Chapter 3

Identifying and characterising germline hypermutators

3.1 Introduction

Germline mutagenesis is a major source of all genetic variation and drives the process of human evolution. The human mutation rate is not a constant, the rate of mutation varies both within and between individuals. For example, parental age explains a large proportion of variance between individuals [111, 60]. The factors influencing variation in germline mutation rates are discussed extensively in Chapter 1. While we have started to explain the general distribution of mutations, little is known about rare outliers with extreme mutation rates. Germline hypermutators are defined here as individuals with an unusually large germline mutation rate. This may be due to environmental factors or could have a genetic basis.

The impact of environmental mutagens has been well established in the soma but this is not as well understood in the germline. Environmental exposures in parents can influence the number of mutations transmitted to offspring. For example, ionising radiation has a mutagenic effect on the germline and offspring of irradiated parents are observed to have an increased number of *de novo* mutations [213]. By comparing nearby populations in regions with differing levels of natural radiation, it has also been observed that radiation increases the rate of mutation in mitochondrial DNA[51]. Exposure to ionising radiation has been confirmed to increase the paternal germline mutation rate in mice *in vivo* [2]. Tobacco smoke has also been hypothesised to increase the number of mutations in the paternal germline. Exposure to tobacco smoke has been observed to increase the number of mutations, specifically at short tandem repeats, in spermatogonial stem cells in mice [242, 139, 11].

Individual mutation rate can also be influenced by genetic background. With regards to somatic mutation, thousands of inherited germline variants have been shown to predispose individual cancer risk [82, 46, 81]. For example, Li-Fraumeni syndrome (LFS) is an autosomal dominant disorder which leads to a large increased risk of early-onset cancers due to inherited germline variants in the gene *TP53*. This elevated rate of mutation, and resulting increased cancer risk, also appears to extend to the germline as families with LFS are also highly enriched for *de novo* CNVs compared to the healthy population [200]. Homozygous and heterozygous germline PTVs in *NTHL1*, a gene involved in the base excision repair pathway, are another example; these variants have been shown to predispose individuals to colorectal cancer[231]. This raises the question of whether other known cancer predisposing variants also impact the rate of mutation in the germline. Pathogenic variants in the gene *MBD4* have been shown to elevate cancer risk, primarily for colorectal cancers. A recent study investigating genetic determinants of cancer identified that patients with germline heterozygous protein truncating variants (PTVs) in the *MBD4* gene have a four-fold increase in C>T mutations at CpG dinucleotides in their tumours [214]. This result agrees with previous studies that showed that Mbd4 knockout (*Mbd4* -/-) in mice was found to accelerate tumorigenesis and mutation analysis of these tumours showed a three-fold increase in the number of C>T mutations at CpGs [147, 232]. *MBD4* is known to play a role in base-excision repair. Specifically it encodes a DNA glycolysase that removes thymidines from T:G mismatches at methyl-CpG sites [79]. Many of these variants associated with elevated cancer risk are in genes encoding components of DNA repair pathways which, when impaired, lead to an increased number of somatic mutations. However it is not known whether variants in known somatic mutator genes can influence germline mutation rates. For example, the CpG mutation signature (Signature 1 in the catalogue of somatic mutations in cancer (COSMIC)) accounts for 16% of *de novo* mutations in the germline which raises the question of whether *MBD*4 PTV germline carriers also show an increased number of C>T germline mutations in their offspring.

There have been several examples where genetic background has been shown to impact the germline mutation rate of a variety of types of genetic variation. As mentioned in Chapter 1, the mutation rate of STRs are known to be affected by both the length of the repeat unit and the repeat number [84, 70, 210]. Variants have also been shown to impact the mutation rate of minisatellites. Through the analysis of single sperm, a variant nearby to minisatellite MS32 has been shown to impact its mutation rate in the male germline[150]. With respect to translocations, an analysis of a recurrent chromosomal translocation demonstrated that the breakpoints occur in the center of a region of palindromic AT-rich repeats (PATRRs). The presence of PATRR-like sequences was also identified at other translocation breakpoints

which suggests that these regions are susceptible to double strand breaks and likely increase the rate of translocation[103]. A recent study has also attempted to identify variants associated with overall germline mutation rate by leveraging a haplotype based approach[197]. This was based on the idea that when a variant increases the germline mutation rate it results in a subset of haplotypes that are more divergent than others at that locus. With this method, the authors identified several candidate mutator loci and found these were enriched for their proximity to genes associated with DNA repair.

An elevated germline mutation rate can have a significant impact on the health of subsequent generations. Increasing germline mutation rate results in an increased risk of offspring being born with a congenital disorder caused by a *de novo* mutation. There are also long-term effects of mutation rate differences. The phenotypic effects of mutation accumulation were examined in homozygous *Pold1* knockout mice[219]. These mice have an ~17 fold increased germline mutation rate due to the lack of the proofreading activity of DNA polymerase delta. Abnormal phenotypes were observed ~4 times more than in controls and after several generations the *Pold1* deficient mice had much lower reproduction rates with lower pregnancy rates, lower survival rates and smaller litter sizes[219]. A recent study examined a set of 41 multi-generational families and observed that a higher germline mutation rate is correlated with higher all-cause mortality and reduced fertility in women [25]. The decrease in fertility is suggested to be due to germline mutation accumulation while the shorter lifespan is hypothesised to be driven by a correlation between germline and somatic mutation rates.

3.1.1 Chapter Overview

In this chapter, I used large cohorts of exome and whole-genome sequenced parent-offspring trios in order to investigate germline hypermutators and the impact of rare genetic variation on individual mutation rates. I focussed on SNV mutation rates specifically in this chapter and tackled this problem from two different angles: a genotype-driven approach and subsequently a phenotype-driven approach.

The genotype-driven approach focused on whether variants in DNA repair genes impact germline mutation rates. For this, I interrogated variants in an established cancer mutator gene, *MBD4*, to investigate if they have a similar effect in the germline.

For the phenotype-driven approach, I aimed to identify germline hypermutators and sought genetic causes for this trait. Germline hypermutators are individuals who have an elevated germline mutation rate and so are likely to have children with an unusually large number of *de novo* mutations (DNMs). A large number of DNMs increases the chance of having a dominant disorder, therefore cohorts of children with rare disease are better powered to identify germline hypermutators. A CNV mutator phenotype has been previously

identified in a cohort of patients with neurodevelopmental phenotypes [131]. I then identified the fraction of variance in germline mutation rate that can be explained by parental age and hypermutation and explore where the remaining fraction of variance may lie by performing analyses of the impact of rare damaging variants in DNA repair genes on an individual's germline mutation rate across the 100kGP dataset.

In this chapter I also take advantage of the large number of WGS trios available in the 100,000 Genomes Project (100kGP) to examine other sources of mutation rate variability such as the effects of parental age.

3.1.2 Contributions

I would like to acknowledge Jan Korbel who initially approached us regarding the possibility that *MBD4* may impact the germline. The *de novo* filtering for the Genomics England dataset was done in collaboration with Patrick Short, Chris Odhams and Loukas Moutsianas. Petr Danecek collated the variants in DNA repair genes in the 100kGP dataset which I used for the analysis looking at the impact of these variants on germline mutation rate. James Stephenson, a post-doc in the group, helped analyse the role of a variant within the context of the protein structure. Raheleh Rahbari and Matthew Neville advised on the mutational signature extraction. This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. All of this work was done under the supervision of Matthew E. Hurles

3.2 Methods

3.2.1 *De novo* calling and filtering in paternal *MBD4* PTV carriers

I identified 14 individuals in the DDD study whose father had a heterozyous protein-truncating variant (PTV) in the gene *MBD4*. This included five stop gained variants, 7 frameshift variants and 2 variants within splice donor sites. There were 11 unique variant sites. All variants were examined in the Integrative Genomics Viewer (IGV) and did not appear to be false positive sites. These 14 parent offspring trios (42 samples) were submitted for whole-genome sequencing PCR-free at >30x mean coverage of Illumina 150 bp paired

end reads via Sanger pipelines. One sample failed at the library creation phase and there was not enough sample left to resubmit. This left me with 13 trios for analysis. The reads were mapped with bwa (v0.7.16). I used GATK (v3.5) HaplotypeCaller best practices to generate a multi-sample VCF and from this created parent-offspring trio VCFs and the input files needed for DNM calling. DNMs were called in these trios using bcftool's trio-dnm. This was a change from how DNMs were called previously in DDD using DeNovoGear (the DNM caller described in Chapter 2 and 4), which no longer functioned efficiently on the Sanger compute cluster. The filters selected here were chosen after inspecting distributions of variant allele fraction (VAF) and examination of these putative DNMs using the Integrative Genomics Viewer (IGV) to estimate true positive rates.

Filters applied:

- Removed DNMs with trio-dnm 'DNM' score < 50, this is the score outputted by trio-dnm. It is the log of the probability of inheriting the variant calculated directly from the genotype likelihoods.
- Removed DNMs that fell within known segmental duplication regions as defined by UCSC (http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab)
- Removed DNMs that fell in highly repetitive regions (http://humanparalogy.gs.washington.edu/build37/
- Removed DNMs with gnomAD allele frequency > 0.01
- Read depth (RD) of child > 7, mother RD > 5, father RD > 5
- Alternative allele depth of child >2
- Fisher exact test on strand bias p-value > 10^{-3}
- Removed DNMs if child RD >98 [173]
- Remove DNMs with >1 alternative read in either parent
- Remove DNMs with > 0.1 parental VAF in either parent
- Test to see if VAF in child is significantly greater than the error rate at that site as defined by error sites estimated using Shearwater. This was calculated by Inigo Martincorena [57].

This resulted in 1,690 DNMs after this stage of filtering (~130 per trio). I examined all of these with IGV and annotated them with whether these appeared true. This resulted in a total of 877 DNMs across the 13 trios (~67 DNMs per person). Due to the small number of trios I

examined here I did not refine my filters based on this annotation but plan to do so in order to improve DNM calling with bcftools for DDD trios in the future.

3.2.2 DNM filtering in 100,000 Genomes Project

I analysed DNMs called in 13,949 parent offspring trios from 12,609 families from the rare disease programme. Sequencing and variant calling for these families was performed via the Genomics England rare disease analysis pipeline which has been extensively documented (https://cnfl.extge.co.uk/display/GERE/10.+Further+reading+and+documentation). DNMs were called by the Genomics England Bioinformatics team using the Platypus variant caller[178]. Filtering of the DNMs was done in collaboration with Patrick Short, Chris Odhams and Loukas Moutsianas. These were selected to optimise various properties including the number of DNMs per person being approximately what we would expect, the distribution of the VAF of the DNMs to be centered around 0.5 and the true positive rate of DNMs to be sufficiently high as calculated from examining IGV plots. The filters applied were as follows:

- Genotype is heterozygous in child (1/0) and homozygous in both parents (0/0)
- Child RD >20, Mother RD>20, Father RD>20
- Remove variants with >1 alternative read in either parent
- VAF>0.3 and VAF<0.7 for child
- Remove SNVs within 20 bp of each other. While this is likely removing true MNVs, the error mode was very high for clustered mutations.
- Removed DNMs if child RD >98 [173]
- Removed DNMs that fell within known segmental duplication regions as defined by UCSC (http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab)
- Removed DNMs that fell in highly repetitive regions (http://humanparalogy.gs.washington.edu/build37/dat
- For DNM calls that fell on the X chromosome these slightly modified filters were used:
  - For DNMs that fell in PAR regions, the filters were unchanged from the autosomal calls apart from allowing for both heterozygous (1/0) and hemizygous (1) calls in males
  - For DNMs that fell in non-PAR regions the following filters were used:

    * For males: RD>20 in child, RD>20 in mother, no RD filter on father
    * For males: the genotype must be hemizygous (1) in child and homozygous in mother (0/0)
    * For females: RD>20 in child, RD>20 in mother, RD>10 in father

3.2.3 DNM filtering for possible DDD hypermutated individuals

Nine trios were selected from the DDD cohort where the offspring has an unusually large number of exome DNMs and submitted along with their parents for whole-genome sequencing PCR-free at >30x mean coverage of Illumina 150bp paired end reads via Sanger pipelines. Reads were mapped with bwa (v0.7.15). DNMs were called from these trios using DeNovoGear[174] (note this analysis was done over a year prior to the *MBD4* analysis which is why DeNovoGear was used here) and were filtered as follows:

- Read depth (RD) of child > 10, mother RD > 10, father RD > 10
- Alternative allele read depth in child >2
- Filtered on strand bias across parents and child (p-value >0.001, Fisher's exact test)
- Removed DNMs that fell within known segmental duplication regions as defined by UCSC (http://humanparalogy.gs.washington.edu/build37/data/GRCh37GenomicSuperDup.tab)
- Removed DNMs that fell in highly repetitive regions (http://humanparalogy.gs.washington.edu/build37/
- Allele frequency in gnomAD < 0.01
- VAF <0.1 for both parents
- Removed mutations if both parents have >1 read supporting the alternative allele
- Test to see if VAF in child is significantly greater than the error rate at that site as defined by error sites estimated using Shearwater. This was calculated by Inigo Martincorena [57].
- Posterior probability from DeNovoGear > 0.00781 [41].
- Removed DNMs if child RD >200 [173].

After applying these filters, this resulted in 1,367 DNMs. I then inspected all of these DNMs using IGV and removed those that appeared to be false positives. I had a final set of 916 DNMs across the 10 trios.

3.2.4 Parental phasing of *de novo* mutations

To phase the DNMs in both 100kGP and DDD I used a custom script which used the following read-based approach to phase a DNM. I first searched for heterozygous variants within 500 bp of the DNM that was able to be phased to a parent (so not heterozygous in both parents and offspring). I then examined the reads or read pairs which included both the variant and the DNM and counted how many times I observe the DNM on the same haplotype of each parent. If the DNM appears exclusively on the same haplotype as a single parent then that was determined to originate from that parent. I discarded DNMs that had conflicting evidence from both parents. This code is available on GitHub (https://github.com/queenjobo/PhaseMyDeNovo).

3.2.5 Analysis of effect of parental age on germline mutation rate

To assess the effect of parental age on germline mutation rate I ran the following regressions. On all (unphased) DNMs I ran two separate regressions for SNVs and indels. I fitted the following model using a Poisson generalized linear model (GLM) with an identity link where Y is the number of DNMs for an individual:

$$E(Y) = \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age \tag{3.1}$$

For the phased DNMs I fit the following two models using a Poisson GLM with an identity link where $Y_{maternal}$ is the number of maternally derived DNMs and $Y_{paternal}$ is the number of paternally derived DNMs:

$$E(Y_{paternal}) = \beta_0 + \beta_1 paternal\_age$$
$$E(Y_{maternal}) = \beta_0 + \beta_1 maternal\_age$$

3.2.6 Identifying hypermutation in 100kGP

To identify hypermutated individuals in the 100kGP cohort I first wanted to regress out the effect of parental age by fitting the following ordinary linear regression model:

$$E(Y) = \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age \tag{3.2}$$

I then looked at the distribution of the studentized residuals and then, assuming these followed a t distribution with N-2-1 degrees of freedon, calculated a t-test p-value for each

individual. I separately did the same approach for the number of indels, except in this case *Y* would be the number of *de novo* indels.

3.2.7 Extraction of mutational signatures

I extracted mutational signatures from maternally and paternally phased DNMs as well as from the 15 hypermutated individuals that I identified. I did this using SigProfiler (v1.0.5) and these signatures are extracted and subsequently mapped on to COSMIC mutational signatures (COMIC v89, Mutational Signature v3) [12, 212].

3.2.8 Defining set of genes involved in DNA repair

I compiled a list of DNA repair genes which were taken from an updated version of the table in Lange et al, Nature Reviews Cancer 2011 (https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html) [116]. These are annotated with the pathways they are involved with (eg. nucleotide-excision repair, mismatch repair). I defined 'rare' variant as those with an allele frequency of <0.001 for heterozygous variants and those with an allele frequency of <0.01 for homozygous variants in both 1000 Genomes as well as across the 100kGP cohort.

3.2.9 Estimating the fraction of variance explained

To estimate the fraction of germline mutation variance explained by several factors, I fit the following Poisson GLMs with an identity link. I would expect data quality to correlate with the number of DNMs detected so to reduce this variation I used a subset of the 100kGP dataset which had been filtered on some base quality control (QC) metrics by the Bioinformatics team at GEL:

- cross-contamination < 5%
- mapping rate > 75%
- mean sample coverage > 20
- insert size <250

I then included the following variables to try and capture as much of the residual measurement error which may also be impacting DNM calling. In brackets I have given the corresponding variable names used in the models below:

- Mean coverage for the child, mother and father (*child_mean_RD*, *mother_mean_RD*, $father\_mean\_RD$)
- Proportion of aligned reads for the child, mother and father (*child_prop_aligned*, *mother_prop_aligned* , $father\_prop\_aligned$)
- Number of SNVs called for child, mother and father (*child_snvs*, *mother_snvs*, $father\_snvs$)
- Median VAF of DNMs called in child (*median_VAF*)
- Median 'Bayes Factor' as outputted by Platypus for DNMs called in the child. This is a metric of DNM quality (*median_BF*).

The first model I fit only included parental age:

$$E(Y) = \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age$$

The second model also included data quality variables as described above:

$$\begin{aligned} E(Y) =& \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age + \\ & \beta_3 child\_mean\_RD + \beta_4 mother\_mean\_RD + \beta_5 father\_mean\_RD + \\ & \beta_6 child\_prop\_aligned + \beta_7 mother\_prop\_aligned + \beta_8 father\_prop\_aligned + \\ & \beta_9 child\_snvs + \beta_{10} mother\_snvs + \beta_{11} father\_snvs + \\ & \beta_{12} median\_VAF + \beta_{13} median\_BF \end{aligned}$$

The third model included a variable for excess mutations in the 14 confirmed hypermutated individuals (*hm_excess*) in the 100kGP dataset. This variable was the total number of mutations subtracted by the median number of DNMs in the cohort (65), $Y_{hypermutated} - median(Y)$ for these 14 individuals and 0 for all other individuals.

$$\begin{aligned} E(Y) =& \beta_0 + \beta_1 paternal\_age + \beta_2 maternal\_age + \\ & \beta_3 child\_mean\_RD + \beta_4 mother\_mean\_RD + \beta_5 father\_mean\_RD + \\ & \beta_6 child\_prop\_aligned + \beta_7 mother\_prop\_aligned + \beta_8 father\_prop\_aligned + \\ & \beta_9 child\_snvs + \beta_{10} mother\_snvs + \beta_{11} father\_snvs + \\ & \beta_{12} median\_VAF + \beta_{13} median\_BF + \beta_{14} hm\_excess \end{aligned}$$

The fraction of variance (F) explained after accounting for Poisson variance in the mutation rate was calculated in a similar way to Kong et al using the following formula[111].

$$F = \frac{pseudo\text{-}R^2}{1 - \frac{\bar{Y}}{Var(Y)}}$$

I used McFadden's $pseudo\text{-}R^2$ as I was fitting a Poisson GLM. I also repeated these analyses fitting an ordinary least squares regression, as was done in Kong et al, using the R^2 from that and got comparable results. To calculate a 95% confidence interval I used a bootstrapping approach. I sampled with replacement 10,000 times and extracted the 2.5% and 97.5% percentiles.

Simulations to explore effect of non-random paternal age sampling

To look at the effect that non-random paternal age sampling has on the fraction of germline mutation rate explained I performed the following simulation:

I first simulated a random sample as follows 5,000 times:

- Randomly sample 78 trios
- Fit OLS of $E(Y) = \beta_0 + \beta_1 paternal\_age$
- Estimated fraction of variance (F) as described above

I then simulated a random sample as follows 5,000 times:

- Sample 78 trios as follows:
  - Sample $\frac{3}{4}$ of the 78 trios from the set of trios were paternal age falls into the top of bottom quartile (paternal age <29 or $\geq$37 years)
  - Sample $\frac{1}{4}$ of the 78 trios from those in the two middle quartiles (29$\leq$ paternal age < 37 years)
- Fit OLS of $E(Y) = \beta_0 + \beta_1 paternal\_age$
- Estimated fraction of variance (F) as described above

3.2.10 Analysis of contribution of rare variants in DNA repair genes

I fit 8 separate regressions to assess the contribution of rare variants in DNA repair genes. These were across three different sets of genes: variants in all DNA repair genes, variants

in a subset DNA repair genes known to be associated with BER, MMR, NER or a DNA polymerase and variants within this subset that have also been associated with a cancer phenotype. For this I downloaded all ClinVar entries as of October 2019 and searched for germline 'pathogenic' or 'likely pathogenic' variants annotated with cancer [115]. I tested both nonsynonymous and PTVs for each set.

To assess the contribution of each of these sets I created two binary variables per set indicating a presence or absence of a maternal or paternal variant for each individual and then ran a poisson regression for each subset including these as independent variables along with hypermutation status, parental age and QC metrics as described in the previous section.

3.3 Results

3.3.1 Examining the effect of PTVs in *MBD4* on germline mutation rate

To investigate genetic variants that may impact germline mutation rate I at first took a genotype-driven approach. I examined the effect of PTVs in the known cancer mutator gene *MBD4* which are associated with a three-fold elevated CpG>TpG mutation rate in tumours. The CpG signature should be seen in both maternally and paternally derived mutations however I would expect to have more power to detect this elevated mutation rate in paternal germlines due to the larger number of paternal mutations. To this end, I identified 13 paternal carriers of *MBD4* PTVs within the DDD study that had a sufficient amount of remaining sample for sequencing and whole genome sequenced them and their parents. DNMs were called and filtered as described in the methods and post-filtering the individuals had an average of 67 DNMs per person. This is not elevated compared to what we would expect under the null and no individual had a significantly large number of DNMs. The mutational spectra looked normal for every individual (Figure 3.1c) and, using the proportion of CpG>TpG mutation expected from previous studies in healthy trios[173], there was no significant increase in the number of CpG>TpG mutations ($p = 0.56$, χ^2-test, Figure 3.1a). The 95% confidence interval around the CpG mutation rate 'multiplier' is 0.90 to 1.22 (ratio of two proportions), so I can confidently exclude that there is more than a 22% increase in the CpG mutation rate. This demonstrates that *MBD4* PTVs are unlikely to have a similar effect in the germline as in the soma.

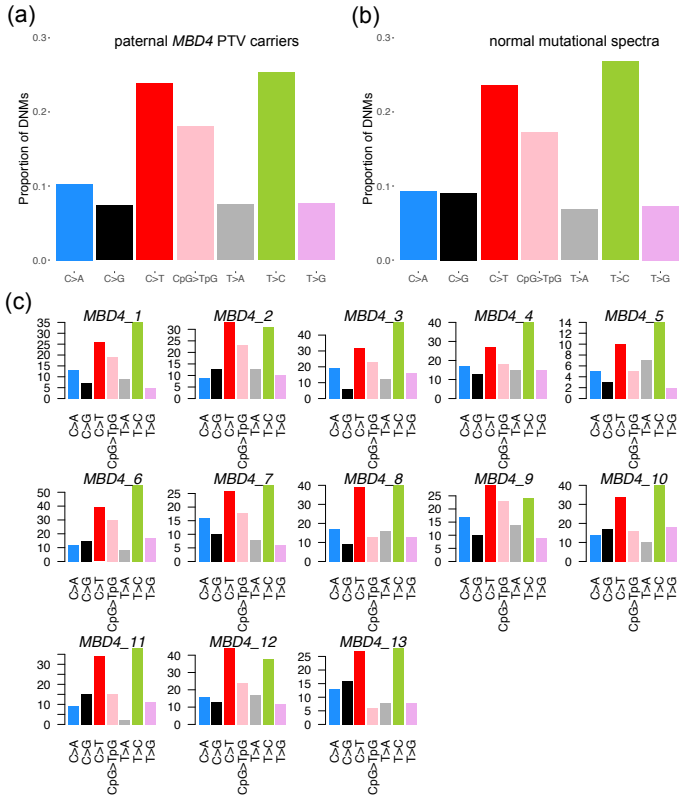

Fig. 3.1 Comparing the mutational Spectra of DNMs across the 13 paternal *MBD4* paternal PTV carriers (a) with the expected proportion of mutations (b) in each mutation type taken from Rahbari et al. [173] (c) The invididual mutational spectra demonstrating that no one individual has an elevated number of CpG>TpG mutations

3.3.2 Identifying germline hypermutators

For the phenotype driven approach I aimed to identify germline hypermutators. For this, I sought to identify offspring with an unusually large number of DNMs in exome-sequenced parent offspring trios in the DDD study and subsequently whole-genome sequenced trios in the rare disease cohort of the 100kGP. For the 100kGP data, this began with extensive DNM filtering that allowed me to explore additional properties of germline mutation variation including parental age. This was an important factor to account for in my downstream analyses. I was also able to explore differences in mutational spectra for maternally versus paternally derived DNMs.

Properties of *de novo* mutations in the 100kGP dataset

DNMs were called in 13,949 parent-offspring trios, across 12,609 families, as part of the rare disease cohort in the 100kGP dataset. After extensive filtering in collaboration with the

bioinformatics team at Genomics England Limited (GEL), this resulted in a total of 999,939 DNMs: 921,433 *de novo* SNVs (dnSNVs) and 78,506 *de novo* indels (dnIndels). IGV examination of 300 random SNVs and 250 random indels demonstrated 95% true positive rate for SNVs and 90% true positive rate for indels. The VAF distribution and mutational spectra of these mutations are as expected (Figure 3.2). The median number of DNMs per individual was 65 for SNVs and 5 for indels, the number of SNVs and indels looked normal apart from some extreme outliers (Figure 3.3).

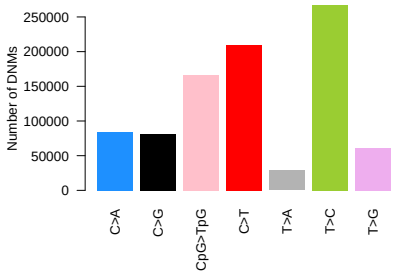

Fig. 3.2 Mutational Spectra of all DNMS called in the 100kGP cohort

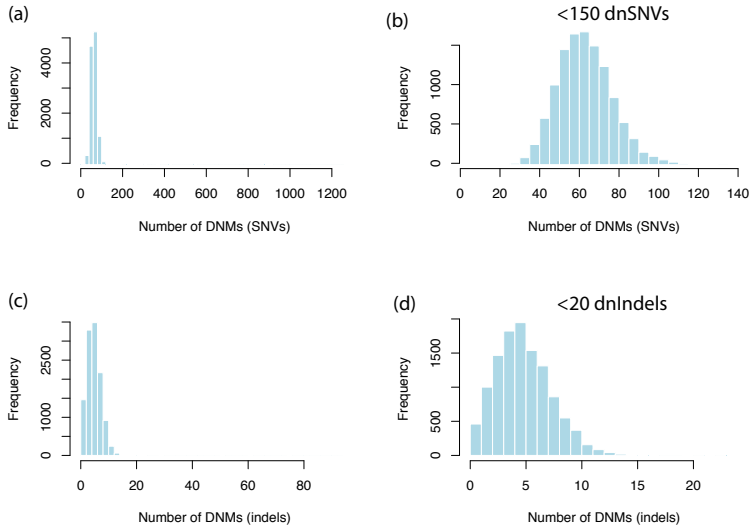

Fig. 3.3 Distribution of number of *de novo* SNVs for all individuals (a) and those with <150 DNMs (b). Distribution of number of *de novo* InDels per person for all individuals (c) and those with <20 indels (d)

Effect of parental age on germline mutation rate and parental differences in mutational spectra

To assess the effect of parental age on the germline mutation rate I ran a Poisson regression of the number of DNMs in the offspring on both maternal and paternal age at birth. This was done separately for SNVs and indels. Both paternal and maternal age were significantly associated with the number of *de novo* SNVs, I found an increase of 1.27 dnSNVs/year of paternal age (CI: 1.24-1.39, p $<10^{-300}$) and an increase of 0.35dnSNVs/year of maternal age (CI: 0.32-0.39, p = 2.8×10^{-80}) (Figure 3.4a). These estimates agree with previous results reported in the literature ([234, 98]).

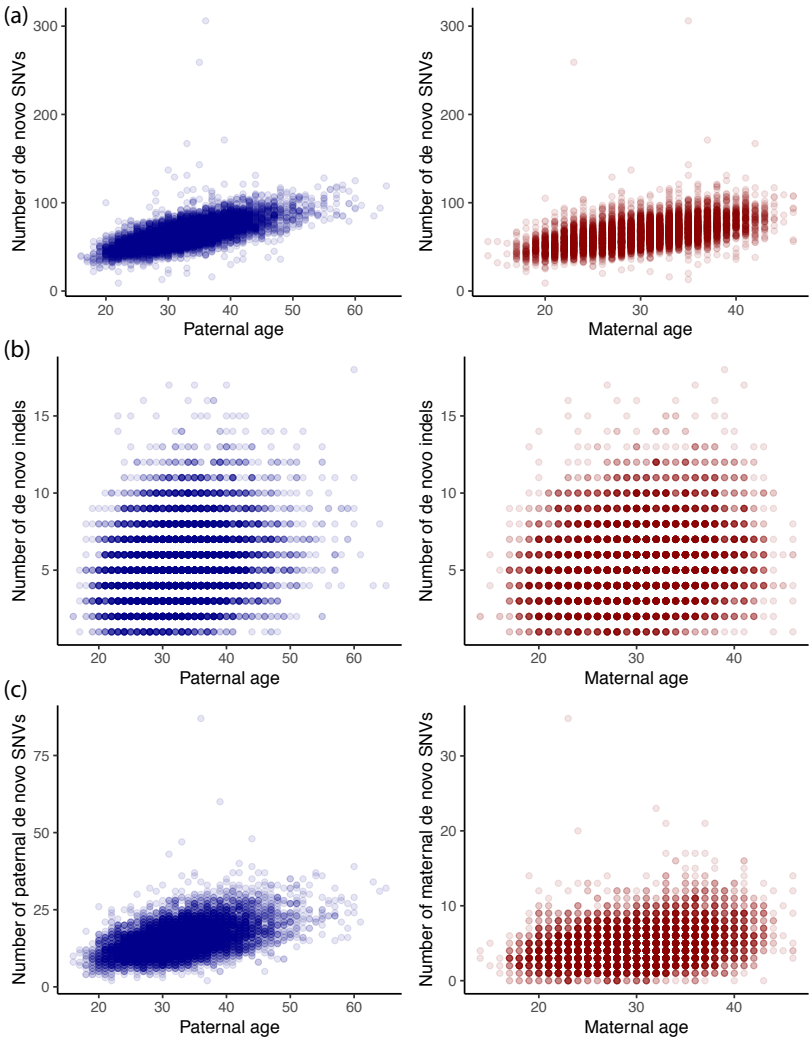

Fig. 3.4 Paternal and maternal age against the number of (a) dnSNVs, (b)dnInDels. (c) Paternal age against number of paternally phased dnSNVs and maternal age against number of maternally phased dnSNVs

I was able to phase 225,854 dnSNVs and the ratio of paternal to maternal DNMs was 3.29 across the dataset, 77% of phased DNMs were paternal in origin which agrees with previous studies [54, 173, 62]. I regressed the number of paternal mutations on paternal age and similarly the number of maternal mutations on maternal age. The effect estimates were not significantly different to the unphased results: 1.24 paternal dnSNVs/year of paternal age (CI:1.20-1.28, p $<10^{-300}$ and 0.38 maternal dnSNVs/year of maternal age (CI: 0.35-0.40, p = 1.6×10^{-211})(Figure 3.4c). Paternal and maternal age were also significantly associated with the number of dnIndels. I found that there was an increase of 0.078 dnIndels/year of paternal age (CI:0.068-0.087, p=1.96×10^{-64}) and a smaller increase of 0.021 dnIndels/year of maternal age (CI: 0.010-0.0031 p = 1.2×10^{-4})(Figure 3.4b). The ratio of paternal to maternal mutation increases for SNVs and indels were very similar, 3.7 for SNVs and 3.6 for indels.

Using the set of phased mutations I was also able to examine differences in properties between paternally and maternally derived DNMs. I found that the proportion of *de novo* mutations that phased paternally increased significantly with paternal age with a proportion increase of 0.0015 for every year of paternal age (p = 2.37×10^{-21}, Binomial regression) (Figure 3.5). This supports the idea that part of the paternal age effect is driven by replication errors as spermatogonial stem cells continue to divide after male puberty while female germ cells do not. However the effect size is small and the proportion of DNMs that phase paternally in the youngest fathers is ~0.75 and so replication errors alone do not fully explain the strong paternal bias.

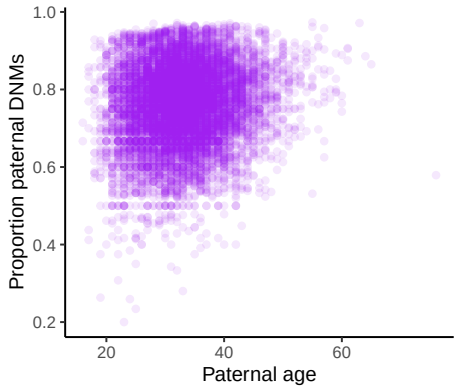

Fig. 3.5 Proportion of paternally phased DNMs against paternal age

I observed significant differences in the mutational spectra of paternally and maternally derived DNMs (Figure 3.6a). Maternally derived DNMs have a significantly higher proportion of C>T mutations while paternally derived DNMs have a significantly higher proportion of C>A,T>G and T>C mutations (p-values: 1.48×10^{-19},2.25×10^{-21},0.002, Binomial test).

These mostly agree with previous studies although the difference in T>C mutations was not previously significant [62]. To further understand the differences in the mutational profile, I extracted mutational signatures for maternally and paternally phased DNMs. These were then mapped on to known mutational signatures from COSMIC and found that the majority of the mutations could be explained by Signature 1 and 5 as has previously been observed in germline mutation (Figure 3.6b) [173]. I found that the proportion of mutations explained by Signature 1 was significantly greater in the paternal compared to maternal mutations although the difference was very slight (0.15 paternal vs 0.14 maternal, chi-sq test $p = 4.53 \times 10^{-6}$).

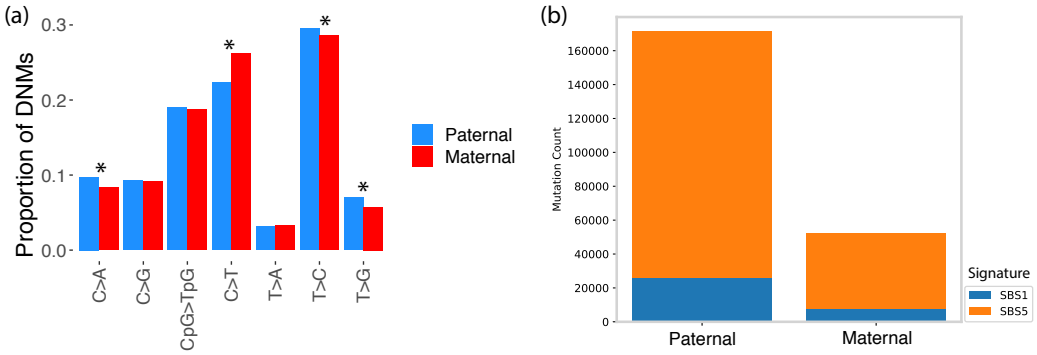

Fig. 3.6 (a) Mutational spectra for maternal vs paternal DNMs across 100kGP cohort. Significant differences (p<0.05/7) are marked with *. (b) Mutational signature decomposition for DNMs in maternally and paternally derived DNMs. Signatures extracted with SigProfiler. Colors correspond to COSMIC signatures.

Identifying hypermutated individuals in DDD and 100kGP

To identify hypermutated individuals in the DDD study, I analysed exome DNMs called in probands from 7,930 parent-offspring trios. This is a slightly larger set than the one described in Chapter 2 but the DNMs were called in the same way and subject to the same filters. To identify probands with an excess of exome DNMs it was important to account for parental age. I fit a Poisson generalized linear model (GLM) with maternal and paternal age as covariates and then looked for individuals that had both a high regression residual and a large absolute number of exome DNMs. After inspection of IGV plots, to ensure the exome DNMs appeared to be real, and ensuring that the child was related to both parents, I narrowed down the list to 10 trios. The 10 probands had 7-17 exome DNMs. It is important to note that not all DNMs detectable from WES fall within exons. The baits overlap with non-coding regions as the exome capture for DDD also had an additional 5MB of non-coding elements. These individuals were then whole genome sequenced to >30 mean depth using Illumina short-read sequencing. Due to a sample fail, I could only analyse 9 of the 10 trios.

ID	Number of SNVs	Number of InDels	Paternal age	Maternal age	SNV p-value	InDel p-value	transcriptional strand-bias	Phase P,M	Phase Ratio p-value	Hypermutation type
GEL_1	425	16	(30,35]	(20,25]	1.78E-68	9.18E-05	7.82E-23	129,1	4.19E-14	paternally_phased
GEL_2	368	6	(25,30]	(25,30]	2.45E-51	0.363	0.219	100,7	1.35E-06	paternally_phased
GEL_3	306	4	(35,40]	(30,35]	3.30E-30	0.745	0.078	87,5	3.86E-06	paternally_phased
DDD_1	277	6	25	37	NA	NA	3.29E-03	72,4	8.06E-07	paternally_phased
GEL_4	259	11	(30,35]	(20,25]	3.91E-21	0.028	0.608	37,35	1.00	post-zygotic
GEL_5	171	7	(35,40]	(35,40]	1.71E-06	0.381	0.096	58,4	2.85E-04	paternally_phased
GEL_6	167	7	(30,35]	(40,45]	1.06E-06	0.330	1	36,4	0.028	other
GEL_7	143	9	(30,35]	(30,35]	1.76E-04	0.129	0.039	23,17	0.998	post-zygotic
GEL_8	137	7	(25,30]	(25,30]	1.11E-04	0.274	0.141	33,11	0.680	other
GEL_9	131	6	(30,35]	(30,35]	9.10E-04	0.448	0.427	47,3	0.001	paternally_phased
GEL_10	131	13	(40,45]	(35,40]	6.35E-03	0.010	0.063	29,15	0.965	post-zygotic
GEL_11	131	9	(40,45]	(35,40]	8.95E-03	0.195	0.268	48,9	0.115	other
GEL_12	129	5	(30,35]	(25,30]	5.23E-04	0.547	0.091	43,0	1.11E-05	paternally_phased
GEL_13	114	3	(30,35]	(30,35]	9.91E-03	0.820	0.001	19,5	0.499	other
GEL_14	111	8	(25,30]	(25,30]	4.96E-03	0.155	2.54E-06	31,1	0.002	paternally_phased

Table 3.1 Properties of hypermutated individuals. Maternal and Paternal age is given in 5 year window for 100kGP as this information was not allowed to be extracted from the research environment due to privacy implications. However the regression was run on the exact ages and the parental age plots also share the exact ages. Phase column de notes the number of DNMs that were phased the paternally (P) and maternally (M).

DNMs were called from these trios using DeNovoGear[174] and were subject to a set of filters described in the Methods. One of these individuals was apparently hypermutated, with 277 DNMs, ~4 fold as many as expected, while the remaining individuals did not have remarkably high numbers of DNMs (median of 81 DNMs).

Identifying hypermutated individuals in 100kGP was more straight forward as the individuals had all been whole-genome sequenced from the outset. After regressing out paternal and maternal age on the number of dnSNVs, 27 individuals had residuals which were larger than the remaining residual distribution using a p-value threshold of 0.01. This threshold was used as opposed to the Bonferroni corrected threshold of 4×10^{-6} as I wanted to capture all possible hypermutated individuals. These individuals had 111-1379 apparent dnSNVs per person. These were extensively followed up to remove false positives. After careful examination of the distribution of these DNMs and their corresponding IGV plots I determined that 14 of these were truly hypermutated (Table 3.1). Here I focused on identifying hypermutated individuals with a large number of dnSNVs rather than dnIndels. This was because I had more confidence in the filtering of SNVs and it was easier to confirm that the supposed hypermutation was not due to a larger structural event that was miscalled. However I did also regress out parental age on the number of dnIndels per individual and calculate a corresponding p-value for whether the residuals were significantly larger than the rest of the cohort (InDel p-value in Table 3.1). Only one of the 14 was significant for indel hypermutation.

There were two main error modes for the 13 individuals that I determined were not truly hypermutated. For ten of these individuals it appears that a somatic deletion in the blood of one of the parents has occurred leading to a very high number of supposed DNMs being called in that region in the offspring. These individuals had some of the highest number of DNMs called (up to 1379 DNMs per individual). For each of these 10 individuals, the DNM calls all clustered to a specific region in a single chromosome. In this same corresponding region in the parent, I observed a loss of heterozygosity when calculating the heterozygous/homozygous ratio (Figure 3.7). In addition, many of these calls appeared

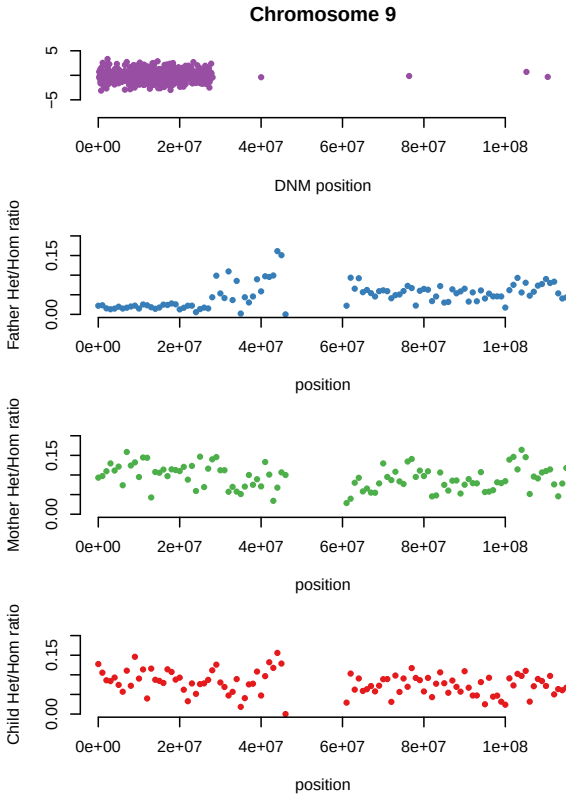

Fig. 3.7 Loss of transmitted allele example leading to false positive DNMs. Top plot shows the location of the called DNMs in the child on chromosome 9. The plots below show the heterozygous/homozygous ratio in the Father, Mother and Child showing a loss of heterozygosity in the father in the same region the DNMs have been called.

to be low level mosaic in that same parent. This type of event has previously been shown to create artifacts in CNV calls and is referred to as a 'Loss of Transmitted Allele' event [175]. I removed two other individuals due to a high false positive rate of called DNMs upon examination of IGV plots and therefore these did not appear to be truly hypermutated. The last individual that I removed had 100 autosomal DNMs and the largest p-value very close to the threshold ($p = 0.0099$). The mutational spectrum was normal, no specific mutation type

was significantly enriched, the VAF distribution was normal and the mutations did not phase more to one parent compared to what we would expect. This led me to believe that this may be an individual on the tail of the DNM count distribution rather than hypermutation.

3.3.3 Characterising hypermutation in 15 individuals

The number of DNMs for each of these 15 hypermutated individuals across both DDD and 100kGP ranged from 111-425 which corresponds to a fold increase of 1.7-6.5 compared to the median number of DNMs per individual across the 100kGP cohort. For each of the 15 hypermutated individuals I explored various characteristics of their DNMs to uncover possible underlying causes of this mutator phenotype (Table 3.1). The mutational spectra varied widely (Figures 3.14,3.15) and I calculated the enrichment of each of these mutation types compared to the average number of mutations observed across the 100kGP cohort (Figure 3.8). I extracted mutational signatures for all of these individuals using SigProfiler (Figure 3.9a)[12]. I found that most of the DNMs mapped on to known mutational signatures in cancer (from COSMIC) however there was also a novel signature extracted (Figure 3.9b)[212]. In addition to mutational spectra, I analysed parental phase of the DNMs, transcriptional strand bias and VAF distributions. Upon examining these properties, I was able to categorise these individuals into three different groups.

Hypermutation due to parental hypermutator

The first of these groups comprised of individuals whose excess DNMs originated from a single parent. I was able to phase ~$\frac{1}{3}$ of DNMs in these individuals and found that for eight of the fifteen the DNMs phased to the father significantly more than what we would expect given the overall ratio of paternal:mutations across all individuals in the 100kGP cohort (p-<0.05/15, Binomial test, Table 3.1). An additional individual was nominally significant (GEL_6 p = 0.028). This implicates the father as a possible germline hypermutator. To try and identify possible genetic causes I searched for rare paternal variants in known DNA repair genes compiled from the literature. Defects in DNA repair are known to increase the mutation rate in the soma and therefore may have a similar effect in the germline. I found possible causal variants in two of these individuals (Table 3.2).

GEL_1 has the largest number of DNMs of all individuals, a ~7 fold enrichment compared to what we would expect. The mutational spectra demonstrates a high enrichment of C>A and T>A mutations (Figure 3.14a,3.8). From extracting mutational signatures I observed a large contribution from Signature 8 in COSMIC (Figure 3.9). This signature is associated with transcription-coupled nucleotide excision repair (TC-NER) and typically presents

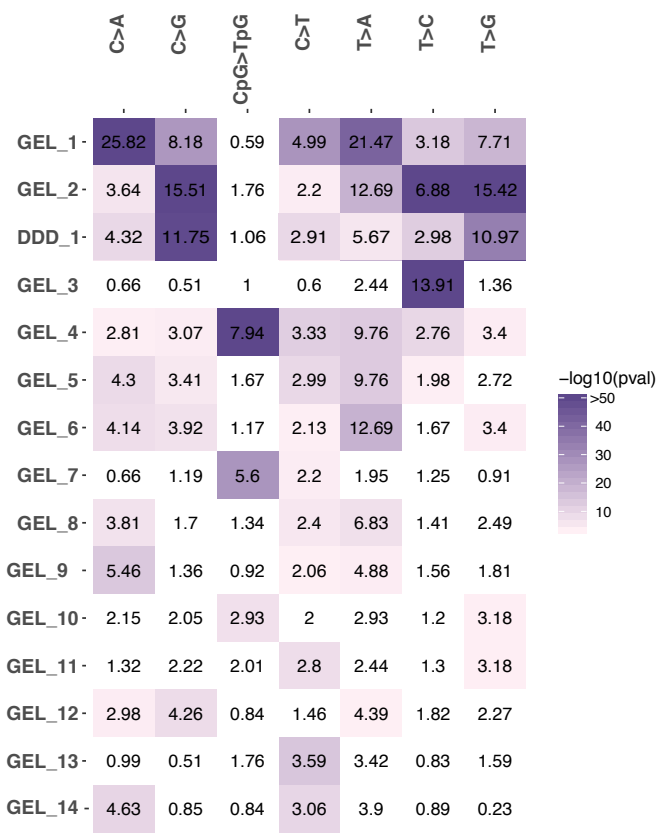

Fig. 3.8 Enrichment (observed/expected) of mutation type for hypermutated individuals. Sample names on the y axis, mutation type on the x axis. The enrichment is colored by the -log10(enrichment p-value) which was calculated using a poisson test comparing the average number of mutations in each type across all individuals in the 100kGP cohort. White coloring indicates no statistically significant enrichment (p-value <0.05/(15 × 7))

with transcriptional strand bias on the untranscribed strand. This agrees with the strong transcriptional strand bias I observed in GEL_1 ($p = 7.8 \times 10^{-23}$, Poisson test, Figure 3.10). This individual was also the only hypermutated individual that also had a significantly increased number of *de novo* indels ($p = 9.18 \times 10^{-5}$, Table 3.1). In my analysis of rare paternal variants in DNA repair genes, I identified a homozgyous stop gained variant in the gene *XPC* (Table 3.2). *XPC* is involved in the early stages of the nucleotide-excision repair (NER) pathway. NER is the main pathway for removing various types of DNA lesions such as those induced by UV light as well as other chemical adducts. There are several rare autosomal recessive syndromes that are a result of defects in NER; these include Cockayne syndrome, trichothiodystrophy and xeroderma pigmentosum [29]. The paternal variant that I identified is annotated as pathogenic for xeroderma pigmentosum in ClinVar and there are no observed homozygotes in the genome aggregation database (gnomAD AF

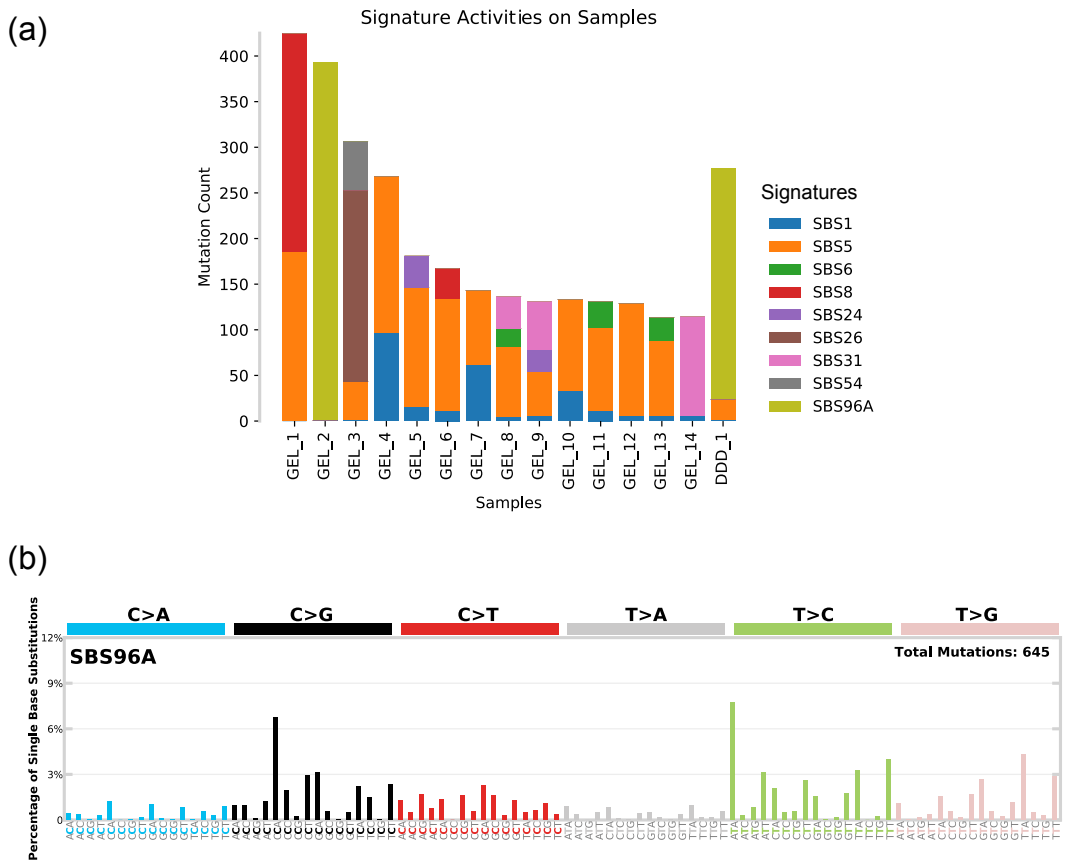

Fig. 3.9 Mutational signature decomposition for DNMs in hypermutated individuals. (a) Signatures extracted with SigProfiler. Colored by signatures number, these numbers correspond to COSMIC mutational signatures apart from SBS96A with is a novel signature. (b) The novel signature extracted which contributes heavily to GEL_2 and DDD_1.

$= 2.2 \times 10^{-5}$)[115, 102]. Upon contact with the corresponding clinician for this patient it was confirmed that the father has been diagnosed with the disorder. Patients with xeroderma pigmentosum have a high risk of developing skin cancer due to their impaired ability to repair UV damage and are also known to be at a higher risk of developing other cancers [123, 169]. *XPC* deficiency has been associated with a similar mutational spectrum to the one we observe in GEL_1. A recent study observed increased Signature 8 mutations in a human intestinal organoid culture in which *XPC* was deleted using CRISPR-Cas9 gene-editing, although transcriptional strand bias was not observed here[92]. The same study observed that genomes of NER-deficient breast tumors show an increased contribution of Signature 8 mutations compared with NER-proficient tumors. There is little previous evidence of the effect of *XPC* deficiency on germline mutation in humans, although a previous study has

ID of child	Chrom	Position (hg38)	Ref	Alt	Csq	Paternal genotype	Gene	DNA repair pathway	Gnomad AF	Pathogenicity evidence
GEL_1	3	14165549	G	A	stop_gained	1/1	*XPC*	NER	2.2e-5	Pathogenic for xeroderma pigmentosum in ClinVar
GEL_5	16	83139	G	A	missense	1/1	*MPG*	BER	9.57e-5	CADD score 27.9; likely interacts with DNA

Table 3.2 Possible paternal mutator variants

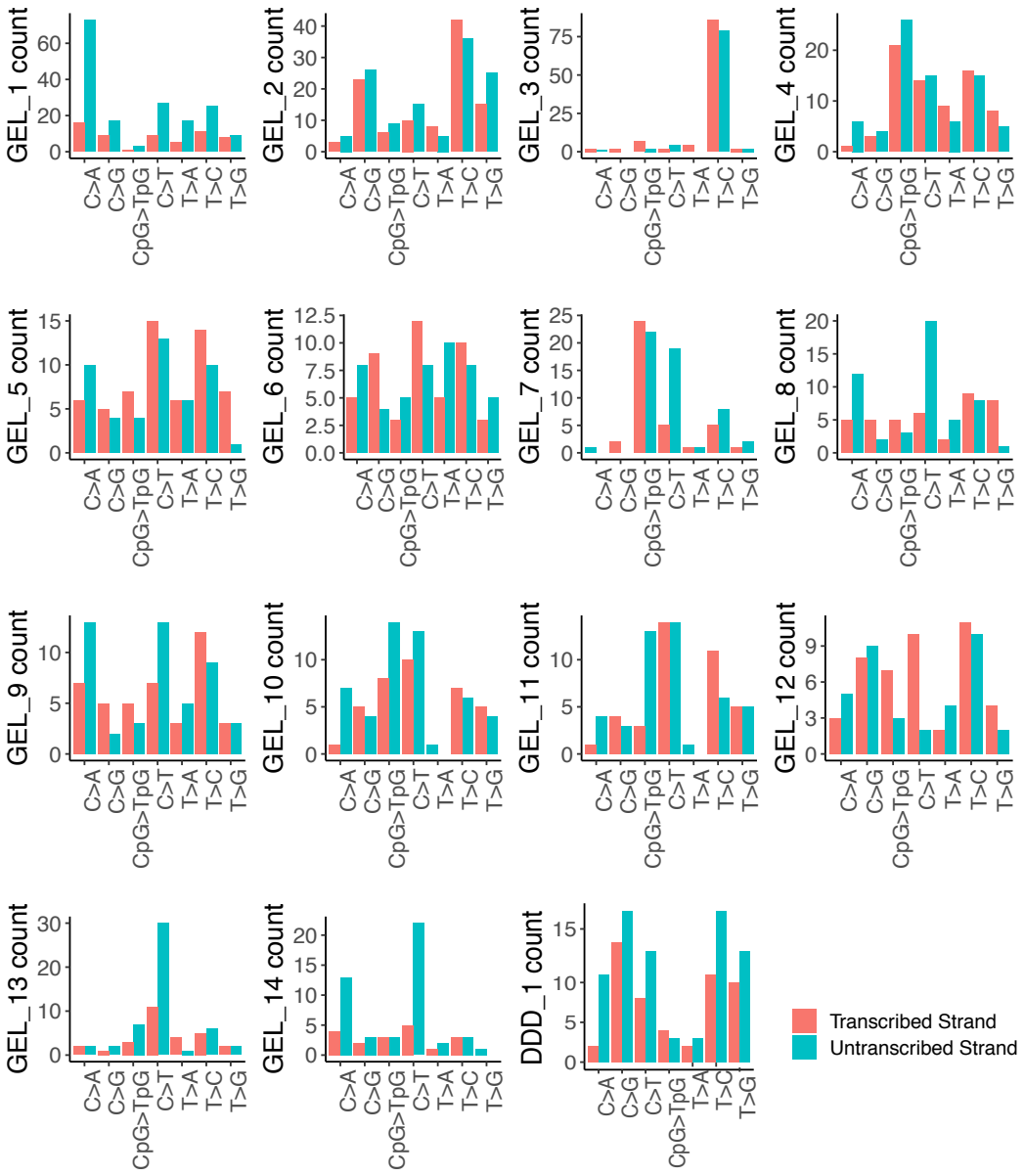

Fig. 3.10 Transcriptional strand bias for DNMs in hypermutated individuals

shown that Xpc deficient (-/-) male mice have a significantly increased germline mutation rate at two STR loci compared to heterozygous *XPC* (+/-) and wild-type (+/+) mice which may indicate a mutator phenotype [145].

GEL_3 has a ~5 fold enrichment of the number of DNMs. These DNMs exhibit a very distinct mutational spectrum with a ~14 fold increase in C>T mutations but no significant enrichment for any other mutation type (Figure3.14d, Figure 3.8). Extraction of mutational signatures revealed that the majority of mutations mapped onto Signature 26 from COSMIC (Figure 3.9a). This signature is associated with defective mismatch repair. In my analysis of paternal variants, I identified a rare homozygous missense mutation in the gene *MPG* (Table 3.2). *MPG* encodes for a DNA glycosylase which is involved in the recognition of base lesions, including alkylated and deaminated purines, and initiation of the base-excision repair (BER) pathway. The paternal variant I identified has an allele frequency of 9.8×10^{-5} in gnomAD with 0 observed homozygotes. The Combined Annotation Dependent Depletion (CADD) score, for this variant is 27.9 and the amino acid residue is highly conserved (conservation = 1 from 172 aligned protein seqs from VarSite) [117]. An analysis of its position in the context of the protein by James Stephenson, a post-doc in the group, revealed it forms part of the substrate binding pocket and is likely interacting with DNA (Figure 3.11) [118]. Studies in yeast have demonstrated that overexpression of *MPG* can lead to a mutator phenotype and that variants that alter other amino acids in the substrate binding pocket, and alter substrate specificity, can result in an increase in the mutation rate of either point mutations or STRs[61, 48]. Another study found that *Mpg*(-/-) mice treated with methyl methanesulfonate resulted in >3 times *hprt* mutations in splenic T lymphocytes compared to wildtype also demonstrating that there can be a mutagenic effect [47].

GEL_2 and DDD_1 have a similar number of DNMs which are significantly more paternal in origin than expected (Table 3.1). The mutational spectra of the DNMs in these individuals are very similar and the cosine similarity between their spectra is 0.79 (Figure 3.14). In my analysis of mutational signatures, a novel signature was extracted which these two individuals share. This does not map onto any known signatures in COSMIC and is characterised by an enrichment of C>G and T>G mutations (Figure 3.9a,b). In my analysis of paternal variants in DNA repair genes I found that the father of DDD_1 has a rare heterozygous missense variant in *BRCA2* and a heterozygous stop gained mutation in the gene *NTHL1*. The *BRCA2* variant has an allele frequency of 0 in gnomAD and is annotated as a variant of uncertain significance (VUS) for breast cancer in ClinVar. It has conflicting interpretations of pathogenicity from different tools (SIFT:'Tolerated', PolyPhen-2:'Probably Damaging'). *BRCA2* is involved in the homologous recombination repair pathway that mends double strand breaks and so defects in *BRCA2* in cancer are known to lead to an increase in

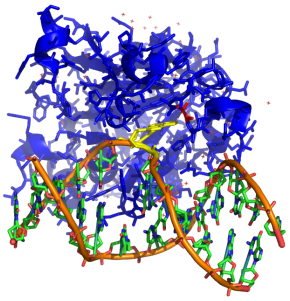

Fig. 3.11 Position of *MPG* missense variant (residue in red) in GEL_3 in the context of the protein (blue). Residue forms part of the binding pocket and image demonstrates its proximity to DNA (orange). Image courtesy of James Stephenson

the number of indel mutations as well as SNVs. DDD_1 does not map on to any mutational signatures associated with defects in *BRCA2* and does not have a significantly increased number of dnIndels and so this variant does not look convincingly causal [159]. *NTHL1* is a gene involved in the BER pathway and germline homozygous mutations in this gene have been associated with multiple cancers [231]. The paternal variant in *NTHL1* has an allele frequency of 1.42×10^{-3} in gnomAD and there were an additional 23 fathers in the DDD study that had this same variant. I examined the mutational spectra across all the DNMs from their offspring and found they were normal and did not have this distinctive signature so this is unlikely to be the sole cause of hypermutation. There were no putative damaging paternal variants in DNA repair genes for GEL_2. Since these two individuals shared this mutational signature I looked for an intersection of genes in which both individuals had rare nonsynonymous paternal variants. For this I looked across all genes, not restricted to DNA repair genes, but found no overlap. In the corresponding clinician's additional notes for patient DDD_1 it has been noted that the father has undergone treatment for Hodgkin's Lymphoma twice. This may be a result of a paternal mutator variant also having an effect in the paternal soma and increasing cancer risk or the hypermutation in the child could be due to damage incurred in the father's germline during cancer treatment. The mutational signature does not resemble known signatures associated with Hodgkin's Lymphoma in cancer or known chemotherapeutic signatures (Table 3.1) [166, 164]. The father does not have any known germline variants that are associated with elevated risk of Hodgkin's Lymphoma although there are other germline *BRCA2* variants that can increase risk[122]. I am currently following up with the corresponding clinician to confirm these cancer treatments occurred

prior to the conception of the child and what these treatments were. I am also following up with GEL_2 to see if their father has had cancer or undergone treatment as well.

For the remaining five individuals that may have a paternal hypermutator, I was not able to identify any putatively causal paternal variants. The mutational signatures in these individuals have various compositions which may indicate the mechanisms in which the DNMs arose. For example the DNMs in GEL_14 map mostly onto Signature 31 which is associated with transcription coupled NER (Figure 3.9). The significant transcriptional strand bias ($p = 2.54 \times 10^{-6}$) in the DNMs would support this mechanism however I did not observe any nonsynonymous rare variants in genes known to be involved in NER. For these five individuals, a paternal mutator variant may fall into a gene not currently associated with DNA repair or may be non-coding. I searched for rare recessive paternal variants in all genes across these five individuals but there was nothing immediately notable. Other explanations may be that the variant may be germline specific and so not detectable in blood, the hypermutation may be due to an environmental mutagen that has impacted the paternal germline or there may be a gene by environment interaction that results in increased mutation rate.

Post-zygotic hypermutation

The second group of hypermutated individuals consists of those where the hypermutation appears to have occurred post-zygotically. I examined the distribution of the VAF in the DNMs for each individual. I found that for three of these individuals (GEL_4, GEL_7 and GEL_10) the VAF distribution was not centered around 0.5 (Figure 3.12). The proportion of DNMs with VAF<0.4 was significantly higher than compared to the distribution of all DNMs across all individuals in GEL_4 ($p = 1.5 \times 10^{-51}$, Binomial test) and GEL_10 ($p = 2.4 \times 10^{-4}$) and nominally significant in GEL_7 ($p = 0.02$). For all three of these individuals, the mutations phased evenly between the maternal and paternal chromosome. This indicates that these mutations most likely occurred post-zygotically and are less likely to be due to a parental hypermutator. All three of these individuals are most strongly enriched for CpG>TpG mutations and have a large contribution of mutations from Signature 1 in COSMIC (Figure 3.9, Figure 3.8).

Other sources of hypermutation

The third group of hypermutated individuals included the remaining 4 hypermutated individuals. The DNMs in these individuals did not phase overwhelmingly to a single parent and the VAF distributions did not indicate a large number of post-zygotic mutations (Figure

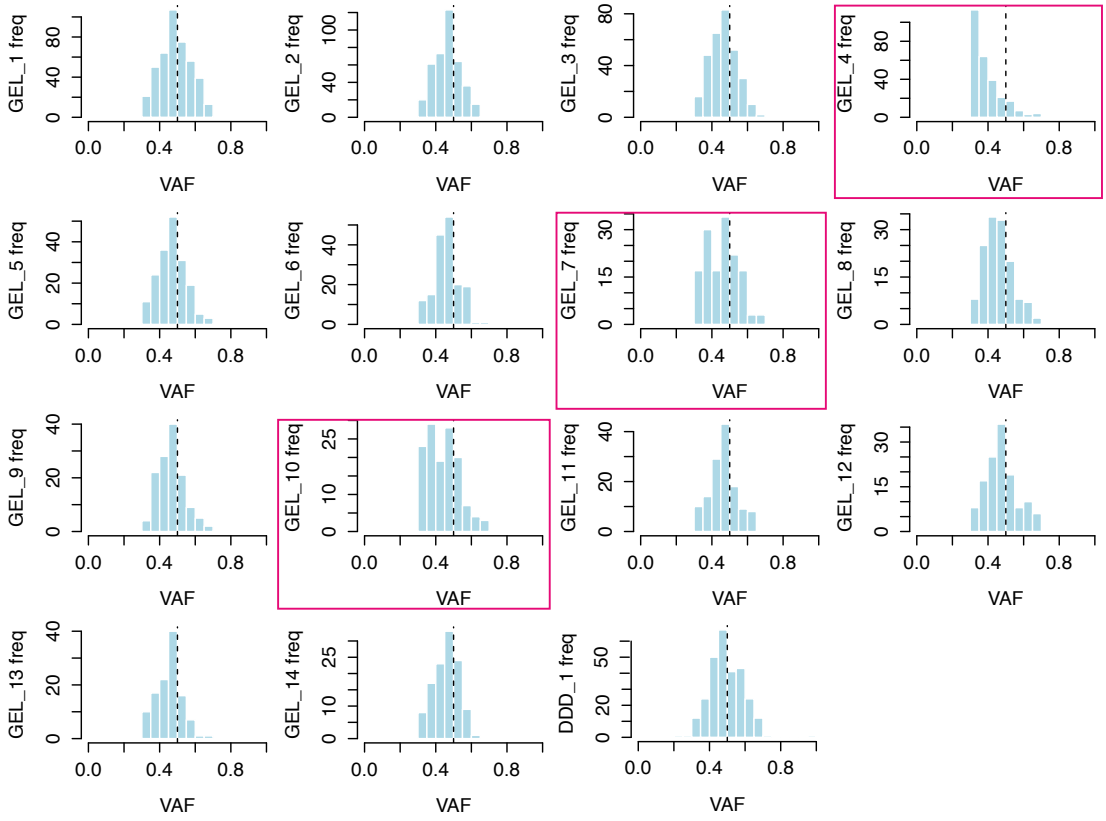

Fig. 3.12 Distribution of variant allele fraction (VAF) for DNMs in hypermutated individuals. The vertical line indicates 0.5 VAF. The three plots highlighted in pink are those where the DNMs appear post-zygotic.

3.12, Table 3.1). They did appear to have mutational spectra that are in different proportions to what we would expect. I observed different levels of enrichment across mutation types compared to expected (the average number of mutations across 100kGP) (Figure 3.8) however these were not as striking. The observed elevated germline mutation rates may be due to a combination of polygenic effects in the parents, shared mutagenic environment for the parents or an interaction between the two.

3.3.4 Fraction of germline mutation rate variation explained

Work from Kong et al. studying 78 trios previously estimated that paternal age accounts for >95% of the variation surrounding germline mutation rate after accounting for Poisson variation [111]. Using a similar approach I fit several GLMs including variables for parental age and hypermutation status and calculated the fraction of variance explained in the 100kGP dataset. To mitigate the effect of data quality this analysis was performed on a subset of

7,700 trios that had been filtered on basic QC metric such as coverage and mapping rate. I also removed the false positive hypermutated individuals that I identified. The details of this can be found in the Methods. I first fit a model that only accounts for parental age and found this explained 70% of the variation of the number of mutations per individual.

This estimate of 70% is considerably lower than the previous estimate from Kong et al and there may be several explanations for this. Firstly, due to the much larger size of the dataset, I was unable to verify the DNMs to the same degree as in the Kong et al paper which was performed on 78 parent offspring trios. I estimated the true positive rate of the called DNMs to be 0.95, therefore the variance may be overestimated. This analysis was done on a subset of higher quality samples to mitigate this but to account for additional measurement error, which may correlate with the number of DNMs called, I also included coverage, mapping and variant calling metrics in my regression models and found this explained ~3% of variation. Secondly, Kong et al. may be slightly underestimating germline mutation rate variation due to the fact that the 78 trios in the paper also included multi-sibling families which we may expect to have more similar number of DNMs than unrelated trios, this would inflate the variation explained. Thirdly, if the trios selected for the Kong et al. analysis were selected non-randomly with respect to paternal age then this could conflate the variance explained by that variable. I performed simulations in the 100kGP dataset where I sampled trios either randomly across the population or more heavily towards the tails (disproportionate amount of young/old fathers) and found that heavier tail sampling significantly increased the median proportion of variation explained from 0.78 to 0.82 ($p = 5.7 \times 10^{-61}$, Wilcox test). While this may contribute to the discrepancy, it is unlikely to fully explain the much higher fraction of variance explained by Kong et al. Finally, by repeated random sampling of 78 trios from the much larger 100kGP data I observed that in such a small dataset estimates of the variance explained varies considerably by chance, and that although the median estimate of variance explained was 0.78, I observed an estimate of variance explained similar or greater to that observed by Kong et al in 7% of simulations. This suggests that Kong et al could have over-estimated the true variance explained by parent age by chance, and that the uncertainty in their estimate was much greater than they estimated.

In addition to parental age and data quality I also included in the regression a variable accounting for the excess number of mutations in individuals I have identified and confirmed as being hypermutated. I found that this accounted for an additional 8% of variation. In total, this means that 20% (17%-22%, Bootstrap 95% confidence interval) of variation remains unaccounted for of which there may be several contributors. Variants in genes involved in DNA repair are implicated here as possible causes of hypermutation therefore they may also

play a role in the remaining germline mutation rate variation. In addition, polygenic effects, environmental mutagens and gene by environment interactions may also contribute.

Impact of variants in DNA repair genes across cohort

To assess whether rare variants in genes known to be involved in DNA repair pathways impact germline mutation rate more generally, I looked across the whole 100kGP cohort. I curated three sets of variants that have increasing likelihoods of impacting germline mutation rate. For all three sets I considered both all nonsynonymous variants and restricting these to just PTVs. The first set was the least stringent set including 186 known DNA repair genes which is the same set described earlier. For this set I also separately considered the impact of rare homozgyous variants in these genes (the counts were too small to assess in the subsequent groups). The second set was restricted to DNA repair genes encoding components of the DNA repair pathways most likely to create SNVs. For this I chose the 66 genes that were known to be associated with the BER, NER and mismatch repair (MMR) pathways as well as DNA polymerases. Again, I looked at the impact of both nonsynonymous and just PTVs on germline mutation rate. The third set were variants within this second set that have also been associated with an increased risk of cancer. This was created by considering variants that are annotated as 'pathogenic' or 'likely pathogenic' germline variants for any cancer phenotype in ClinVar. I found that for all eight regressions that I ran there was no statistically significant effect after Bonferroni correction (Table 3.3, Figure 3.13). The only effect that

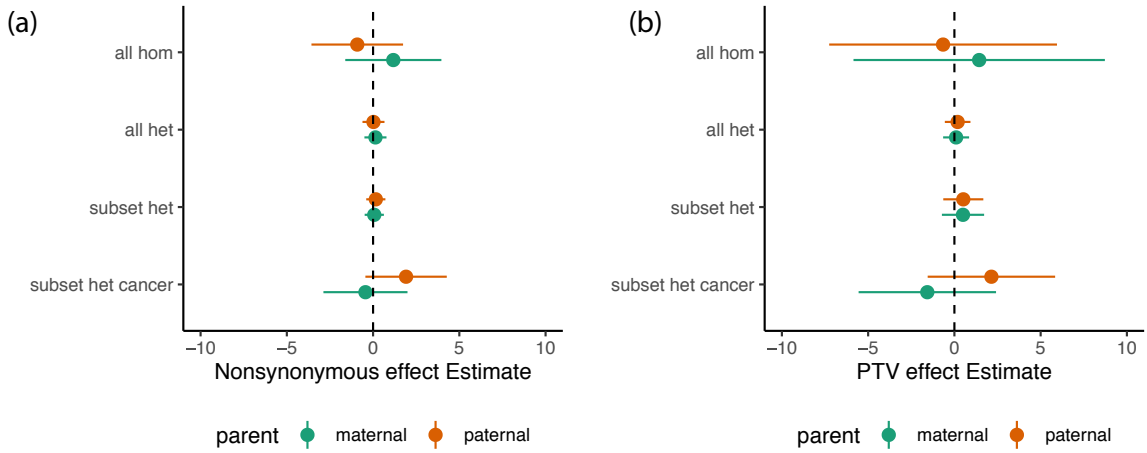

Fig. 3.13 Impact of rare variants in DNA repair genes on germline mutation rate. Poisson regression effect estimates for binary variables of having a parental variant in genes known to be involved in DNA repair. (a) considered all nonsynonymous variants in the subsets (b) is restricted to PTVs.

was nominally significant was for paternal nonsynonymous variants known to be associated with cancer phenotypes (p = 0.018) and this only explained an additional 0.03% of variance. This demonstrates that rare variants in DNA repair genes do not explain a large amount of the remaining variation in germline mutation rate. To detect more subtle effects of these variants other analytical approaches will need to be explored. The role of genetic variation, not restricted to these genes also needs to be investigated.

Variant subset	Consequence	Genotype	Paternal count	Paternal Effect	Paternal p-value	Maternal count	Maternal Effect	Maternal p-value
all DNA repair	nonsynonymous	het	5865	0.023	0.915	5916	0.137	0.526
	PTV	het	1203	0.187	0.456	1153	0.099	0.697
	nonsynonymous	hom	78	-0.917	0.307	71	1.174	0.213
	PTV	hom	13	-0.657	0.769	11	1.437	0.560
subset DNA repair	nonsynonymous	het	3076	0.159	0.398	2928	0.069	0.715
	PTV	het	434	0.516	0.189	391	0.498	0.229
germline cancer	nonsynonymous	het	103	1.912	0.017	97	-0.442	0.592
	PTV	het	41	2.145	0.086	35	-1.570	0.244

Table 3.3 Impact of parental rare variants in DNA repair genes on germline mutation rate

3.4 Discussion

Germline hypermutation is an uncommon but important phenomenon which can impact the health of subsequent generations. In this chapter, I identified 15 individuals from ~20,000 parent-offspring sequenced trios in the DDD study and 100kGP with a significant 2-7 fold increased number of DNMs compared to expected. For 3 of these individuals the excess mutations appear to have occured post-zygotically however for the majority (8) of these hypermutated individuals, the excess DNMs phased paternally implicating the father as a potential germline hypermutator. I identified possible paternal mutator variants in two of these individuals. These were rare nonsynonymous homozygous variants in two genes known to be involved in DNA repair, *XPC* and *MPG*. The missense variant in *MPG* is likely damaging however functional follow up is necessary here to assess whether and how it may disrupt the BER pathway and create such a distinctive mutational spectrum. A collaborator is currently carrying out functional assays to interrogate the impact of the change in this residue. The father carrying the *XPC* PTV has been diagnosed with xeroderma pigmentosum (XP) which carries a very high risk of skin cancer as well as an increased risk of other cancers.

It is well established that defects in DNA repair genes can increase the somatic mutation rate and elevate cancer risk [105]. The findings in this chapter imply that the germline can be similarly affected and that defects in DNA repair can lead to a dramatic increase in germline mutation rate. However defects in DNA repair pathways do not always appear to behave similarly in the soma and the germline. I interrogated protein-truncating variants in

an established cancer mutator gene, *MBD4*, and found they did not have a detectable effect in the germline [232]. I also looked at the impact of parental nonsynonymous variants in DNA repair genes on the number of DNMs in offspring across the 100kGP cohort and did not find a significant difference. Paternal variants that have previously been associated with a cancer phenotype were nominally significant but having one of these variant only amounted to an estimated increase of approximately ~2 DNMs in the child. If only a subset of these variants have an impact in the germline this would dilute our power to detect an effect and it is likely we will need both larger sample sizes as well as a more stringently curated set of variants to investigate this further. There are also likely to be pathways that impact the germline more than the soma and uncovering the genes and associated variants in these genes will be more challenging.

A limitation to the approach I took in this chapter is that I used DNMs of a single offspring as a proxy for the germline mutation rate of both parents. Aside from sequencing large families, directly sampling the germline would be more reliable in estimating individual mutation rate. Sequencing oocytes is difficult to do at a large scale due to the invasive and costly procedure needed to sample only a few eggs. Moreover, I did not observe a significant maternal bias in any of the hypermutated individuals. Since the mother contributes only a quarter of a child's DNMs on average, I may be less powered to detect an increase in maternal DNMs. The maternal germline may also be more protected to mutator variants as oocytes stop replicating during gestation while spermatagonial stem cells continue to replicate throughout a male's life and may be more vulnerable to impaired repair processes due to uncorrected replication errors. Sperm is more feasible to sample at scale and would be an important resource to estimate individual male mutation rate variation. At a smaller scale we are currently following up with Genomics England Limited in order to recontact the likely paternal hypermutators to collect sperm for single-cell sequencing. This will allow us to interrogate whether all sperm are affected equally by the hypermutation, the presence of mutator variants that are only present in the germline and improve our ability to extract mutational signatures on a larger number of mutations. Another useful next step would be to follow up more directly with parents with different DNA repair disorders, including those with pathogenic variants in *XPC*. Sequencing sperm or families of other male XP patients would allow us to see if germline hypermutation is observed in those with the same and other pathogenic variants in this gene. Variants in other other genes associated with XP (*XPA*, *XPB* etc.) might also be worth investigating. This information may be clinically useful for these patients as germline hypermutation means future children are at a higher risk of having a genetic disorder caused by a DNM.

It is important to note that to identify hypermutators I fit an ordinary linear regression of the number of DNMs on parental age and then applied a threshold on the studentized residuals to capture those with an unusually large number of DNMs. In part, this was for comparability to the Kong et al study, which used the same regression approach. The studentized residuals are expected to follow a t distribution with $N-p-1$ degrees of freedom where N is the sample size and p is the number of parameters included in the model. On examining the residuals I found that they had a much narrower variance than expected and thus the threshold of 0.01 was much more stringent that I was anticipating. This also explains why for ~12,000 individuals I only had a few individuals pass the threshold. On fitting a Poisson GLM, as I have done in other parental age analyses in this chapter, I found the variance of these studentized pearson residuals was inflated and so may also not be the correct approach. In my next steps I aim to improve this methodology (for example using quasi-Poisson or negative binomial regression) to ensure I am using the most appropriate model and capturing all possible hypermutated individuals. Although I would note that the rank order of hypermutated individuals is barely altered under these different models, only the p values change.

I found that germline hypermutation explained 8% of the variance in germline mutation rate in 100kGP. The fact that this is evaluated in a cohort that consists of offspring with genetic disorders may mean this is an overestimate of how much variance is explained by hypermutation in the general population. *De novo* mutations are a major cause of DD and cohorts of children with developmental disorders are enriched for DNMs overall and so would be more likely to contain hypermutated individuals [41]. In a healthy population this variance explained may be smaller. However we would still expect to see hypermutation in a healthy population. The absolute risk of a germline hypermutator having a child with a genetic disease is still low. The population average risk is estimated to be 1 in 300 births and so a 4 fold increase in DNMs in a child will amount to the risk of a genetic disease is just over 1% [41]. I found that parental age explained ~70% of the germline mutation rate variance which is substantially smaller than a previous estimate of 95% [111] based on a sample of families ~100x smaller than the one I analysed. This may be due to several factors such as differences in measurement error, non-random selection of parental age or by chance. Another possible contributing factor may be that in 100kGP the variance of the number of DNMs is larger than it would be in a healthy population. The remaining ~20 % of germline mutation variation remains unexplained in this analysis. Part of this may be attributable to additional hypermutated individuals that may be identified upon improving my model although this is unlikely to amount to a substantial additional fraction. Rare coding and non-coding variants in DNA repair genes or genes currently not known to be associated with

germline mutation rate may also explain more variance. However even with thousands of whole genome sequenced trios we may not be powered to identify these across the genome. Another source of variation may be explained by polygenic effects on germline mutation rate. Previous work has demonstrated that there are differences in germline mutation rate between populations and that there are loci in the genome that may be associated with a higher germline mutation rate [75, 197]. A genome wide association study (GWAS) approach using the DNMs as a proxy for germline mutation rate in the parents requires parent-offspring trio sequencing just to measure the phenotype. This means sequencing 3x as many individuals as you expect to test which is costly especially considering that a very large sample size would be needed. Another possibility may be to conduct an association study on male germline mutation rate by using estimates of individual mutation rates from single cell sequencing of sperm. This would also allow interrogation of the within variation of individual mutation rate. This may be feasible as single cell technology and methodology improves and sequencing costs decrease however large sample sizes would be needed and a similar interrogation of the female germline mutation rate would not be feasible. Environmental effects are also likely to contribute to germline mutation rate variation so including deep phenotyping and details of possible exposures would be important to include in a large germline mutation rate study and may also help reveal gene by environment interactions.

The analyses in this chapter provide new insights into the role of genetic variation on the human germline mutation rate. I have demonstrated the existence of germline hypermutators as well as possible genetic causes. I have shown that hypermutation explains a significant proportion of germline mutation rate variation in addition to parental age but also that there is residual variance that still needs to be explored.

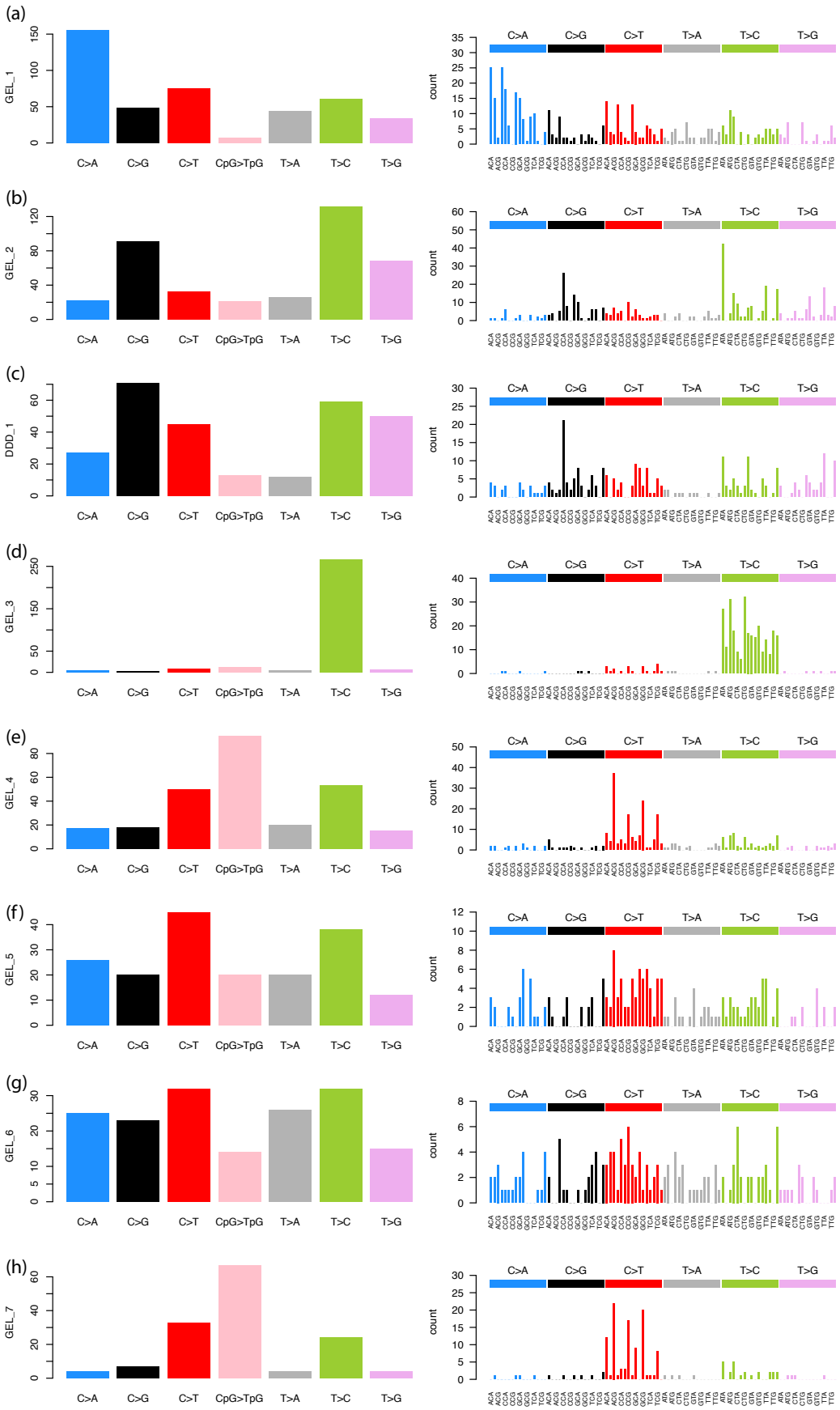

Fig. 3.14 Mutational spectra of DNMs from hypermutated individuals (A)

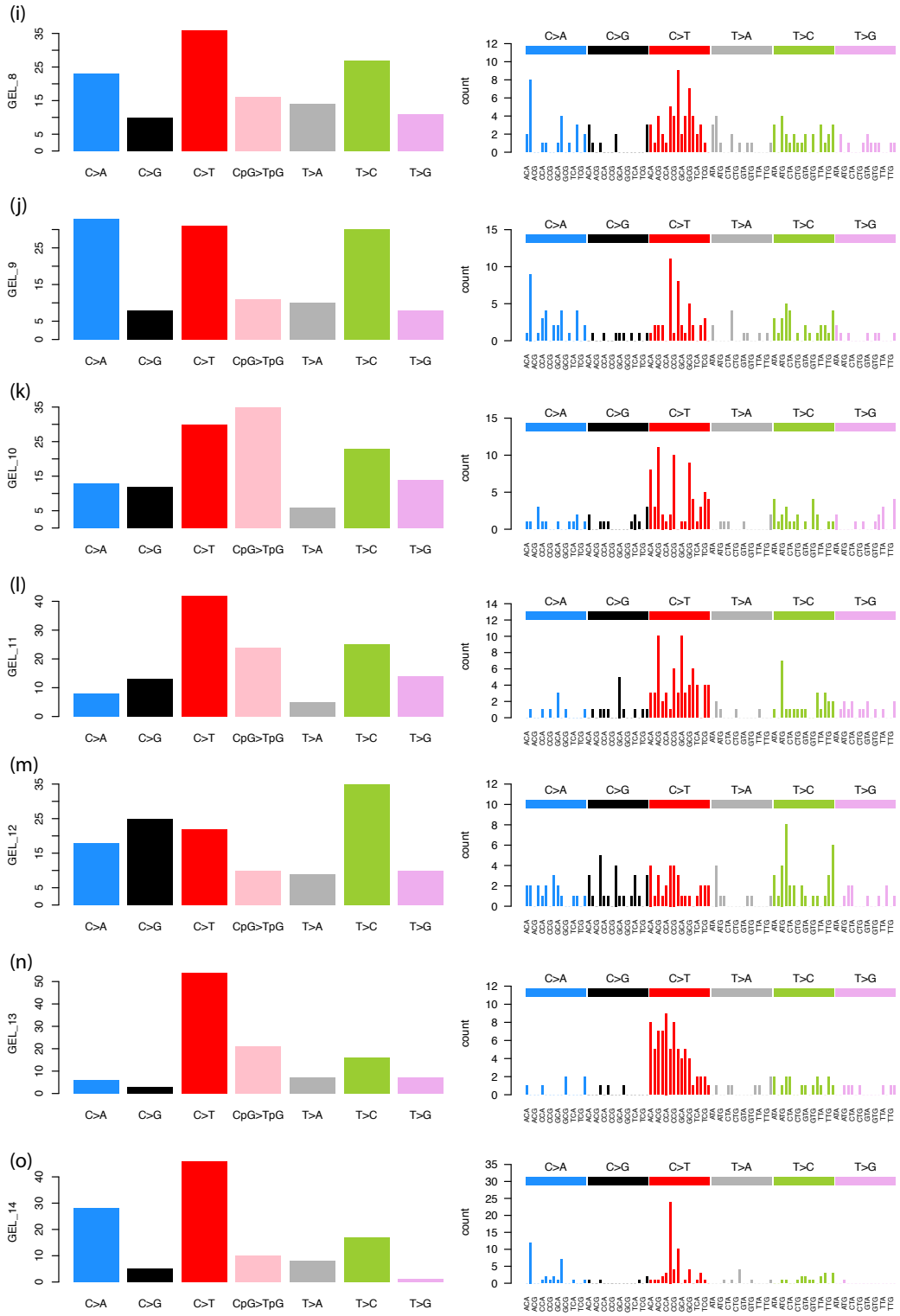

Fig. 3.15 Mutational spectra of DNMs from hypermutated individuals (B)