Chapter 3

Analysis of populations of *S. cerevisiae* strains carrying simple polymerase mutations

Overarching hypothesis

DNA polymerase mutations can contribute to tumour progression likely by elevating the mutation rate.

Aims:

- To assess all candidate DNA polymerase mutations for the number of mutations acquired in the same amount of time
- To identify the type of mutations caused by mutated DNA polymerases
- To determine whether candidate mutations leave a distinct mutation pattern on the genome
- To compare mutations acquired in mutated polymerase strains to those resulting from mismatch repair deficiency

3.1 Introduction

In the previous chapter, a list of mutations in DNA polymerases identified in colorectal and endometrial cancer was collated and after alignment to the yeast proteins, where possible, those

mutations were introduced into diploid *Saccharomyces cerevisiae* as heterozygous mutations. Determining the effect of these mutations on a genome - for instance to identify mutations in DNA polymerases that raise the mutation rate - will assist in differentiating between passenger mutations and those that promote tumourigenesis. To this end, strains carrying these mutations were subjected to mutation accumulation experiments and whole-genome sequenced before and after the experiment to collect information about acquired mutations in this timeframe. Analysis protocols for budding yeast whole-genome sequencing data were developed and tested and then used to analyse mutation accumulation experiment data.

3.2 Increased mutation rates for strains heterozygous diploid: *pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K* and *pol3-S483N*

3.2.1 Increased number of single-nucleotide variants for a subset of polymerase variants

After propagation, the polymerase mutant strains were sequenced at the Wellcome Trust Sanger Institute (see Chapter 6.8) and the sequencing data aligned to the yeast reference genome. Variant calling for single-nucleotide variants and small insertions and deletions was performed to detect any changes in mutation accrual. First, samples were checked for their polymerase genotype: any sample not identified to carry the expected polymerase mutation was discarded from the dataset. While it is possible, that a missing polymerase mutation is a case of a false negative, these samples were discarded in case of contamination or early reversion of the mutation. The remaining samples were analysed as described in Chapter 1.4.3: filtered samples were intersected with the initial starting strains to remove any background mutations and only retain those mutations acquired during the experiment (for a detailed description of the analysis workflow, the software, scripts, and commands used see Chapter 6.9.5 and 6.9.6).

Strains not carrying a mutation in the DNA polymerases acquired a mean of 6.8 single-nucleotide variants (SNVs) during the course of the experiment translating to roughly 5.5×10^{-10} SNV mutations per generation per base (assuming 500 generations), which is consistent with recently published mutation rates in vegetative diploid *S. cerevisiae*[875]. The exonuclease deficient *pol2-4* mutants acquired on average 11.8 SNVs, meaning 1.7× the number of mutations as the wild-type(Fig. 3.1). Of the *pol2* candidate mutations four showed significant

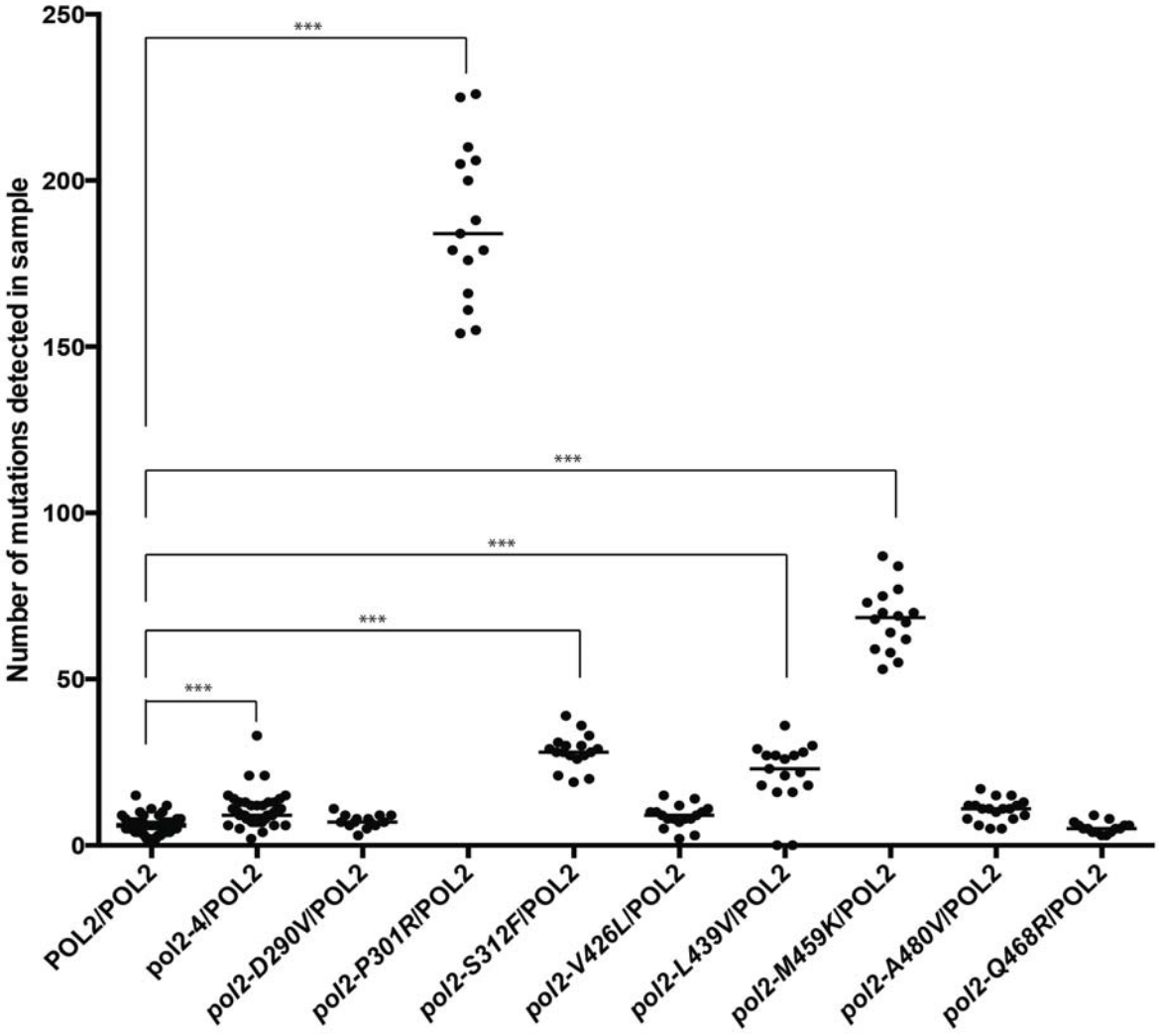

Figure 3.1: Number of single-nucleotide variants per sample in *pol2* mutant strains
Pol2 mutant strains were whole-genome sequenced before and after a three-month propagation on rich medium. The number of acquired single-nucleotide variants was determined for each parallel line that was propagated. The number of samples for each strain are as follows: n=65 (*POL2*), n=49 (*pol2-4*), n=18 (*pol2-A480V*), n=17 (*pol2-S312F, pol2-V426L, pol2-L439V*), n=16 (*pol2-M459K*), n=15 (*pol2-P301R, pol2-Q468R*), n=14 (*pol2-D290V*). All strains are heterozygous diploid for the mutation in question. The median is denoted by a black line. Student's T-test was used to determine which samples are significantly different from wild-type at *** $p < 0.001$.

increases in mutation accrual: *pol2-P301R* (27.9×), *pol2-S312F* (4.3×), *pol2-L439V* (3.4×) and *pol2-M459K* (10.3×). Strikingly, their increase in mutation number exceeds that observed for the exonuclease deficient strain. Also, interestingly, while *POLE p.Val411Leu* is one of the most frequently observed mutations in *POLE* in sequenced cancer samples (Fig. 2.3), the equivalent budding yeast variant, *pol2-V426L,* does not lead to an increase in SNV accumulation. Whether this holds true for the human mutation remains unclear.

In the case of *pol3* mutants, the exonuclease deficient strain also shows increased mutation accumulation when compared to wild-type with an average of 22.3 SNVs per strain (3.3×, Fig. 3.2). Of the *pol3* candidate mutations tested, one, the *pol3-S483N* strain, accumulated a mean of 230 SNVs per strain, meaning it accumulated 33.3× the number of SNVs as the wild-type strains. Again, this is a striking increase compared to the mutational increase observed in exonuclease deficient cells. Why these mutations in the exonuclease domain produce an effect stronger than mutating the catalytic residues of this domain is currently unclear.

3.2.2 Single-nucleotide variants in haploid polymerase mutant strains

In the heterozygous diploid strains, the effects of the polymerase mutations are likely mitigated by the presence of a wild-type copy of the polymerase on the other chromosome. In a haploid setting, only the mutated polymerase would be present and the genome would be half the size. Theoretically, if the polymerases - the mutated and wild-type one - are available in cells at similar levels and share the burden of copying the genome equitably in the heterozygous strains, then in the haploid strain the mutant polymerase would copy roughly the same amount of DNA each division. To examine mutation numbers in haploid strains, four of the *pol2* mutant strains - *pol2-P301R, pol2-S312F, pol2-A480V* and *pol2-M459K* - were propagated as haploids alongside the wild-type and exonuclease deficient strain for 13 passages using single colony bottlenecks as described in Chapter 2.3.2.1.

Fig. 3.3 depicts the SNVs per haploid genome accumulated in each line after the propagation for the haploid and heterozygous diploid strains. While the number of SNVs accumulated in wild type strains is fairly similar between haploid and heterozygous diploid strains, a difference can be seen for all *pol2* mutant strains (see Fig. 3.3 and Table 3.1). For some, such as *pol2-S312F*, the fold change in mutation numbers compared to wild type is 14× bigger in the haploid than in the heterozygous diploid, suggesting the increase in mutation accrual is not likely simply due to a wild type polymerase replicating half the genome with high fidelity. Similarly, while *pol2-4* and *pol2-A480V* show similar mutation numbers in a heterozygous diploid setting, the absence of a wild type polymerase and half the genome have a markedly

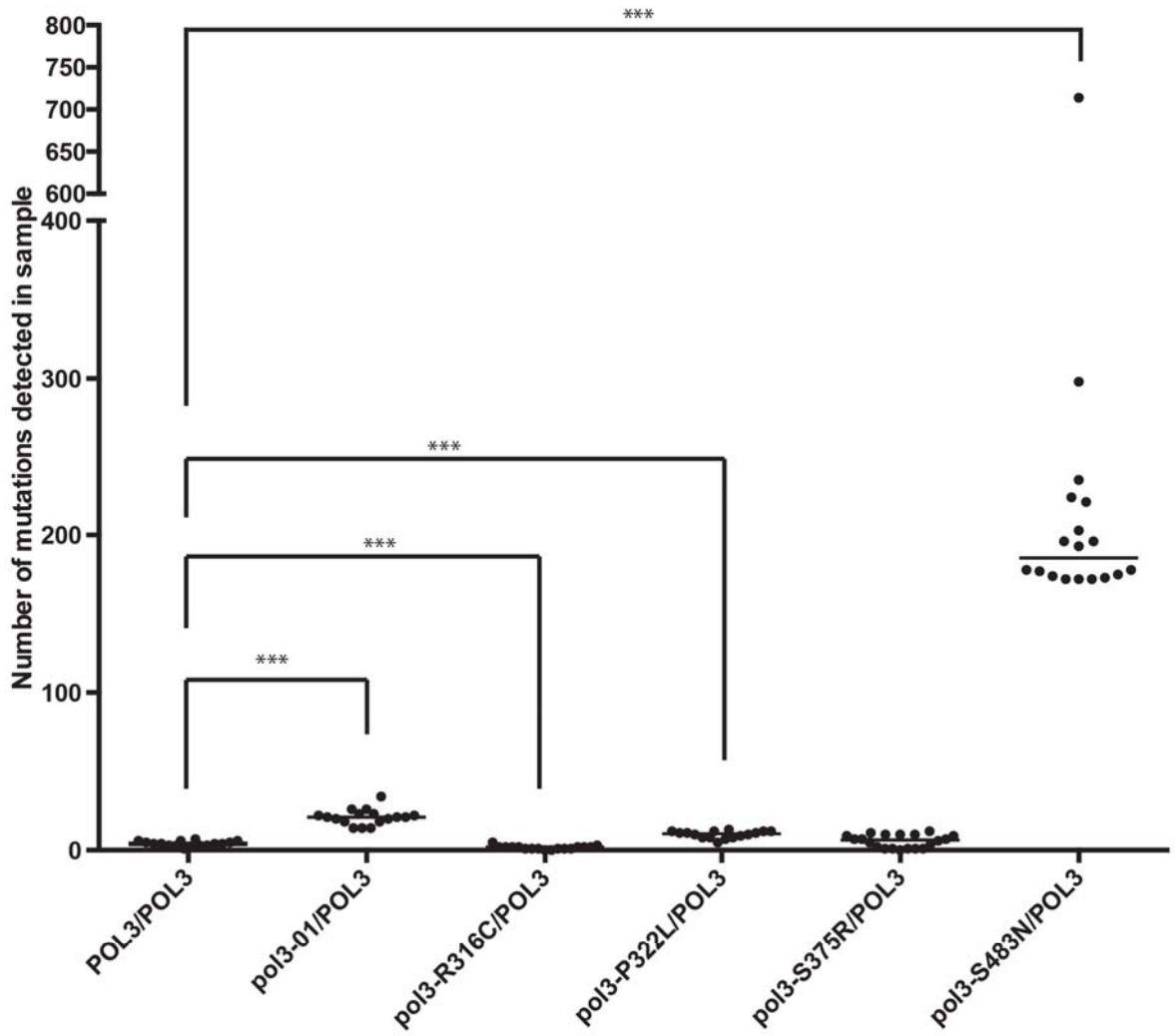

Figure 3.2: Number of single-nucleotide variants per sample in *pol3* mutant strains
Pol3 mutant strains were whole-genome sequenced before and after a three-month propagation on rich medium. The number of acquired single-nucleotide variants was determined for each parallel line that was propagated. The number of samples for each strain are as follows: n=65 (*POL3*), n=18 (*pol3-S483N, pol3-S375R*), n=17 (*POL3, pol3-01, pol3-R316C*), n=16 (*pol3-P322L*). All strains are heterozygous diploid for the mutation in question. The median is denoted by a black line. Student's T-test was used to determine which samples are significantly different from wild-type at *** $p < 0.001$.

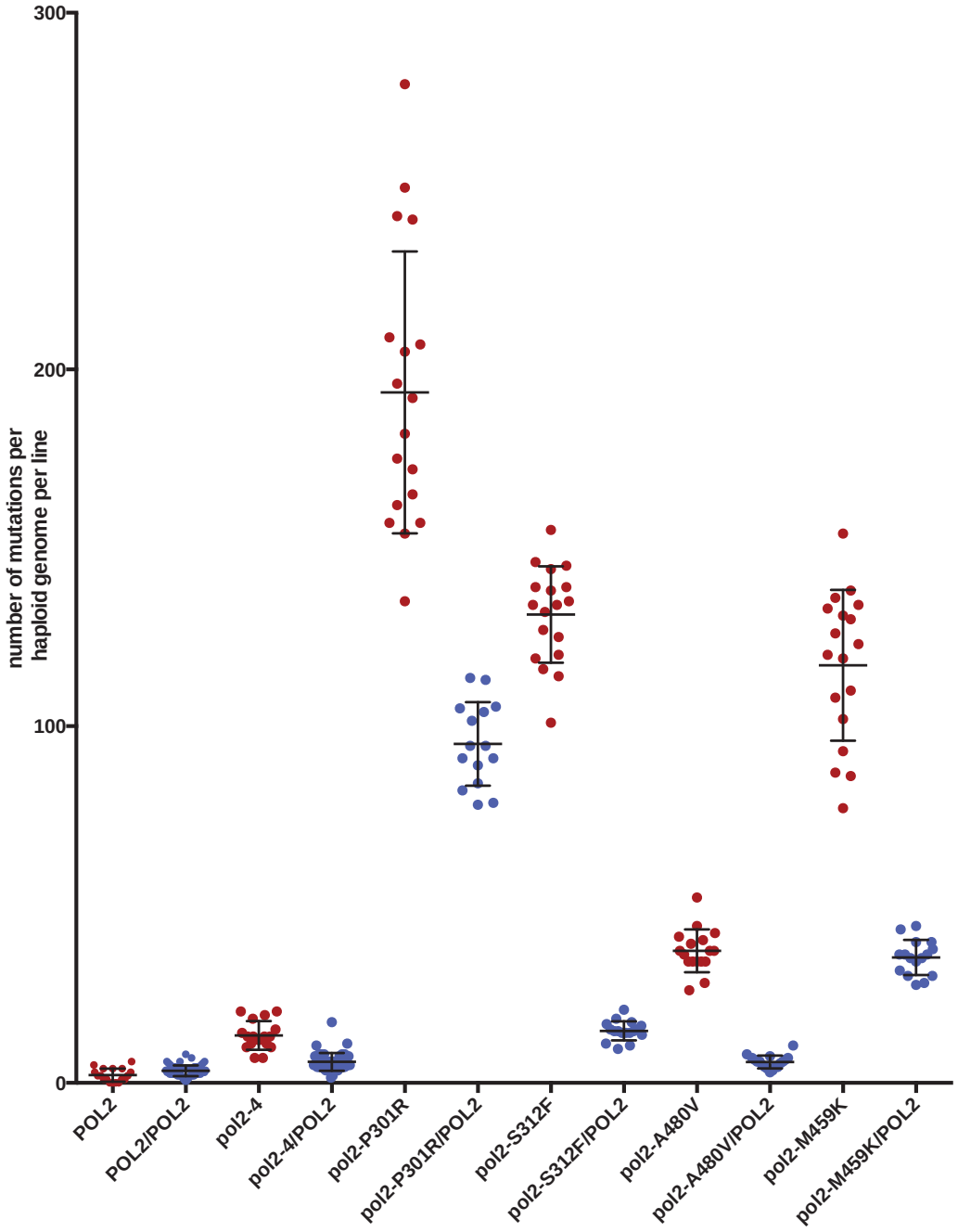

Figure 3.3: Number of single-nucleotide variants per line per haploid genome for selected haploid and heterozygous diploid *pol2* mutant strains
The number of detected single-nucleotide variants (SNVs) is plotted for heterozygous diploid strains similar to in Fig. 3.1: each dot is the measurement of an independent line. To account for differences in genome size, measurements were normalised to a haploid genome. For haploid strains the number of detected SNVs after 13 passages is depicted in the same manner. The black bars indicate mean and standard deviation.

	HAPLOID	DIPLOID
pol2-4	6.1x	1.7x
pol2-A480V	17.1x	1.7x
pol2-M459K	54x	10.3x
pol2-P301R	89.3x	27.9x
pol2-S312F	60.6x	4.3x

Table 3.1: Mutation number fold change of *pol2* haploid and heterozygous diploid mutant strains when compared to the *POL2* strain
After propagation, the number of detected single-nucleotide variants in the *pol2* mutant strains is normalised to the number detected in the wild-type *POL2* strain of the same ploidy.

different effect on mutation numbers.

3.2.3 *pol2* mutants grow at a similar rate to wild type strains

To correlate acquired mutation numbers over time with a mutation rate, the cells would need to grow at a similar rate and undergo a similar number of divisions in a given amount of time. In order to test whether the heterozygous diploid polymerase mutant strains grow at similar rates to the wild type, I monitored cell growth rates by measuring the absorbance at 595nm wavelength culturing cells in rich medium from stationary phase for 450 minutes.

All *pol2* heterozygous diploid polymerase mutant strains grow similar to the *POL2* wild-type strain suggesting that the mutation numbers obtained at the end of the mutation accumulation experiments can be compared and that any increases in mutation rate cannot be explained by differences in proliferation speed (Fig. 3.4).

3.2.4 Correlation of mutation rate estimates with mutations accrual

Additional to the propagation experiments, mutation rate estimates were also obtained using the resistance to thialysine (Thia^r) (Table 3.2.4). While Thia^r measurements are much more variable than mutation accumulation experiments (see Table 3.2.4 and Fig. 3.1 & 3.2), there is a positive linear relationship ($R^2 = 0.912$) between mutation rate estimates using Thia^r and the number of SNVs detected by mutation accumulation experiments followed by NGS (Fig. 3.5).

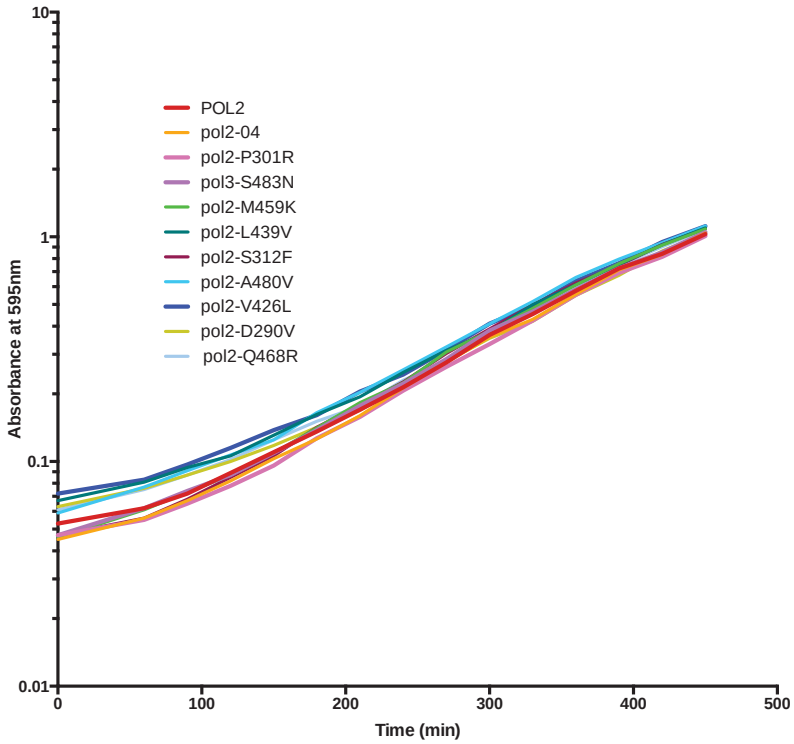

Figure 3.4: Growth of *S. cerevisiae* mutant strains in rich medium
Cell growth of heterozygous diploid mutant strains in rich medium was monitored by measuring absorbance at 595nm in a spectrophotometer. Cells were grown to saturation overnight at 30°C and released them into fresh rich medium at a dilution of 1:200. Measurements were taken every 30 minutes for 450min. Data from one experiment shown.

	MEDIAN (CI)	FOLD CHANGE
POL2 wt	7.499E-08 (3.7E-08 - 5.4E-07)	1x
pol2-4	8.482E-07 (1.6E-07 - 1.8E-06)	11.3x
pol2 A480V	4.732E-06 (1.1E-06 - 1.5E-05)	63.1x
pol2 D290V	6.315E-07 (1.9E-07 - 1.4E-06)	8.4x
pol2 L439V	2.979E-06 (2.3E-06 - 7.0E-06)	39.7x
pol2 Q468R	1.233E-07 (4.5E-08 - 2.6E-07)	1.6x
pol2 S312F	3.377E-07 (1.6E-07 - 2.7E-05)	4.5x
pol2 V426L	3.615E-07 (1.1E-07 - 1.4E-06)	4.8x
pol2 P301R	2.815E-05 (5.2E-06 - 4.2E-05)	375.5x
pol2 M459K	1.032E-05 (4.5E-06 - 1.4E-05)	137.6x
pol3 S375R	2.839E-07 (4.4E-08 - 1.3E-06)	3.8x
pol3 S483N	9.834E-06 (2.0E-06 - 0.00027)	131.2x

Table 3.2: Estimates of mutation rate increases using resistance to Thialysine
Numbers of resistant colonies were obtained from seven independent cultures for each tested strain. Fluctuation analysis was used to determine median mutation rate estimates as well as a 95% confidence interval (in brackets). Fold change with respect to the wild type is given.

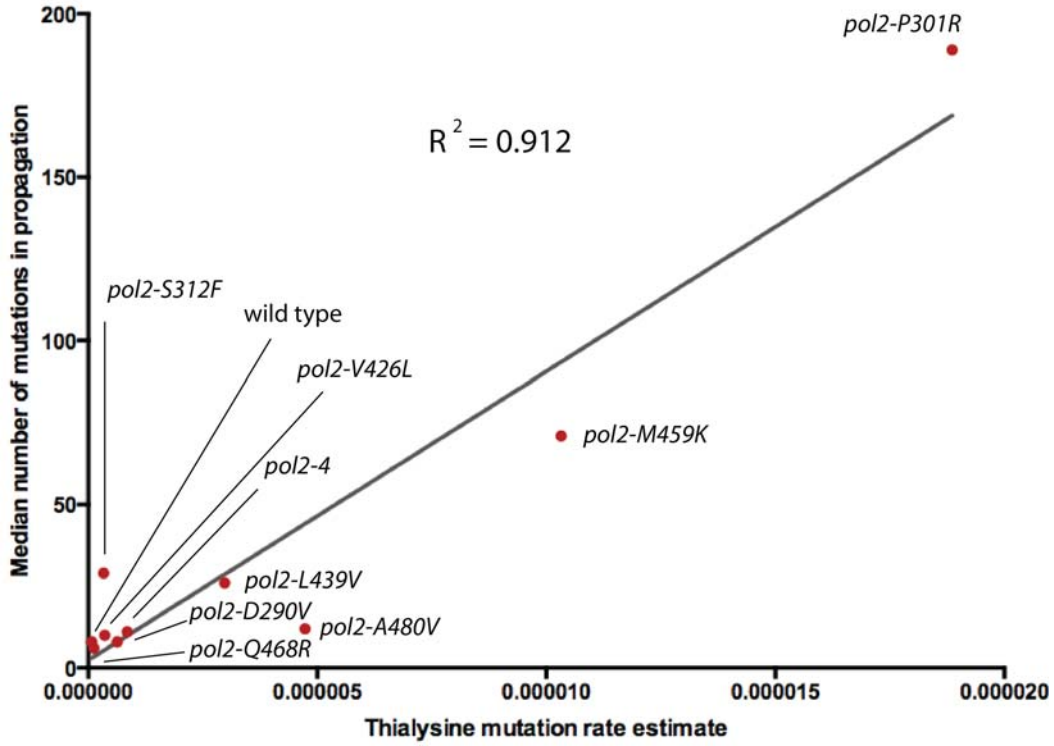

Figure 3.5: Correlation of mutation rate estimates and mutation accrual for *pol2* mutant strains For all *pol2* mutant strains and a wild type control, mutation rate estimates for haploid mutants obtained using thialysine resistance were plotted against the number of mutations detected after propagating heterozygous diploids for 25 passages. A regression line was drawn and the regression coefficient R^2 is given.

3.3 Patterns of single-nucleotide variants

The recent advances in identifying mutational processes and the patterns that they leave in cancer cells - mutational signatures - have focussed on base substitutions considering the affected base as well as the ones immediately upstream and downstream. There are six different kinds of single-nucleotide mutations (Fig. 1.26), which leads to 96 different triplet changes.

The mutations identified in the polymerase mutants can be visualised in the same format to show whether any preferences for certain mutations exist. The mutation pattern for the wild-type and *pol2-P301R* and *pol3-S483N* is shown, because of the high mutation accumulation in those samples (Fig. 3.6). These patterns are further normalised to the occurrence of triplets in the genome to show mutational preference independent of abundance of each triplet (Fig. 3.7). The *pol2-P301R* strain shows a stark preference for CTC>CAC mutations as well as for ACA>AAA and TCT>TAT. Adjustment to triplet occurrence in the genome makes the latter two less prominent, but highlights the enrichment for CTC>CAC mutations considering the low abundance of CTC triplets in the *S. cerevisiae* genome (Fig. 3.7-A). While the *pol3-S483N* pattern is a lot more similar to the one observed in the wild-type, adjustment to the genome-wide triplet distribution highlights a preference for T>C mutations, especially ATC>ACC and CTC>CCC changes.

Cancer cell mutational profiles are usually much more complex than this in that they are composites of often many mutational processes acting at different times, for varying lengths with diverse intensities. To deconvolute this multidimensional dataset and identify common underlying patterns several mathematical approaches have been employed, among them principal component analysis (PCA) and non-negative matrix factorization (NMF). NMF is a method from linear algebra allowing the deconstruction of a matrix into two smaller matrices, whose product approximates the original matrix, with the property that all values be non-negative[876]. One of its most well-known uses is to use NMF to decompose an object into its parts: NMF can be successfully used to represent faces as a composite of eyes, mouths, noses and so on or can be used to find semantic features in an encyclopedia and recombine them to reconstruct encyclopedic features[876]. In many ways, identifying underlying patterns of mutations in sequenced cancer samples is analogous to these examples and NMF has been used successfully to extract mutational signatures from cancer data and can also estimate the relative contribution of each mutational signature to a particular cancer[700]. NMF can be computationally expensive when the dataset is very large in which case PCA is more suitable. Considering the size of my dataset, NMF is a suitable method for this analysis.

To extract the mutational signatures from the sequencing data, the SomaticSignatures[763]

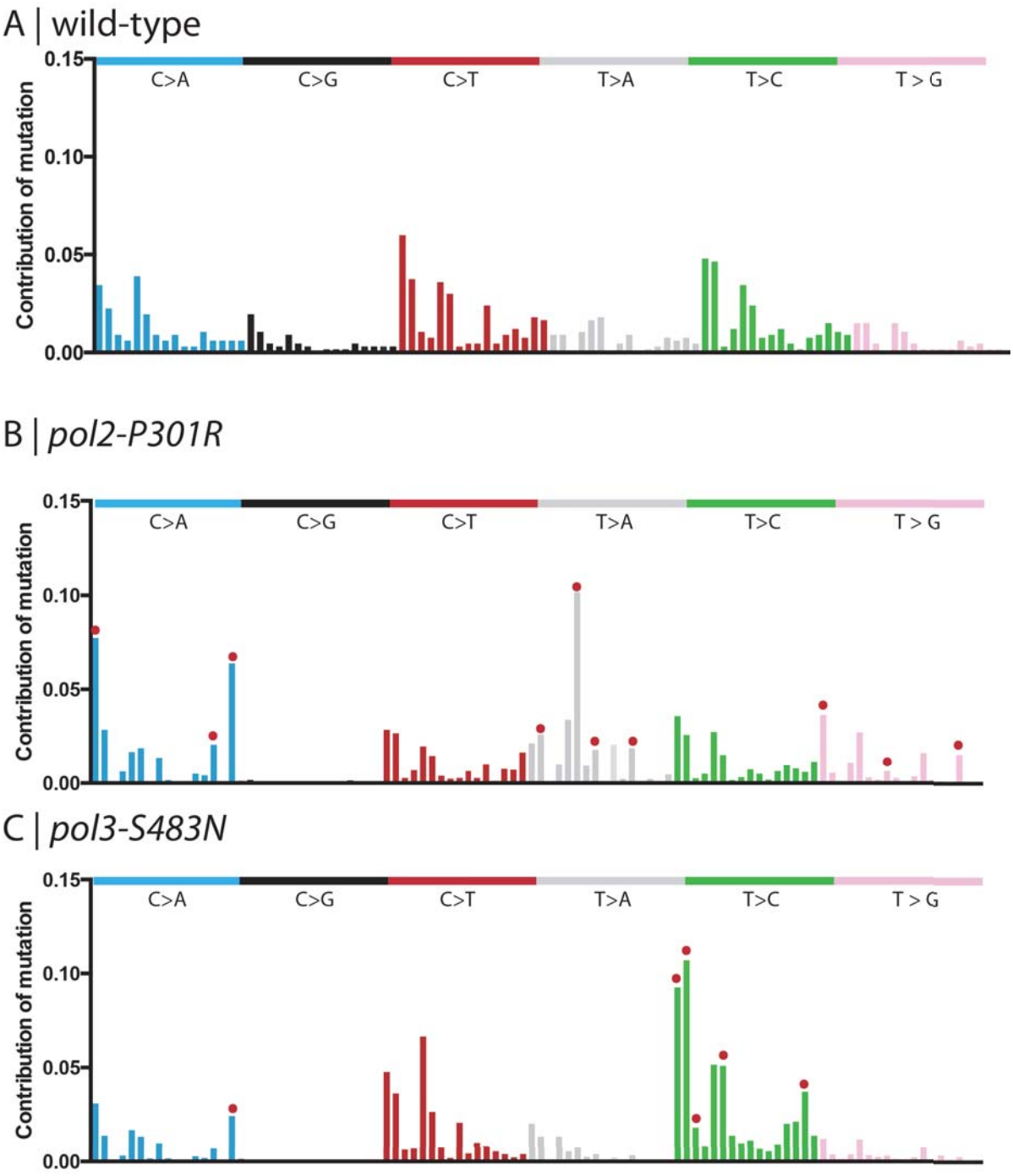

Figure 3.6: Single-nucleotide variant patterns

The mutational landscape of single-nucleotide variants found after propagation in wild-type strains **(A)**, *pol2-P301R* strains **(B)** and *pol3-S483N* **(C)** strains was visualised by considering the base change itself (all changes were transformed to have a pyrimidine base) as well as the immediately flanking residues. For all variants the sequence context was extracted based on the genomic location within the reference sequence. The trinucleotide changes that diverge most from the wild-type as tested by χ^2 are marked by a "red dot".

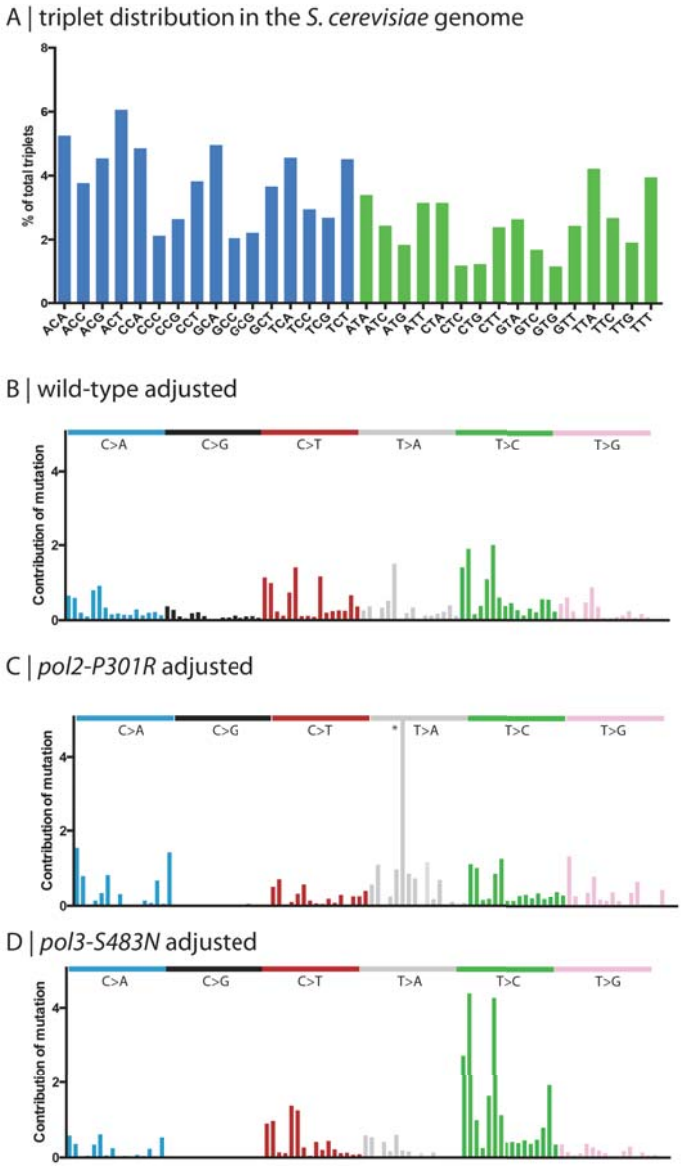

Figure 3.7: Single-nucleotide variant patterns adjusted to frequencies of trinucleotides in *S. cerevisiae*

A| The abundance of each triplet in the W303 genome was determined. Triplets with a central cytosine are shown in blue, those with a central thymidine are shown in green. **B-D|** The mutational landscape of single-nucleotide variants found after propagation adjusted for the genome wide occurrence of triplets in budding yeast. * The scale ends at 0.05, but the CTC>CAC mutation was measured to contribute at an adjusted value of 0.085.

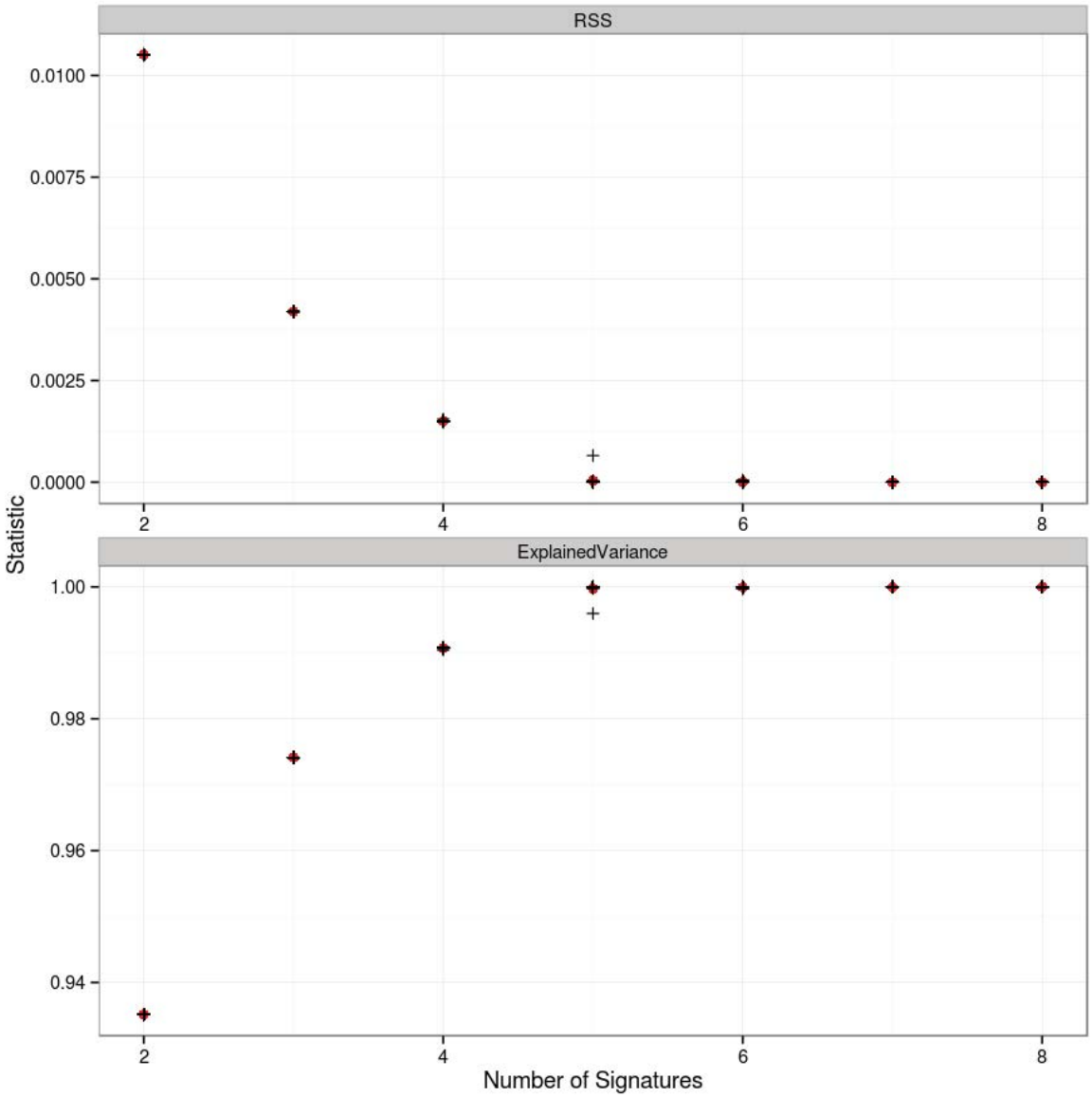

Figure 3.8: SomaticSignatures: Determining the numbers of signatures
The residuals sum of squares (RSS) and the explained variance between the observed matrix and fitted mutational spectrum for 2 to 8 signatures. The number of signatures can be chosen so that the addition of one more signature does not yield a sufficiently better approximation to the data. The first inflexion point has been suggested as an appropriate measure for the number of signatures [877].

package was used to apply NMF to the data and plot the results (see 6.9.4). The algorithm determined the residuals sum of squares (RSS) and the explained variance for 2 to 8 signatures (Fig. 3.8). The RSS is a measure of the discrepancy between the data and the model and a small RSS and a large value for explained variance indicate a tight fit of the model to the data. The likely number of signatures can be chosen by looking for the number where little improvements in RSS and explained variance are made by adding another signature. The first inflexion point has also been proposed as a measure to determine the number of signatures[877]. The values for RSS and explained variance obtained for the mutator strains displayed in Fig. 3.8 indicates that two signatures explain nearl 94% of the variance. Adding a third signature would explain roughly another 3.5% of the variance, while adding a fourth signature only improves the explained variance from 97.5% to 99%. The algorithm thus extracted two-three signatures from the aggregated data of all sequenced *pol2* and *pol3* heterozygous diploid mutant strains and the wild-type control. In the case of two signatures, Signature 1 shows a striking peak in the C>A mutations (the same TCT>TAT also preferred in *pol2-P301R* samples), while Signature 2 is very similar to the pattern observed for the wild-type and the *pol3-S483N* strain(Fig. 3.9-A). Contributions of each of the two signatures to the mutation patterns observed in each strain are also estimated by SomaticSignatures(Fig. 3.10-A). In accordance with the SNV patterns displayed in Fig. 3.6 and the mutation accrual displayed in Fig. 3.1 and Fig. 3.2, Signature 1 is estimated to mainly contribute to mutations in *pol2* mutant strains. For the four strains with significantly increased mutation numbers, *pol2-P301R, pol2-S312F, pol2-L439V* and *pol2-M459K*, Signature 1 has an estimated contribution of almost 100%. For *pol3-S483N*, a strain with a significant increase in mutation numbers, the Signature 1 contribution is approximately 10%, in line with the observed mutation pattern in these strains. When determining three signatures, Signature 1 remains unchanged, while the previous Signature 2 is split into two distinct signatures (Fig. 3.9-B). In this case Signature 2 is the main contributor to mutations in the *pol3-S483N* strain, while Signature 3 is the main contributor to the wild-type strain, which is consistent with it contributing to half the mutations acquired by the *pol2-4* strain(Fig. 3.10-B).

To obtain further evidence for the number of likely signatures, another signature extraction algorithm EMu was applied to the collection of acquired mutations identified in Section 3.2. EMu identifies the number of mutational signatures using expectation-maximization (EM) and model selection criteria, such as the Bayesian information criterion (BIC)[764]. EMu identified three Signatures in the data. When comparing the output of EMu and SomaticSignatures, Signature 1 and Signature B are similar and show similar contributions to sample mutations, while Signature 2 and Signature A as well as Signature 3 and Signature C show similarities

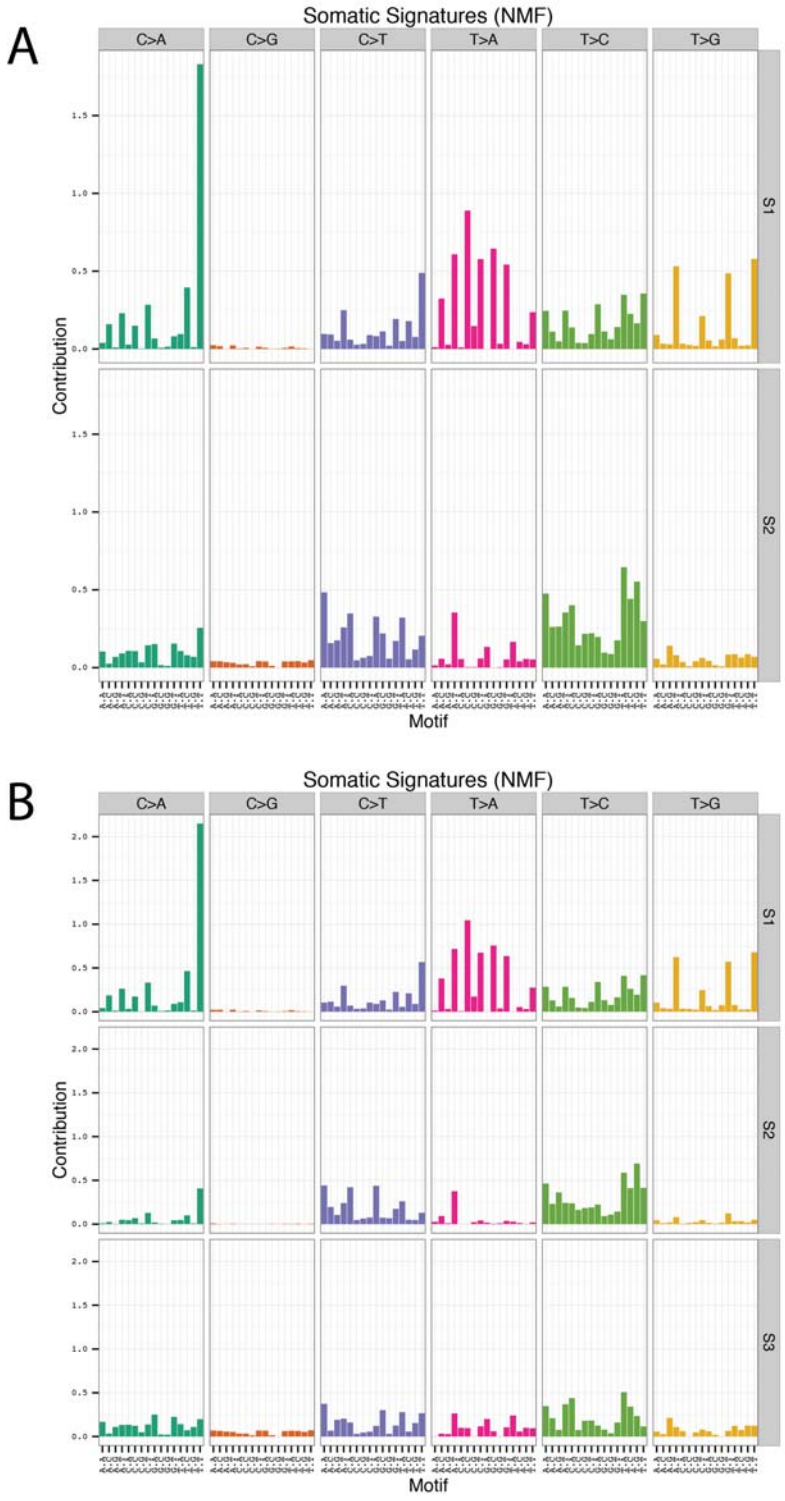

Figure 3.9: 2-3 signatures are determined using Non-negative matrix factorization
Two and three signatures were extracted from mutation data of all mutator strains and the wild type combined using SomaticSignatures with NMF. The signatures extracted are displayed in the 96 trinucleotide-change channel format indicating each mutation's contribution to the overall pattern. **A|** Two signatures extracted from mutator strains. **B|** Three signatures extracted from mutator strains.

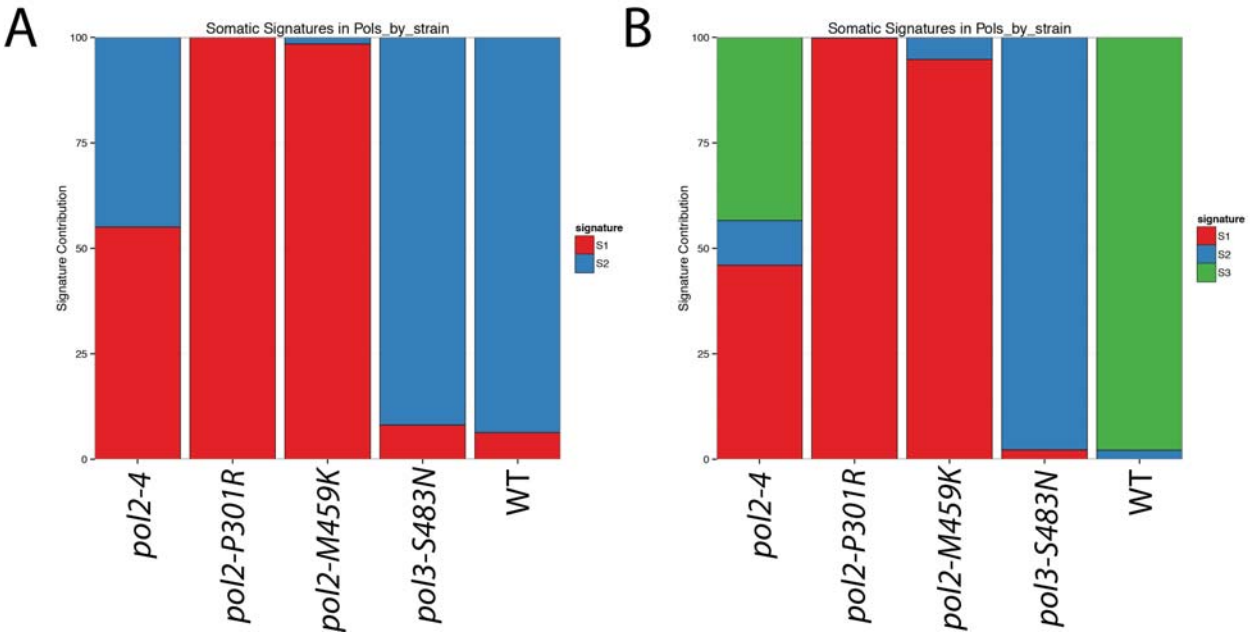

Figure 3.10: Contribution of the signatures to the variant pattern
Two and three signatures were extracted from mutation data of all mutator strains and the wild type combined using SomaticSignatures with NMF. The contribution of each of the two and three signatures to the mutational landscape of each strain is displayed. **A|** Contributions of two signatures to mutations in both samples. **B|** Contributions of three signatures to mutations in both samples.

(Fig. 3.3).

As part of the Catalogue of somatic mutations in cancer (COSMIC), human exome and whole-genome sequencing data from cancers was subjected to mutational signature extraction [700, 738, 761, 762, 878]. Currently, the collection holds 30 Signatures assembled from "an analysis of 10,952 exomes and 1,048 whole-genomes across 40 distinct types of human cancer"[879]. To understand how these signatures from human cancers compare to those extracted from the yeast samples in this work, the similarity between Signature 1, Signature 2 and Signature 3 with all 30 COSMIC human signatures was assessed. Signature 1 (the most common contributor to *pol2* mutated samples) was found to be most similar to COSMIC Signature 10 (cosine similarity = 0.63) and COSMIC Signature 8 (cosine similarity = 0.62). COSMIC Signature 10 was found most commonly in colorectal and uterine cancer and is statistically associated with the presence of *POLE* mutations, notably *Pro286Arg* and *Val411Leu*. Both signatures feature C>A mutations at TpCpT, but discrepancies in C>T and T>A mutations. COSMIC Signature 8 shows similarities to the yeast Signature 1 for C>A mutations at TpCpT and T>A mutations, however many peaks seen in COSMIC Signature 8 are absent in the yeast signature. Signature 3 (the signature observed in the wild-type yeast strains) is most similar to COSMIC Signature 5 (cosine similarity = 0.82), which is of unknown aetiology. And Signature 2 (the signature observed in the *pol3-S483N* strain) is most similar to COSMIC Signature 12 (cosine similarity = 0.78), Signature 5 (cosine similarity = 0.77) and Signature 16 (cosine similarity = 0.75). This signature and Signature 12 both show increased amounts of T>C mutations compared to all other mutation types.

3.4 Geographical mutation patterns

Apart from mutation numbers or mutational signatures, where in the genome mutations are located can provide more information on the mutagenic process at work and its possible effects. Do mutations cluster? Are there regions of the genome particularly prone to mutation? How do mutations occur with respect to features of the genome such as genes or origins of replication? In this next section, I have attempted to identify striking differences between wild type strains and polymerase mutants when it comes to the locations of the mutations within the genome.

Kataegis Kataegis describes localised hypermutation, sometimes observed in cancer samples [738]. While kataegis is linked to the APOBEC deaminases[880], we can nonetheless

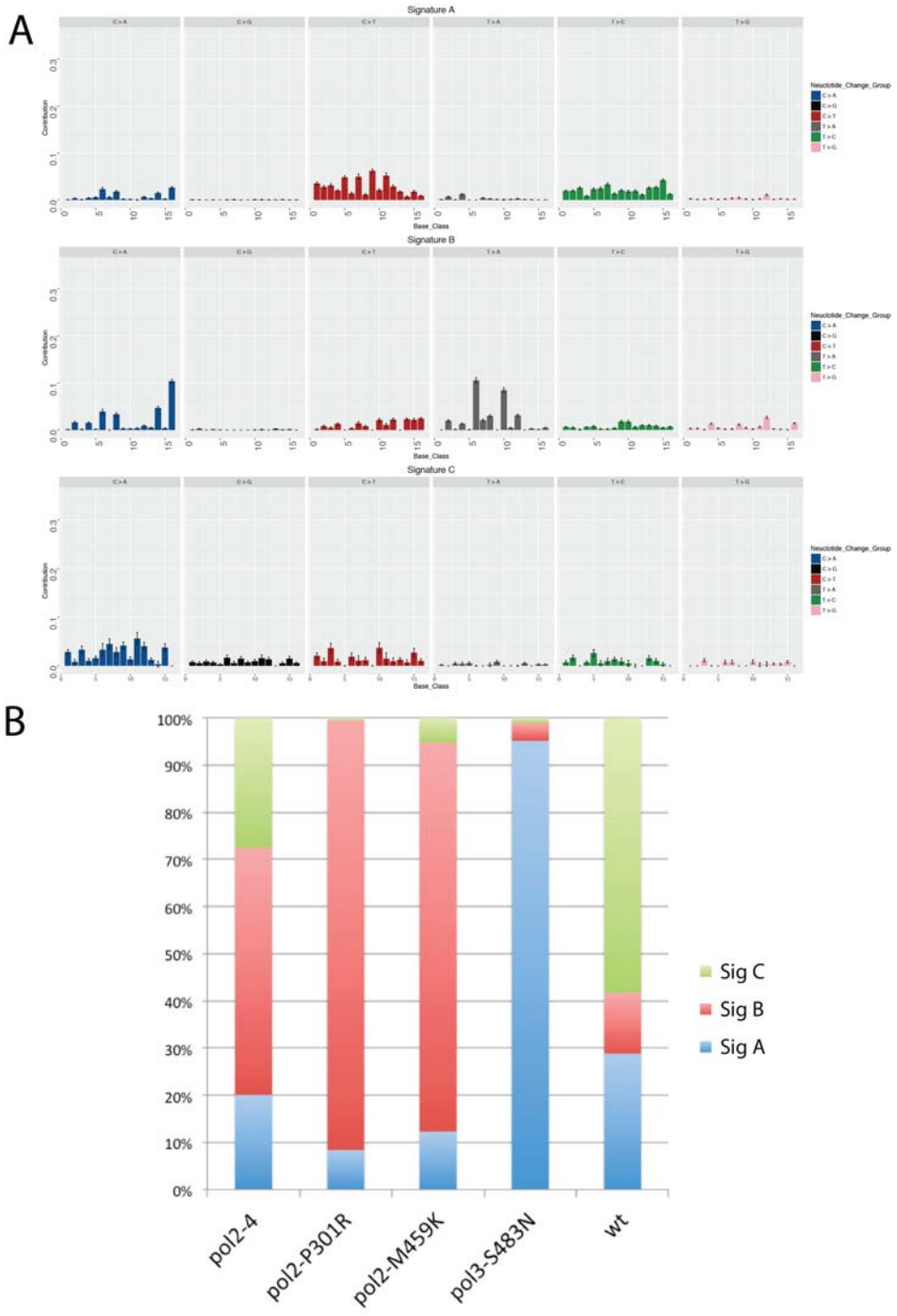

Figure 3.11: EMu: Validating Signature Analysis

Signature analysis was also performed using EMu. Using model selection criteria, such as the Bayesian information criterion (BIC), EMu determined that a model with three signatures has the strongest statistical support. **A|** Signatures displayed in their trinucleotide mutation pattern. **B|** Contribution of signatures displayed in **A** to the total mutations observed in mutator strains.

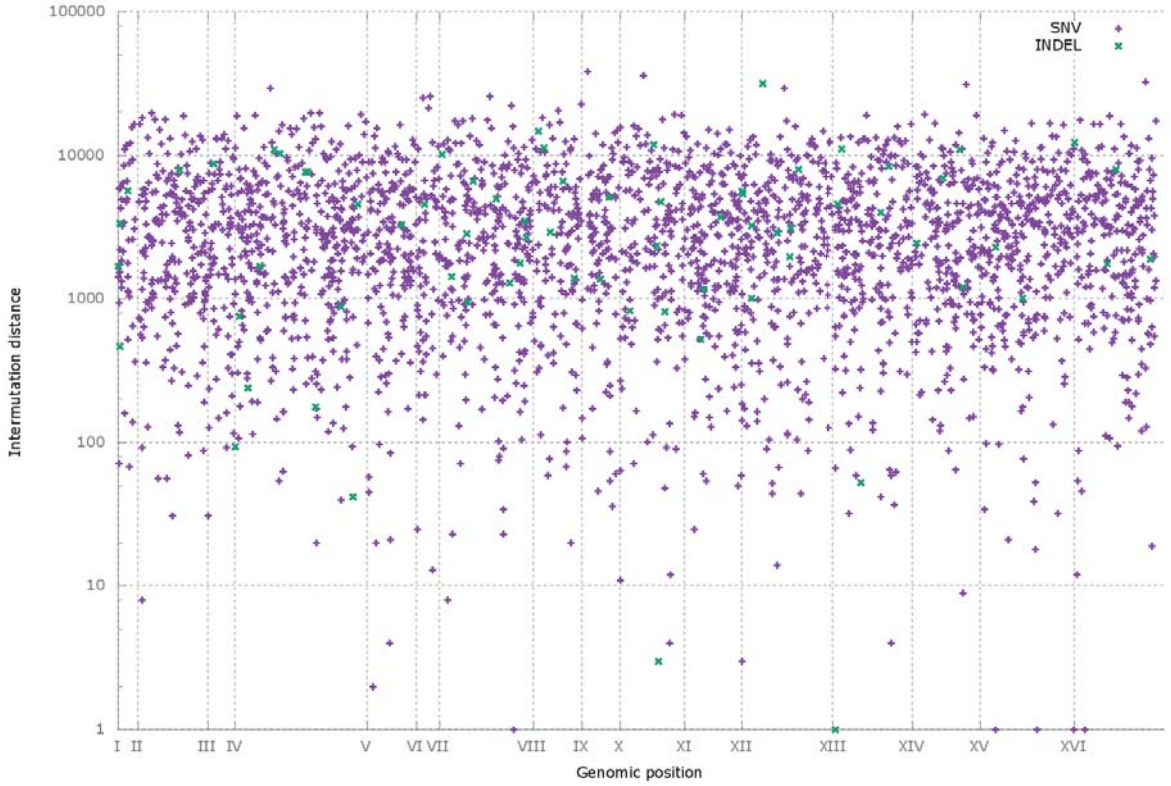

Figure 3.12: No observed clustering of mutations acquired by *pol2-P301R* strains
Mutations acquired across 15 parallel lines of propagated *pol2-P301R* strains were pooled, sorted and inter-mutation distances were determined and plotted.

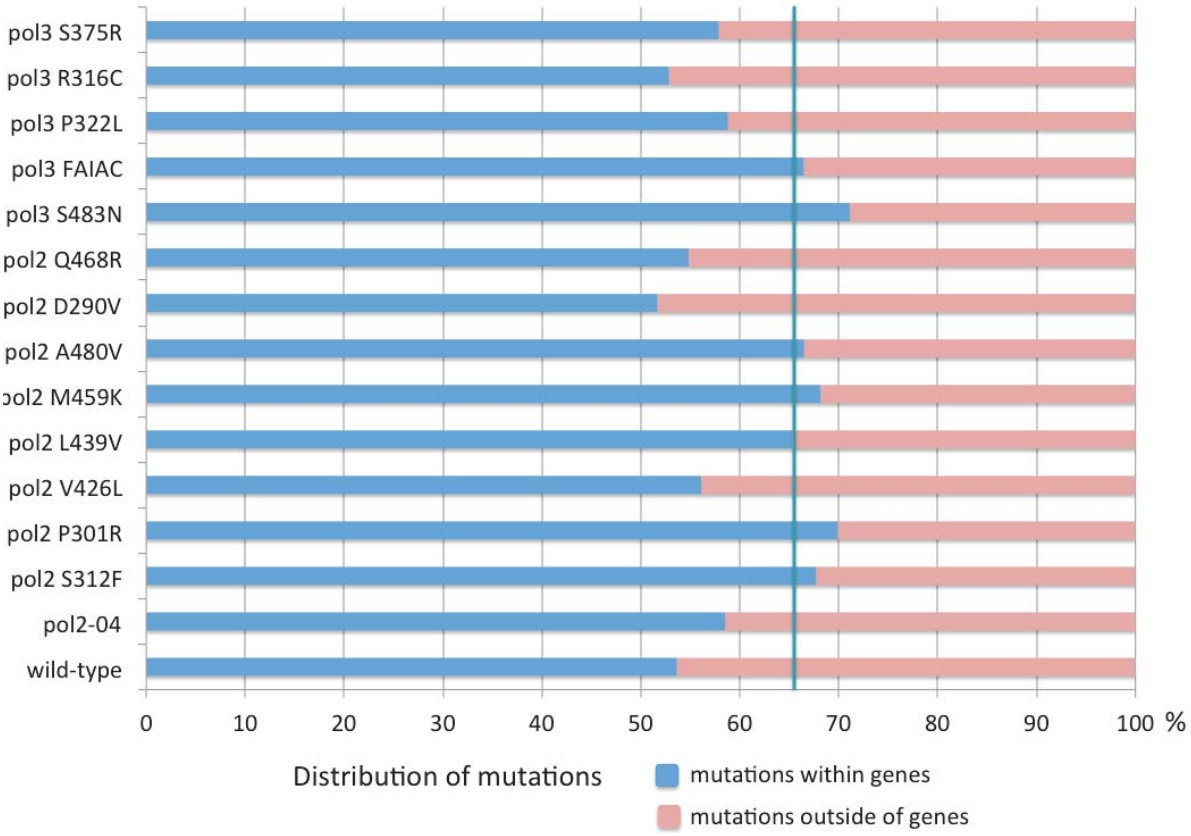

Figure 3.13: Mutations falling inside and outside of genes in heterozygous diploid polymerase mutant strains
Each mutation acquired during the three month propagation of the heterozygous diploid polymerase mutant strains was scored for their presence inside one of the 5133 verified *Saccharomyces cerevisiae* open reading frames (ORFs). Here the percentage of how many mutations fall inside a gene versus outside a gene for each strain is given. The vertical, teal coloured line at 65.5% represents the genome wide percentage of nucleotides that form ORFs.

look for regions of the genome where mutations are clustered or common. To that end, mutations across all parallel lines for each strain were pooled and sorted. The distance between consecutive mutations was calculated and plotted. This results in a "rainfall plot", where most mutations will appear as a "cloud" at the value of the mean inter-mutation distance. Where mutations cluster the inter-mutation distance will be significantly smaller and data points will appear as "raindrops" making them easy to spot. Across all heterozygous diploid polymerase mutant strains no striking examples of mutation clustering was observed (see Fig. 3.12 for an example, Appendix B for all plots).

Genic versus intergenic locations of mutations If mutations were acquired randomly across the genome, regardless of the effect on genomic information, we would expect the fraction of mutations within gene sequences to be concordant with the fraction of the genome covered by genes. The *S. cerevisiae* reference genome contains 12071326 nucleotides. The 5133 verified open reading frames (ORFs) listed in the Saccharomyces Genome Database cover 7905244 nucleotides or 65.49% of the genome. Each of the mutations acquired in the propagation experiment of heterozygous polymerase mutant strains was checked against all verified ORFs to determine whether it falls within a gene or outside them. If mutations were acquired in a truly random fashion, then one would expect roughly 65% of acquired mutations to fall into a gene. Interestingly, for wild-type propagated strains this percentage is quite low with 53.7%, while for strains with a significantly raised mutation number the percentage approaches or surpasses 65.5% (Fig. 3.13). The percentage of mutations that fall within genes is expected to be lower in haploid samples, due to the absence of a second copy for genes. Indeed, the fraction of mutations observed within genes in wild-type samples does go down to approximately 44.8% (Fig. 3.14). However, for strains that show an increased mutation accrual, the difference between haploid and heterozygous diploid strains in this respect is almost negligible.

Mutations around origins of replication Because leading and lagging strand switch at origins of replication and because polymerases ε and δ are thought to replicate each, respectively, the mutational patterns in polymerase mutant strains could show interesting behaviours around origins of replication (ARS elements in *S. cerevisiae*).

ARS sequences and their locations were obtained from the Saccharomyces Genome Database. To include only high confidence DNA replication origin sites this list was compared to the OriDB database and only confirmed ARS elements were retained. The location data between the two databases varies as origin location was determined differently. Here, the SGD origin locations were used. For each origin, the center was determined as well as the coordinates 500bp upstream and downstream. For each mutation acquired by polymerase mutants, that falls within that window, the nucleotide change and distance from origin center was determined. Using the *pol2-P301R* samples, 2782 different SNVs were tested for their proximity to 352 different ARS elements. Within 500bp of the center of the origin only 4 mutations were identified, the same number for the 4015 mutations acquired by the *pol3-S483N* strains is 3 mutations, only. If the window is extended to 2500bp either side, 8 mutations are detected in the case of both strains.

Thus, while potentially, an interesting feature of the acquired mutations, currently, the number of mutations is not sufficient to determine patterns of mutations around origins of

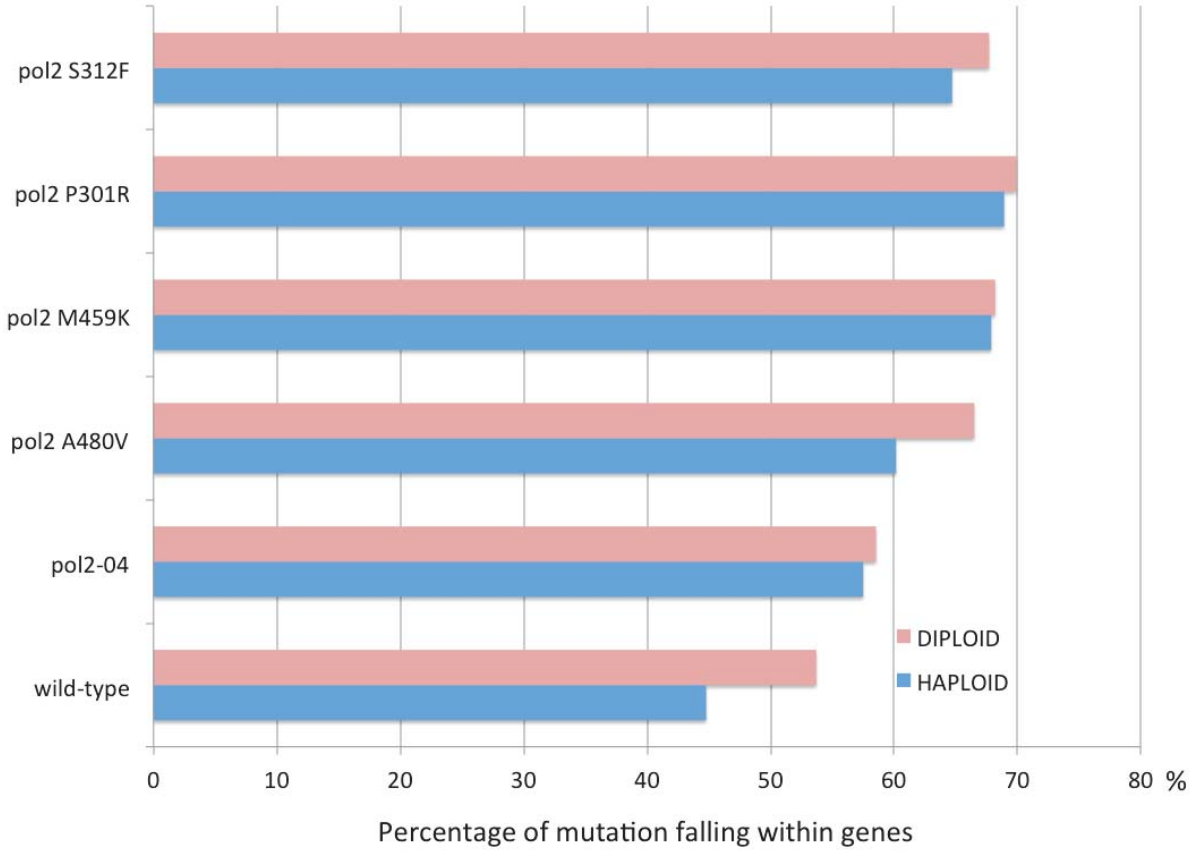

Figure 3.14: Percentage of mutations within genes in haploid strains
The percentage of mutations that fall within genes is shown for both haploid and heterozygous diploid strains for a selection of *pol2* mutants.

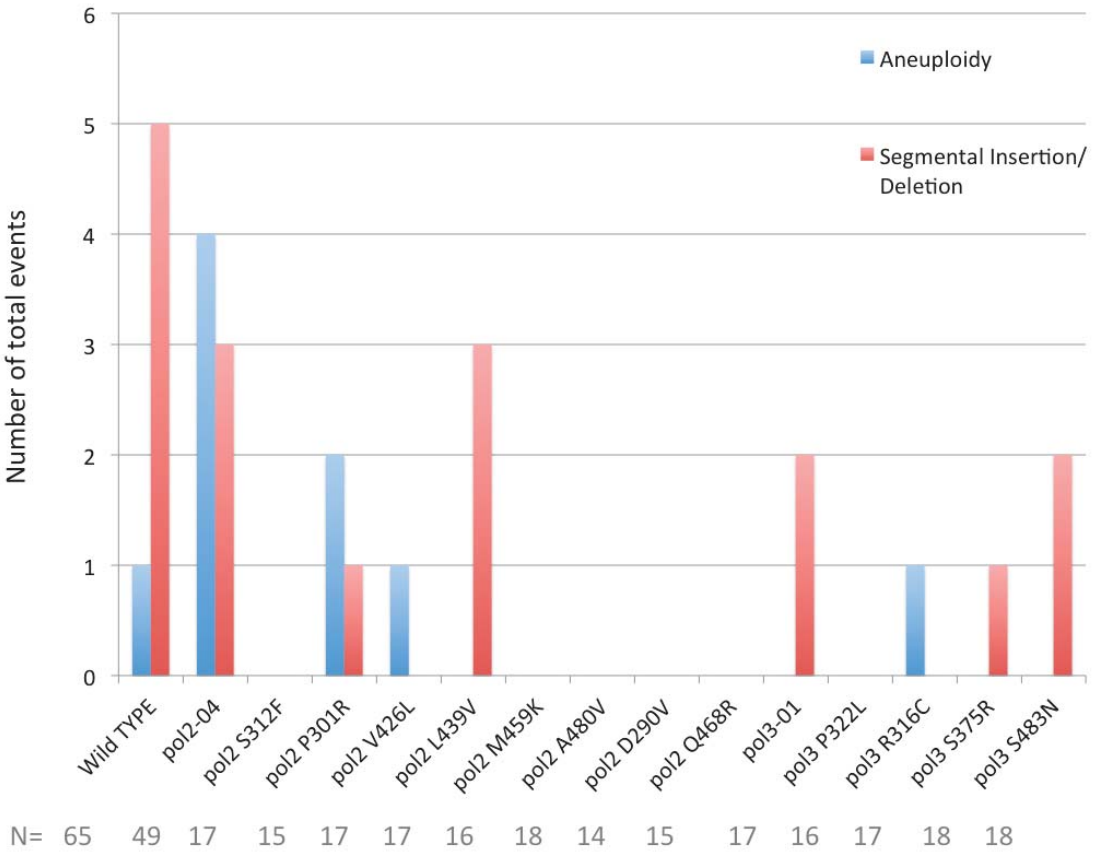

Figure 3.15: Number of total aneuploidy and segmental insertions/duplications identified The number of events of aneuploidy and large deletions/amplification was determined from the coverage data of all strains and counted. The number of samples available for each strain is listed below the figure legend.

replication.

3.5 Large-scale variation: aneuploidy, CNVs and rDNA copy number

Aneuploidy and large insertions and deletions All samples pre- and post-propagation were checked for variations in chromosome number and large segmental deletions as well as amplifications. Aneuploidies and segmental deletions are relatively rare. In 65 wild-type samples, one variation in chromosome number (1.5%) and 5 instances of segmental deletions/amplifications (7.7%) were detected. Small increases in number of aneuploidies can be seen for the strains *pol2-04, pol2-P301R* and *pol2-V426L* (Fig. 3.15). However, with

mutations that occur this rarely, the sample size would need to be bigger to make conclusive statements about significant increases. Examples of aneuploidy and segmental deletions/amplification are shown in Fig. 3.16 and Fig. 3.17, respectively, and the full set of figures can be found in Appendix B.

rDNA copy number To screen for changes in rDNA copy number, rDNA copy number estimate analysis was performed for all *pol2* and *pol3* heterozygous mutants after their three month propagation. All strains were derived from the same starting cells and any increase or decrease in rDNA copy number could likely be due to the polymerase mutation, though that would have to be reconfirmed by an independent introduction of the mutation into a new wild-type strain. Because data from the start of the experiment are not available due to data corruption, these numbers are only indications, but increases in mean rDNA copy number for strains like *pol3-S483N* will be followed up with later experiments.

3.6 No increase in INDELs (compared to MMR mutants)

While there are clear differences in mutation accrual with respect to SNVs, no striking increases in INDELs were detected (Fig. 3.19). After three months most strains (including the wild-type) will have accumulated a mean 1.18 ± 0.4 INDELs. The only increase can be seen for *pol3-S483N* with a mean of 4.45 INDELs for each strain. However, compared to the average of 200 SNVs this strain accumulates the increase is minor.

This is in stark contrast to other mutations that affect DNA replication fidelity found in cancer. As discussed in Chapter 1.4.2, loss-of-function mutations in mismatch repair proteins are known to predispose to colorectal cancer[803, 804] and as a pilot experiment using the automated robot propagation set-up (see 6.7), 150 strains carrying mutations in known DNA repair genes were propagated for 3 months by our group and the Warringer group in Sweden as described. These included strains deleted for mismatch repair genes: *MSH2, MSH6, MLH1* and *PMS*1. Ten colonies each had their DNA extracted, were sequenced and the data was aligned as described in Chapter 2 and 6. One Mlh1 sample did not pass sequence quality control.

Figure 3.20 shows that in the case of most mismatch repair mutants, the mutation increase is fairly evenly split between SNVs and INDELs with the exception of of Msh6 (part of MutSα), whose absence, as expected, results mainly in single-nucleotide mismatches. Loss of Msh2 (part of both MutS$\alpha$and MutSβ), Mlh1 (MutL homolog) and Pms1 (which forms a heterodimer with Mlh1) leads to high numbers of SNVs and small INDELs.

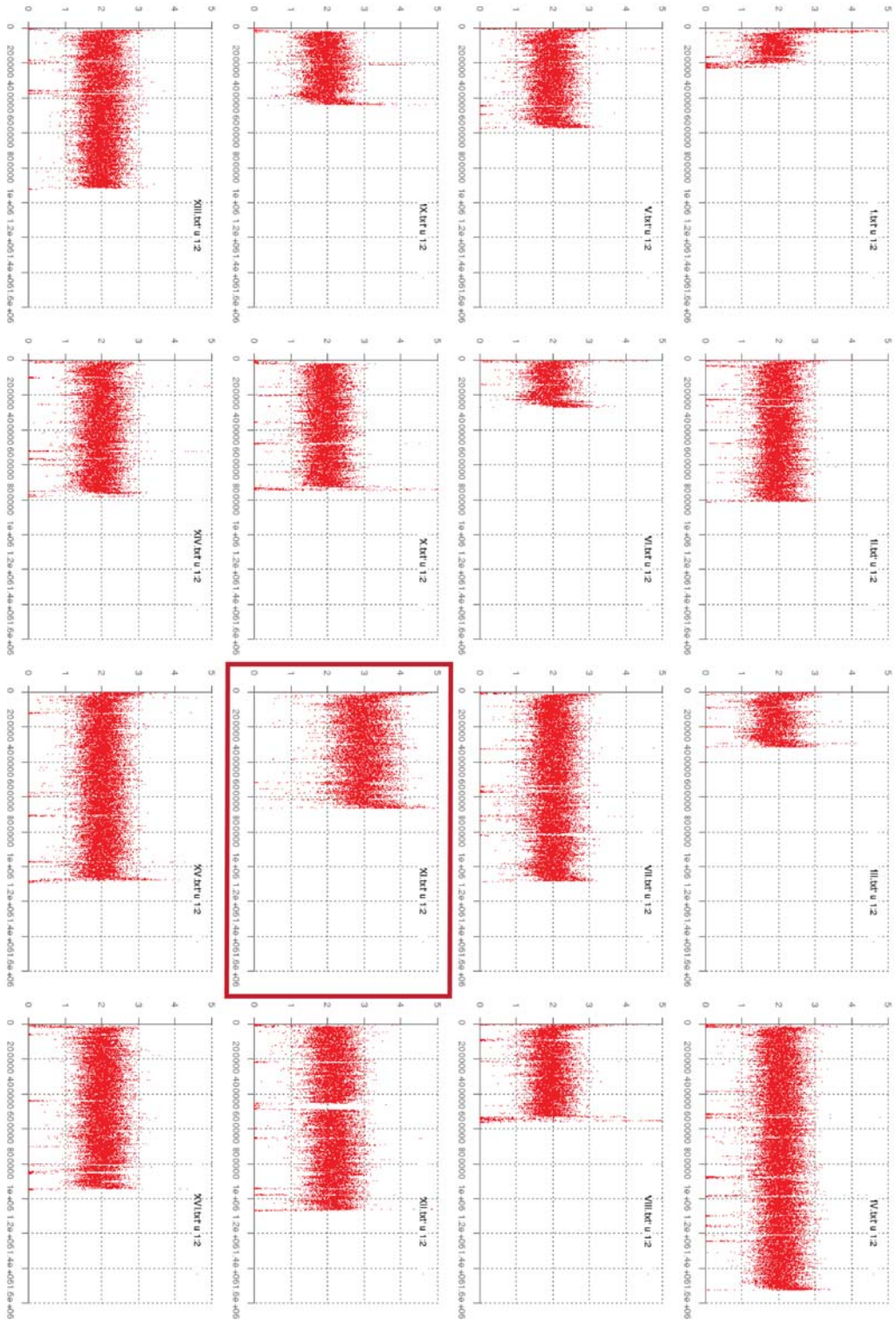

Figure 3.16: Example of aneuploidy in *S. cerevisiae*
This example shows the coverage profile of a *pol2-04* propagated heterozygous diploid strain with an aneuploidy for chromosome XI (3n). At repetitive regions the coverage drops to 0 (for instance see chromosome XII) due to mapping quality thresholds in the program.

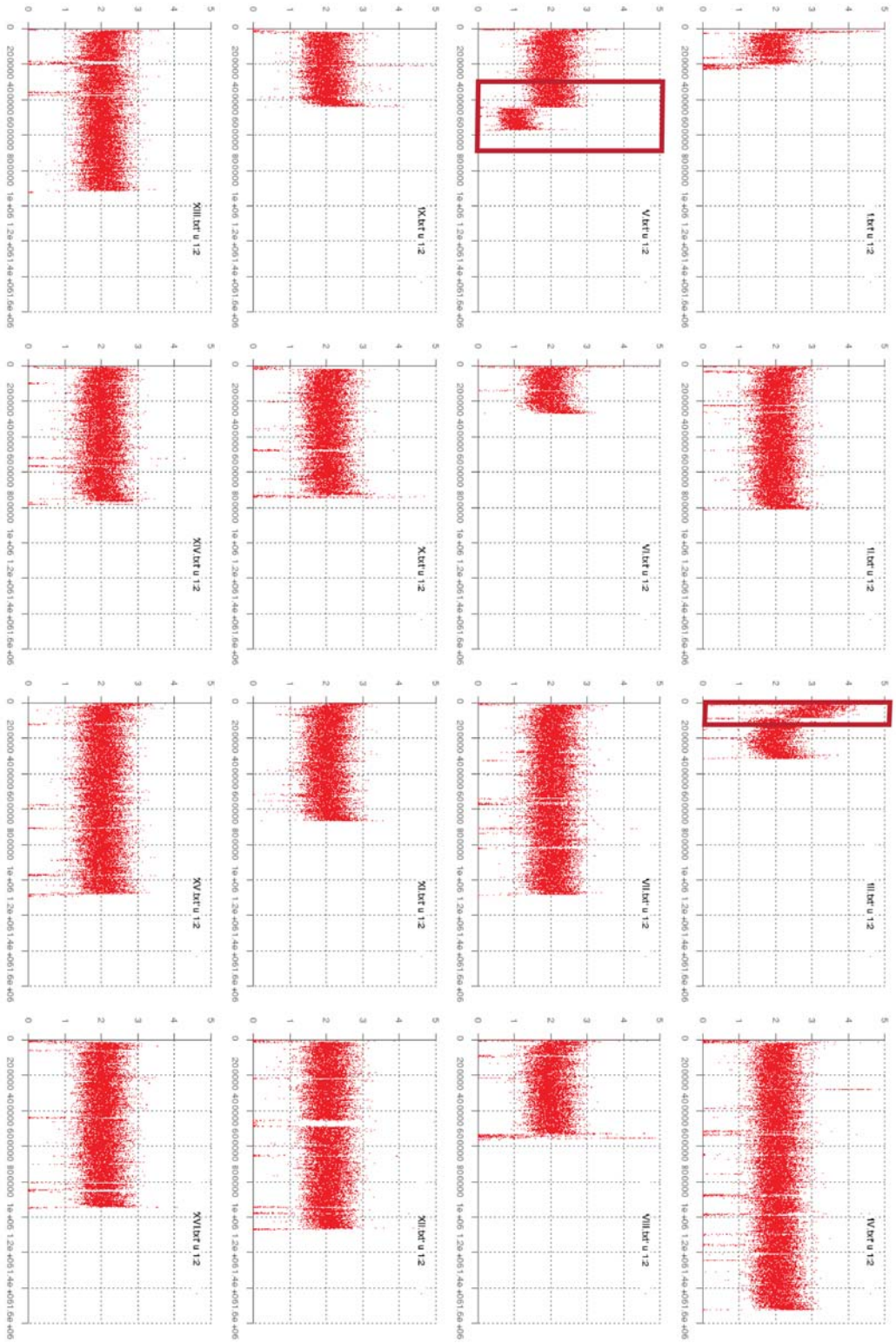

Figure 3.17: Example of segmental deletions and amplifications in *S. cerevisiae*
This example shows the coverage profile of a *pol3-01* propagated heterozygous diploid strain with a segmental deletion on chromosome V and an amplification of a region of chromosome III. At repetitive regions the coverage drops to 0 (for instance see chromosome XII) due to mapping quality thresholds in the program.

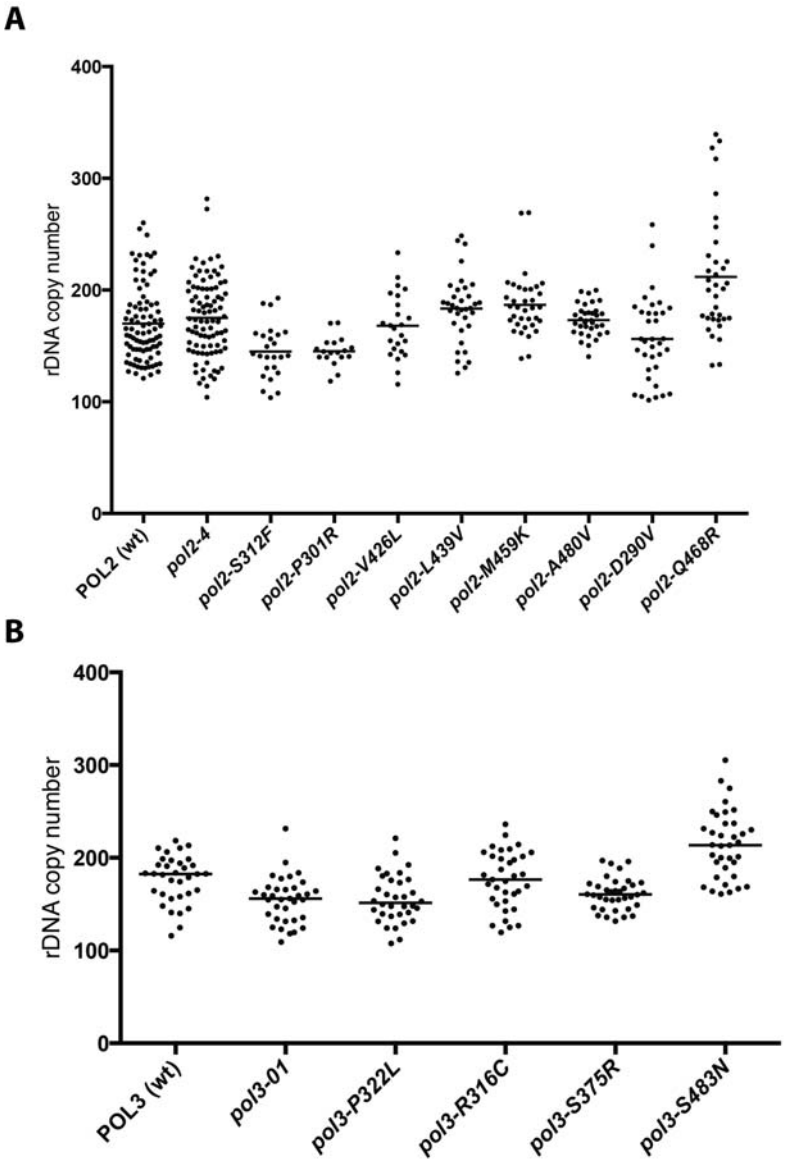

Figure 3.18: rDNA copy number changes in polymerase mutants
rDNA copy number estimates of all post-propagation samples is shown. The median is denoted by a black line. Each dot represents a post-propagation sample. **A|** rDNA repeat number for *POL2* wild type samples and all heterozygous diploid *pol2* mutant strains **B|** rDNA repeat number for *POL3* wild type samples and all heterozygous diploid *pol3* mutant strains.

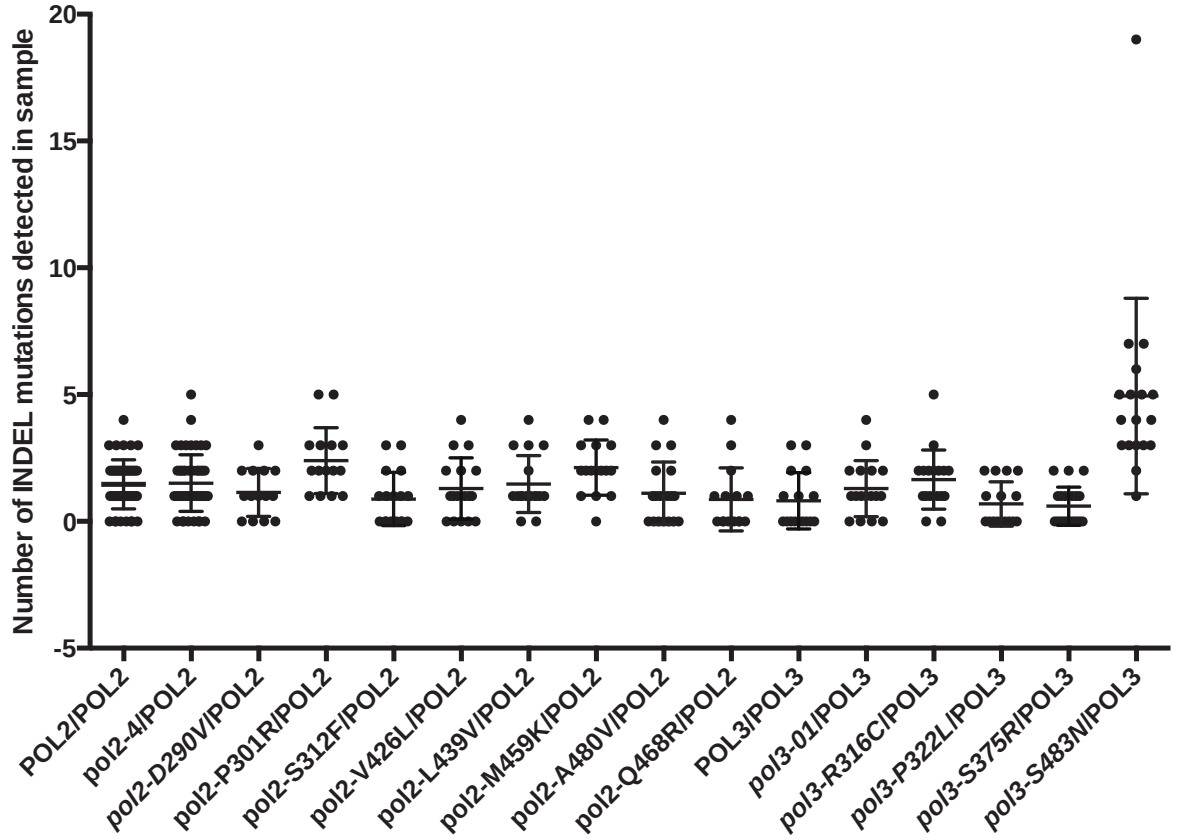

Figure 3.19: No increase in the number of INDELs detected per sample across strains
The number of INDELs per sample was determined from the heterozygous diploid polymerase mutant strains that were propagated for three months using single colony bottlenecks. The back bars indicate mean and standard deviation.

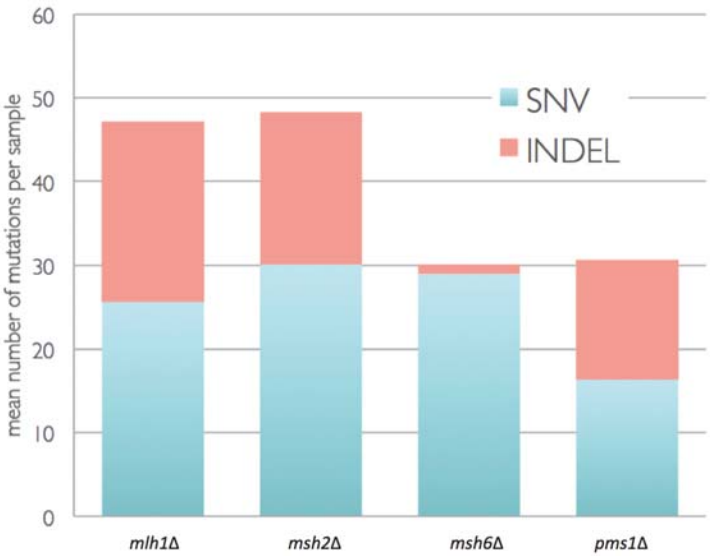

Figure 3.20: Mutation accrual in strains with mismatch repair deficiencies
Strains with deletions in mismatch repair genes were propagated for three months using population bottlenecks. Ten colonies each were sequenced and single-nucleotide variants as well as INDELs were identified. Results for strains carrying deletions in *MLH1, MSH2, MSH6* and *PMS1* are shown. Unique mutations across all colonies were counted and divided by number of sequenced colonies to obtain mean mutation numbers per sample.

These differences in mutational patterns between polymerase mutants and MMR deficient cells, is also expected to be reflected in the respective cancer genomes. Indeed, MMR deficient tumours (most commonly deficient for Mlh1 due to hypermethylation of the MLH1 promoter) commonly show high frequency of mutations, either mismatches in single bases or in regions of short tandem DNA repeats (microsatellites), the former leading to SNVs and the latter to INDELs[881]. Microsatellite instable (MSI) tumours show mutation loads ranging from 10 to 100 mutations per Mb. Polymerase epsilon mutated tumors, on the other hand, often show a mutation incidence exceeding 100 mutations/Mb and are mostly microsatellite stable (MSS)[882], characterised by mostly point mutations. The data we have collected in *S. cerevisiae*, has thus been confirmed by the data collected in human tumours.

3.7 Summary

In this part of this work, I have assessed the mutagenic potential of all candidate polymerase mutations in heterozygosity *in vivo* using the budding yeast system. Strains carrying the mutations *pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K* or *pol3-S483N* show significant

increases in single-nucleotide variants and, intriguingly, these increases are more pronounced than those resulting from mutating catalytic residues of the polymerase exonuclease domains. We further show that the pattern of SNVs is distinct from the wild-type and differs between *pol2* and *pol3* mutant strains. Striking geographical patterns or increases in large-scale mutations such as aneuploidy were not detected. Furthermore, in contrast to most mismatch-repair deficient cells, polymerase mutated strains show no increase in INDEL incidence. This disparity is also reflected in comparisons between MMR deficient tumours and those carrying polymerase epsilon mutations.

Evaluating hypotheses

Aims:

- To assess all candidate DNA polymerase mutations for the number of mutations acquired in the same amount of time

  Candidate DNA polymerase mutations lead to varying degree of mutation rate increases in budding yeast. Significant increases in mutation rates were identified for strains carrying: pol2-P301R, pol2-S312F, pol2-L439V, pol2-M459K and *pol3-S483N. The increase in mutation numbers in these cases exceeds that observed in exonuclease deficient control strains.*

- To identify the type of mutations caused by mutated DNA polymerases

  Across all mutation types examined, striking increases in the number of mutations were shown for single-nucleotide variants. Comparatively, other types of mutations were acquired rarely by polymerase mutant strains tested here.

- To determine whether candidate mutations leave a distinct mutation pattern on the genome

  Where the numbers of mutations allow, trinucleotide mutation patterns were plotted, adjusted to genome wide trinucleotide frequencies and mutation signature extraction was attempted. While the pattern of mutations in pol3-S483N strains looks similar to that observed in wild-type samples, it is subtly distinct with an increase in T>C mutations. Additionally, the pol2-P301R mutation results in a distinctive mutation pattern with key peaks in C>A and T>A mutations.

- To compare mutations acquired in mutated polymerase strains to those resulting from mismatch repair deficiency

It is well-documented that mismatch repair deficiency predisposes to colorectal cancer and polymerase mutations are implicated in predisposition to colorectal cancer. Here, mutations in yeast accumulated to both are explored. While a near-complete mismatch repair deficiency results in roughly equal amounts of single-nucleotide variants and insertions/deletions, insertions/deletions make no or negligible contributions to any increases in mutation accrual due to a mutated DNA polymerase.