Chapter 2

Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations

2.1 Introduction

In genomic analyses, single nucleotide variants (SNVs) are often considered independent mutational events. However SNVs are more clustered in the genome than expected if they were independent [146, 196, 8]. On a finer scale, there is an excess of pairs of mutations within 100 bp that appear to be in perfect linkage disequilibrium in population samples [195, 209, 76]. While some of this can be explained by the presence of mutational hotspots, natural selection or compensatory variants, it has been shown that multi-nucleotide mutations play an important role [191]. Recent studies found that 2.4% of *de novo* SNVs were within 5 kb of another *de novo* SNV within the same individual [14], and that 1.9% of *de novo* SNVs appear within 20 bp of another *de novo* SNV [191]. Multi-nucleotide variants (MNVs) occurring at neighbouring nucleotides are the most frequent of all MNVs [14]. Moreover, analysis of phased human haplotypes from population sequencing data also showed that nearby SNVs are more likely to appear on the same haplotype than on different haplotypes [191].

The mutational origins of MNVs are not as well understood as for SNVs, however different mutational processes leave behind different patterns of DNA change which are dubbed mutational 'signatures'. Distinct mutational mechanisms have been implicated in creating MNVs. Polymerase ζ is an error-prone translesion polymerase that has been shown to be the predominant source of *de novo* MNVs in adjacent nucleotides in yeast[76, 14]. The

most common mutational signatures associated with polymerase ζ in yeast have also been observed to be the most common signature among MNVs in human populations[76], and were also found to be the most prevalent in *de novo* MNVs in parent-offspring trios [14]. It has been suggested that translesion DNA polymerases play an important role in the creation of MNVs more generally [27]. A distinct mutational signature has also been described that has been attributed to the action of APOBEC deaminases [5].

Although MNVs are an important source of genomic variability, their functional impact and the selection pressures that operate on this class of variation has been largely unexplored. In part, this is due to many commonly used workflows for variant calling and annotation of likely functional consequence annotating MNVs as separate SNVs [188]. When the two variants comprising an MNV occur within the same codon – as occurs frequently given the propensity for MNVs at neighbouring nucleotides – interpreting MNVs as separate SNVs can lead to an erroneous prediction of the impact on the encoded protein. The Exome Aggregation Consortium (ExAC) systematically identified and annotated over 5,000 MNVs that occurred within the same codon in genes, including some within known disease-associated genes[125]. Although individual pathogenic MNVs have been described [115], the pathogenic impact of MNVs as a class of variation is not yet well understood.

2.1.1 Chapter overview

In this chapter I analysed 6,688 exome sequenced parent-offspring trios from the Deciphering Developmental Disorders (DDD) Study to evaluate systematically the strength of purifying selection acting on MNVs in the population sample of unaffected parents, and to quantify the contribution of pathogenic *de novo* MNVs to developmental disorders in the children.

2.1.2 Publication and contributions

The results described in this chapter were published in 2019 [100]. I briefly summarise the various contributions to this project. Giuseppe Gallone performed the upstream variant calling for the DDD project, Jeremy McRae called and filtered the *de novo* mutations (DNMs) in DDD and Elena Prigmore experimentally validated the *de novo* MNVs. All of this work was done under the supervision of Matthew E. Hurles. The parts of the publication reproduced in this Chapter are all my original work.

2.2 Methods

2.2.1 Variant and *De Novo* calling in DDD

The analysis in this chapter was conducted using exome sequencing data from the DDD study of families with a child with a severe, undiagnosed developmental disorder. The recruitment of these families with developmental disorders has been described previously[236]. 7,833 parent-offspring trios from 7,448 families and 1,791 singleton patients (without parental samples) were recruited at 24 clinical genetics centres within the United Kingdom National Health Service and the Republic of Ireland. Families gave informed consent to participate, and the study was approved by the UK Research Ethics Committee (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). In this analysis, I only included trios from children with apparently unaffected parents in our analysis to avoid bias from pathogenic inherited MNVs. This was defined as those trios where the clinicians did not report any phenotypes for either parent. This resulted in a total of 6,688 complete trios. Sequence alignment and variant calling of single nucleotide variant and insertions/deletions were conducted as previously described. DNMs were called using DeNovoGear and filtered as described in a previous DDD publication [41, 174].

2.2.2 Estimating the MNV mutation rate

I estimated the MNV mutation rate by scaling the SNV mutation rate estimate of 1.1×10^{-8} mutations per base pair per generation by the ratio of MNV segregating sites/ SNV segregating sites observed in our data set[180]. This approach is based on a rearrangement of the equation for the Watterson estimator[226]. This is outlined below where θ is the Watterson estimator, μ is the mutation rate, K denotes the number of segregating sites, N_e is the effective population size, n is the sample size and a_n is the $n-1$th harmonic number.

$$\begin{aligned}
\hat{\theta} &= \frac{K_{SNV}}{a_n} = 4N_e\mu_{SNV} \\
\mu_{SNV} &= \frac{K_{SNV}}{a_n 4N_e} = 1.1 \times 10^{-8} \\
a_n 4N_e &= \frac{K_{SNV}}{1.1 \times 10^{-8}} \\
\mu_{MNV} &= \frac{K_{MNV}}{a_n 4N_e} \\
&= \frac{K_{MNV}}{K_{SNV}} 1.1 \times 10^{-8}
\end{aligned}$$

To avoid any potential bias from selection I excluded variants that fell into potentially constrained genes (pLI>0.1).

To ensure the validity of this method, I also estimated the SNV missense mutation rate in the same way by scaling the overall SNV mutation rate by the ratio of the number of missense SNVs in unconstrained genes compared to all SNVs and obtained an estimate of the missense mutation rate across coding regions to be 1.07×10^{-8} per coding base pair per generation which agrees with the estimate of 1.09×10^{-8} per coding base per generation which was calculated using the trinucleotide context mutational model as described by Samocha et al[187].

2.2.3 Estimating the enrichment of *de novo* MNVs

To test for the enrichment of *de novo* MNVs I used a Poisson test for three categories of genes: all genes, genes known to be associated with developmental disorders and genes that are not known to be associated with developmental disorders. Genes known to be associated with developmental disorders, in which *de novo* mutations can be pathogenic, were defined as those curated on the Gene2Phenotype website (htttp://www.ebi.ac.uk/gene2phenotype/) and listed as monoallelic that were 'confirmed' and 'probable' associated with DD. I did the same tests for synonymous, missense and protein-truncating variants using gene-specific mutations rates for each consequence type derived by Samocha et al, 2014 [187]. Significance of these statistical tests was evaluated using a Bonferroni corrected p-value threshold of 0.05/12 to take into account the 12 tests across all three subsets of genes, SNV consequence types and MNVs. To correct for sequence context when comparing DD genes and non-DD genes, I adjusted the expected number of MNVs in the DD genes category based on the excess of polymerase ζ dinucleotide contexts.

2.2.4 Estimating the number of clinically reported MNVs

I downloaded all clinically reported variants from the website ClinVar and subsetted these variants to those that fell into autosomal dominant DDG2P genes and those that were annotated as 'definitely pathogenic' or 'likely pathogenic'. This set was then subsetted to 321 genes with at least one pathogenic missense mutation. This was to ensure that missense mutations cause disease in these genes. I then counted the numbers of SNV missense variants and used this to estimate the number of expected missense MNVs across those genes. This was scaled using the ratio of the SNV to MNV missense mutation rate across these genes. The MNV mutation rate used for this calculation was specifically for $\text{MNV}_{1\text{bp}}$ ($\mu_{MNV_{1bp}} = 8.76 \times 10^{-11}$ mutations per base pair per generation). The $\text{MNV}_{1\text{bp}}$ missense

mutation rate was calculated as:

$$\mu_{\text{DDG2P MNV missense}} = \mu_{MNV_{1bp}} \times \frac{2}{3} \times 0.97 \times \sum \text{coding bp in DDG2P genes}$$

Where 2/3 is the probability of a coding MNV falling within a codon and 0.97 is the probability that a within-codon MNV results in a missense change. The probability of an MNV falling within a codon was calculated from the properties of codons and the probability of a within codon MNV resulting in a missense change was calculated by looking at the consequences of all possible within-codon MNVs and calculating the proportion that result in a missense. The expected number of missense MNVs in DDG2P genes was then calculated as follows:

$$\text{E(reported pathogenic missense MNVs)} = n_{\text{reported missense SNVS}} \frac{\mu_{\text{DDG2P MNV missense}}}{\mu_{\text{DDG2P SNV missense}}}$$

This assumes that the enrichment of MNV and SNV missense mutations in these genes are comparable, as I have observed in the DDD study. This yielded an expected number of 25.67 reported pathogenic MNVs compared to 22 observed reported pathogenic MNVs. To test if this difference was significant I performed a Poisson test (p = 0.55).

2.3 Results

2.3.1 Identifying and categorising MNVs

I defined MNVs as comprising two variants within 20 bp of each other that phased to the same haplotype across >99% of all individuals in the dataset in which they appear (Figure 2.1a). This definition encompasses both MNVs due to a single mutational event and MNVs in which one SNV occurs after the other. To identify all possible candidate MNVs I searched for two heterozygous variants that were within 100bp of each other in the same individual across 6,688 DDD proband VCFs and had a read depth of at least 20 for each variant. The variants were phased using trio-based phasing, which meant that the ability to phase the variants was not dependent on the distance between them. I was able to determine phase for approximately 2/3 of all possible MNVs across all individuals. Those that could not be phased were discarded. The condition of phasing impacts the allele frequency spectrum of the MNVs I could identify since the probability of being able to phase a rare MNV in at least one individual is smaller than a more common MNV. There was a significantly smaller allele frequency in the variants that could not be phased compared to those that can be phased ($p = 1.8 \times 10^{-46}$,Wilcox test). I do not expect the mutational mechanisms to

differ at different allele frequencies although the rare MNVs may be more likely to be more damaging; this would make the assessment of pathogenicity more conservative. Phasing also provided an additional layer of quality assurance by requiring that the variant was called in both parent and child. MNVs tend to have lower mapping quality scores than SNVs and so traditional variant filtering criteria based on quality metrics could potentially miss a substantial number of MNVs. This also enabled me to use the same filtering criteria for different classes of variants to ensure comparability. The threshold distance of 20 bp between variants was selected as I observed that pairs of SNVs that define potential MNVs are only enriched for phasing to the same haplotype within this distance (Figure 2.1b).

De novo MNVs were defined as two *de novo* SNVs within 20bp of each other that were confirmed to be on the same haplotype using read based phasing. To identify *de novo* MNVs I looked within a set of 51,942 putative DNMs for pairs of *de novo* variants within 20 bp of each other. This set of DNMs had been filtered requiring a low minor allele frequency (MAF), low strand bias and low number parental alt reads. I did not impose stricter filters at this stage as true *de novo* MNVs tend to have worse quality metrics than true *de novo* SNVs. I found 301 pairs, approximately 1.2% of all candidate DNMs. A third of these were 1-2 bp apart (Figure 2.2a). For analysis of mutational spectra I did not filter these further however when looking at functional consequences of these *de novo* MNVs I wanted to be more stringent and examined IGV plots for all *de novo* MNVs of which 91 passed IGV examination. Ten of the *de novo* MNVs fell within genes previously associated with dominant developmental disorders. These were all validated experimentally using MiSeq or capillary sequencing. The experimental validation was done by Elena Prigmore.

In total, I identified 69,940 MNVs transmitted from the 13,376 unaffected trio parents as well as 91 *de novo* MNVs in the trio children. A set of 693,837 coding SNVs was obtained from the DDD probands with the exact same ascertainment as those for MNVs (read depth >20, phased to confirm inheritance). These were used when comparing MNV properties to SNVs to reduce any ascertainment bias.

Different mutational mechanisms are likely to create MNVs at different distances. To capture these differences, I stratified analyses of mutational spectra based on distance between the variants. The distance between the two variants that make up an MNV will be denoted as a subscript. For example, adjacent MNVs will be referred to as MNV_{1bp}. MNVs can be created by either a single mutational event or by consecutive mutational events. For MNVs that were created by a single mutational event, the pair of variants are likely to have identical allele frequencies as they are unlikely to occur in the population separately (I assume recurrent mutations and reversions are rare). The proportion of nearby pairs of SNVs with identical allele frequencies that phase to the same haplotype remains close to

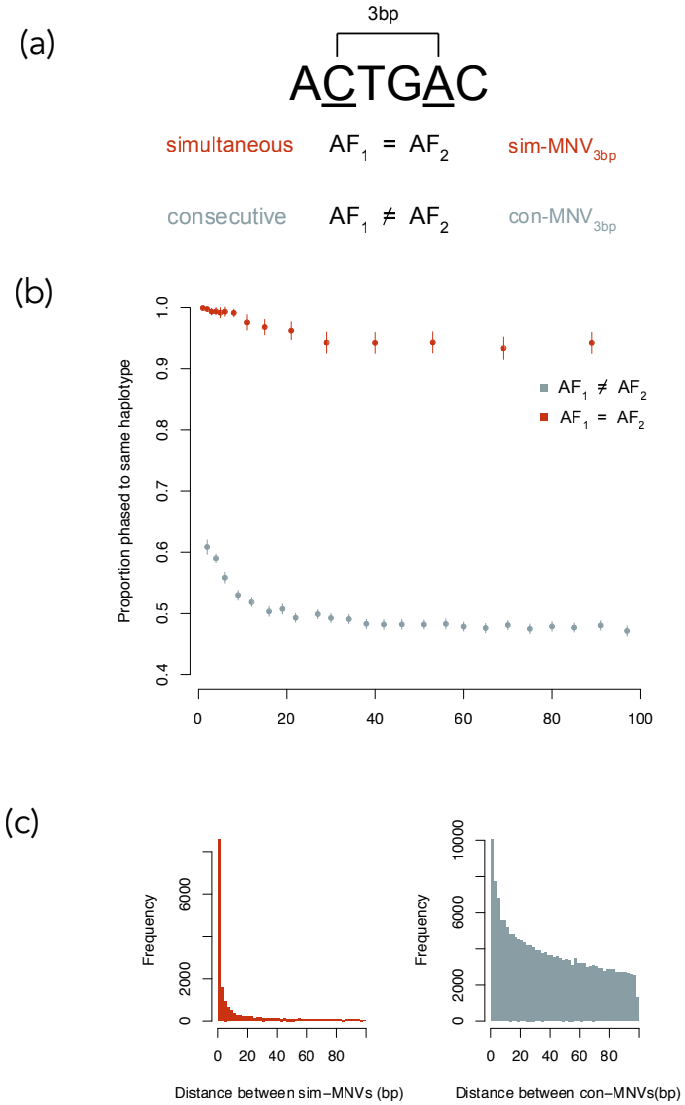

Fig. 2.1 Properties of MNVs (a) Schematic showing how sim-MNVs, two variants that occur simultaneously, are defined as having two variants with identical allele frequencies and con-MNVs, two variants that occur consecutively, as having different allele frequencies (b) Proportion of pairs of heterozygous variants (possible MNVs) that phase to the same haplotype as a function of distance separated by sim and con. (c) The number of sim-MNVs and con-MNVs by distance between the two variants.

100% even at a distance of 100 bp apart (Figure 2.1b). These variants most likely arose simultaneously and will be referred to as sim-MNVs. The proportion of pairs of SNVs with different allele frequencies that phase to the same haplotype approaches 50% at around 20bp. These SNVs probably arose consecutively and will be referred to as con-MNVs. I observed

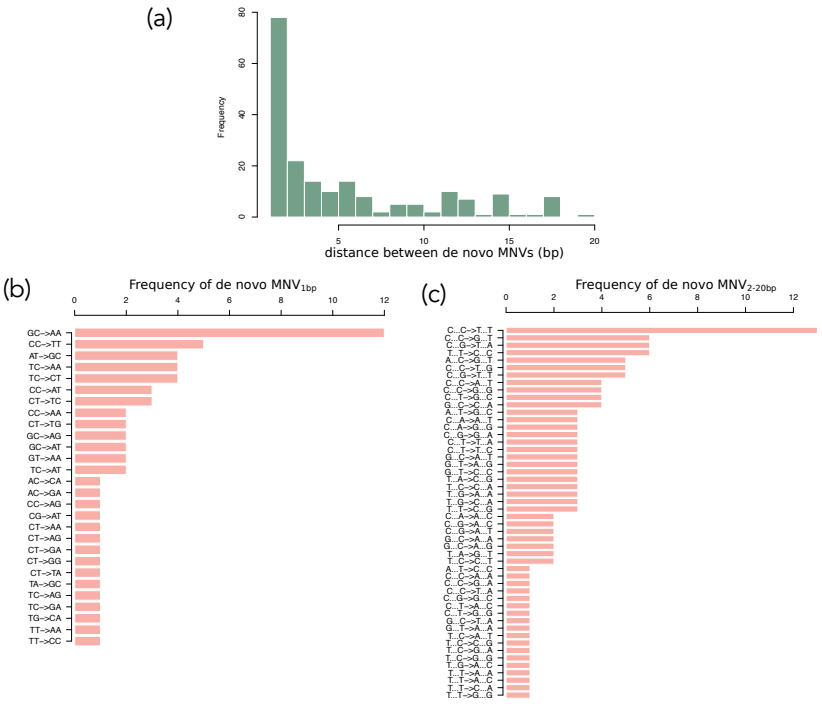

Fig. 2.2 Mutational spectra of *de novo* MNVs (a) Frequency of *de novo* MNVs according to the distance between the two variants in base pairs (b) Frequency of different mutation types for *de novo* MNV_{1bp} (c) Frequency of different mutation types for *de novo* $MNV_{2\text{-}20bp}$

that sim-MNVs account for 19% of all MNVs and 53% of MNV_{1bp}. All *de novo* MNVs are, by definition, sim-MNVs as they occurred in the same generation.

I identified 888 trinucleotide variants (trinucleotide sim-MNVs) which I defined as three SNVs within 20 bp with identical allele frequencies. One hundred and fourteen of these occurred in three adjacent nucleotides. I observed one *de novo* trinucleotide MNV.

2.3.2 Analysis of MNV mutational spectra

Differences in mutational spectra across different subsets of MNVs can reveal patterns or signatures generated by the underlying mutational mechanism. I analysed the spectra of both simultaneous and consecutive MNV_{1bp}, MNV_{2bp} and $MNV_{3\text{-}20bp}$. For sim-MNVs the proportion of variants that fell into these groups were 51%, 12% and 37% respectively. For con-MNVs, most variants were further away with the proportions being 10%,7% and 83% (Table 2.1, Figure 2.1c). There were significant differences between the mutational spectra of sim-MNVs and con-MNVs (Figure 2.3a,c)

DNA polymerase ζ, a translesion polymerase, is a known frequent source of *de novo* MNVs and has been associated with the mutational signatures GC->AA and TC->AA

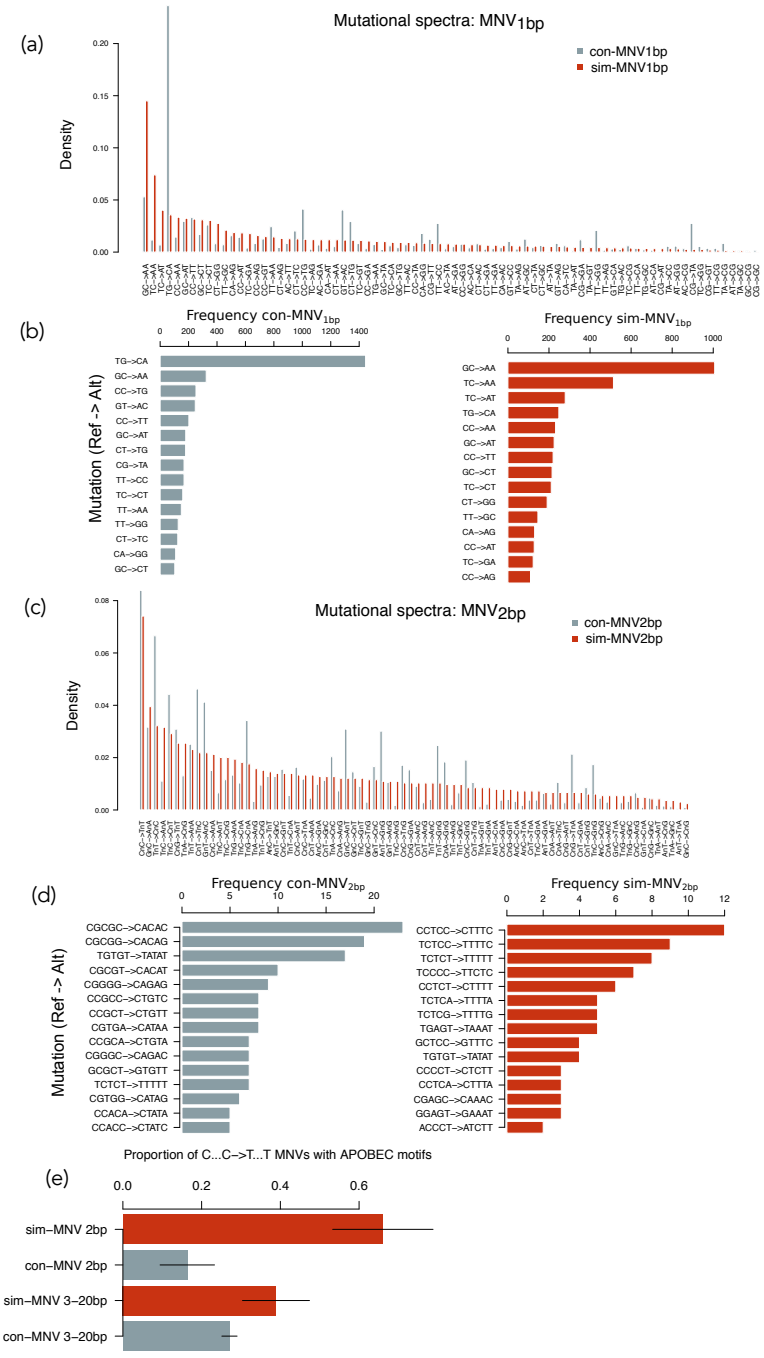

Fig. 2.3 Mutational Spectra of MNVs (a) The frequency of mutational spectra for sim-MNV_{1bp} and con-MNV_{1bp} (b) The 15 most common mutations for sim-MNV1bp and con-MNV_{1bp} (c) The frequency of mutational spectra for sim-MNV_{2bp} and con-MNV_{2bp} (d) The 15 most common mutations for sim-MNV_{2bp} and con-MNV_{2bp} (e) The proportion of C...C->T...T MNVs that have motifs associated with mutations caused by APOBEC.

MNV type	Distance (bp)	Intra Codon	Inter Codon	Non-coding	TOTAL (% of all MNVs)
sim	1	1893	863	3850	6606 (9.4%)
	2	243	350	975	1568 (2.2%)
	3-20	-	1832	2970	4802 (6.9%)
con	1	1155	735	3923	5813 (8.3%)
	2	449	685	2649	3783 (5.4%)
	3-20	-	15316	32052	47368 (67.7%)
TOTAL (% of all MNVs)		3740 (5.3%)	19781 (28.2%)	46419 (66.4%)	69940

Table 2.1 Numbers of MNVs in each category type

[76, 14]. These signatures, and their reverse complements are the most common dinucleotide changes that I observed and account for 22% of all sim-MNV_{1bp}s (Figure 2.3b). These two signatures made up 18% of the *de novo* sim-MNV_{1bp}s which is comparable to the 20% of observed *de novo* MNVs in a recent study (Figure 2.2b) [14]. In the remaining 78% of sim-MNV_{1bp}s I observed sixteen other mutations, after Bonferroni multiple correction, that were significantly more prevalent in sim-MNV_{1bp}s compared to con-MNV_{1bp}s. This suggests that there are other unidentified mechanisms that are specific to creating sim-MNVs. The most prevalent sim-MNV_{1bp} that is not attributed to polymerase ζ is TC>AT which accounts for 4% of all sim-MNV_{1bp}s. There were two *de novo* sim-MNV_{1bp}s with this signature however an extensive literature search, including somatic mutational signatures, has not yielded any possible mechanism behind this mutation [212].

APOBEC are a family of cytosine deaminases that are known to cause clustered mutations in exposed stretches of single-stranded DNA. These mutational signatures are commonly found in cancer and more recently discovered in germline mutations [181, 168]. The most common mutation for sim-MNV_{2bp} is CnC->TnT where n is the intermediate base between the two mutated bases and is 8% of the mutations (Figure 2.3c). They are found primarily in a TCTC>TTTT or CCTC>CTTT sequence context (Figure 2.3d). CC and TC are known mutational signatures of APOBEC[77, 5, 168]. However, the APOBEC signature described previously in germline mutations was found in pairs of variants that were a larger distance apart (10-50bp). C...C -> T...T was also the most prolific mutation in sim-$MNV_{3\text{-}20bp}$ and had a significantly larger proportion of APOBEC motifs in both variants compared to con-$MNV_{3\text{-}20bp}$bp (p value 0.0056) (Figure 2.3e). The mutation C...C -> T...T was the most frequent *de novo* $MNV_{2\text{-}20bp}$ (Figure 2.2c). However only three of the twelve *de novo* $MNV_{2\text{-}20bp}$ had APOBEC motifs.

There were 6 other mutations that are significantly more common in sim-MNV_{2bp} compared to con-MNV_{2bp}. The most prevalent of these is CnG>TnT which accounts for 3% of sim-MNV_{2bp}. I did not observe any *de novo* MNVs with this mutation and I was not able to attribute a mutational mechanism after reviewing the literature.

I analysed the mutational signatures of the set of 114 adjacent trinucleotide sim-MNVs and found that the most prevalent mutation was AAA>TTT (Figure 2.4) however was not able to establish a possible mutational mechanism for this.

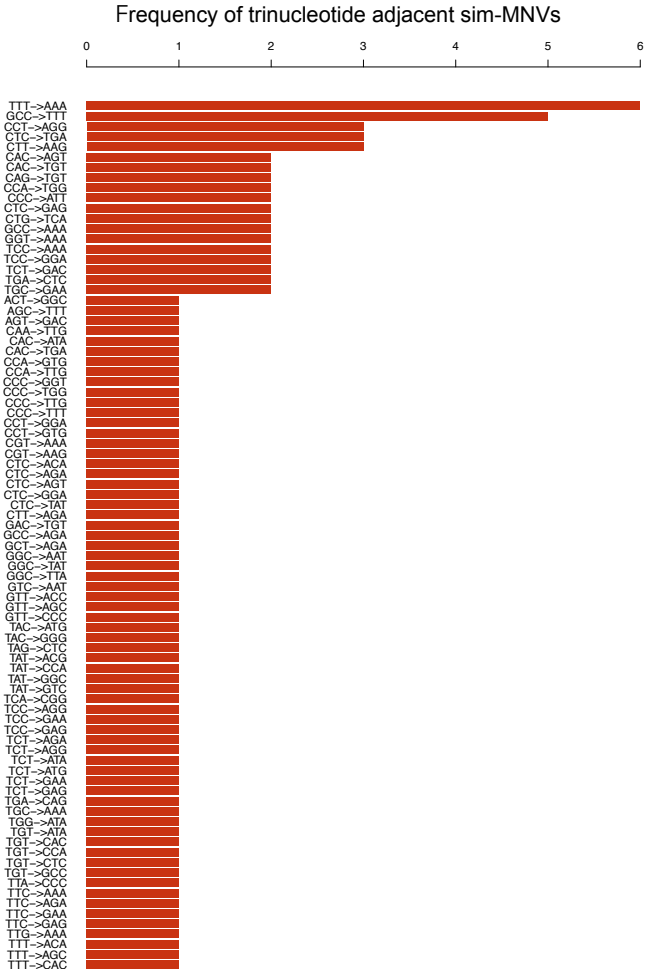

Fig. 2.4 Mutational spectra of all adjacent trinucleotide MNVs (N=114)

Mutational signatures in con-MNVs were primarily driven by hypermutability of CpG sites. In humans, the 5' C in a CpG context is usually methylated and has a mutation rate that is approximately ten-fold higher than any other context[45]. For con-$MNV_{2\text{-}20bp}$ the most common mutation is C...C->T...T and is driven by two mutated CpG sites CG...CG> TG...TG (Figure 2.3d). For con-MNV_{1bp}, 24% are accounted for by the mutation CA->TG, and its reverse complement (Figure 2.3b). These adjacent consecutive mutations most likely

came about due to a creation of a CpG site by the first mutation. If the first mutation creates a CpG then the mutations would be expected to arise in a specific order: CA>CG>TG. In this scenario, the A>G mutation would likely happen first and that variant would have a higher allele frequency than the subsequent C>T. This was the case for 96% of the 1,445 CA>TG con-MNV_{1bp}s. This was also the case for 96% and 92% of the other less common possible CpG creating con-MNVs CC>TG and AG>CA. CA>TG is probably the most common variant as it relies on a transition mutation A>G happening first which has a higher mutation rate compared to the transversions C>G and T>G. I identified 255 *de novo* con-MNVs and 26 of these were *de novo* con-MNV_{1bp}s. In half of these, the inherited variant created a CpG site which was then mutated *de novo* in our data.

I also observed that for con-MNV_{3bp}s that were not as a result of CpG creating sites, the first variant increases the mutability of the second variant more than expected by chance. I compared the median difference in mutation probability of the second variant based on the heptanucleotide sequence context before and after the first variant occurred using a signed Wilcoxon Rank Test[3]. The median increase in mutation probability of the second variant was 0.0002 ($p = 9.8 \times 10^{-17}$,signed Wilcoxon rank test).

2.3.3 Misannotation of MNVs

When an MNV occurs within a single codon, the consequence of this MNV can be different to the consequences when the two comprising variants are annotated separately. For 98% of the intra-codon MNVs, the consequence class (synonymous, missense, stop-gained etc.) of the MNV was the same as at least one of the SNVs annotated separately. For only 1% of the intra-codon MNVs was the consequence class of the MNV more severe than the separate SNVs. For almost all of these the MNV caused a stop-gain. Most intra-codon MNVs result in a missense change (Table 2.2) and so even though one of the comprising variants is most likely annotated as a missense separately as well, the MNV can create a different amino acid change.

MNV Consequence	Sim- MNV (% of all sim-MNVs)	Con-MNV (% of all con-MNVs)
Synonymous	10 (0.5%)	5 (0.3%)
1-step missense	815 (38.2%)	814 (50.7%)
2-step missense	1265 (59.2%)	757 (47.2%)
Stop Loss	2 (0.1%)	4 (0.2%)
Stop Gain	44 (2.0%)	24 (1.5%)

Table 2.2 Numbers and proportions of consequence types for MNVs within same codon

2.3.4 Functional Consequences of MNVs

The structure of the genetic code is not random. The code has evolved such that the codons that correspond to amino acids with similar physiochemical properties are more likely to be separated by a single base change [7, 233]. SNVs that result in a missense change will only alter one of the bases in a codon, however MNVs that alter a single codon (‘intra-codon’ MNVs) will alter two of the three base pairs. Therefore, they are more likely to introduce an amino acid that is further away in the codon table and thus less similar physicochemically to the original amino acid. Most intra-codon MNVs result in a missense change (Table 2.2). Intra-codon missense MNVs can be classified into two groups: ‘one-step’ and ‘two-step’ missense MNVs. One-step missense MNVs lead to an amino acid change that could also have been achieved by an SNV, whereas two-step MNVs generate amino-acid changes that could only be achieved by two SNVs. For example if we consider the codon CAC which codes for Histidine (H) then a single base change in the codon can lead to missense changes creating seven possible amino acids (Y,R,N,D,P,L,Q) (Figure 2.5a). There are one-step missense

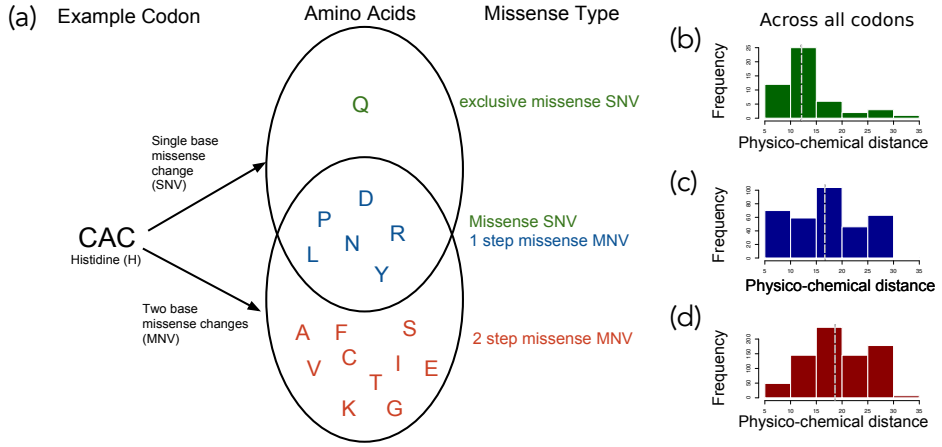

Fig. 2.5 Classification of intra-codon MNV missense mutations (a) Example of how one-step missense MNVs and two-step missense MNVs are classified using a single codon ’CAC’. Venn diagram shows amino acids that can be created with either a single base change or a two base change in the codon ’CAC’.(b-d) Across all codons the distribution of physiochemical distances for the amino acid changes caused by different types of missense variants, dashed line indicates the median of the distribution (b) exclusive SNV missense (c) one-step MNV missense (d) two-step MNV missense

MNVs within that codon that can lead to most of the same amino acids (Y,R,N,D,P,L). However two-step missense MNVs could also lead to an additional eleven amino acids that could not be achieved by an SNV (F,S,C,I,T,K,S,V,A,E,G). For some codons there are also amino acid changes that can only be created by a single base change, for this Histidine

codon this would be Glutamine (Q). These will be referred to as exclusive SNV missense changes. For this analysis I only considered sim-MNVs that most likely originated from the same mutational event. This is because I was primarily interested in the functional effects of mutations occurring simultaneously and where the amino acid produced would have changed directly from the original amino acid to the MNV consequence and not via an intermediate amino acid.

2.3.5 MNVs can create a missense change with a larger physico-chemical distance compared to missense SNVs

I assessed the differences in the amino acid changes between exclusive missense SNVs, one-step MNVs and two-step MNVs by examining the distribution of physicochemical distance for each missense variant type across all codons (Figure 2.5b). I used a distance measure between quantitative descriptors of amino acids based on multidimensional scaling of 237 physical-chemical properties[223]. I chose this measure as it does not depend on observed substitution frequencies which may create a bias due to the low MNV mutation rate making these amino acid changes inherently less likely. The median amino acid distance was significantly larger for two-step missense MNVs when compared to one-step missense MNVs ($p = 1.10 \times 10^{-7}$, Wilcoxon test). The median distance for one-step missense MNVs was also significantly larger from exclusive SNV missense changes (p = 0.0008, Wilcoxon test) (Figure 2.5b-d).

2.3.6 Missense MNVs are on average more damaging than missense SNVs

If the physico-chemical differences between these classes of missense variants resulted in more damaging mutations in the context of the protein, then I would expect to see a greater depletion of two-step missense MNVs compared to one-step missense MNVs or missense SNVs in highly constrained genes. I looked at the proportion of variants of different classes that fell in highly constrained genes, as defined by their intolerance of truncating variants in population variation, as measured by the probability of loss-of-function intolerance (pLI) score (Figure 2.6a). Highly constrained genes were defined as those with a pLI score >=0.9 [125]. MNVs that impact two nearby codons (inter-codon MNVs) are likely to have a more severe consequence on protein function, on average, than an SNV impacting on a single codon. I observed that the proportion of inter-codon $MNV_{1\text{-}20bp}$s that fall in highly constrained genes (pLI>0.9) is significantly smaller compared to missense SNVs (p = 0.0007, proportion test)

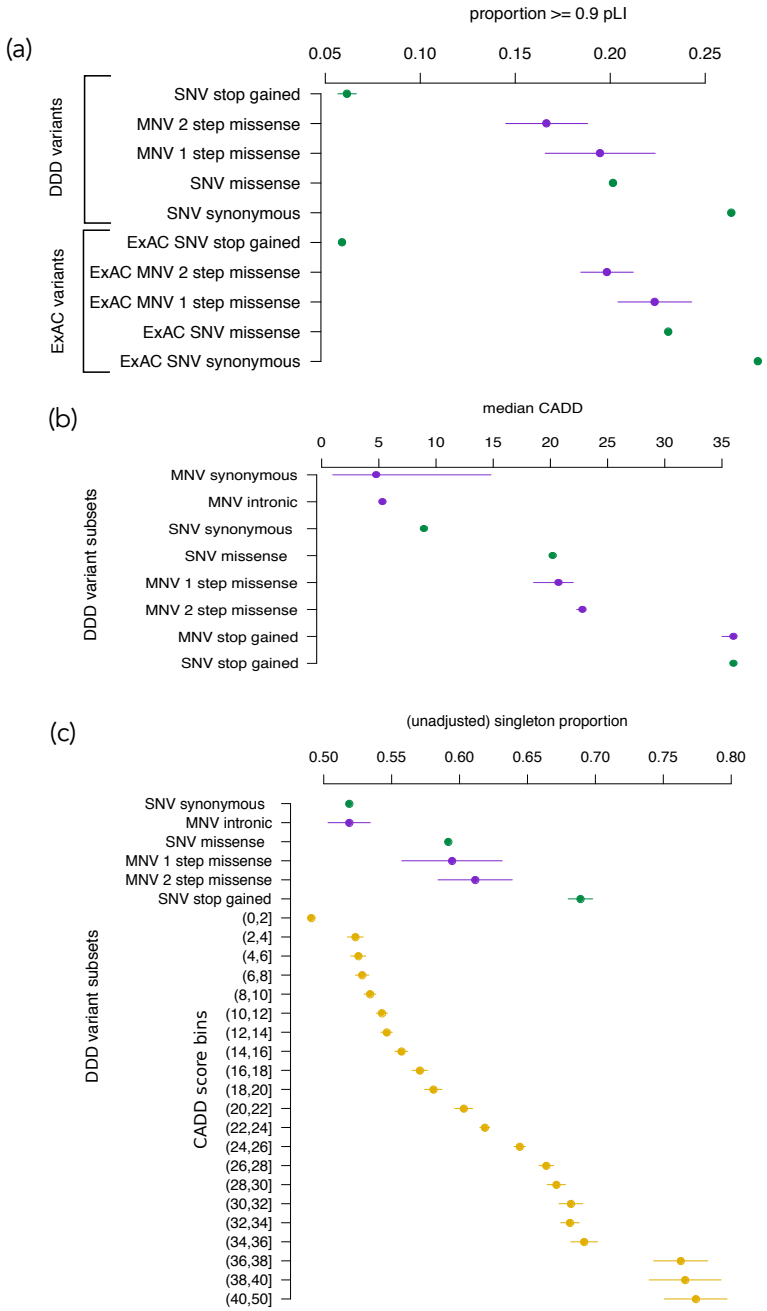

Fig. 2.6 Quantifying the pathogenicity of MNVs.(a) Proportion of variants that fall in genes with pLI >= 0.9 over different classes of variants for both DDD and ExAC datasets. Green are SNVs, Purple are MNVs. Lines are 95% confidence intervals.(b) The median CADD score over different classes of variants identified from DDD data with bootstrapped 95% confidence intervals.(c) Singleton proportion for different classes of DDD variants. In yellow are SNVs stratified by binned CADD scores with their corresponding singleton proportions. Lines are 95% confidence intervals.

(Figure 2.6a). For intra-codon MNVs, the proportion of two-step missense MNVs observed in highly constrained genes was also significantly smaller than for missense SNVs (p = 0.0016, Proportion test). The proportion of one-step missense MNVs was not significantly different from either missense SNVs or two-step missense MNVs. The analysis was repeated using SNVs and MNVs that were identified by the Exome Aggregation Consortium (ExAC) that were subject to different filtering steps [125]. The same relationship was observed, the proportion of ExAC two-step MNVs in high pLI genes was significantly smaller than for ExAC missense SNVs ($p = 9.84 \times 10^{-6}$).

I then compared variant deleteriousness across the variant classes using Combined Annotation Dependent Depletion (CADD) score that integrates many annotations such as likely protein consequence, constraint and mappability[107](Figure 2.6b). I found that the median CADD score for two-step missense MNVs was significantly higher than both one-step missense MNVs (p = 0.00017, Wilcoxon test) and missense SNVs ($p = 2.70 \times 10^{-8}$, Wilcoxon test). Two-step MNV missense had a median CADD score of 22.8 compared to a one-step missense median CADD score of 20.7 and a SNV missense median CADD score of 20.2.

The proportion of singletons across variant classes is a good proxy for the strength of purifying selection acting in a population[125]. The more deleterious a variant class is, the larger the proportion of singletons. We found that the singleton proportion for two-step missense MNVs was nominally significantly higher compared to missense SNVs (p = 0.02, proportion test) (Figure 2.6c). This increase in singleton proportion corresponded to an increase of about two in the interpolated CADD score. This is concordant with the increase in CADD scores that was computed directly above.

2.3.7 Estimation of the MNV mutation rate

I estimated the genome-wide mutation rate of sim-$MNV_{1\text{-}20bp}$s to be 1.78×10^{-10} mutations per base pair per generation by scaling the SNV mutation rate based on the relative ratio of segregating polymorphisms for MNVs and SNVs [226], see Methods. For this estimate I only used variants that fell into non-constrained genes (pLI<0.1) and non-coding regions to avoid any bias from ascertainment bias. I assumed that recurrent mutation is insufficiently frequent for both classes of variation to alter the proportionality between the number of segregating polymorphisms and the mutation rate. This estimate is ~1.6% the mutation rate estimate for SNVs and accords with the genome-wide proportions of SNVs and MNVs described previously [191]. I was concerned that the selective pressure on MNVs and SNVs might still be different in non-constrained genes and this could affect the mutation rate estimate. To see if this was the case, I applied the same method to estimate the SNV missense

mutation rate across coding region and found that the estimate was concordant with that obtained from using an SNV tri-nucleotide context mutational model[187]. I also estimated the MNV mutation rate using the set of *de novo* MNVs that fell into non-constrained genes (pLI<0.1) that have not previously been associated with dominant developmental disorders and obtained a concordant mutation rate estimate of 1.79×10^{-10} (confidence interval: 0.88×10^{-10}, 2.70×10^{-10}) mutations per base pair per generation, very similar to the estimate based on segregating polymorphisms described above.

2.3.8 Contribution of *de novo* MNVs to developmental disorders

I identified 10 *de novo* MNVs within genes known to be associated with dominant developmental disorders (DD-associated) in the DDD trios (Table 2.3), which is a significant (p = 1.03×10^{-3},Poisson test) 3.7 fold enrichment compared with what we would expect based on our estimated MNV mutation rate. This enrichment is similar in magnitude to that observed for *de novo* SNVs in the same set of DD-associated genes (Figure 2.8).

Decipher ID	Distance between variants	Chr	Positions	Gene	Ref	Alt	Consequence (first variant/ second variant)	MNV falls within/between codon	Clinician pathogenicity annotation on Decipher
261423	1	5	161569244, 161569245	*GABRG2*	CC	TT	missense (two step)	Within codon	Likely pathogenic (Full)
292136	1	14	29237129, 29237130	*FOXG1*	TC	CT	missense (one step)	Within codon	Likely pathogenic (Full)
280956	1	19	13135878, 13135879	*NFIX*	GC	TT	missense (one step)	Within codon	Likely pathogenic (Partial)
270803	1	3	49114312, 49114313	*QRICH1*	GC	AA	stop gain/missense	Between codon	Likely pathogenic (Partial)
258688	1	5	67591021, 67591022	*PIK3R1*	TA	GC	missense/missense	Between codon	Likely pathogenic (Full)
274482	1	16	30749053, 30749054	*SRCAP*	GG	AT	synonymous/ stop gain	Between codon	Definitely pathogenic (Full)
274606	1	9	140637863, 140637864	*EHMT1*	GA	TT	missense/stop gain	Between codon	Likely pathogenic (Full)
274453	1	9	140637863, 140637864	*EHMT1*	GA	TT	missense/stop gain	Between codon	Definitely pathogenic (Full)
260753	13	6	157454286, 157454297	*ARID1B*	G..C	T..G	missense/stop gain	Between codon	Definitely pathogenic (Full)
270916	3	1	7309651, 7309654	*CAMTA1*	G..G	A..A	missense/missense	Between codon	Likely pathogenic (partial)

Table 2.3 *De Novo* MNVs that fall in genes associated with developmental disorders

To assess the sensitivity of this enrichment to the estimate of the MNV mutation rate I recalculated this by using an MNV mutation rate estimate based on all variants, as opposed to excluding those that fall in DDG2P genes, as well as a more stringent estimate just using variants that fell into non-coding regions. When I redid the enrichment analysis using these mutation rate estimates of varying stringency, the enrichment of *de novo* MNVs in DD-associated genes remained significant (all variants p = 2.7×10^{-4}, non coding control regions p = 4.9×10^{-3}, Figure 2.7a). The SNV mutation rate estimate varies across studies therefore I also recalculated the MNV mutation rates using SNV mutation rate estimates of 1.0×10^{-8} and 1.2×10^{-8} mutations per base pair per generation [195]. These were also recalculated across the three different variant subsets (all variants, excluding variants in

genes with pLI>0.1, variants in non-coding control regions). The enrichment ratio of *de novo* MNVs that fall into DD genes ranged from 2.7 to 4.8 however always remained significantly greater than 1 and the confidence intervals consistently overlapped with that of the SNV missense enrichment ratio (Figure 2.7b).

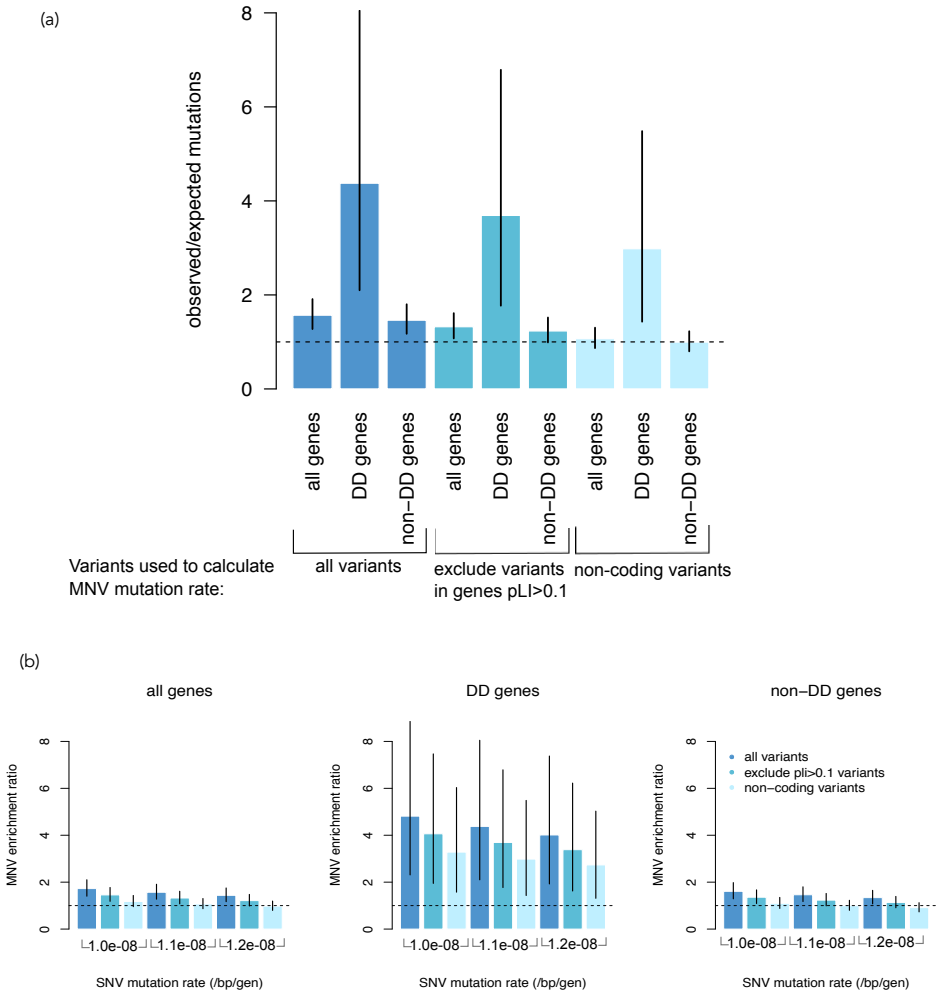

Fig. 2.7 Sensitivity of MNV enrichment analysis to MNV mutation rate estimates (a) The impact of varying the subsets of variants used to estimate the MNV mutation rate estimate on the enrichment of *de novo* MNVs in different subcategories of genes as in Figure 2.8. These were all calculated using an SNV mutation rate estimate of 1.1×10^{-8} /bp/generation. (b) Using three different estimates of the SNV mutation rate estimate and the subcategories of variants as in (a) looking at the difference in enrichment ratios across the same subcategories of genes as in (a).

To ensure this observed enrichment was not driven by differences in sequence context, I also evaluated whether DD-associated genes are enriched for the primary mutagenic dinucleotide contexts associated with the signatures of polymerase ζ. I found that DD-

associated genes had a small (1.02 fold) but significant ($p = 1.9 \times 10^{-59}$,proportion test) enrichment of polymerase ζ dinucleotide contexts compared to genes not associated with DD. However, this subtle enrichment is insufficient to explain the four-fold enrichment of *de novo* MNVs in these genes. The enrichment for *de novo* MNVs remains significant after correcting for this sequence context ($p = 2.28 \times 10^{-3}$, Poisson test).

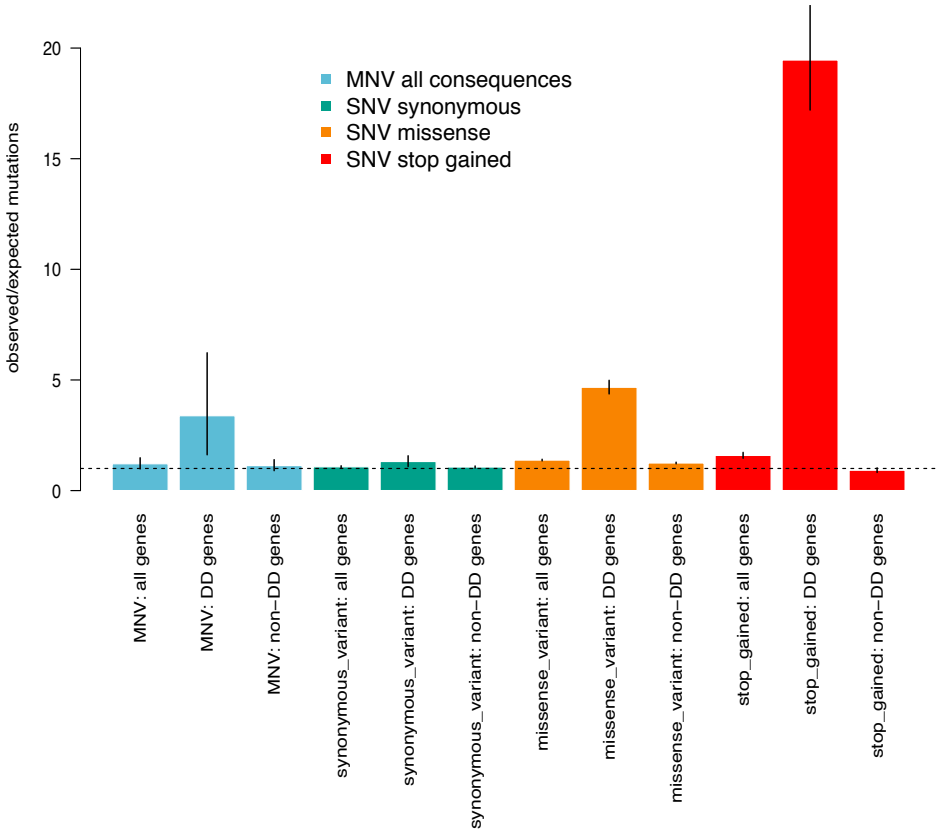

Fig. 2.8 Enrichment of *de novo* MNVs in DDD study. Ratio of observed number of *de novo* MNVs vs the expected number of *de novo* MNVs based on the estimate of the MNV mutation rate. Compared to enrichment of SNVs in DD genes in consequence classes synonymous, missense and stop gain.

Eight of the 10 *de novo* MNVs in DD-associated genes were 1 bp apart while the other two were 3 and 13 bp apart. All of these *de novo* MNVs were experimentally validated in the child (and their absence confirmed in both parents) using either MiSeq or capillary sequencing. This validation was done by Elena Prigmore. All ten MNVs were thought to be pathogenic by the child's referring clinical geneticist. Seven of the MNVs impacted two different codons while three fell within the same codon, one of which created a two-step missense change. Of those MNVs that impacted two codons, five caused a premature stop

codon. I found a recurrent *de novo* MNV in the gene *EHMT1* in two unrelated patients that bore the distinctive polymerase ζ signature of GA>TT.

2.3.9 Clinically reported MNVs in DD-associated genes

To assess whether MNVs are being underreported in genes associated with DD, I downloaded all clinically reported variants in DD-associated genes from ClinVar (accessed September 2017, [115]). I looked at the number of intra-codon missense MNVs in genes that have at least one reported pathogenic missense mutation. This was to ensure that missense mutations in that gene would likely cause DD. I focused on intra-codon MNVs as it is the interpretation of this class of MNV that is most impacted by failing to consider the variant as single unit. I calculated the expected number of pathogenic MNVs in these genes based on the MNV mutation rate and the number of pathogenic SNV missense variants reported. There were 22 reported pathogenic MNVs compared to the expected number of 26 across 321 genes which was not significantly different (p = 0.55, Poisson test). I also looked for clinically relevant SNVs in ClinVar that overlapped with inherited sim-MNVs that I identified in our data. I found one SNV that had been reported as a nonsense variant in the gene *AGPAT2*. The variant had been reported as pathogenic and of uncertain significance for congenital generalised lipdystrophy type 1 by two contributors. However I observe this variant as an MNV in our dataset in three individuals. The MNV falls within the same codon and causes a missense as opposed to a stop gain which decreases its likelihood of pathogenicity, especially since it was also observed 70 times in ExAC (Allele Frequency 3.96^{-4}).

2.3.10 MNV mutator phenotype

Five DDD children had more than one *de novo* sim-MNVs. This is significantly greater than what we would expect assuming these MNVs arose independently. Using our estimated MNV mutation rate, the probability of seeing five or more individuals in our data set with more than one MNV is 5.8×10^{-7}. The number of MNVs per person range from 2-5 *de novo* MNVs. These mostly appear on different chromosomes and have different distances between the pair of variants. None of the MNVs share the same mutation and the number of mutations is too small to pick up on more subtle similarities in the mutational signatures. A comparable mutator phenotype has been observed in other classes of genetic variation such as CNVs but, similarly, a relevant mutational mechanism has not yet been discovered [131]. A larger number of *de novo* MNVs will help to uncover possible mechanisms behind this apparent mutator phenotype.

2.4 Discussion

MNVs constitute a unique class of variant, both in terms of mutational mechanism and functional impact. I found that 18% of segregating MNVs were at adjacent nucleotides and estimated that 19% of all MNVs represent a single mutational event, increasing to 53% of MNV_{1bp}. I estimated the sim-MNV germline mutation rate to be 1.78×10^{-10} mutations per base pair per generation, roughly 1.6% that of SNVs. Most population genetics models assume that mutations arise from independent events [76]. MNVs violate that assumption and this may affect the accuracy of these models. Recent studies suggest that certain phylogenetic tests of adaptive evolution incorrectly identify positive selection when the presence of these clustered mutations are ignored [222]. Correcting these population genetic models will require knowledge of the rate and spectrum of MNV mutations. The observation of a possible MNV mutator phenotype complicates this correction further. In the future it would be of interest to whole-genome sequence those individuals with a potential MNV mutator phenotype to uncover the causal underlying mechanisms. I replicated the observations from previous studies that several different mutational processes underlie MNV formation, and that these tend to create MNVs of different types. Error-prone polymerase ζ predominantly creates sim-MNV_{1bp} [76, 14]. APOBEC-related mutation processes have been described to generate MNVs in the range of 10-50bp [181, 5, 77], but here I show that an enrichment for APOBEC motifs can be detected down to MNV_{2bp}. Nonetheless, there remain other sim-MNVs that cannot be readily explained by either of these mechanisms, and it is likely that other, less distinctive, mutational mechanisms remain to be delineated as catalogs of MNVs increase in scale. These future studies should also investigate whether these MNV mutational signatures differ subtly between human populations as has been recently observed for SNVs [75]. Consecutive MNVs, by contrast, exhibit greater similarity with known SNV mutation processes, most notably with the creation and subsequent mutation of mutagenic CpG dinucleotides. The non-Markovian nature of this consecutive mutation process challenges Markovian assumptions that are prevalent within standard population genetic models [179].

These findings validated the intuitive hypothesis that MNVs that impact upon two codons within a protein are likely, on average, to have a greater functional impact than SNVs that alter a single codon. I evaluated the functional impact of intra-codon MNVs using three complementary approaches: (i) depletion within genes under strong selective constraint, (ii) shift towards rarer alleles in the site frequency spectrum and (iii) enrichment of *de novo* mutations in known DD-associated genes in children with DDs. I demonstrated that intra-codon MNVs also tend to have a larger functional impact than SNVs, and that MNV missense changes that cannot be achieved by a single SNV are, on average, more deleterious than those

that can. This is most likely due to the fact that they are on average more physico-chemically different compared to amino acids created by SNVs and are not as well tolerated in the context of the encoded protein. These 'two-step' missense MNVs make up more than half of all sim-MNVs that alter a single codon. I also identified 10 pathogenic *de novo* MNVs within the DDD study, including both intra-codon and inter-codon MNVs. With larger trio datasets we will have more power to tease apart more subtle differences in pathogenic burden and purifying selection between different classes of MNVs and SNVs, for example, to test whether two-step missense *de novo* MNVs are more enriched than missense SNVs or one-step missense MNVs in developmental disorders. More data will also allow us to assess the population genetic properties of inter-codon MNVs.

These findings emphasise the critical importance of accurately calling and annotating MNVs within clinical genomic testing both to improve diagnostic sensitivity and to avoid misinterpretation. While MNVs are not underrepresented in reported clinically reported variants in ClinVar, we did observe that pathogenic *de novo* MNVs can be mis-annotated, indicating that current analytical workflows may not be calling these correctly. In a recent comparison of eight different variant calling tools it was noted that only two callers, FreeBayes and VarDict, report two mutations in close proximity as MNVs. The others reported them as two separate SNVs [188]. Both FreeBayes and VarDict are haplotype aware callers which is necessary for MNV detection [56, 114]. Even if variant callers do not identify MNVs directly, software also exists that can correct a list of previously called SNVs to identify mis-annotated MNVs [228]. To further our understanding of the role of MNVs in evolution and disease, calling and annotating these variants correctly is a vital step.