Extensive variation in the mutation rate between and within human genes associated with Mendelian disease

Article (Accepted Version)

Extensive Variation in the Mutation Rate Between and Within Human Genes Associated with Mendelian Disease.

Thomas Smith[1]

Gladys Ho[2]

John Christodoulou[2,3]

Elizabeth Ann Price[4]

Zerrin Onadim[4]

Marion Gauthier-Villars[5]

Catherine Dehainault[5]

Claude Houdayer[5,6,7]

Beatrice Parfait[8,9]

Rick van Minkelen[10]

Dietmar Lohman[11]

Adam Eyre-Walker[1]*

1. School of Life Sciences, University of Sussex, Brighton, BN1 9QG, United Kingdom

2. NSW Centre for Rett Syndrome Research, Western Sydney Genetics Program, Children’s Hospital at Westmead, Sydney, Australia

3. Disciplines of Paediatrics and Child Health and Genetic Medicine, Sydney Medical School, University of Sydney, Sydney, Australia

4. Retinoblastoma Genetic Screening Unit, Barts Health NHS Trust, The Royal London Hospital, 80 Newark Street, London, E1 2ES, United Kingdom

5. Service de Génétique, Institut Curie, Paris, France

6. INSERM U830, centre de recherche de l'Institut Curie, Paris, France

7. Université Paris Descartes, Sorbonne Paris Cité, Paris, France

8. EA7331, Faculté de Pharmacie de Paris, Université Paris Descartes, Sorbonne Paris Cité, Paris, France,

9. Service de Biochimie et de Génétique Moléculaire, Hôpital Cochin, AP-HP, Paris, France

10. Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, Netherlands

11. Institut für Humangenetik, Universitätsklinikum Essen, Universität Duisburg-Essen, 45122 Essen, Germany

*Corresponding author: a.c.eyre-walker@sussex.ac.uk

Abstract

We have investigated whether the mutation rate varies between genes and sites using *de novo* mutations (DNMs) from three genes associated with Mendelian diseases. We show that the relative frequency of mutations at CpG dinucleotides relative to non-CpG sites varies between genes and relative to the genomic average. In particular we show that the rate of transition mutation at CpG sites relative to the rate of non-CpG transversion is substantially higher in our the disease genes than amongst DNMs in general; the rate of CpG transition can be several hundred-fold greater than the rate of non-CpG transversion. We also show that the mutation rate varies significantly between sites of a particular mutational type, such as non-CpG transversion, within a gene. We estimate that for all categories of sites, except CpG transitions, there is at least a 30-fold difference in the mutation rate between the 10% of sites with the highest and lowest mutation rates. However, our best estimate is that the mutation rate varies by several hundred-fold variation. We suggest that the presence of hypermutable sites may be one reason certain genes are associated with disease.

Introduction.

There is evidence that the mutation rate varies substantially across the human genome in the germ-line from studies of *de novo* mutations (DNMs) (Francioli, et al., 2015; Michaelson, et al., 2012) and from comparative genomics (reviewed in (Hodgkinson and Eyre-Walker, 2011)). Although this occurs at a number of different scales the most dramatic variation is seen at the single nucleotide level (Hodgkinson and Eyre-Walker, 2011). In part this variation is due to context - the identity of the nucleotides surrounding a site (Bulmer, 1986; Cooper and Krawczak, 1990; Gojobori, et al., 1982; Hwang and Green, 2004; Nachman and Crowell, 2000). The most well known example of a context effect is that of CpGs; C followed by G is often methylated in mammals, and methylated cytosine undergoes a high rate of deamination to generate T (Bulmer, 1986; Cooper and Krawczak, 1990; Coulondre, et al., 1978; Gojobori, et al., 1982; Hwang and Green, 2004; Nachman and Crowell, 2000). It has been estimated that CpGs undergo rates of mutation 10-15 fold higher than other sites in the human genome (Hwang and Green, 2004; Nachman and Crowell, 2000) and generate ~20% of all mutations (Fryxell and

Moon, 2005). There are also other context effects, but these lead to variation in the mutation rate of only 2 to 3-fold (Hwang and Green, 2004).

In addition to variation associated with context, there also appears to be variation at the single nucleotide level that does not depend upon the identity of the adjacent nucleotides, at least not in a simple manner, variation that has been termed cryptic (Hodgkinson, et al., 2009). The evidence for this variation initially came from the observation that there is at least a 50% excess of sites in which humans and chimpanzees share a single nucleotide polymorphism (SNP), even when the influence of context on the mutation rate is taken into account (Hodgkinson, et al., 2009). Such an excess could be due to sequencing error, assembly error of paralogous duplications or ancestral polymorphism. However, several lines of evidence suggest that these explanations do not explain the excess of coincident SNPs. First, the distribution of allele frequencies amongst coincident SNPs is identical to non-coincident SNPs (Johnson and Hellmann, 2011); if coincident SNPs were due to assembly errors or ancestral polymorphisms we would expect them to be more frequent in the population than other SNPs. Second, sequencing coverage is no greater at coincident SNPs than other sites (Johnson and Hellmann, 2011). And third, there is also an excess of coincident SNPs between human and macaque (Hodgkinson, et al., 2009), two species which are very unlikely to share ancestral polymorphisms. There is also an excess of sites with substitutions in two independent pairs of primate species (Johnson and Hellmann, 2011). These lines of evidence therefore suggest that the excess of coincident SNPs most likely arises from variation in the mutation rate that is not associated with context, at least not sequence contexts that are close to the site in question. It has been estimated that cryptic variation may generate more variance in the mutation rate than simple contexts, such as the CpG effect (Hodgkinson, et al., 2009).

Although, variation in the mutation rate is most conspicuous at a single nucleotide scale it has also been known for sometime that the mutation varies at larger scales in the human genome (Matassi, et al., 1999; Michaelson, et al., 2012; Spencer, et al., 2006). The scale of this variation remains poorly characterised but a recent analysis of where DNMs occur suggest that the variation is probably at a scale of 10,000s of base pairs (Michaelson, et al., 2012). The variation in the rate of CpG and non-CpG mutations appears to be at least partly independent, because the variation correlates to different genomic variables (Tyekucheva, et al., 2008), but no systematic analysis of the relative rates of CpG and non-CpG mutation has been performed to our knowledge.

Here we investigate two aspects of variation in the mutation rate. First, does the relative frequency of transition and transversion mutations at CpG and non-CpG sites differ between genes, and second, is there variation in the mutation rate for transition and transversion mutations within CpG and non-CpG sites (e.g. does the rate of transition mutation differ between CpG sites within a single gene). We address these questions using a dataset of *de novo* mutations (DNMs) that have been discovered during clinical screening in three genes associated with Mendelian diseases. In each case the DNMs were discovered in an unbiased manner – the causative gene was sequenced in a patient with the disease and their parents who did not have the disease.

Methods

Data

DNMs were discovered as part of routine clinical screening for individuals suffering from bi-lateral retinoblastoma, neurofibromatosis type I and Rett's syndrome; these diseases are caused by mutations in *RB1*, *NF1* and *MECP2* respectively. All data were collected after Ethics committee approval at each of the institutions involved. The *MECP2* data were gathered from RettBASE, International Rett Syndrome Foundation *MECP2* Variation Database (http://mecp2.chw.edu.au), a curated database for *MECP2* variants from research and clinical laboratories {Christodoulou, 2003 #1369}. Variants included in this study were limited to those for which parental testing had been carried out, with both parents tested for female patients, or maternal testing for male patients, since the gene is X-linked. Only variants from studies in which exons 2-4 had been sequenced were included, and our analysis was restricted to this part of the gene. The *NF1* data were gathered from the *NF1* LOVD database (https://grenada.lumc.nl/LOVD2/mendelian_genes/home.php?select_db=NF1). Both parents were tested for the pathogenic mutation and the father was tested for paternity. We only included studies in which all exons had been sequenced in transcript variant 2, this differs from transcript vatiant 1 in missing exon 23a. The *RB1* data came from three laboratories. Mutations were identified using a number of approaches including sequencing, single strand conformational polymorphism, heteroduplex analysis and high resolution melt analysis. Mutations were confirmed in each case by direct sequencing. There is alternative start codon in *RB1* in the first exon {Sanchez-Sanchez, 2007 #1372} and so exon 1 was ignored in the analysis. Both parents were tested for all *RB1* variants. Some of the *RB1* data has been previously published in {Price, 2014 #1368}. The transcript numbering that we use is from NM_004992.3 for *MECP2*, NM_000267 for *NF1* and NM_000321.2 for *RB1*. We focus our analysis on nonsense mutations since nonsense mutations are more likely to have consistent phenotypic effects (see results section for further discussion).

Testing for mutation rate variation between genes

We performed to two tests of mutational rate heterogeneity. First we tested whether the relative rates of CpG transitions, CpG transversions, non-CpG transitions and non-CpG transversions were significantly different between genes and between the genes and the background rate. To do this we performed a chisquare goodness-of-fit test, in which we calculated the expected number of CpG transitions and transversions, and non-CpG tranistions and transversions, assuming that the ratios between the various mutational types were the same in the two genes, by finding the parameters of a simple model which minmised the chi-square statistic. We assumed that each gene has its own "mutation rate", which reflects both the intrinsic mutation rate and the probability that the mutation comes to clinical attention; let this be μ_i. If we assume that the relative rates of the different mutation categories are the same in two genes then without loss of generality we can let the rate of non-CpG transversions rate be μ_i and the rates of CpG transition, CpG transversion and non-CpG transition be $\mu_i r_{cts}$, $\mu_i r_{ctv}$, and $\mu_i r_{nts}$, where r_{cts}, r_{ctv} and r_{nts} are shared between the two genes. To test whether the patterns of mutation are the same we find the values of μ_i, r_{cts}, r_{ctv} and r_{nts} that minmise the chi-square value, comparing the observed and expected values. Having found the parameters that minimise the chi-square value we performed a goodness of fit test using the ch-square value with 3 degrees of freedom (because we have eight observations, the number of CpG and non-CpG transitions and transversions for each gene, and we have estimated 5 parameters).

Testing for mutation rate variation within genes

Second, we tested whether the rate of mutation varied within a mutational category (e.g. CpG transitions). If the rate of mutation is the same across all sites of a particular type then DNMs should be randomly distributed across those sites. To test whether DNMs tend to recur at sites more often than by chance we generated the expected number of sites hit recurrently by DNMs by randomly distributing the observed number of DNMs of the required type (e.g. CpG transitions) across the sites of that type that could generate a nonsense mutation. For each randomized dataset we tabulated the number of sites a site was hit zero, once, twice…etc by a DNM. By repeating this randomization 10,000 times we derived the expected distribution of DNMs (i.e. the number of times a site is expected to have been hit by one, two…etc DNMs). We compared the observed to the expected using a chi-square test. However, the test statistic is unlikely to be chi-square distributed because some of the expected values can be very small, We therefore empirically determined the distribution of the chi-square statistic by calculating the chi-square statistic for each simulated dataset using the expected values estimated across all simulated datasets (as we did for the observed data). We then compared the observed chi-square statistic to this distribution. The p-value was the proportion of simulated datasets that had a chi-square value greater than observed chi-square value plus half the simulated datasets that had an identical chi-square value; this latter condition prevents the test being overly conservative when there are few DNMs. We performed simulations to check that this method did not generate excessive levels of type I error. For a given number of DNMs and sites we randomly allocated DNMs across sites and tabulated the number of sites that had been hit 0,1,2…etc times. We then performed the analysis as though this was real data, and repeated this 1000 times for a given combination of sites and DNMs. Simulations confirm that it does not increase the level of type I error, although it can decrease it when there are very few DNMs.

To combine probabilities from the heterogeneity tests we used the unweighted z-method (Whitlock, 2005); in this method we find the z-value of the a normal distribution, with a mean of zero and variance of one, that would yield the corresponding p-value. These z-values can be added to yield a z-value with an expected mean of zero and a variance equal to the number of tests that have been combined. The overall p-value is then obtained by converting the combined z-value into a p-value. We set p-values in which no simulated data had a greater chi-square value to 0.0001, and no simulated data had a smaller chi-square to 0.9999.

Parameter estimation

We estimated the variation in the mutation rate within a mutational category as follows. Let us assume that the mutation rate at a site is $\mu\alpha$ where μ is the mean mutation rate and α is a deviation from the mean that is taken from some distribution $D(\alpha)$, which has a mean of 1; in our analysis we assume that $D(\alpha;\beta)$ is a gamma distribution with a shape parameter β. The number of mutations at a site can be modelled as a Poisson process because mutations are rare events, and hence the number of mutations at a site is Poisson distributed. The probability of observing *x* mutations at a site is therefore

$$G(x;u,\beta)=\int_0^\infty D(\alpha;\beta)\frac{e^{-u\alpha}(u\alpha)^x}{x!}d\alpha=\frac{1}{x!\Gamma(\beta)}\left(\frac{1}{\beta}\right)^{-\beta}u^x(\beta+u)^{-x-\beta}\Gamma(x+\beta) \qquad (1)$$

which is the negative binomial distribution, where *u*=*μk* and *k* is a parameter proportional to the chance of observing a DNM; this is dependent upon the incidence and interest in the disease. The number of sites with *x* mutations is multinomially distributed and hence the likelihood of observing n_x sites with *x* mutations is

$$L(u,\beta)=n!\prod_x\frac{G(x;u,\beta)^{n_x}}{n_x!} \qquad (2)$$

where *n* is the total number of sites. We found the maximum likelihood values of the distribution using the Nelder-Mead algorithm as implemented in the NMaxmize function in Mathematica. The model above is described for a single mutational category in a single gene. However, it is straightforward to expand the analysis across multiple mutational categories and genes. In each analysis each mutational category in each gene is allowed its own *u* parameter reflecting the fact that the chance of observing a mutation varies between genes, and that the rate of mutation varies between mutational categories. Confidence intervals on parameters were derived from the likelihood surface – i.e. by finding the parameter values that decreased the log-likelihood by 2 units.

Results

Data

We have analysed DNMs in three genes that are associated with Mendelian disease. The genes are *RB1*, mutations in which cause retinoblastoma; we only consider mutations causing bilateral retinoblastoma since this disease is almost exclusively caused by a *de novo* germ-line mutation, whereas unilateral retinoblastoma is usually caused by somatic mutations. The second gene we consider is *NF1*, mutations in which cause neurofibromatosis type I. And the third gene is *MECP2*, mutations in which cause Rett's syndrome.

It is critical to our analysis that all mutations in a gene have similar penetrance, otherwise any apparent variation in the mutation rate might be due to variation in penetrance (i.e. a site with multiple recurrent DNMs might have a high mutation rate or the mutation might be partly penetrant). As a consequence we only consider nonsense mutations and in *RB1* and *NF1* we only consider sites at which nonsense mutations are predicted to be caught by nonsense mediated decay (NMD). Furthermore, in *RB1* we ignore data from the first exon because mutations in the first exon may display variable levels of penetrance due to alternative transaltion initiation sites {Sanchez-Sanchez, 2007 #1372}. All the nonsense mutations we consider in *RB1* and *NF1* should therefore have the same probability of being detected. The analysis of *MECP2* is more complex because the vast majority of sites that could generate a nonsense mutation are in the last exon and hence would not be caught by NMD; hence some nonsense mutations, particularly those towards the end of the gene could be less penetrant than those earlier in the gene. Furthermore, it is possible that nonsense mutations in the second and third exons (first and second coding exons) are lethal and therefore not routinely observed. Figure 1 shows the distribution of DNMs along the *MECP2* gene. It is conspicuous that almost all pathogenic mutations occur between the start of the final exon and the end of the transcription repression domain. As a consequence we analysed two datasets for *MECP2* – all sites at which a nonsense mutations could occur, and all sites at which nonsense mutations could occur between the first and last sites that have multiple DNMs (sites 423 to 889 inclusive). Reducing the dataset in this manner does not alter the relative rates of mutation greatly, but it does reduce the evidence for heterogeneity within mutational categories (see below); this reduced dataset can therefore be considered a conservative dataset.

Table 1 gives the number of DNMs in each of four mutation categories, transitions and transversions at CpG sites, and transitions and transversions at non-CpG sites, and Table S1 gives the number of sites hit by 0,1, 2…etc DNMs. We divided the data in this way because there are large differences in the rate of mutation of these mutational types (reviewed in (Hodgkinson and Eyre-Walker, 2011). For each of our genes we have large numbers of nonsense DNMs. These are dominated by CpG transitions but we have substantial numbers of non-CpG transitions and transversions.

Heterogeneity between genes

It is of interest to know whether the frequencies of different types of mutation vary substantially between genes. Unfortunately, because of the way in which our data have been sampled we cannot answer this question directly – the rate at which DNMs are detected in our genes depends upon the frequency of the disease, the severity of phenotype and the interest of clinicians. However, we can compare the relative frequency of different types of mutation between genes and compare those between genes and to the genomic average. We test for differences between genes, and between genes and the genomic average using a chi-square goodness of fit test, fitting a model in which we assume the relative rates of mutation in the four mutational categories are the same in the two genes (or genes and genome) (see the Materials and Methods section).

Two recent studies have obtained substantial numbers of DNMs from the complete genome sequencing of trios (Kong, et al., 2012; Michaelson, et al., 2012). Surprisingly the relative frequencies of the four mutation types differ significantly between these studies (Table 2)(Chi-square goodness of fit test p = 0.045). The difference seems to be largely a consequence of a higher relative rate of CpG transitions in the data of Kong et al. (Kong, et al., 2012) compared to the data of Michaelson et al. (Michaelson, et al., 2012) (28x the rate of non-CpG transversions versus 19x) (Table 2). The reason for this discrepancy is not clear; it may be due to different ages amongst the two cohorts, or different biases in the sequencing methods, as other analyses seem to suggest (Eyre-Walker and Eyre-Walker, 2014).

However, more striking than the difference between the two trio datasets are the differences in the relative rates of mutation between these datasets and the three genes for which we have DNMs (Table 2); each of the disease genes shows higher rates of mutation relative to the rate of non-CpG transversion than trio datasets. The most dramatic difference is the relative rate of CpG transition mutation in the *MECP2* gene where the mutation rate is estimated to be over 240x higher than the rate of non-CpG transversion (640x for the complete *MECP2* dataset and 236x for the restricted dataset). Pairwise comparisons show that the patterns of mutation are highly significantly different between MECP2 and the other two genes (Chi-square goodness of fit tests: p<0.001 in both cases), but not between *RB1* and *NF1*. The patterns are also highly significantly different between each of the three genes and both the datasets of Kong et al. and Michaelson et al (Chi-square goodness of fit tests: p<0.0001). Unfortunately, it is not possible to say from these data whether the large relative rates are due to a low rate of mutation at non-CpG sites or a high rate at CpG sites.

Heterogeneity within genes

The analyses above show that the relative frequency of different types of mutation varies between genes. We can also test whether the rate of mutation within each of these mutational types varies between sites within a gene. Using a chi-square test of heterogeneity (deriving the null distribution by randomisation) we find highly significant evidence of heterogeneity over the entire dataset whether we consider all sites in the *MECP2* gene or the restricted *MECP2* dataset ($p < 10^{-5}$). Surprisingly we significant homogeneity, not heterogeneity, for CpG transitions sites in the *RB1* gene (i.e. mutations are more evenly distributed between sites than one would expect by chance alone). The data for this gene comes from three different labs. None of these datasets shows this excessive homogeneity individually and in fact the dataset from Barts Hospital shows marginally significant evidence of heterogeneity (p=0.081) (Table S2). The datasets are not significantly different to each other (chi-square = 23.6, df = 20, p = 0.26) (Table S3). It therefore remains unclear why the combination of the three datasets leads to significant homogeneity.

The strongest evidence for heterogeneity comes from non-CpG tranversions in the *MECP2* gene. In the restricted dataset there are 38 sites at which a non-CpG transversion will generate a nonsense mutation and there are 10

DNMs that have occurred at these sites. However, 6 of the DNMs have occurred at one site (site 423); all of these are C>G changes even though a C>A would also generate a stop codon.

Using the heterogeneity analysis we can identify 3 sites that have mutation rates that are significantly above background levels (Table 4). The mutation rate at these sites relative to all other sites, of the same mutational type, in the respective genes are given in Table 4. For two of these sites the mutation rates are only modestly above background levels; this reflects the power that we have to detect significantly hyper-mutable sites in CpG transition sites because we have more data than in other mutational categories. However, in *MECP2* we estimate that site 423, the site which has been hit by 6 non-CpG transversion DNMs has a mutation rate at least 150x (or 56x in the restricted dataset) higher than the background rate of non-CpG transversion in this gene. There is no obvious context effect associated with these sites (Table 4).

Quantification

The estimates of the mutation rate at different sites are crude; one would expect that as more data accumulate so more sites will be found to be significantly hyper-mutable and hence the estimates of the rates will increase as sites are excluded from the background level. Therefore to better quantify the variation in the mutation rate we used maximum likelihood to fit a model in which mutation rates were distributed according to a gamma distribution. We fit several models in which the distribution of rates (i) was shared across all genes and mutational categories, (ii) in which it was shared across genes, (iii) across mutational categories and (iv) finally a model in which every gene and mutational category combination had its own distribution. Using likelihood ratio tests we find the best supported model is one in which the gamma distribution is specific to a mutational category but is shared across genes (Table S4).

If we consider the shape parameter estimates for each mutational category it seems that CpG transitions have a much lower level of variation than the other mutational categories (the higher the values of the shape parameter the lower the level of variation) (Table 5). In contrast the other three categories show substantial variation. To quantify this variation we calculated the ratio of the rates from the 90th and 10th percentiles. Whereas CpG transitions show just 1.4 fold variation between the deciles the ratio for all the other categories is very substantial; for non-CpG transitions there is 36-fold variation but CpG transversions and non-CpG transversions we infer more than a 1000-fold variation. However, the confidence intervals on these individual estimates are large and are also compatible with modest levels of heterogeneity; this is due a lack of data. We therefore combined data from CpG transversions, non-CpG transitions and non-CpG transversions. Our estimate of the shape parameter is 0.39 (0.21, 0.85) and this corresponds to a ratio of deciles of 550 (i.e. the top 10% of sites mutate at least 550x faster than the bottom 10% of sites) with 95% CIs of 32x, 56,000x. In other words there appears to be very substantial variation in the mutation rate within each mutational category, with the exception of CpG transitions.

Discussion

We have provided evidence for two types of mutational heterogeneity. First, we have demonstrated that there is substantial variation in the relative rates of CpG and non-CpG mutations. The most conspicuous pattern is the very high rate of CpG transitions relative to non-CpG transversions. Whereas on average CpG dinucleotides undergo transition mutations between 18-30 fold the rate of non-CpG transversions, in our three Mendelian disease genes they undergo 90 to more than 200-fold higher rates of mutation. It is not possible to infer whether this is due to a low rate of non-CpG transversion or a high rate of CpG transition. Second, we have shown that there is significant heterogeneity in the mutation rate between sites within each mutational category. This is particularly evident for all categories other than CpG transitions; we estimate amongst the other categories that the mutation rate may vary by 100-fold or more.

Our conclusions are conditional on the assumption that all the mutations, which we have considered, both those that have occurred and those that could occur in a gene, are equally likely to be sampled. Variation in sampling might arise through three processes – variation in penetrance, alternative splicing and ascertainment bias.

In an attempt to ensure that all mutation were equally penetrant we restricted our analysis to nonsense mutations, and in the case of *RB1* and *NF1*, to nonsense mutations that are predicted to be caught by NMD. In the case of

MECP2 most sites that could cause a nonsense mutation are found in the last exon and hence would not be caught by NMD. To mitigate against this, we have analysed the pattern of mutation both amongst all DNMs and amongst a subset of DNMs between the first and last recurrently hit sites. We have found similar patterns. If we remove *MECP2* from our analysis we still find evidence that the ratio of CpG to non-CpG mutations varies between the two disease genes and the background rate, and there is still significant heterogeneity in the mutation rate within a mutational category (Table 3). Never-the-less it is difficult to completely rule-out variation in the level of penetrance as an explanation for our results; if the variation in the density of DNMs is due to variation in penetrance then our results suggest that penetrance varies considerably between mutational categories and between sites within a gene.

The apparent variation in the mutation rate could also be due to ascertainment bias. Although we restricted our analysis to data that had come from studies in which the same part of the gene had been analysed it is possible that the causative mutation was not ascertained and these cases discarded. If some mutations are more likely to escape detection than others then it will appear as though there is mutation rate variation.

The variation could also potentially be due to alternative splicing (or alternative translation start sites) since nonsense mutations in constitutive exons might be more penetrant (or more lethal) than nonsense mutations in alternatively spliced exons. However, this seems an unlikely explanation for our results. *RB1* is known to have an alternative translation start site

{Sanchez-Sanchez, 2007 #1372} and as a consequence this exon was removed from the analysis. There are two major splice forms of MECP2 which differ in both their translation start site and the inclusion of exon 2 – variant 1 includes exon 2, in which translation starts, whereas variant 2 excludes exon 2 with translation starting in exon 1 {Kriaucionis, 2004 #1370}. We have analysed data mapped to variant 1, which differs from variant 2 in the first 26 bp, so the variation we observe is unlikely to be a consequence of alternative splicing associated with the major splice variants in this gene; as such the vast majority of the data we have analysed comes from exons that are found in the major splice forms. There are multiple splice forms of NF1, although most of them yield products that are removed by NMD or result in highly truncated proteins {Barron, 2012 #1371}. We have used data that maps to transcript variant 2, which is one of two major splice forms. This differs from variant 1 in missing exon 23a, a 21 amino acid exon, found in variant 1. So again it seems unlikely that alternative splicing can be responsible for our results because we have analysed data only from the the exons present in both of the two major splice forms.

Another potential explanation for our results is positive selection in the germ-line. It has been found that some pathogenic mutations are advantageous within the male germ-line leading to an increased prevalence of diseases such as Apert’s syndrome, which is caused by mutations in the gene FGFR2 (Goriely and Wilkie, 2012). None of the genes that we have studied are known to have mutations that are positively selected in the germ-line and it seems unlikely that the heterogeneity amongst nonsense sites could be caused by this process, since all nonsense mutations are predicted to have the same or similar phenotypes.

The variation in the density of DNMs is therefore most likely due to variation in the mutation rate. It has previously been shown that mutation rates vary at a regional scale (reviewed in (Hodgkinson and Eyre-Walker, 2011)). However, it has not been noted before that the relative rates of CpG and non-CpG mutation can vary substantially. The magnitude of the variation that we have observed might in part be due to the fact that genes with the highest mutation rates are those most likely to be associated with disease, assuming that the high CpG to non-CpG mutation rate reflects a high CpG mutation rate and not a low non-CpG rate. There is some evidence for this effect; a recent model of the mutation rate at sites in the human genome, based on the analyses of DNMs and where they occur, predicts that disease genes have higher rates of mutation than non-disease genes (Michaelson, et al., 2012).

It has also been noted that the mutation rate can vary within a mutational category because of context (Hodgkinson and Eyre-Walker, 2011) however the effects within CpG or non-CpG categories have been inferred to be quite modest. For example, Hwang and Green (Hwang and Green, 2004) estimated, using the divergence between primate species, that on average CCG, ACG, GCG and TCG mutate 24, 29, 18 and 23-fold faster than the genomic average. Context effects at non-CpG sites are also fairly modest, typically showing 2-3 fold variation when the immediately adjacent nucleotides are considered (Hwang and Green, 2004), getting progressively weaker as sites further away from the focal site are considered (Zhao and Boerwinkle, 2002). The level of variation within all mutational categories except CpG transitions seems to be considerably greater than this. The substantial variation in the mutation rate within each mutational category, except CpG transitions, is consistent with the cryptic variation in the mutation rate, which was first identified in nuclear DNA from the coincidence of SNPs in humans and chimpanzees (Hodgkinson, et al., 2009; Johnson and Hellmann, 2011). As we have found here, Hodgkinson et al. (Hodgkinson, et al., 2009) estimated that there was more variation in the mutation rate within non-CpG sites, than within CpG-sites, and estimated that a gamma distribution with a shape parameter of 0.85 (0.83, 0.87) fitted the data at non-CpG sites. This is not significantly different to the estimate obtained here, 0.39 (0.20, 0.91).

The variation at CpG sites could potentially be a consequence of variation in methylation. Methylated CpGs are expected to mutate faster than non-methylated CpGs due to the instability of methyl-cytosine (Coulondre *et al.* 1978, Bird. 1980, Sved and Bird. 1990). None of the sequences that we have analysed contain CpG islands, regions of the genome in which CpGs are not methylated. However, some of the variation may be due to residual variation in methylation.

In summary we have shown that there is significant variation in the mutation rate both between and within genes. Some of this variation might explain why these genes are associated with disease; they have high mutation rates, either overall or at specfic sites that can cause disease, and this makes it more likely that pathogenic mutations will recurr in the human population and cause disease.

References

Bulmer M. 1986. Neighbouring base effects on substitution rates in pseudogenes. Mol. Biol. Evol. 3:322-329.

Cooper DN, Krawczak M. 1990. The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. Human genetics 85(1):55-74.

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274:775-780.

Eyre-Walker A, Eyre-Walker YC. 2014. How Much of the Variation in the Mutation Rate Along the Human Genome Can Be Explained? G3-Genes Genomes Genetics 4(9):1667-1670.

Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, van Duijn CM, Swertz M, Wijmenga C, van Ommen G, Slagboom PE, Boomsma DI, Ye K, Guryev V, Arndt PF, Kloosterman WP, de Bakker PI, Sunyaev SR. 2015. Genome-wide patterns and properties of de novo mutations in humans. Nature genetics 47(7):822-6.

Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol 22(3):650-8.

Gojobori T, Li W-H, Graur D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18:360-369.

Goriely A, Wilkie AO. 2012. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. American journal of human genetics 90(2):175-200.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. Nature Reviews Genetics 12(11):756-66.

Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic Variation in the Human Mutation Rate. PLoS Biol 7(2):e27.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A 101(39):13994-4001.

Johnson PLF, Hellmann I. 2011. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. Genome Biol. Evol. 3:842-850.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WSW, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K. 2012. Rate of de novo mutations and the importance of father's age to disease risk. Nature 488(7412):471-475.

Matassi G, Sharp PM, Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. Curr. Biol. 9:786-791.

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A, Gore A, Kang S, Lin GN, Estabillo J, Gadomski T, Singh B, Zhang K, Akshoomoff N, Corsello C, McCarroll S, Iakoucheva LM, Li Y, Wang J, Sebat J. 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151(7):1431-42.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156:297-304.

Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. PLoS Genet 2(9):e148.

Tyekucheva S, Makova KD, Karro JE, Hardison RC, Miller W, Chiaromonte F. 2008. Human-macaque comparisons illuminate variation in neutral substitution rates. Genome Biol 9(4):R76.

Whitlock MC. 2005. Combining probability from independent tests: the weighted Z-method is siperior to Fisher's approach. J. Evol. Biol. 18:1368-1373.

Zhao Z, Boerwinkle E. 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. Genome Res 12(11):1679-86.

	CpG ts	CpG tv	non-CpG ts	non-CpG tv
RB1	97	5	15	32
NF1	52	4	24	20
MECP2	253	0	6	12
MECP2 – restricted	252	0	2	10
Kong	855	73	2489	1516
Michaelson	70	10	282	185

Table 1. The numbers of nonsense DNMs in each gene and mutational category for three disease genes. The *MECP2* – restricted figures are for DNMs between positions 421 and 888. The Kong and Michaelson data are all the DNMs reported in Kong et al. (2012) and Michaelson et al. (2012) respectively.

	CpG ts	CpG tv	non-CpG ts	non-CpG tv
RB1	90	5.7	3.5	1.0
NF1	120	14	5.5	1.0
MECP2	640	0	4.8	1.0
MECP2 – restricted	240	0	1.5	1.0
Kong	28	2.4	1.6	1.0
Michaelson	19	2.7	1.5	1.0

Table 2. The rates of mutation expressed relative to the rate of transversion at non-CpG sites. These are derived by dividing the numbers of DNMs by the number of sites (Table S1), and then dividing the rate by the rate for non-CpG transversions.

	CpG ts	CpG tv	non-CpG ts	non-CpG tv
RB1	0.9990	0.0629	0.0969	0.0412
NF1	0.0052	0.0344	0.1312	0.6017
MECP2	<0.0001	-	0.1281	<0.0001
MECP2 – restricted	<0.0001	-	0.0997	<0.0001
Overall – without *MECP2*	0.65	0.0089	0.044	0.15
Overall (row above)	0.0087			
Overall – *MECP2* all	0.032	0.0089	0.020	0.0013
Overall (row above)	2.0×10^{-6}			
Overall – *MECP2* restricted	0.032	0.0089	0.016	0.0013
Overall (row above)	1.6×10^{-6}			

Table 3. The probability of observing the data under the null hypothesis that sites are equally mutable. Probabilities were combined using the unweighted Z-method (Whitlock 2005). The data from individual genes are combined in a number of different combinations – with and without the *MECP2* data, and with the restricted *MECP2* data. The probabilities are combined for each mutational type, but also across genes and mutational types.

Gene	Mutation type	Rate	Context	HG19 coordinates
MECP2	CpG ts	2.0 (1.5)	CCCCTCCCGGCGAGAGCAGAA	chrX:153,296,777
MECP2	non-CpG tv	150 (56)	TGATTGCGTACTTCGAAAAGG	chrX:153,296,856
NF1	CpG ts	4.0	TGTTGGAAGACGACCTTTTGA	chr17:29,588,751

Table 4. Significantly hypermutable sites. Numbers in parentheses in the rate column are estimates from the restricted *MECP2* data. The nucleotide underlined in the context column is the hypermutable site.

Mutation type	Shape	Ratio of the rates of first and last deciles
cts	63 (13, infinity)	1.4 (2.1, 1.0)
ctv	0.39 (0.069, infinity)	550 (7.7 x 10^{13}, 1.0)
nts	0.81 (0.26, infinity)	36 (8000, 1.0)
ntv	0.24 (0.11, 0.64)	16,000 (6.1 x 10^{14}, 70)
ctv+ntv	0.26 (0.13, 0.64)	8000 (3.0 x 10^{7}, 70)
nts+ntv	0.39 (0.20, 0.91)	550 (93,000, 27)
ctv+nts+ntv	0.39 (0.21, 0.85)	550 (56,000, 32)

Table 5. Estimates of the shape parameter of the gamma distribution and the ratio of the upper and lower deciles of the distribution. 95% confidence intervals, as inferred from the likelihood surface, are given in brackets.

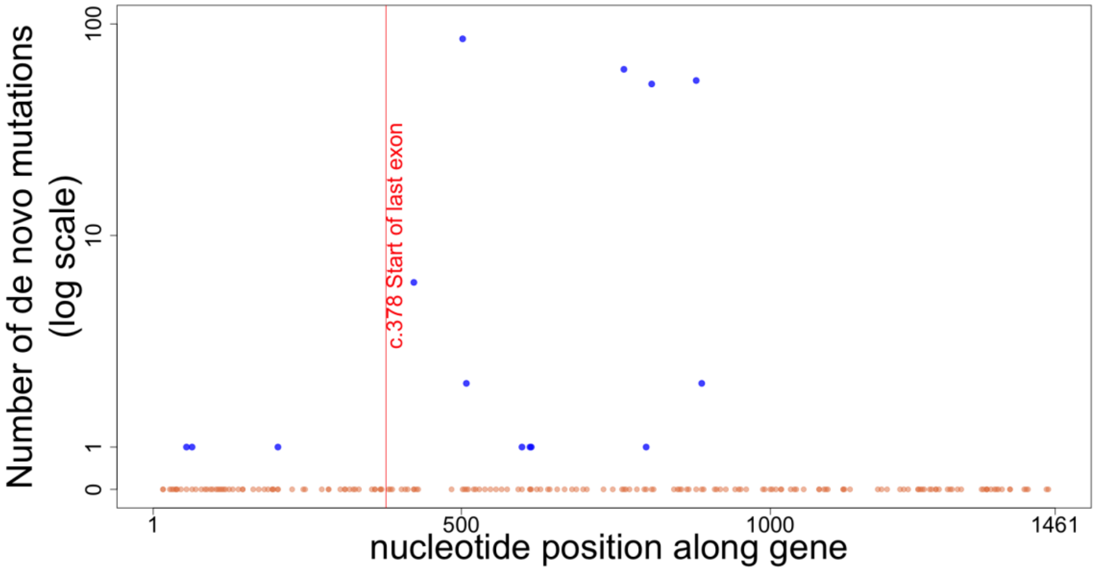

Figure 1. The distribution of nonsense DNMs and sites at which a mutation can cause a nonsense mutation in the *MECP2* gene. The start of the last exon is marked.