Exome-wide assessment of the functional impact and pathogenicity of multi-nucleotide mutations

Joanna Kaplanis[1], Nadia Akawi[5], Giuseppe Gallone[1], Jeremy F. McRae[1], Elena Prigmore[1], Caroline F. Wright[2], David R. Fitzpatrick[3], Helen V. Firth[1,4], Jeffrey C. Barrett[1], Matthew E. Hurles[1*] on behalf of the Deciphering Developmental Disorders study

Affiliation

1. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
2. Institute of Biomedical and Clinical Science, University of Exeter Medical School, RILD Level 4, Royal Devon & Exeter Hospital, Barrack Road, Exeter, EX2 5DW, UK
3. MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK
4. Department of Clinical Genetics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
5. Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

*Correspondence to: meh@sanger.ac.uk

Abstract

Approximately 2% of *de novo* single nucleotide variants (SNVs) appear as part of clustered mutations that create multinucleotide variants (MNVs). MNVs are an important source of genomic variability as they are more likely to alter an encoded protein than a SNV, which has important implications in disease as well as evolution. Previous studies of MNVs have focused on their mutational origins and have not systematically evaluated their functional impact and contribution to disease. We identified 69,940 MNVs and 106 de novo MNVs in 6,688 exome sequenced parent-offspring trios from the Deciphering Developmental Disorders Study comprising families with severe developmental disorders. We replicated the previously described MNV mutational signatures associated with DNA polymerase zeta, an error-prone translesion polymerase, and the APOBEC family of DNA deaminases. We found that most MNVs within a single codon create a missense change that could not have been created by a SNV. MNV-induced missense changes were, on average, more physico-chemically divergent, more depleted in highly constrained genes (pLI>=0.9) and were under stronger purifying selection compared to SNV-induced missense changes. We found that *de novo* MNVs were significantly enriched in genes previously associated with developmental disorders in affected children. This demonstrates that MNVs can be more damaging than SNVs even when both induce missense changes and are an important variant type to consider in relation to human disease.

Main Text

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 (Michaelson et al. 2012; Seidman et al. 1987; Amos 2010). 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(Segurel, Wyman, and Przeworski 2014; Stone et al. 2012; Harris and Nielsen 2014). While some of this can be explained by the presence of mutational hotspots, natural selection or compensatory mechanisms, it has been shown that multi-nucleotide mutations play an important role (Schrider, Hourmozdi, and Hahn 2011). Recent studies found that 2.4% of *de novo* SNVs were within 5kb of another *de novo* SNV within the same individual (Besenbacher et al. 2016), and that 1.9% of *de novo* SNVs appear within 20bp of another *de novo* SNV (Schrider, Hourmozdi, and Hahn 2011). Multi-nucleotide variants (MNVs) occurring at neighbouring nucleotides are the most frequent of all MNVs (Besenbacher et al. 2016). 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 (Schrider, Hourmozdi, and Hahn 2011).

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 zeta is an error-prone translesion polymerase that has been shown to be the predominant source of *de novo* MNVs in adjacent nucleotides in yeast (Harris and Nielsen 2014; Besenbacher et al. 2016). The most common mutational signatures associated with polymerase zeta in yeast have also been observed to be the most common signature among MNVs in human populations (Harris and Nielsen 2014), and were also found to be the most prevalent in *de novo* MNVs in parent-offspring trios (Besenbacher et al. 2016). A distinct mutational signature has also been described that has been attributed to the action of APOBEC deaminases (Alexandrov et al. 2013).

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 (Sandmann et al. 2017). 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 in genes, including some within known disease-associated genes(Lek et al. 2016). Although individual pathogenic MNVs have been described ('ClinVar'), the pathogenic impact of MNVs as a class of variation is not yet well understood.

Here we 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.

Results

Identifying and categorising MNVs

We identified 69,940 MNVs transmitted from the 13,376 unaffected trio parents as well as 106 *de novo* MNVs in the trio children. We defined MNVs as comprising two variants within 20bp of each other that phased to the same haplotype across >99% of all individuals in the dataset in which they appear (Figure 1a). This definition encompasses both MNVs due to a single mutational event and MNVs in which one SNV occurs after the other. 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 and it also provided an additional layer of quality assurance by conditioning on the variant being called in both parent and child. MNVs tend to have lower mapping quality than SNVs and so traditional variant filtering criteria based on quality metrics would potentially miss a substantial number of MNVs. This also enabled us to use the same filtering criteria for different classes of variants to ensure comparability. The distance of 20bp between variants was selected as we observed that pairs of SNVs that define potential MNVs are only enriched for phasing to the same haplotype at this distance (Figure 1b). *De novo* MNVs were defined as two *de novo* SNVs within 20bp of each other and were confirmed to be on the same haplotype using read based phasing. Due to the small numbers we were able to filter these by manually inspecting these variants using the Integrative Genomics Viewer (IGV). 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.

Figure 1

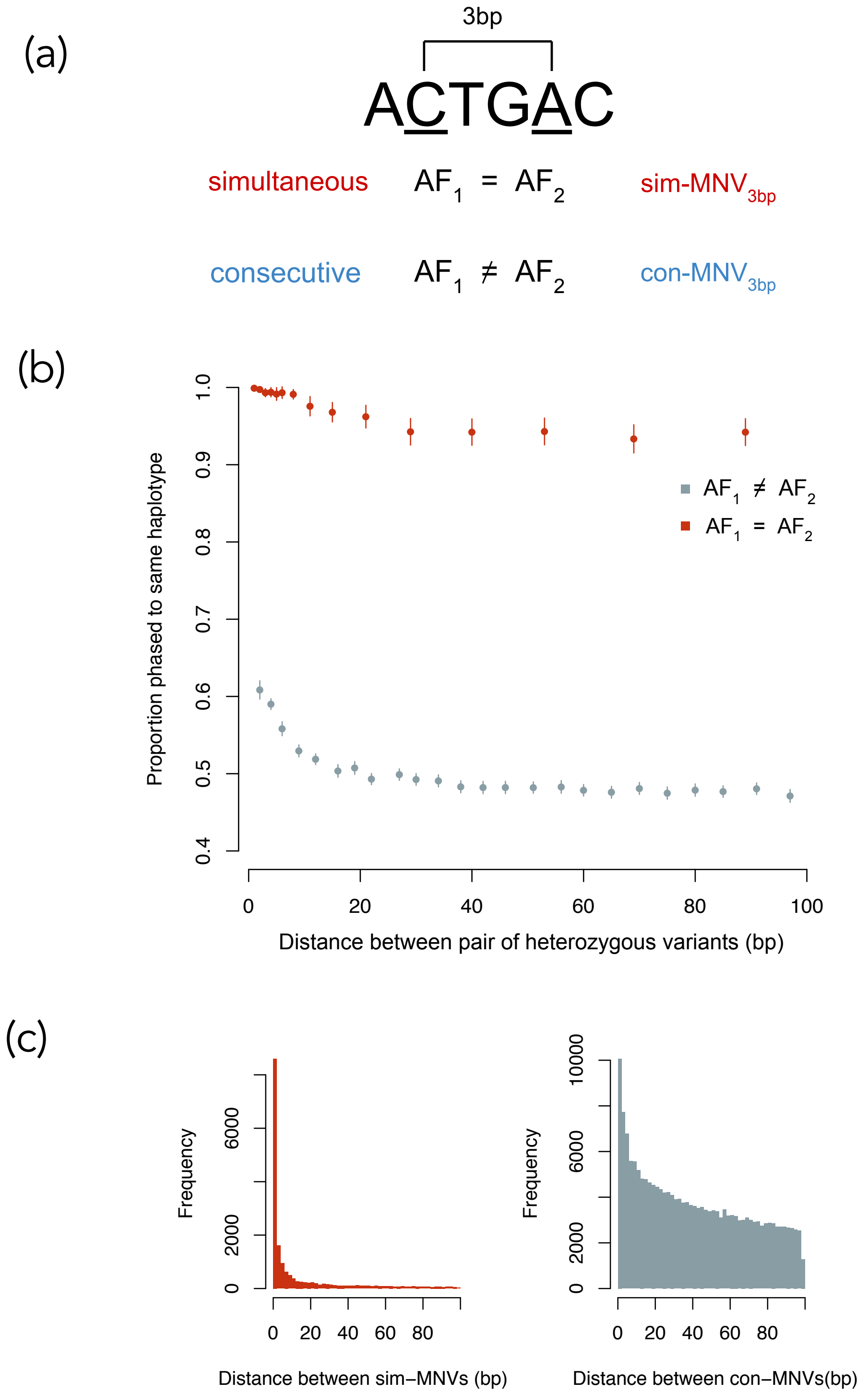

Figure 1: Properties of MNVs (a) Schematic showing how sim-MNVs, two variants that occur simultaneously, are defined as having two variants with identical allele frequencies and con-MNVs, two variants that occur consecutively, as having different allele frequencies (b) Proportion of pairs of heterozygous 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.

Different mutational mechanisms are likely to create MNVs at different distances. To capture these differences, we 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 (we 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 100% even at a distance of 100bp apart (Figure 1b). We can assume that 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 probably arose consecutively and will be referred to as con-MNVs. We observed 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.

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 1: Numbers of MNVs in each category type

Analysis of MNV mutational spectra confirms mutational origins.

Differences in mutational spectra across different subsets of MNVs can reveal patterns or signatures left by the underlying mutational mechanism. We analysed the spectra of both simultaneous and consecutive MNV_{1bp}, MNV_{2bp} and $\mathrm{MNV}_{3\text{-}20bp}$.

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 1). We observed significant differences between the mutational spectra of sim-MNVs and con-MNVs (Figure S1a, S1c).

DNA polymerase zeta, a translesion polymerase, is a known frequent source of *de novo* MNVs and has been associated with the mutational signatures GC->AA and GA->TT (Harris and Nielsen 2014; Besenbacher et al. 2016). These signatures, and their reverse complements, account for 22% of all sim-MNV$_{1bp}$s (Figure S1b). 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 S2b) (Besenbacher et al. 2016).

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 (Roberts et al. 2013; Pinto et al. 2016). The most common mutation for sim-MNV$_{2bp}$ is CnC->TnT where n is the intermediate base between the two mutated bases and for ~8% of the mutations (Figure S1c). They are found primarily in a TCTC>TTTT or CCTC>CTTT sequence context (Figure S1d). CC and TC are known mutational signatures of APOBEC(Harris 2013; Alexandrov et al. 2013; Pinto et al. 2016). However, the APOBEC signature described previously in germline mutations were 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$_{3\text{-}20bp}$ (p value 0.0056) (Figure S1e). The mutation C…C -> T…T was the most frequent *de novo* MNV$_{2\text{-}20bp}$ (Figure S2c). However only three of the twelve *de novo* MNV$_{2\text{-}20bp}$ had APOBEC motifs.

Mutational signatures in con-MNVs were primarily driven by 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(Duncan and Miller 1980). For con-MNV$_{2\text{-}30bp}$ the most common mutation is C…C->T…T and is driven by two mutated CpG sites CG…CG> TG…TG (S1d). For con-MNV$_{1bp}$s, 24% are accounted for by the mutation CA->TG, and its reverse complement (S1b). 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. We would therefore expect that the A>G mutation would 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. We also observed that for con-MNV$_{1\text{-}3bp}$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. We 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 (Aggarwala and Voight 2016). The median increase in mutation probability of the second variant was 0.0002 (signed Wilcoxon rank test p-value 9.8×10^{-17}).

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(Amirnovin 1997; Wong 1975). 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). 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 2a). There are one-step 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 we only considered sim-MNVs that most likely originated from the same mutational event. This is because we were 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.

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: Numbers and proportions of consequence types for MNVs within same codon

Figure 2

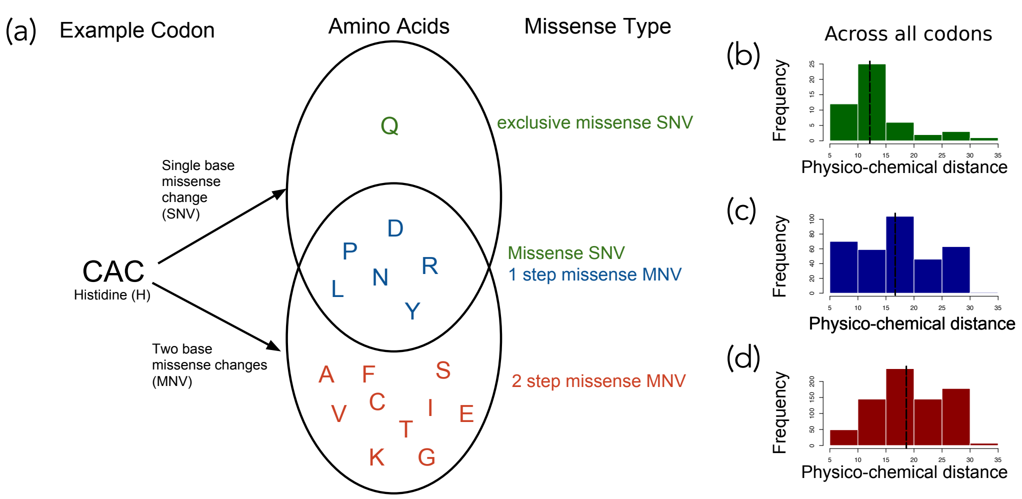

Figure 2: Classification of intra-codon MNV missense mutations (a) Example of how one-step missense MNVs and two-step missense MNVs are classified using a single codon 'CAC'. Venn diagram shows amino acids that can be created with either a single base change or a two base change in the codon 'CAC'. (b-d) 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 can create a missense change with a larger physico-chemical distance compared to missense SNVs

We 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 2b). We used a distance measure between quantitative descriptors of amino acids based on multidimensional scaling of 237 physical-chemical properties(Venkatarajan and Braun 2001). We 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. We found that the median amino acid distance was significantly larger for two-step missense MNVs when compared to one-step missense MNVs (Wilcoxon test, p-value 1.10e-07). The median distance for one-step missense MNVs was also significantly larger from exclusive SNV missense changes (Wilcoxon test, p-value 0.0008) (Figure 2 b-d).

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 we 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. We 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 3a). Highly constrained genes were defined as those with a pLI score >=0.9 (Samocha et al. 2014). 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. We observed that the proportion of inter-codon MNV_{1-20bp} that fall in highly constrained genes (pLI>0.9) is significantly smaller compared to missense SNVs (p-value 0.0007) (Figure 3a). For intra-codon MNVs, we saw that the proportion of two-step missense MNVs observed in highly constrained genes was also significantly smaller than for missense SNVs (p-value: 0.0016). 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(Lek et al. 2016). 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-value: 9.84e-06).

We 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 (Kircher et al. 2014). We found that the median CADD score for two-step missense MNVs was significantly higher than both one-step missense MNVs (Wilcoxon test, p value 0.00017) and

Figure 3

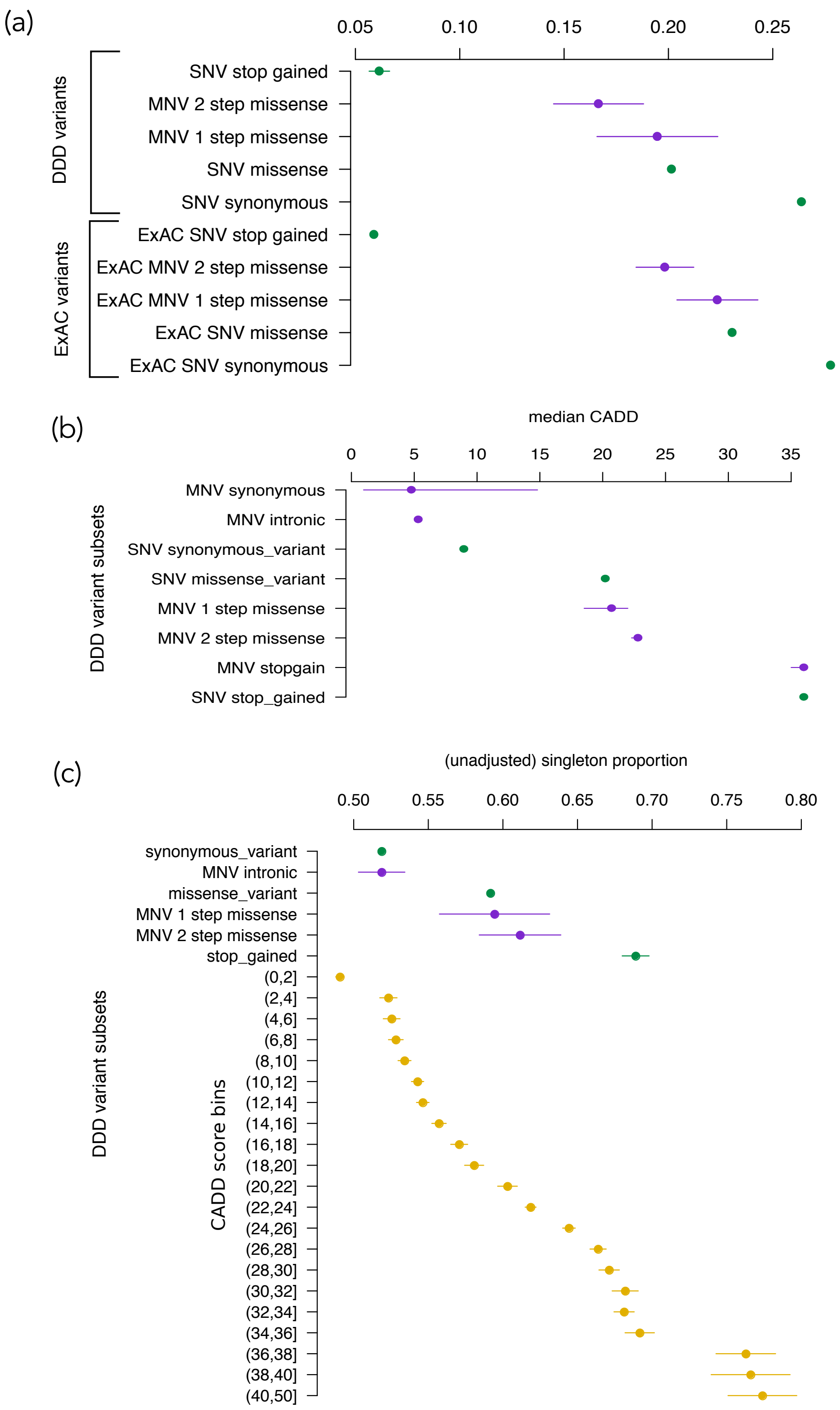

Figure 3: Quantifying the pathogenicity of MNVs (a) Proportion of variants that fall in genes with pLI >= 0.9 over different classes of variants for both DDD and ExAC datasets. Green are SNVs, Purple are MNVs. Lines are 95% confidence intervals (b) The median CADD score over different classes of variants 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.

missense SNVs (Wilcoxon test, p value $2.70x10^{-8}$). 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 (Lek et al. 2016). The more deleterious a variant class, 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-value 0.02). The increase in 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.

Contribution of *de novo* MNVs to developmental disorders

We estimate the genome-wide mutation rate of sim-$MNV_{1\text{-}20bp}$ to be $1.78x10^{-10}$ mutations per base pair per generation by scaling the SNV mutation rate based on the relative ratio of segregating polymorphisms for MNVs and SNVs (Watterson 1975). For this estimate we only used variants that fell into non-constrained genes (pLI<0.1) and non-coding regions to avoid any bias from selection. We assume 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 (Schrider, Hourmozdi, and Hahn 2011). We were concerned that the selective pressure on MNVs and SNVs would still be different in non-constrained genes and this might affect our mutation rate estimate. To see if this was the case, we applied the same method to estimate the SNV missense mutation rate across coding region and found that our estimate was concordant with that obtained from using an SNV tri-nucleotide context mutational model (Samocha et al. 2014).

We 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 1.80x10^{-10} (confidence interval 0.88, 2.70 x 10^{-10}) mutations per base pair per generation, very similar to the estimate based on segregating polymorphisms described above.

We identified 10 *de novo* MNVs within genes known to be associated with dominant developmental disorders (DD-associated) in the DDD trios (Table 3), which is a significant (Poisson test, p value 1.03 x 10^{-3}) 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 4). We evaluated whether DD-associated genes are enriched for the primary mutagenic dinucleotide contexts associated with the signatures of polymerase zeta to ensure this observation was not driven by sequence context. We found that DD-associated genes had a small (1.02 fold) but significant (proportion test, p-value 1.9x10^{-59}) enrichment of polymerase zeta 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 sequence context (Poisson test, p value 2.28 x 10^{-3}).

Eight of the 10 *de novo* MNVs in DD-associated genes were 1bp apart while the other two were 3 and 13bp 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. All ten MNVs are 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. Interestingly we found a recurrent *de novo* MNV in the gene *EHMT1* in two unrelated patients that bore the distinctive polymerase zeta signature of GA>TT.

Figure 4

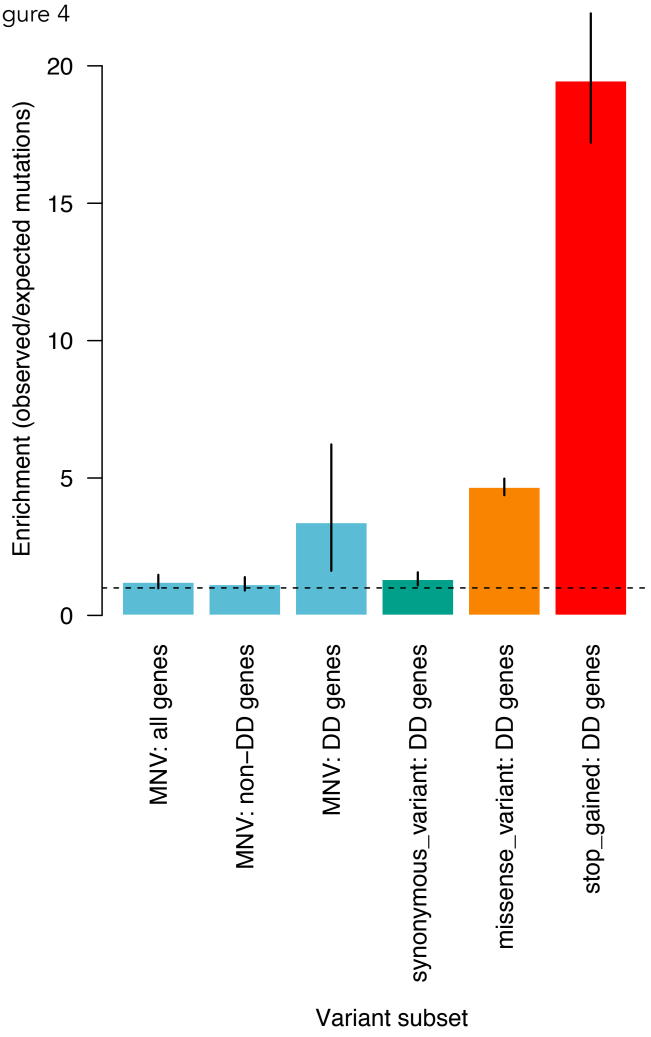

Figure 4: Enrichment of de novo MNVs in DDD study Ratio of observed number of de novo MNVs vs the expected number of de novo MNVs based on the estimate of the MNV mutation rate. Compared to enrichment of SNVs in DD genes in consequence classes synonymous, missense and stop gain.

De Novo MNVs are underrepresented in clinically reported variants in DD-associated genes

To assess whether de novo MNVs are being underreported in genes associated with DD, we downloaded all clinically reported variants in DD-associated genes from ClinVar (accessed September 2017). We 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. We 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. We calculated the expected number of pathogenic de novo MNVs in these genes based on the MNV mutation rate and the number of pathogenic SNV missense variants reported. We observed a significant depletion of only 24 reported pathogenic de novo MNVs compared to the expected number of 52 across 321 genes (p-value 2.8x10^{-5}, Poisson test).

Decipher ID	Distance between variants	Chrom	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 3: De Novo MNVs that fall in genes associated with developmental disorders

Discussion

MNVs constitute a unique class of variant, both in terms of mutational mechanism and functional impact. We found that 18% of segregating MNVs were at adjacent nucleotides. We estimated that 19% of all MNVs represent a single mutational event, increasing to 53% of MNV_{1bp}. We estimated the sim-MNV germline mutation rate to be $1.78\text{x}10^{-10}$ mutations per base pair per generation, roughly 1.6% that of SNVs. Most population genetics models assume that mutations arise from independent events (Harris and Nielsen 2014). 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 (Venkat, Hahn, and Thornton 2017). Correcting these population genetic models will require knowledge of the rate and spectrum of MNV mutations. We 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 zeta predominantly creates sim-MNV_{1bp} (Harris and Nielsen 2014; Besenbacher et al. 2016). APOBEC-related mutation processes have been described to generate MNVs in the range of 10-50bp (Roberts et al. 2013; Alexandrov et al. 2013; Harris 2013), but here we 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 (Harris 2015). 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 (Rizzato, Rodriguez, and Laio 2016).

Our 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. We 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. We 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. We 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.

Our findings emphasise the critical importance of accurately calling and annotating MNVs within clinical genomic testing both to improve diagnostic sensitivity and to avoid misinterpretation. We observed that pathogenic de novo MNVs are significantly underrepresented among reported pathogenic clinical variants in ClinVar, indicating that current analytical workflows have diminished sensitivity for identifying pathogenic MNVs. 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 (Sandmann et al. 2017). Both freeBayes and VarDict are haplotype aware callers which is necessary for MNV detection (Garrison and Marth 2012; Lai et al.

2016). 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(Wei et al. 2015). To further our understanding of the role of MNVs in evolution and disease, calling and annotating these variants correctly is a vital step.

Subjects and Methods

Variant and *De Novo* Calling in DDD

The analysis in this report 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 has been described previously(Wright et al. 2015). 7833 trios from 7448 families and 1791 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, we only included trios from children with 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. *De novo* mutations were called using DeNovoGear and filtered as before (Deciphering Developmental Disorders 2017).

Identifying MNVs

MNVs were defined as two nearby variants that always appear on the same haplotype. To identify all possible candidate MNVs we 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. These pairs of variants were classified as MNVs if both variants appeared on the same haplotype for more that 99% of individuals in which they appear. This was determined by phasing variants using parental exome data. We were able to determine phase for approximately 2/3 of all possible MNVs across all individuals. Those that could not be phased were discarded. Read based phasing for these variants proved to be more error-prone than trio-based phasing and so was not performed. After examining the properties of these MNVs we restricted the analyses to those that were 1-20bp of each other. We identified 69,940 unique MNVs.

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.

To identify *de novo* MNVs we looked within a set of 51,942 putative DNMs for pairs of *de novo* variants within 20bp 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. We did not impose stricter filters at this stage as true *de novo* MNVs tend to have worse quality metrics than true *de novo* SNVs. We found 301 pairs, approximately 1.2% of all candidate DNMs. A third of these were 1-2bp apart (Figure 3a). For analysis of mutational spectra we did not filter these further however when looking at functional consequences of these *de novo* MNVs we wanted to be more stringent and examined IGV plots for all *de novo* MNVs of which 106 passed IGV examination.

Estimating the MNV mutation rate

We estimated the MNV mutation rate by scaling the SNV mutation rate estimate of $1.1\text{x}10^{-8}$ mutations per base pair per generation by the ratio of MNV segregating sites/ SNV segregating sites observed in our data set(Roach et al. 2010). This approach is based on a rearrangement of the equation for the Watterson estimator(Watterson 1975). 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 n-1th harmonic number.

$$\hat{\theta} = \frac{K_{SNV}}{a_n} = 4N_e\mu_{SNV}$$

$$\mu_{SNV} = \frac{K_{SNV}}{a_n 4N_e} = 1.1\times10^{-8}$$

$$a_n 4N_e = \frac{K_{SNV}}{1.1\times10^{-8}}$$

$$\mu_{MNV} = \frac{K_{MNV}}{a_n 4N_e}$$

$$= \frac{K_{MNV}}{K_{SNV}} 1.1\times10^{-8}$$

To avoid any potential bias from selection we excluded variants that fell into potentially constrained genes (pLI>0.1). The MNV mutation rate was estimated to be 1.78 x 10^{-10} mutations per base pair per generation.

We 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 1.07x10^{-8} per coding base pair per generation which agrees with the estimate of 1.09x10^{-8} per coding base per generation which was calculated using the trinucleotide context mutational model as described by Samocha et al(Samocha et al. 2014).

Enrichment of *de novo* MNVs

To test for the enrichment of *de novo* MNVs we 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(EBI 2017) and listed as monoallelic that were 'confirmed' and 'probable' associated with DD. We 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 (Samocha et al. 2014) (Figure S3). 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 (Figure S3). To correct for sequence context when comparing DD genes and non-DD genes, we adjusted the expected number of MNVs in the DD genes category based on the excess of polymerase zeta dinucleotide contexts. We also estimated the MNV mutation rate using all variants as well as a more stringent estimate just using variants that fell into non-coding regions. When we 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-value: 2.7x10^{-4}, non coding control regions p-value: 4.9x10^{-3} , Figure S4a). The SNV mutation rate estimate varies across studies therefore we also recalculated the MNV mutation rates using SNV mutation rate estimates of 1.0x10^{-8} and 1.2x10^{-8} mutations per base pair per generation (Segurel, Wyman, and Przeworski 2014). 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 S4b).

Analysis of the number of clinically reported de novo MNVs

We 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. We 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 missense mutation rate calculated as:

$$\mu_{DDG2P\ MNV\ missense} = \mu_{MNV} * \frac{2}{3} * 0.97 * \sum coding\ bp\ in\ DDG2P\ genes ==$$

Where 2/3 is the probability of an MNV falling within a codon and 0.97 is the probability that a within-codon MNV results in a missense change. The expected number of missense MNVs in DDG2P genes was then calculated as follows:

$$Expected\ \#reported\ pathogenic\ missense\ MNVs$$
$$= \#reported\ missense\ SNVs * \frac{\mu_{DDG2P\ MNV\ missense}}{\mu_{DDG2P\ SNV\ missense}}$$

This assumes that the enrichment of MNV and SNV missense mutations in these genes are comparable as we have observed in DDD. This yielded an expected number of 51.94 reported pathogenic MNVs compared to 24 observed reported pathogenic MNVs. To test if this difference was significant we performed a poisson test (p-value 2.8x10^{-5}).

References

Aggarwala, V., and B. F. Voight. 2016. 'An expanded sequence context model broadly explains variability in polymorphism levels across the human genome', *Nat Genet*, 48: 349-55.

Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Borresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjord, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jager, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. Lopez-Otin, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. Tutt, R. Valdes-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Initiative Australian Pancreatic Cancer Genome, Icgc Breast Cancer Consortium, Icgc Mmml- Seq Consortium, Icgc PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton. 2013. 'Signatures of mutational processes in human cancer', *Nature*, 500: 415-21.

Amirnovin, R. 1997. 'An analysis of the metabolic theory of the origin of the genetic code', *J Mol Evol*, 44: 473-6.

Amos, William. 2010. 'Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence?', *Proceedings of the Royal Society B: Biological Sciences*.

Besenbacher, S., P. Sulem, A. Helgason, H. Helgason, H. Kristjansson, A. Jonasdottir, A. Jonasdottir, O. T. Magnusson, U. Thorsteinsdottir, G. Masson, A. Kong, D. F. Gudbjartsson, and K. Stefansson. 2016. 'Multi-nucleotide de novo Mutations in Humans', *PLoS Genet*, 12: e1006315.

'ClinVar'. Accessed September 2017. (http://www.ncbi.nlm.nih.gov/clinvar/).

Deciphering Developmental Disorders, Study. 2017. 'Prevalence and architecture of de novo mutations in developmental disorders', *Nature*, 542: 433-38.

Duncan, B. K., and J. H. Miller. 1980. 'Mutagenic deamination of cytosine residues in DNA', *Nature*, 287: 560-1.

EBI. 2017. 'Gene2Phenotype'. https://www.ebi.ac.uk/gene2phenotype/downloads/DDG2P.csv.gz.

Garrison, Erik, and Gabor Marth. 2012. 'Haplotype-based variant detection from short-read sequencing', *arXiv preprint arXiv:1207.3907*.

Harris, K. 2015. 'Evidence for recent, population-specific evolution of the human mutation rate', *Proc Natl Acad Sci U S A*, 112: 3439-44.

Harris, K., and R. Nielsen. 2014. 'Error-prone polymerase activity causes multinucleotide mutations in humans', *Genome Res*, 24: 1445-54.

Harris, R. S. 2013. 'Cancer mutation signatures, DNA damage mechanisms, and potential clinical implications', *Genome Med*, 5: 87.

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. 2014. 'A general framework for estimating the relative pathogenicity of human genetic variants', *Nat Genet*, 46: 310-+.

Lai, Z., A. Markovets, M. Ahdesmaki, B. Chapman, O. Hofmann, R. McEwen, J. Johnson, B. Dougherty, J. C. Barrett, and J. R. Dry. 2016. 'VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research', *Nucleic Acids Res*, 44: e108.

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, and Consortium Exome Aggregation. 2016. 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*, 536: 285-91.

Michaelson, J. J., Y. J. Shi, M. Gujral, H. C. Zheng, D. Malhotra, X. Jin, M. H. Jian, G. M. Liu, D. Greer, A. Bhandari, W. T. Wu, R. Corominas, A. Peoples, A. Koren, A. Gore, S. L. Kang, G. N. Lin, J. Estabillo, T. Gadomski, B. Singh, K. Zhang, N. Akshoomoff, C. Corsello, S. McCarroll, L. M. Iakoucheva, Y. R. Li, J. Wang, and J. Sebat. 2012. 'Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation', *Cell*, 151: 1431-42.

Pinto, Y., O. Gabay, L. Arbiza, A. J. Sams, A. Keinan, and E. Y. Levanon. 2016. 'Clustered mutations in hominid genome evolution are consistent with APOBEC3G enzymatic activity', *Genome Res*, 26: 579-87.

Rizzato, Francesca, Alex Rodriguez, and Alessandro Laio. 2016. 'Non-Markovian effects on protein sequence evolution due to site dependent substitution rates', *BMC bioinformatics*, 17: 258.

Roach, J. C., G. Glusman, A. F. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas. 2010. 'Analysis of genetic inheritance in a family quartet by whole-genome sequencing', *Science*, 328: 636-9.

Roberts, S. A., M. S. Lawrence, L. J. Klimczak, S. A. Grimm, D. Fargo, P. Stojanov, A. Kiezun, G. V. Kryukov, S. L. Carter, G. Saksena, S. Harris, R. R. Shah, M. A. Resnick, G. Getz, and D. A. Gordenin. 2013. 'An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers', *Nat Genet*, 45: 970-6.

Samocha, K. E., E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnstrom, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur,

S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, Jr., R. A. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, and M. J. Daly. 2014. 'A framework for the interpretation of de novo mutation in human disease', *Nat Genet*, 46: 944-50.

Sandmann, S., A. O. de Graaf, M. Karimi, B. A. van der Reijden, E. Hellstrom-Lindberg, J. H. Jansen, and M. Dugas. 2017. 'Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data', *Sci Rep*, 7: 43169.

Schrider, D. R., J. N. Hourmozdi, and M. W. Hahn. 2011. 'Pervasive multinucleotide mutational events in eukaryotes', *Curr Biol*, 21: 1051-4.

Segurel, L., M. J. Wyman, and M. Przeworski. 2014. 'Determinants of mutation rate variation in the human germline', *Annu Rev Genomics Hum Genet*, 15: 47-70.

Seidman, M. M., A. Bredberg, S. Seetharam, and K. H. Kraemer. 1987. 'Multiple Point Mutations in a Shuttle Vector Propagated in Human-Cells - Evidence for an Error-Prone DNA-Polymerase-Activity', *Proceedings of the National Academy of Sciences of the United States of America*, 84: 4944-48.

Stone, J. E., S. A. Lujan, T. A. Kunkel, and T. A. Kunkel. 2012. 'DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in Saccharomyces cerevisiae', *Environ Mol Mutagen*, 53: 777-86.

Venkat, Aarti, Matthew W. Hahn, and Joseph W. Thornton. 2017. 'Multinucleotide mutations cause false inferences of positive selection', *bioRxiv*.

Venkatarajan, M. S., and W. Braun. 2001. 'New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties', *Journal of Molecular Modeling*, 7: 445-53.

Watterson, G. A. 1975. 'On the number of segregating sites in genetical models without recombination', *Theor Popul Biol*, 7: 256-76.

Wei, L., L. T. Liu, J. R. Conroy, Q. Hu, J. M. Conroy, C. D. Morrison, C. S. Johnson, J. Wang, and S. Liu. 2015. 'MAC: identifying and correcting annotation for multi-nucleotide variations', *BMC Genomics*, 16: 569.

Wong, J. T. 1975. 'A co-evolution theory of the genetic code', *Proc Natl Acad Sci U S A*, 72: 1909-12.

Wright, C. F., T. W. Fitzgerald, W. D. Jones, S. Clayton, J. F. McRae, M. van Kogelenberg, D. A. King, K. Ambridge, D. M. Barrett, T. Bayzetinova, A. P. Bevan, E. Bragin, E. A. Chatzimichali, S. Gribble, P. Jones, N. Krishnappa, L. E. Mason, R. Miller, K. I. Morley, V. Parthiban, E. Prigmore, D. Rajan, A. Sifrim, G. J. Swaminathan, A. R. Tivey, A. Middleton, M. Parker, N. P. Carter, J. C. Barrett, M. E. Hurles, D. R. FitzPatrick, H. V. Firth, and D. D. D. study. 2015. 'Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data', *Lancet*, 385: 1305-14.

Data Access

The raw exome sequencing data from this study have been submitted to the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/) under accession number EGAS00001000775, and are available following Data Access Committee (DAC) approval.

Acknowledgements

We thank the families for their participation and patience. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the UK Department of Health, and the Wellcome Sanger Institute (grant WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the UK Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South Research Ethics Committee and GEN/284/12, granted by the Republic of Ireland Research Ethics Committee). The research team acknowledges the support of the National Institutes for Health Research, through the Comprehensive Clinical Research Network. We thank the Sanger Human Genome Informatics team, the Sample Management team, the Illumina High-Throughput team, the New Pipeline Group team, the DNA pipelines team and the Core Sequencing team for their support in generating and processing the data.