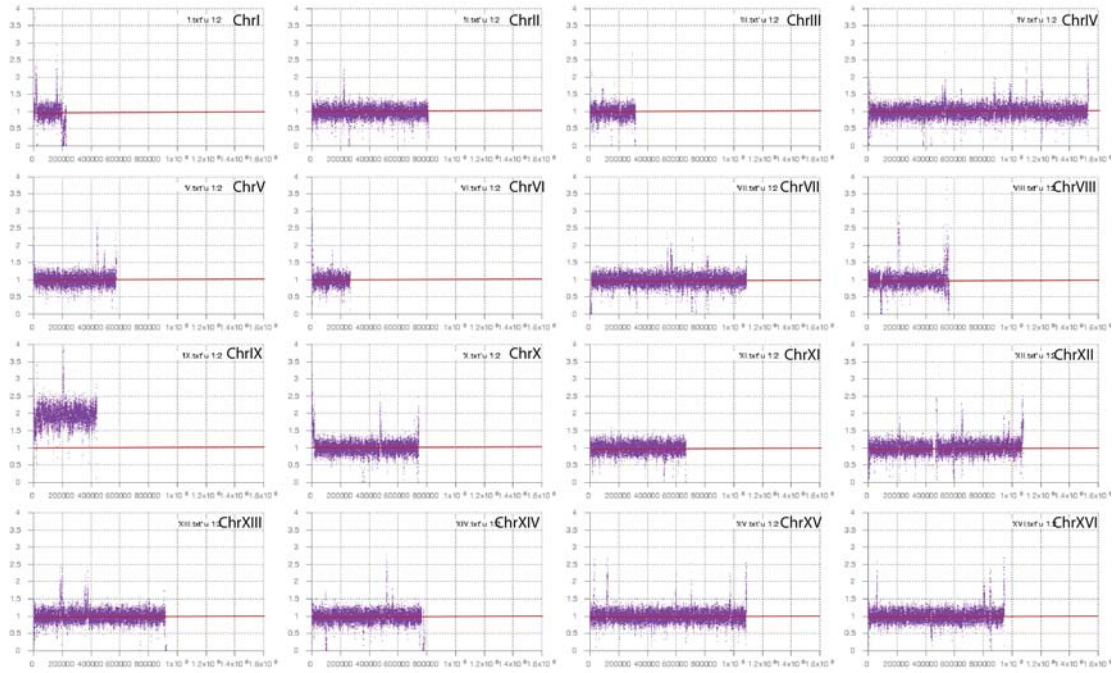


Figure 2.20: Visualising aneuploidy in budding yeast

Output of script to visualise aneuploidy: normalised coverage plotted by position for each of *S. cerevisiae's* 16 chromosomes. Here the DNA content of a haploid strain diploid for Chromosome IX is shown.

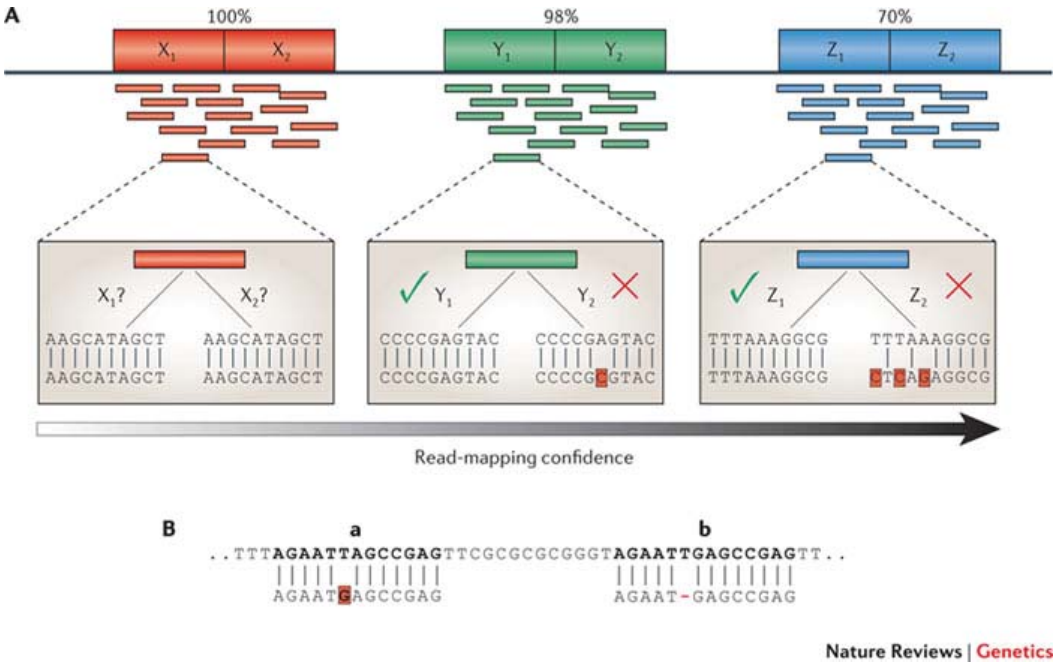


Figure 2.21: Ambiguities in read mapping

”**A** | Read-mapping confidence versus repeat-copy similarity. As the similarity between two copies of a repeat increases, the confidence in any read placement within the repeat decreases. At the top of the figure, we show three different tandem repeats with two copies each. Directly beneath these tandem repeats are reads that are sequenced from these regions. For each tandem repeat, we have highlighted and zoomed in on a single read. Starting with the leftmost read (red) from tandem repeat X, we have low confidence when mapping this read within the tandem repeat, because it aligns equally well to both X1 and X2. In the middle example (tandem repeat Y, green), we have a higher confidence in the mapping owing to a single nucleotide difference, making the alignment to Y1 slightly better than Y2. In the rightmost example, the blue read that is sequenced from tandem repeat Z aligns perfectly to Z1, whereas its alignment to Z2 contains three mismatches, giving us a high confidence when mapping the read to Z1. **B** | Ambiguity in read mapping. The 13 bp read shown along the bottom maps to two locations, a and b, where there is a mismatch at location a and a deletion at b. If mismatches are considered to be less costly, then the alignment program will put the read in location a. However, the source DNA might have a true deletion in location b, meaning that the true position of the read is b.” Figure and Text reproduced from [860] with permission from the publisher.

Species	Copy number
Saccharomyces cerevisiae	150
Caenorhabditis elegans	55
Drosophila melanogaster	~240
Xenopus laevis	~600
Gallus domesticus (chicken)	200
Mus musculus (mouse)	100
Homo sapiens	350
Arabidopsis thaliana	570
Pisum sativum (pea)	3,900
Triticum aestivum (wheat)	6,350

Table 2.6: Haploid copy number of rDNA repeats across Eukaryotic species
Selection of rDNA repeats observed in different eukaryotic species [865].

when aligned. There are cases when the aligner will distribute members of the same pair on different chromosomes when the mapping quality is 0, suggesting a translocation where there is none. Also, suppose another example of these two TEs: the genome was sequenced to a mean depth of 30x and the two TEs show a coverage of about 60x. One may suppose that this means instead of two, this sample contains four copies of the TE. However, the coverage varies considerably across the genome, making the distinction between N and N+1 a low confidence proposition[860]. To cope with multi-reads (those with a reported mapping quality of 0) some prefer to discard them with unmapped read pairs and many SV detection programs ignore them in their analysis (though some allow the manual setting of mapping quality thresholds).

The budding yeast *S. cerevisiae* contains three major repetitive regions: ribosomal DNA (rDNA), Ty retrotransposons and telomeres. The rDNA genes encode ribosomal RNAs, major components of ribosomes, and rRNA makes up about 80% of RNA in budding yeast cells[862]. To cope with the high biosynthetic demand, eukaryotic cells tend to have hundreds of rDNA copies organised into clusters. In budding yeast, they exist in a single cluster located on chromosome XII (accounting for almost 2/3 of the chromosome's length and 10% of the entire genome)[863]. Their highly repetitive nature makes the rDNA locus a highly fragile region of the genome and copies are continuously lost for example due to recombination events[862]. However, under normal conditions, cells can maintain a characteristic number of repeats and counteract loss by gene amplification (Table 2.6; see [864] for a review).

The maintenance of rDNA clusters involves many factors that are required generally for genome maintenance (such as replication, DNA repair and chromatin dynamics) and the de-

mand the rDNA cluster places on these factors means that perturbations in rDNA stability and copy number affect the availability of these factors in other regions of the genome[862]. Additionally, rDNA instability has been linked to aging in budding yeast[862] and a reduced copy number of rDNA repeats was shown to increase sensitivity to DNA damage[863]. Apparently, cells require a copy number of rDNA genes in excess of transcriptional demand to allow for DNA repair to proceed effectively[866]. In low-rDNA-copy-number cells the locus reportedly shows more genetic instability and this instability extends to other parts of the genome[862]. While the exact contributions and mechanism of the rDNA locus and its effects on genome instability and aging are still under active investigation, it is clear that reductions in rDNA copy number are detrimental to genomic stability and the cell as a whole and should be assessed when assessing effects of polymerase mutations on genome stability.

In *Saccharomyces cerevisiae*, the rDNA locus on Chromosome XII consists of approximately 150 repeats of a 9.1kbp unit (Fig. 2.22-A)[860] and when one plots the Illumina sequencing coverage along the chromosome the locus can be clearly identified as a sharp peak(Fig. 2.22-B). The rDNA unit contains genes for the 5S rRNA and the 35S rRNA, which are separated by two intergenic spacers (IGS1, 2). IGS2 contains the rARS, an origin of replication and IGS1 contains EXP, an expansion sequence made up of the replication fork barrier (RFB) and E-pro, a bi-directional promoter for non-coding RNAs that functions in regulating the rDNA repeat number[860]. The RFB ensures the unidirectionality of replication forks by the association with the protein Fob1[867], preventing head-on collisions between the replication and the transcription machinery in this highly transcribed region[867–869]. In the *S. cerevisiae* S288c reference genome assembly (R64-1-1/EF4) contains two copies of the 9.1kb rDNA repeat unit separated by the IGS1 to indicate the repetitive nature of the rDNA locus(Fig. 2.22-C). To estimate the amount of rDNA repeats present Dr. Fabio Puddu and I wrote a script, that measures the sequencing coverage in the first of the two rDNA unit copies and compares it to coverage upstream of the locus (Fig. 2.22-C). The upstream region was chosen because four copies of the 5Svariant, four copies of the *ASP3*, and a transposon are located downstream of the rDNA locus. In biology, the efficacy of any measurement method depends on two things: (1) when measuring the same sample more than once, does it give the same answer (technical reproducibility) and (2) how accurately does it measure the thing it purports to measure (how does it compare to other widely used measurement methods)? To answer the first question, Dr. Fabio Puddu and I sequenced 116 yeast strains twice starting from the same genomic DNA. The results show a strong correlation between the two independent measures (Fig. 2.22-D), with the divergence between the two measures increasing with the size of the rDNA locus but remaining almost always contained within +/-5% of the

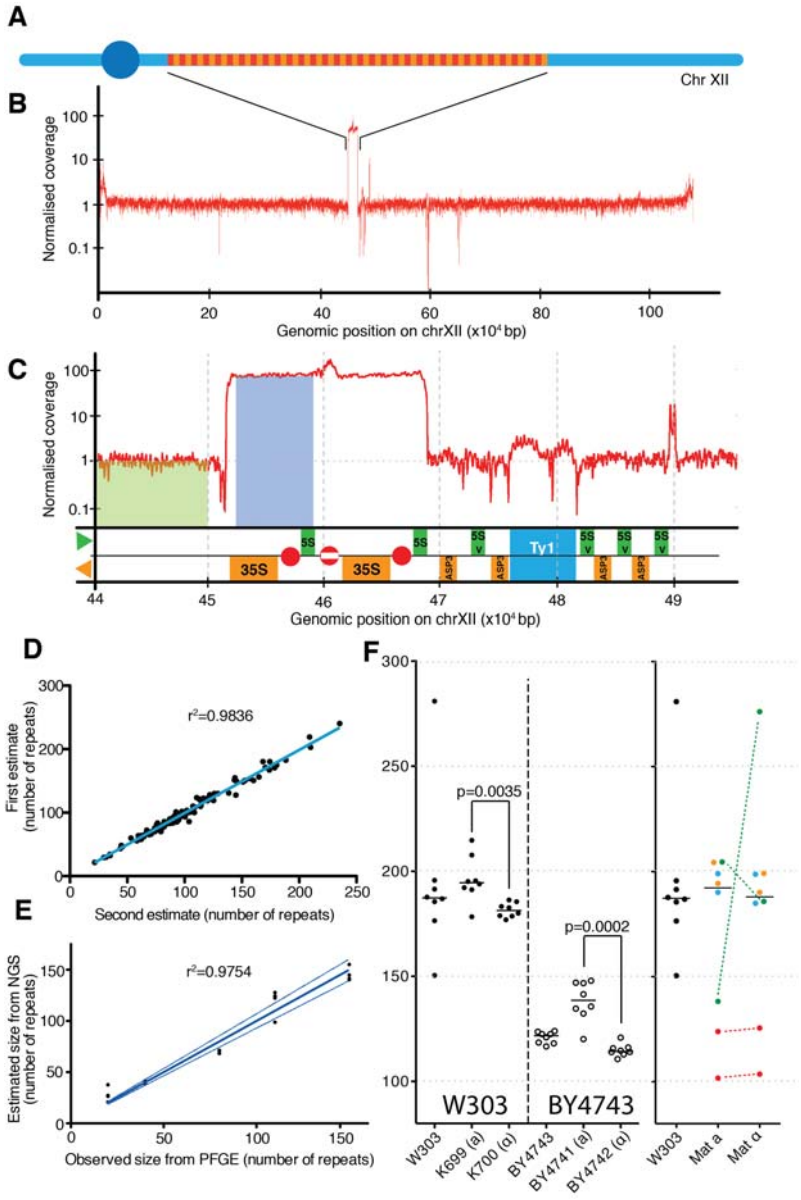


Figure 2.22: Next-generation sequencing data can be used to estimate rDNA copy number reliably

A | rDNA exist in ~150 repeats of a 9.1 Kbp unit on Chromosome XII. **B** | The coverage of Illumina sequencing reads across Chromosome XII. **C** | The *S. cerevisiae* S288c assembly (R64-1-1/EF4) contains two copies of the rDNA unit with an origin of replication each (red circle) separated by one copy of the replication fork barrier (RFB, red circle with white bar). Measurements of rDNA enrichment are derived from coverage over rDNA repeats (blue box) over the average genomic coverage. **D** | Reproducibility of the measurement using the same sample. **E** | Accuracy of the measurement: strains stable for their rDNA copy number with estimates of rDNA copy number by pulse-field gel electrophoresis (20, 40,60,110 and 150 repeats)[866] had their rDNA copy number estimated using NGS coverage. **F** | rDNA copy number measured in wild-type W303 and BY4743 laboratory strains that are diploid, MATa and MATα. The W303 diploid strain was sporulated, four tetrads (biological replicates) were dissected and the resulting haploid cells sequenced and assessed for rDNA copy number (fall four spores of the same tetrad are labelled in the same colour).

average of the two measures. To answer the second question and assess the precision of this method four colonies each derived from 5 yeast strains carrying stable rDNA loci of known length were sequenced[866]. The rDNA copy number was estimated in these strains by Ide et al. using pulsed-field gel electrophoresis (PFGE) to be 20, 40,60,110 and 150 repeats, respectively. Fig. 2.22-E shows that estimating rDNA loci size using whole-genome sequencing produces results in agreement with PFGE and considering that PFGE also produces estimates at best, WGS estimates of rDNA copy number perform just as accurately. By employing this method, I found that wild-type laboratory strains in the W303 background have consistently bigger rDNA loci (~180 units) compared to wild-type strains in the BY4743 background (~120 units) (Fig. 2.22-G, left). We also observed that in both backgrounds, haploid strains of the mating type a, seemed to have slightly bigger loci than the corresponding Matα strains (Fig. 2.22-G, left). To determine if this is generally true, we sporulated a wild-type W303 strain and analysed the rDNA length in the progeny. In these conditions, Mat a and Matα strains did not show any significant difference between each other, but they showed a greater variability in rDNA length. When the four spores coming from a single meiotic event (marked with the same color in Fig. 2.22-G, right) show rDNA loci of different size, the data observed are compatible with Mendelian inheritance of this trait, in the presence or in the absence of unequal sister chromatid exchange. In sum, this tool can be used as a read-out in a screen identifying genes that regulate and maintain rDNA copy number.

Beyond the rDNA locus we have investigated the other two repetitive DNA regions in the yeast genome: Ty elements and telomeres. Ty elements are retrotransposons pervasive in the yeast genome (3.1% of the genome) characterised by their flanking long terminal repeats (LTRs). There are five distinct retrotransposon families (Ty1–Ty5)[870]. Their success at colonising the yeast genome varies greatly and while the numbers observed can vary greatly, the consensus is that Ty1 elements occur most and Ty5 elements least often[870–872]. Considering these fluctuations and that, in principle, Ty elements are very similar to the rDNA locus in that their length greatly exceeds read length, we extended our approach to measuring Ty element copy number. Because, unlike rDNA, Ty elements are spread throughout the genome, a custom "Ty reference genome", a fasta file principally made up of the Ty element sequences and surrounding control sequences, was constructed and NGS reads aligned to it. This redistributes Ty element reads back onto a single locus allowing estimates of Ty element number within the genome (Fig. 2.23). In principle this approach works, but as of yet we have not completed sequencing of strains to show that our approach accurately determines Ty element number (akin to Fig. 2.22-E), but the general trend of Ty elements reported in the literature is reflected in our measurements. Telomeres provide a greater challenge to copy

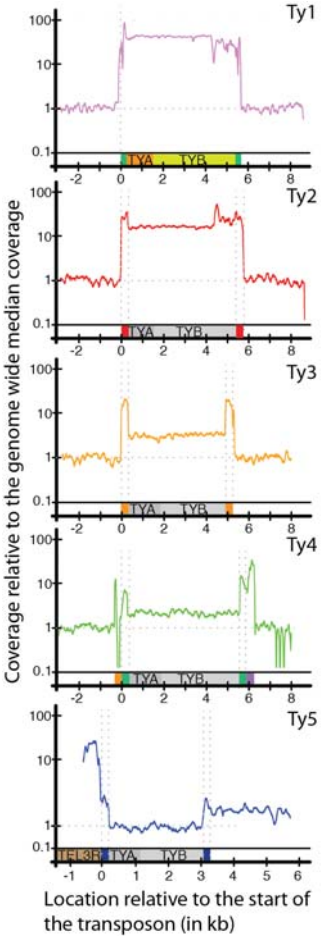


Figure 2.23: Next-generation sequencing data could also be used to assess Ty element copy number

To estimate copy number of Ty elements a custom "reference genome" was built mainly consisting of a single copy of each Ty element and control surrounding sequences. NGS data was aligned to this references and coverage plots normalised to the genomic median are shown.

number estimation. Their repeat size is smaller than a single read and, thus, requires another approach for telomere length estimates. While a program exists for the estimation of human telomere length from NGS data[873], it relies entirely on the fact that the human repeat is invariable (a TTAGGG tandem repeat). In contrast, the *S. cerevisiae* telomere repeat is degenerate with the consensus sequence $G_{2–3}(TG)_{1–6}$[874]. This poses a great challenge to budding yeast telomere length measurements using next-generation sequences. Together with Zhihao Ding, I have been trying to adjust his program to measure *S. cerevisiae* telomere repeat number, but so far we are still underestimating yeast telomere length, likely because we don't capture the degenerate nature adequately yet.

In summary, together with Dr. Puddu, I developed a simple program to measure rDNA repeat number in the budding yeast *S. cerevisiae* and we can show that our method is a suitable alternative to classic laboratory approaches. We are now employing this method as a tool in our array of methods to document genomic changes, but are also using it to identify factors involved in rDNA copy number maintenance (see Chapter 4). Work to accurately estimate Ty element number and telomere length in budding yeast is ongoing.

2.5 Summary

During this phase of my work, I compiled a list of mutations in DNA polymerases delta and epsilon identified in sequencing of human cancer samples. After assessment of the occurrence of these in the wider population, the corresponding residues in the budding yeast *S. cerevisiae's* replicative polymerases were identified and those that affect residues that are evolutionarily conserved were retained. These mutations were then introduced into yeast cells and mutation accumulation experiments performed where cells were propagated to let any effects of polymerase mutations manifest in the genome. DNA was extracted at the beginning and end of these experiments and sent for whole-genome sequencing. In the meantime, I developed and tested a sequencing analysis strategy using existing datasets that had the advantages of positive controls and follow-up validation. This allowed the development of accurate protocols and tools for identifying SNVs, small INDELs, changes in rDNA repeat number and to a lesser extent structural variants. In the process, I contributed to projects unraveling complex genetic interactions in budding yeast and the proof-of-concept application of our yeast synthetic viability screens to mouse genetics.

Evaluation of hypotheses

Aims:

- To compile a list of relevant mutations in DNA polymerases identified in cancer samples

  A list of DNA polymerase mutations found in colorectal and endometrial cancers was assembled from the literature.

- To prioritise mutations in DNA polymerases and determine their *Saccharomyces cerevisiae* equivalents

  Recurrence in cancer sample, bioinformatic predictions and the alignment of the human and yeast protein sequences identfied a list of mutations with priority for the variants POLE S297F, POLE P286R and POLE V411L, which were tested for effects such as mutation rate increases first.

- To conduct mutation accumulation experiments to identify the consequences of DNA polymerase mutations on a genome wide scale

  After construction of all remaining mutations in budding yeast, they were subjected to mutation accumulation experiments for three months in several parallel lines. Starting and final yeast colonies were sent for whole-genome sequencing to identify acquired mutations and characterize any changes in numbers, locations and patterns compared to wild-type.

- To establish sequence analysis protocols for budding yeast whole-genome sequencing data

  Whole-genome sequencing data analysis in budding yeast was developed for single nucleotide variants, insertions/deletions, aneuploidy and copy number changes in repetitive regions. Determinations of false negative and false positive rates were estimates and measuring rDNA repeat copy number was validated with published southern blot data.

- To show that these sequence analysis protocols are functional and can be applied beyond this project

  My whole-genome sequencing analysis protocols were applied to suppressor screens in budding yeast and identified both expected mutations (for instance mutations in TOP1 as suppressors for camptothecin sensitivity), which act as a positive control, and previously unknown mutations, which were shown to be biologically relevant by further

experiments. Taking this work successfully into suppressor screens with haploid mouse embryonic stem cells shows that overall this analysis protocol is robust, produces validated results and can be used for applications beyond its initial conception.