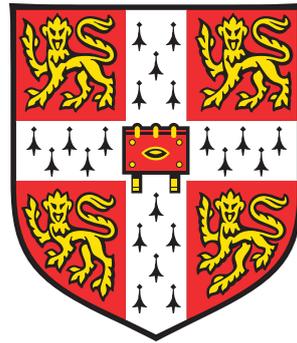


DNA polymerase mutations as drivers of genome instability and cancer



Mareike Herzog

Wellcome Trust Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

“Think, think, think.”
- Winnie-the-Pooh

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 60,000 words excluding appendices, bibliography, footnotes, tables and equations and has less than 150 figures.

Mareike Herzog
2016

Acknowledgements

First I would like to thank my supervisors Dave Adams, Steve Jackson and Thomas Keane for the opportunity to be a member of their teams, learn from them and earn my PhD with them. I want to thank them for the warm welcome I received, their patience, their kindness and their support. I am grateful to Dave for his many thoughts and advice on my work, for all his encouragement in stressful times and his trust in me, to Steve for never being too busy to talk to me and for the independence in exploring my projects and to Thomas for putting himself through the gruelling process of introducing me to bioinformatics.

I would like to thank all members of the Experimental Cancer Genetics Team at the Sanger Institute, for their friendship and their enthusiastic willingness to help me with anything I need: Daniela, James, Manu, Sofia, Mamun, Vivek, Martín, Clara, Stefan, Marco, Nicky, Chi, Louise, Rebecca, Marcela, Richard and Alistair. I want to especially thank James for supporting me with everything I could imagine: from getting me a desk, supplying me with reagents to getting me started in the world of tissue culture. Clara Alsinet Armengol and Manu Supper have my gratitude for teaching me how to take care of embryonic stem cells and for being great "babysitters" when I had to be away from the institute. My mouse work was greatly aided by Louise van der Weyden, who is the goddess of mouse work and a super friendly one at that. I am indebted to Mamun Rashid for his help with the EMu signature extraction and to Vivek Iyer for helping me manage import and storage of my mouse sequencing data. Martin Del Castillo Velasco Herrera and Stefan Dentro were providers of great advice for all things statistics and bioinformatics. And I cannot thank Daniela Robels Espinoza enough. Not only was she the patient and kind with all my questions, taught me endless programming tricks and helped me fix some tricky bugs, she also put up with me as a roommate. She has been an endless source of motivation, reassurance and encouragement for me.

I would also like to thank all of the members of the Jackson Group at the Gurdon Institute: Fabio, Israel, Josep, Yaron, Gabi, Rimma, Serena, Delphine, Paco, Paul, Will, Matt, Donna, Natasha, Pallavi, Christine, Andy, Sati, David, Carlos, Ryotaro, Jessica, Jon, Abdul, Linda, Nicola, Julia, Helen, Kate, Gopal, Matylda, Siyue, Muku and Ana. To highlight but a few, I will always remember Rimma Belotserkovskaya, because she was the first person from the lab

I met when she lectured me in my undergraduate days and got me interested in DNA repair, Julia Coates, because she shared a "bench" with me for years and we suffered the institute's temperamental air conditioning together and because she helped me move house no questions asked, Serena Bologna, for involving me in her project, being a great friend and speaking Italian to me, Linda Baskcomb, for being such a competent lab manager, and Nicola Geisler for supervising me during my rotation project, which convinced me to do my PhD in this group. Special thanks goes to Josep Forment, who invited me to collaborate on a rewarding project with him. He is a joy to work with and I will miss him greatly. I am also deeply grateful to Israel Salguero Corbacho, who helped me a lot with designing my plasmids and my cloning for the yeast strain construction. He is kind and funny and always happy to help me with anything. My deepest gratitude goes to Fabio Puddu, who, more than anyone, has been my daily mentor on this journey. From teaching me lab techniques (pulse-field gels are fiddly) to discussing and planning experiments (why would we run this analysis?), he has tried his best to make me a better scientist and I can only hope that somewhere along the way I was able to teach him something in return. I am grateful for all the criticism, encouragement, laughter, patience, praise and support.

I would like to acknowledge the Wellcome Trust for funding my PhD. Without their generous support this work would not have been possible and their funding has allowed me the privilege of not having to worry about feeding myself, paying the bills and having a roof over my head.

My friends have filled my days with joy and were always there with advice when needed and ultimately contributed much, that is intangible, to this project. I am very happy to count Chrissey, Böcki, Stephie, Monica and Julia as some of my best friends. Finally, I want to thank my family. My parents, Martina and Stefan, and my "baby" brother Gunnar, for their never ending support, motivation and their infinite love. And to Löffelchen I would just like to say with all my love "I couldn't have done it without you. Ich habe dich über alles lieb."

Abstract

Genomic stability is essential to preserve the genetic information encoded in DNA, and many biochemical pathways are devoted to repair DNA damaged by external factors, or during the course of essential cellular processes such as transcription and DNA replication. Malfunctioning of these processes may alter the DNA, leading to abnormal cellular behaviour or cell death, which in multicellular organisms may be associated with disease. For this reason, the machineries that safeguard the integrity of eukaryotic genomes are of prime interest to research in the areas of ageing, rare disease and cancer. Every time a cell divides, duplication of the genome is principally carried out by two DNA polymerases — Pol δ and Pol ϵ — which are highly processive and accurate. Together with polymerase gamma, which is active in mitochondria, these are the only human polymerases known to possess "proofreading" activity, making them extremely accurate. In parallel, cells have also evolved a repair system for base mismatches, to identify and correct mispaired bases occasionally produced by DNA polymerases. While it has been known that defects in mismatch repair promote carcinogenesis, mutations in replicative DNA polymerases driving tumorigenesis in mismatch repair proficient cells have only been recently identified. Here, I report the interrogation of twelve such DNA polymerase mutations for their potential to alter genetic information and contribute to genomic instability using the budding yeast *Saccharomyces cerevisiae* as model system. Of all the polymerase mutations tested, a subset caused significant increases in mutation accrual, and a shift in the observed mutation patterns/signatures. Most intriguingly, I observed that these increases are more severe than those caused by mutations disrupting the proofreading activity of the corresponding DNA polymerase, with my results further indicating that in some cases the high mutagenic potential depends on the proofreading activity. These strong increases in mutation rates do not likely result from inhibition of mismatch repair, as combination of these mutations with loss of mismatch repair factors results in synthetic sickness or lethality. My results point to these DNA polymerase mutations as driving extensive alterations of the genetic information, and are consistent with them being drivers of colorectal and endometrial cancer. Future work will be required to determine the exact mechanisms by which these mutations impair the fidelity of DNA replication.

Contents

Contents	ix
List of Figures	xv
List of Tables	xix
1 Genomic integrity and instability	1
1.1 Genome stability and maintenance	1
1.1.1 Genome replication	2
1.1.1.1 Structure of DNA, semiconservative replication and prokaryotic replication	2
1.1.1.2 Replication initiation and prevention of re-replication in eukaryotes	5
1.1.1.3 DNA replication in eukaryotes	14
1.1.1.4 DNA polymerases	21
1.1.2 DNA repair and Translesion Synthesis	31
1.1.2.1 Direct Damage Reversal	31
1.1.2.2 Damage to one strand of the DNA	32
1.1.2.3 Double stranded breaks (DSBs) in the DNA	37
1.1.2.4 Translesion synthesis (TLS)	39
1.1.2.5 Pausing the cell cycle: checkpoints	40
1.1.3 Dividing up the genome: chromosome segregation	41
1.2 Genome variation	42
1.2.1 Large-scale genomic variation	42
1.2.1.1 Whole-genome, segmental and gene duplications	44
1.2.1.2 Aneuploidy	46
1.2.1.3 Chromosomal translocation and chromoanagenesis	50

1.2.1.4	Mobile elements	54
1.2.1.5	Exon/domain shuffling	54
1.2.1.6	Acquisition of foreign DNA	58
1.2.2	Small-scale mutations	58
1.2.2.1	Point mutation instability (PIN)	58
1.2.2.2	Small insertions/deletions (INDELS)	60
1.3	Causes of mutations	61
1.3.1	Endogenous causes of mutation	61
1.3.2	Exogenous causes of mutations	65
1.4	Mutational processes and human disease	68
1.4.1	DNA repair deficiencies	68
1.4.2	Cancer	69
1.4.3	Mutational signatures	70
1.4.4	DNA polymerase defects in cancer	72
1.5	DNA Sequencing	72
1.6	The budding yeast <i>Saccharomyces cerevisiae</i> as a model organism	80
2	Analysis of cancer-associated polymerase mutations	85
2.1	Introduction	87
2.2	Identification of polymerase mutations	89
2.2.1	Literature search for DNA polymerase mutations in cancer	89
2.2.2	Query of COSMIC database, discarding single nucleotide polymorphisms and unconserved residues	89
2.3	Generation and propagation of polymerase mutants in <i>S. cerevisiae</i>	96
2.3.1	Constructing single mutant polymerase strains	96
2.3.2	Mutation accumulation experiment: Propagation of single mutant polymerase strains	97
2.3.2.1	Single-colony bottleneck propagation of mutant polymerase strains	100
2.3.2.2	Population bottleneck propagation of mutant polymerase strains	100
2.4	Establishing sequence analysis practices	102
2.4.1	Automating genomic DNA extraction and whole-genome sequencing of <i>Saccharomyces cerevisiae</i> strains	102
2.4.2	Establishing sequencing analysis protocols for the identification of SNVs and INDELS	103

2.4.3	Testing analysis protocol on <i>Saccharomyces cerevisiae</i> genetic screens	108
2.4.4	Applying analysis protocols to mouse genetic screens	113
2.4.5	Establishing a sequencing analysis protocols for large genomic changes	120
2.4.6	Analysing repetitive DNA regions in the yeast genome	122
2.5	Summary	131
3	Analysis of populations of <i>S. cerevisiae</i> strains carrying simple polymerase mutations	135
3.1	Introduction	135
3.2	Increased mutation rates for strains heterozygous diploid: <i>pol2-P301R</i> , <i>pol2-S312F</i> , <i>pol2-L439V</i> , <i>pol2-M459K</i> and <i>pol3-S483N</i>	136
3.2.1	Increased number of single-nucleotide variants for a subset of polymerase variants	136
3.2.2	Single-nucleotide variants in haploid polymerase mutant strains . . .	138
3.2.3	<i>pol2</i> mutants grow at a similar rate to wild type strains	141
3.2.4	Correlation of mutation rate estimates with mutations accrual	141
3.3	Patterns of single-nucleotide variants	144
3.4	Geographical mutation patterns	151
3.5	Large-scale variation: aneuploidy, CNVs and rDNA copy number	157
3.6	No increase in INDELs (compared to MMR mutants)	158
3.7	Summary	163
4	Polymerase mutations in mammalian systems and in combination with other mutations	167
4.1	Introduction	168
4.2	Synthetic lethality with mismatch repair deficiency	168
4.3	Epistatic relationship of mutations with exonuclease deficiency	170
4.4	Observed mutagenesis in <i>pol2-P301R</i> strains is not due to increased participation of Pol ζ in DNA replication	171
4.5	Examining polymerase mutations in other organisms	176
4.5.1	The <i>Pole</i> and <i>Pold1</i> mutations in mouse models	176
4.5.2	Human <i>POLE</i> P286R mutant cell lines	178
4.6	Summary	178
5	Discussion and future directions	181
5.1	Whole-genome sequencing as a flexible tool to address problems in cell biology	181

5.2	Polymerase mutations as drivers of mutagenesis	182
5.3	Future directions	187
6	Materials and Methods	189
6.1	Growth Medium	189
6.1.1	<i>Escherichia coli</i> Growth Media	189
6.1.2	<i>Saccharomyces cerevisiae</i> Growth Media	190
6.2	Other solutions	193
6.3	Microbial Strains	195
6.3.1	<i>Escherichia coli</i> strains	195
6.3.2	<i>Saccharomyces cerevisiae</i> strains	196
6.4	Oligonucleotides	198
6.5	Solutions	201
6.6	Protocols	202
6.7	Automated serial propagation platform	207
6.8	Illumina sequencing	207
6.9	Sequencing analysis	207
6.9.1	Quality control of DNA sequencing	207
6.9.2	Alignment of sequencing reads to the reference genome	208
6.9.3	Variant Calling of SNPs and INDELS, Annotation and Filtering	208
6.9.4	Extracting mutational signatures	208
6.9.5	Scripts written for this work	208
6.9.6	Step-by-step workflow of variant analysis	211
	References	215
A	List of Abbreviations	309
B	Supplementary Tables, Electronic Files and Articles Published	313
B.1	Supplementary figures, tables and notes	313
B.1.1	Software tools and parameters used	313
B.1.1.1	Software tools and parameters used for simulated genomes and capillary sequencing analysis	313
B.1.1.2	Software tools and parameters used for sequencing analysis of <i>S. cerevisiae</i>	314
B.1.2	Strains used in mutation accumulation (MA) experiments experiments	315

B.1.2.1	Manual propagation of strains heterozygous diploid for candidate polymerase mutations	315
B.1.2.2	Automated propagation of strains haploid and heterozygous diploid for candidate polymerase mutations	315
B.1.3	6-Thioguanine suppressor screen of haploid mouse cells	316
B.1.4	Custom filters for DNA sequencing Filters	316
B.2	Electronic files of supplementary information	317
B.2.1	Supplementary files for the mouse synthetic lethality screens	317
B.2.1.1	6TG_mouse_Sup1.xlsx	317
B.2.1.2	6TG_mouse_Sup2.xlsx	317
B.2.1.3	6TG_mouse_Sup3.xlsx	317
B.2.1.4	6TG_mouse_Sup4.xlsx	318
B.2.2	Supplementary files for the mouse synthetic lethality screens	318
B.2.2.1	MA_SampleNames.pdf	318
B.2.2.2	S1-3.experiment_merge.vcf	318
B.2.2.3	S4.experiment_merge.vcf	318
B.2.2.4	S5.experiment_merge.vcf	318
B.3	Articles published during my PhD	318

List of Figures

1.1	Structure of DNA	3
1.2	Replication initiation in <i>E. coli</i>	4
1.3	Lagging strand DNA synthesis in <i>E. coli</i>	5
1.4	Overlapping replication cycles in <i>E. coli</i>	6
1.5	The eukaryotic cell cycle	7
1.6	Licensing of eukaryotic origins of replication	10
1.7	Degradation of Cdt1 during the cell cycle and in response to DNA damage	13
1.8	Regulation of Cdt1 by association with Geminin	14
1.9	Structure of DNA polymerase δ and DNA polymerase ϵ	18
1.10	Structure and representation of replicative DNA polymerases	22
1.11	Comparison of primer-template DNA bound to four DNA polymerases.	23
1.12	Mechanism of DNA polymerization	25
1.13	Replication fidelity	28
1.14	Fidelity of different DNA polymerases	29
1.15	Base excision repair (BER) of oxidized DNA base lesions	33
1.16	Nucleotide excision repair (NER)	35
1.17	A general outline of the DNA damage signal transduction pathway	41
1.18	The mitotic spindle	43
1.19	Gene duplications: the duplication-degeneration (DDC) model	46
1.20	Uniparental Disomy - A special case of aneuploidy	49
1.21	Consequences of chromosomal translocations	51
1.22	Chromothripsis	53
1.23	Chromoplexy and Chromothripsis	54
1.24	Classes of DNA transposons	55
1.25	Blood clotting cascade	57
1.26	Transitions and Transversions	60

1.27	Codon table	61
1.28	Replication slippage	62
1.29	Unequal crossovers result in chromosome rearrangements	64
1.30	Mutational signatures leave their marks on the genome	71
1.31	Summary of known mutational signatures	73
1.32	Early sequencing techniques: Gilbert and Sanger	74
1.33	Sequencing by synthesis	76
1.34	Solid-phase bridge amplification and sequencing by synthesis (Illumina)	77
1.35	Third-generation Sequencing Techniques	79
1.36	Life cycle of the budding yeast <i>Saccharomyces cerevisiae</i>	81
1.37	The budding yeast mating type locus	83
2.1	Methodology of the work carried out during my PhD	86
2.2	Locations of DNA polymerase mutations within the proteins	91
2.3	Prevalence of polymerase mutations of interest in COSMIC	92
2.4	Alignment of polymerase residues of interest to the yeast proteins	95
2.5	Rationale for plasmid construction	98
2.6	Exonuclease domains conserved in B family polymerases	99
2.7	Mutation accumulation experiment: manual propagation of mutated <i>S. cerevisiae</i> strains	101
2.8	DNA extracted using a high-throughput protocol produces high quality sequencing data	104
2.9	The number of variants in W303 strains compared to the S288c reference genome	105
2.10	Experimental strategy to identify acquired mutations	107
2.11	Sequencing analysis identifies mutations capable of suppressing <i>sae2Δ</i> DNA damage hypersensitivity	109
2.12	Mutations in <i>SIR3</i> and <i>SIR4</i> identified as the cause for the hypersensitivity of <i>tof1Δ</i> cells to camptothecin	111
2.13	Generation of mutagenized libraries	112
2.14	Identification of suppressor mutations	114
2.15	Using multiple controls and multiple variant callers to enrich for high confidence variants	116
2.16	Clinically-relevant and newly-identified suppressor mutations	117
2.17	EMS mutagenic action	119

2.18	Relationship between read pairs and structural variants	121
2.19	An overview of the SVMerge pipeline	123
2.20	Visualising aneuploidy in budding yeast	124
2.21	Ambiguities in read mapping	125
2.22	Next-generation sequencing data can be used to estimate rDNA copy number reliably	128
2.23	Next-generation sequencing data could also be used to assess Ty element copy number	130
3.1	Number of single-nucleotide variants per sample in <i>pol2</i> mutant strains	137
3.2	Number of single-nucleotide variants per sample in <i>pol3</i> mutant strains	139
3.3	Number of single-nucleotide variants per line per haploid genome for selected haploid and heterozygous diploid <i>pol2</i> mutant strains	140
3.4	Growth of <i>S. cerevisiae</i> mutant strains in rich medium	142
3.5	Correlation of mutation rate estimates and mutation accrual for <i>pol2</i> mutant strains	143
3.6	Single-nucleotide variant patterns	145
3.7	Single-nucleotide variant patterns adjusted to frequencies of trinucleotides in <i>S. cerevisiae</i>	146
3.8	SomaticSignatures: Determining the numbers of signatures	147
3.9	2-3 signatures are determined using Non-negative matrix factorization	149
3.10	Contribution of the signatures to the variant pattern	150
3.11	EMu: Validating Signature Analysis	152
3.12	No observed clustering of mutations acquired by <i>pol2-P301R</i> strains	153
3.13	Mutations falling inside and outside of genes in heterozygous diploid polymerase mutant strains	154
3.14	Percentage of mutations within genes in haploid strains	156
3.15	Number of total aneuploidy and segmental insertions/duplications identified	157
3.16	Example of aneuploidy in <i>S. cerevisiae</i>	159
3.17	Example of segmental deletions and amplifications in <i>S. cerevisiae</i>	160
3.18	rDNA copy number changes in polymerase mutants	161
3.19	No increase in the number of INDELs detected per sample across strains	162
3.20	Mutation accrual in strains with mismatch repair deficiencies	163
4.1	Tetrad dissection to generate double mutants and detect synthetic lethality	169

4.2	The mutagenesis observed in strong mutator strains is partially rescued by mutating critical residues in the exonuclease domain active site	172
4.3	Synergistic effects on mutation number between <i>rev3Δ</i> and <i>pol2-P301R</i> . . .	174
4.4	Mutational patterns observed in <i>pol2-P301R</i> cells and <i>pol2-P301R rev3Δ</i> cells are highly similar	175
4.5	Constructs used for conditional knock-in mutations in mice	177

List of Tables

1.1	Eukaryotic replicative DNA polymerases	15
1.2	Families of DNA polymerases	22
1.3	Error rates of DNA polymerases from different families	26
1.4	Incidence of aneuploidy during development	47
1.5	The origin of human trisomy	50
1.6	Standard Nomenclature for <i>S. cerevisiae</i> genetics using <i>POL2</i> as an example.	83
1.7	A selection of Nobel Prizes awarded for work using <i>S. cerevisiae</i> as a model organism.	84
2.1	Polymerase exonuclease domain mutations in <i>S. cerevisiae</i>	88
2.2	Genomic locations of mutations in DNA polymerases in different human genome assemblies	90
2.3	Checking DNA polymerase mutations for common variants	93
2.4	Polymerase mutations identified from the literature with predicted consequences	94
2.5	Budding yeast equivalents of human DNA polymerase mutations of interest .	96
2.6	Haploid copy number of rDNA repeats across Eukaryotic species	126
3.1	Mutation number fold change of <i>pol2</i> haploid and heterozygous diploid mutant strains when compared to the <i>POL2</i> strain	141
3.2	Estimates of mutation rate increases using resistance to Thialysine	142
4.1	Synthetic lethality of polymerase mutants and mismatch repair deficiency . .	171

