Linked mutations at adjacent nucleotides have shaped human population differentiation and protein evolution

James GD Prendergast[1*], Carys Pugh[1,2], Sarah E Harris[2,3], David A Hume[4], Ian J Deary[2], Allan Beveridge[5],

1 The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, United Kingdom
2 Centre for Cognitive Ageing and Cognitive Epidemiology, Department of Psychology, The University of Edinburgh, Edinburgh, United Kingdom
3 The University of Edinburgh Centre for Genomic and Experimental Medicine and MRC Institute of Genetics and Molecular Medicine, Edinburgh, United Kingdom
4 Mater Research Institute-University of Queensland, Woolloongabba, Qld 4160, Australia
5 Glasgow Polyomics, College of Medical, Veterinary and Life Science, University of Glasgow, Glasgow, United Kingdom

* james.prendergast@roslin.ed.ac.uk

Abstract

Despite the fundamental importance of single nucleotide polymorphisms (SNPs) to human phenotypes there are still large gaps in our understanding of the forces that shape their distribution across the genome. SNPs have been shown to not be distributed randomly, with directly adjacent SNPs found more often than expected by chance. Why this is the case is unclear. In this study we illustrate how neighbouring SNPs are driven by distinct mutational processes and selective pressures. By characterising multi-nucleotide polymorphisms (MNPs) across multiple populations, including a novel cohort of 1,358 Scottish genomes, we show that, like SNPs, MNPs display distinct mutational spectra across populations. These biases are though not only different to those observed among SNPs, but more clearly define population groups. The changes that make up MNPs are not independent, with $CA \rightarrow CG \rightarrow TG$ changes observed an order of magnitude more often than other changes involving the gain and subsequent deamination of CpG sites, suggesting these changes are driven by a distinct mutational process. In coding regions these particular biases have favoured the creation of single codon amino acids, offsetting the low frequency with which they are created by SNPs. Intriguingly selection has favoured particular pathways through the amino acid code, with epistatic selection appearing to have disfavoured sequential non-synonymous changes.

Introduction

Single nucleotide polymorphisms (SNPs) are the pillars of modern genetics studies [1, 2]. From their use in genome-wide association studies to map the genetics of diseases [3], to studying patterns of evolution [4], SNPs are widely used and studied across fields. A common implicit assumption across these studies is that SNPs are independent [5], with each substitution assumed to have resulted from a distinct mutational event. However, the distribution of SNPs in the genome has been known for some time to not be random,

with not only polymorphisms [6, 7] but also fixed differences between species clustering in the genome [8]. Where multiple base substitutions are found in the same genomic region, the derived alleles are more often found on the same haplotype [5], suggesting that the two changes have not occurred independently.

Directly neighbouring polymorphisms are particularly enriched in the human genome [7], though why this is the case is unknown. Previous studies of trios have shown that many of these multi-nucleotide polymorphisms (MNPs) have probably arisen in a single generation [5, 9], and are particularly enriched with $GA \rightarrow TT$ and $GC \rightarrow AA$ changes, suggesting that they are linked to error prone replication by polymerase zeta [10]. However, not all MNPs can be explained by a single mutational event. Many neighbouring polymorphisms occur at different frequencies in the population suggesting they have arisen from two distinct mutations [7]. These clustered changes may simply reflect mutational hotspots but may also be the result of selection, with the impact of an initial deleterious SNP being at least partly corrected by a second neighbouring change [11]. One obvious circumstance is where the impact of a deleterious nonsynonymous change is offset by a second polymorphism nearby. For example, nonsynonymous changes are less often found on highly expressed haplotypes, suggesting selection has favoured particular combinations of coding and regulatory alleles [12]. Likewise there is evidence that selection has favoured particular combinations of nonsynonymous changes spanning different amino acids in the same protein [13].

In this study we focused on compound MNPs comprising two neighbouring changes (sometimes referred to as dinucleotide polymorphisms [14]) that cannot be readily explained by a single mutational event, and investigate whether they simply reflect two independent but neighbouring polymorphisms or whether the second change depends on the first. The human mutation spectrum has diverged between human populations, with particular SNPs in particular sequence contexts more common in different continental groups [15]. Accordingly, we also explore whether MNP frequencies have diverged between population groups and whether any divergence simply mirrors the previously observed differences in SNP frequencies. We also characterise the functional importance of these MNPs by studying the selective pressures acting upon them and whether they have favoured particular pathways through the amino acid code.

Results

To investigate the frequency of neighbouring mutations in different populations we defined MNPs in the 1000 genomes phase 3 dataset [16] using the approach illustrated in Fig 1. Although MNPs cover a broad class of linked polymorphisms, in this study we focused specifically on those comprising two neighbouring nucleotide changes that, due to being found at different frequencies, cannot be readily explained by a single mutational event. All of these MNPs were annotated with respect to their ancestral alleles, triplet nucleotide context and occurrence in each individual (see methods), allowing us to infer the order of nucleotide changes. SNPs were defined in the same way to enable direct comparisons of their frequencies to MNPs.

We observed that a typical individual carries over 43,000 of these compound MNPs (mean of 43,014, minimum of 37,218 and maximum of 50,676) of which on average 27 fall within a protein coding region. The distribution of MNPs across the genome broadly follows that of SNPs (S1 Fig to S5 Fig), with the exception of the major histocompatibility complex on chromosomes 6, a region under strong positive selection due to its role in antigen presentation [17], that carries an unusually high proportion of MNPs despite its high SNP density (S6 Fig).

MNPs show biases between humans populations distinct to those among SNPs

Despite being a reduced summary statistic compared to individual genotypes, comparisons of the frequency of SNPs found in different triplet contexts between individuals has been shown to separate out the major human population groups [15]. This may be due to differences between populations in the large number of genes that control DNA mutation and repair. Characterisation of the individual SNPs that make up MNPs recapitulates the patterns observed in this previous study. Principal component analysis of the frequencies of the first and second changes in compound MNPs, defined by their sequence context, show highly similar patterns to that observed in [15] (Fig 2A and B). To explore whether the separation between populations in these two PCA was being driven by the same mutational biases, we calculated the differences in the frequencies of the changes occurring as the first or second change in MNPs. If the first and second change in MNPs showed the same biases between populations, then continental groups should not still separate in a PCA of these differences. However, as seen in S7 Fig the major populations are as effectively separated as before, indicating that different populations show different biases for the first and second mutations in MNPs.

Consequently the individual changes that make up MNPs display similar abilities to separate population groups as SNPs. However, despite their more limited numbers, the frequency of MNPs can even more effectively separate out the major human sub-populations (Fig 2C); with certain types of MNPs in specific nucleotide contexts enriched in different populations. Surprisingly, unlike the PCAs of individual SNPs, the American continental group separates in this analysis, in part driven by a depletion for $CG \rightarrow TG \rightarrow TA$ MNPs among these individuals (Fig 3A). In contrast MNPs are relatively depleted at AT rich triplets in African populations, and in particular MNPs where the mutations have the effect of switching AT base composition between strands (i.e. that involve multiple neighbouring $A{:}T \rightarrow T{:}A$ or $T{:}A \rightarrow A{:}T$ mutations; Fig 3B).

Fig 2D illustrates it is possible to also effectively separate population groups using an even simpler summary statistic than SNP or MNP frequencies classified by their triplet context. Plotting the number of MNPs against the number of SNPs in each individual also effectively separates the major continental groups, including the relatively admixed Americas group. This suggests the net rate of accumulation of MNPs relative to that of SNPs has not been the same across human populations, with African populations carrying a relative excess of MNPs, while Asian populations carry a lower number than expected compared to other groups.

Consequently, MNPs show different mutational spectra across populations that are distinct and more pronounced than those among SNPs, which can effectively separate major population groups.

CpG deamination is not sufficient to explain MNP variation

Intriguingly, the results in S7 Fig imply that the first and second mutations in MNPs are not driven by the same processes, with different biases between populations. To investigate this further we characterised the types of mutations observed as the first and second change in MNPs classified again by base change and triplet context. To ensure that batch effects and sequencing artefacts did not confound this analysis we replicated it across two independent datasets. The global set of 2504 genomes sequenced by the 1000 genomes consortium used above, as well as a novel dataset of 1358 Scottish genomes from the Lothian Birth Cohorts 1921 and 1936 [18, 19], whole genome sequenced at a mean depth of 36X.

Comparison of the frequency with which particular changes occur as the first and second mutation in an MNP indicates that each shows biases for different types of

change. As shown in Fig 4 a clear role of CpG dynamics in shaping MNPs is observed. Methylated cytosines immediately followed by a guanine (i.e. CpG sites) are known to be particularly prone to deaminate to a thymine, with mutation rates at these sites up to 18 times higher than at other dinucleotides [20]. As shown in Fig 4, the first mutation in MNPs is more likely to create a new CpG site, with the second change more likely to lead to the loss of one. This suggests that a dominant factor underlying MNPs is an initial mutation that creates a new CpG site, which subsequently rapidly mutates. This signature is observed across both the 1000 genomes and Lothian Birth Cohorts S8 Fig.

This raises the question as to whether MNPs simply reflect the known mutational profiles of SNPs (S9 Fig). To explore this we determined the expected frequency of each MNP in the genome given the observed frequencies of its constituent changes among SNPs (see methods for more details). Although moderate correlations were seen between these expected and observed counts (Fig 5), background single nucleotide mutation rates (as indicated by intergenic SNP frequencies) explain only just over half of the variation in counts between MNPs (negative bionomial regression McFadden's psuedo-R^2: 0.57 (Lothian Birth Cohort), 0.61 (1000 Genomes Cohort)). Substantial outliers were observed, where the frequencies of MNPs in the genome do not match what would be expected from the rates of changes among SNPs. In particular the frequencies of MNPs involving $CA \rightarrow CG \rightarrow TG$ changes (and their complement $TG \rightarrow CG \rightarrow CA$) occur over an order of magnitude more frequently than expected given the frequency of the constituent changes among SNPs (Fig 5). Notably, they also occur over an order of magnitude more often than other changes that involve the creation and subsequent deamination of CpG sites. Whereas 10633 intergenic $CAG \rightarrow CGG \rightarrow TGG$ changes were observed in the 1000 genomes population there are only 1038 intergenic $CTG \rightarrow CGG \rightarrow TGG$ changes, despite both changes leading to a similar creation and loss of CpG sites. This implies that a distinct process where an initial $A : T \rightarrow G : C$ change is favoured has led to the comparatively high frequency of these changes, and not simply CpG dynamics. Modelling the interaction between these two factors (original base change and the impact of changes on CpG sites) using regression analysis confirms that the importance of different types of CpG site change depends on the form of the original base change (S10 Fig).

Further mutational biases specific to MNPs are observed. For example TTA to TTT polymorphisms are more common as the second change in an MNP than the first, and MNPs containing two consecutive $A{:}T \rightarrow T{:}A$ changes are more common than expected given the frequency of the same changes among SNPs (Fig 4 and Fig 5). Consequently although the turnover of CpG sites drives the creation of a large proportion of MNPs, further mutational processes appear to be contributing to different biases not only between MNPs and SNPs, but also between the first and second changes of neighbouring polymorphisms.

The sequence of changes observed at MNPs depends on their ancestral states

To explore the relationship between the first and second change in neighbouring polymorphisms further, we examined MNPs that passed through the same intermediate nucleotide triplet. If the two changes that comprise an MNP are independent, then these intermediate triplets should show similar frequencies of downstream changes irrespective of the ancestral triplet from which they derived.

The dominant feature underlying downstream changes at many of these intermediate triplets was the biased deamination of CpG sites. MNPs passing through an intermediate triplet containing a CpG site are dominated by subsequent CpG to TpG changes irrespective of the original ancestral nucleotides. However, intermediate triplets that

contain no C/G base pairs often show biased downstream changes depending on their original ancestral state. For example, 94% of MNPs passing through an intermediate TTA triplet go on to become TAA if the original ancestral codon was TTT. This number is though only 17% if the original ancestral codon was TTG (Fig 6A and Table 1). To minimise the potential impact of batch effects and sequencing artefacts we again sought replication for this observation in the independent Lothian Birth Cohort collection of Scottish genomes and this difference was found to be highly significant in both datasets (Fig 6B).

There appear therefore to be constraints on the second change in MNPs dependent on their original ancestral state. This phenomenon is not exclusively restricted to intermediate triplets containing just adenines and thymines. For example, ACA intermediate triplets are more likely to be linked to an ATA final triplet if the ancestral triplet was CCA than if it was ACG (Fig 6A+B).

Coding MNPs have favoured the creation of single codon amino acids

We next investigated the impact of these particular mutational biases of MNPs on coding regions. To do this we compared the observed number of amino acid changes among MNPs to the corresponding number among coding SNPs (Fig 7). The particular biases of MNPs noted above has meant that their impact on coding regions is substantially different to that of coding SNPs, with different amino acid changes favoured (Fig 7, S11 Fig). Nine out of ten of the most common coding MNPs involve the previously described $CA \rightarrow CG \rightarrow TG$ change enriched among MNPs, and as a result of the layout of the amino acid code the bias for this change has led to the preferential creation of the single codons that code for methioinine and tryptophan (logistic regression of frequencies of changes that create these single codon amino acids to the frequencies of all other changes: $P = 6.4x10^{-04}$). $Thr_{ACA} \rightarrow Thr_{ACG} \rightarrow Met_{ATG}$ is the most common coding MNP, with $Gln_{CAG} \rightarrow Arg_{CGG} \rightarrow Trp_{TGG}$ the fourth most common change (Fig 7). The mutational biases of MNPs and organisation of the amino acid code have therefore combined to favour the creation of single codon amino acids. This is particularly intriguing given these amino acids are difficult to create from individual mutations, which is therefore partially compensated for by the particular mutational biases of MNPs.

To investigate the potential impact of selection acting on coding MNPs we compared the frequencies of coding and intergenic MNPs, under the assumption that intergenic MNPs are under comparatively little selection and provide a baseline of background mutation rates. A noticeable difference to the previous comparison with coding SNPs is that the first change of coding MNPs is depleted with a number of the changes that create a CpG site (Fig 7, S12 Fig). MNPs where the first mutation is non-synonymous (missense, stop gained or lost) are depleted among coding regions, likely reflecting their removal by selection, with the notable exception of the $Gln_{CAG} \rightarrow Arg_{CGG} \rightarrow Trp_{TGG}$ change leading to the creation of the tryptophan codon that is enriched in coding regions (S13 Fig). Accounting for background MNP rates the $Thr_{ACA} \rightarrow Thr_{ACG} \rightarrow Met_{ATG}$ pathway through the amino acid code also remains unusually enriched among coding MNPs.

To explore the mutational profiles of coding MNPs further, we compared the normalised frequencies with which MNPs occur on the coding and non-coding strands of genes. Selection on MNPs in genes is expected to depend on the changes with respect to the coding strand, suggesting that differences in the occurrence of particular changes on the two strands is attributable to stronger selection on the change when on the coding strand than its reverse complement. Although the majority

of MNPs show similar frequencies across both strands, as shown in Fig 8 a subset of MNP changes are comparatively enriched on the coding strand of genes, including both the $Thr_{ACA} \rightarrow Thr_{ACG} \rightarrow Met_{ATG}$ and $Gln_{CAG} \rightarrow Arg_{CGG} \rightarrow Trp_{TGG}$ changes. Of the five MNPs significantly enriched on the coding strand, only one; $Pro_{CCG} \rightarrow Pro_{CCA} \rightarrow Leu_{CTA}$ does not involve a $CA \rightarrow CG \rightarrow TG$ change. The reverse complement of this change is $Arg_{CGG} \rightarrow Trp_{TGG} \rightarrow Stop_{TAG}$ suggesting that this MNP is comparatively infrequent on the coding strand due to it leading to the introduction of a deleterious stop codon.

We conclude that although particular changes are favoured by MNP mutation rates, selection has preferentially removed changes passing through stop or a subset of missense codons, leaving changes that create single codon amino acids comparatively unaffected. Together, this has led to MNPs shaping coding regions differently to SNPs.

Evidence for epistatic selection at neighbouring coding polymorphisms

We finally investigated whether there is evidence of epistatic selection acting across the polymorphisms that make up coding MNPs i.e. whether the selective pressure acting on a nucleotide change depends on neighbouring changes. The amino acid code is thought to have been optimised so that physically similar amino acids have been brought together at neighbouring positions [21]. This ensures that the negative effect of a single mutation is minimised. By extension though, this means that two successive missense mutations in a codon are potentially more deleterious than one, despite the net effect still being one amino acid change. If the impact of an original missense mutation on fitness depends on the form of subsequent neighbouring polymorphisms (missense or synonymous), then this would suggest epistatic selection has partly shaped the frequencies of coding MNPs.

To explore this hypothesis we focused on MNPs where the initial mutation led to a missense change, and characterised whether the strength of selection acting upon this initial change depends upon subsequent neighbouring mutations. We used multiple linear regression to control for potential confounders (background rates of MNPs at intergenic regions, the number of different codons that encode the final amino acid and the impact on CpG sites of each change in the MNP). S14 Fig shows that MNPs made up of two successive missense changes are less frequently observed in the genome when compared to missense changes followed by a subsequent synonymous change (P=0.0044, false discovery rate=0.0084). The frequency of missense-synonymous MNPs in the genome is on average 35% higher than missense-missense MNPs after accounting for intergenic MNP rates, CpG mutability and codon frequency.

Although single codon amino acids appear to have been generally favoured by MNPs, having accounted for this effect, the creation of a methionine codon following an original missense change in fact occurs particularly infrequently relative to synonymous changes (P=0.00052, false discovery rate=0.0013). Consequently the impact of missense changes on fitness appears to depend on the form of subsequent neighbouring mutations, and epistatic selection between neighbouring coding polymorphisms has helped shape the frequency of MNPs in the human genome.

Discussion

Although the changes that make up MNPs have been traditionally categorised as SNPs, we have shown they appear to be driven by distinct mutational processes. Although the creation and subsequent deamination of CpG sites underlies a large proportion of MNPs, we show, we believe for the first time, that $CA \rightarrow CG \rightarrow TG$ changes are substantially overrpresented relative to other changes that involve the gain and deamination of a

CpG site. This suggests a distinct mutational process is driving this bias for these specific changes. The creation of new G:C base pairs is often attributed to biased gene conversion (BGC) that favours weak (A:T) to strong (G:C) basepair changes. However, previous studies have found little evidence of a strong effect of BGC on genome-wide mutational profiles [15, 22], and, for example, $T:A \rightarrow G:C$ changes are not similarly enriched among the first change of MNPs, suggesting a distinct mutational process is potentially driving the elevated rate of these polymorphisms.

As with SNPs, MNP mutational profiles appear to differ between human groups, but, unlike SNPs, MNP rates can more clearly define populations. In contrast to SNP mutational profiles, the MNP profiles of the Americas populations distinguishes them from other continental groups, in part driven by a depletion of $CG \rightarrow TG \rightarrow TA$ changes among these individuals. The larger spectrum of changes among MNPs, and the fact that the changes in MNP are not independent, and do not simply reflect underlying SNP rates, likely provides the greater resolution in defining populations.

Not all common MNPs are associated with CpG sites. In particular, changes at AT rich triplets show population differences and enrichment among African populations. The second change in AT-rich MNPs also often appears dependent on the original ancestral sequence. One potential explanation for this is a role for homologous recombinational repair among these MNPs. Recombination based repair mechanisms transfer nucleotide sequence information between chromosome copies, and as a result between ancestral and derived haplotypes. The second changes in these MNPs may therefore reflect errors in this repair process, potentially arising due to the pre-existing mismatch between chromosomal copies. Alternatively constraints on the sequence composition of regions may lead to biases towards those maintaining local nucleotide composition.

An intriguing consequence of the bias for the creation and subsequent loss of CpG sites among MNPs is the fact that this favours the creation of methionine and tryptophan codons due to the specific arrangement of the amino acid code. As both codons contain a TG dinucleotide they are readily created by the preference for $CA \rightarrow CG \rightarrow TG$ changes. This therefore partly offsets the fact that these amino acids are more rarely created by single mutations due to being only encoded by a single codon. Among coding MNPs the creation of new methionine codons is the most common of all double mutations, even when accounting for background mutation rates.

Nonadditive, i.e. epistatic, genetic interactions have been proposed to underlie a range of phenomenon such as the missing heritability of phenotypes, but detecting such interactions in humans has proven difficult [23]. Previous studies have suggested that the strength of selection acting upon a coding variant may depend on the alleles carried at other variants nearby. For example [12] showed that putatively functional coding variants are less often observed on more highly expressed regulatory haplotypes. In this study we showed that, when accounting for factors such as background mutation rates and the impact of CpG deamination, an original non-synonymous change is less often than expected followed by a non-synonymous change in the same codon. This suggests that the strength of selection acting upon the original missense mutation depends on the type of change that follows it, despite the net effect still being a single amino acid change. This result is broadly consistent with a previous study of mutations across different codons in proteins that suggested the fitness of a given non-synonymous change can depend on other nonsynonymous changes elsewhere in the protein [13]. However, an assumption made to varying degrees by all these studies is that missense changes can be grouped, and that their impact on fitness is broadly similar. Even larger sequencing cohorts, such as those being generated as part of the UK Biobank [24], would help refine this analysis and the epistatic selection acting upon individual types of intermediate missense changes.

Consequently human mutation profiles appear more complex than previously thought,

with neighbouring polymorphisms driven by distinct mutational processes. These MNPs are under unusually strong selective pressure and have played an important and distinct role in shaping human protein evolution.

Materials and methods

Variant calling

1000 genomes consortium version 3 phased haplotypes along with information on their ancestral alleles were obtained from `http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`. After excluding variants with missing or low confidence ancestral allele annotations, neighbouring variants where both derived alleles were observed together on the same haplotype were flagged as potential MNPs. As the focus of this analysis was on MNPs originating from two neighbouring mutation events, only MNPs where haplotypes also existed which carried the derived allele at one but not the other variant were kept, i.e. the two changes were onserved at different frequencies. See Fig 1 for more details on how MNPs were defined in this study. Due to the very low probability of recombination events occurring between neighbouring bases in the human genome, we also excluded MNPs where both combinations of one derived and one ancestral allele were observed. Following this filtering 377,766 neighbouring pairs of polymorphisms remained.

Illumina HiSeq X paired-end sequencing data for 1370 Lothian Birth Cohort [18, 19] individuals (mean sequencing depth of 36X) were aligned to the build 38 version of the human reference genome using BWA [25]. Variants were called using GATK [26] according to its recommended best practices. This included the use of GATK's HaplotypeCaller software that implements read based phasing of nearby alleles. After checking identities with previous array data, excluding duplicate individuals and those displaying excessive levels of heterozygosity 1358 individuals remained. All SNP coordinates were then lifted over to build 37 so as to match the 1000 genomes dataset using Crossmap [27]. Ancestral sequences available at `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/` were used to determine the ancestral alleles of each SNP, and as with the 1000 genomes data only those variants with high-confidence calls were kept (see [28] for more details). MNPs were subsequently called using the same approach as the 1000 genomes data.

All SNPs were annotated using variant effect predictor (VEP, [29]) with gene models from the version 85 release of Ensembl. MNPs were subsequently re-annotated using custom python scripts to correctly annotate the consequence of the double base change in coding regions.

All non-coding SNPs and MNPs were defined with respect to their triplet context on the reference strand. This meant that all non-coding MNPs were recorded twice, once with respect to their immediate 5' neighbour and once with respect to their 3' neighbour (Fig 1). non-coding SNPs were recorded with respect to all three frames within which they fell. On the other hand coding polymorphisms were recorded with respect to just the actual codon within which they occurred. This enabled the direct comparison of SNP and MNP frequencies inside annotated codons to matching triplet bases outside of coding regions.

Calculating observed and expected frequencies

The frequencies of particular mutation types were calculated as in previous studies of SNPs [15]. First, the number of distinct changes in a particular triplet context (m) was counted in each population (P). Then this count, $C_p(m)$ was converted to a frequency by

dividing it by the count of all observed changes. For MNPs, an individual was a carrier if at least one of their haplotypes carried both derived alleles at the corresponding pair of nucleotides.

In various analyses the two changes that comprise an MNP were separated into those that came first (i.e. the derived allele at just one of the two nucleotides is observed by itself in the population, see Fig 1) and those that came second. Their frequencies were also calculated as above.

The expected MNP frequencies were derived from background SNP frequencies by calculating the conditional probability of observing an MNP comprising the corresponding two changes (Eq 1).

$$P(Change1\ then\ Change2) = P(Change1) \cdot P(Change2|Change1) \quad (1)$$

Where $P(Change1)$ is the proportion of SNPs displaying the corresponding change when defined by their triplet context. $P(Change2|Change1)$ is the proportion of SNPs displaying the same change as change 2 among all SNPs with the same ancestral triplet and where the change is at a position in the triplet neighbouring the location of the first change. For example, if the first change was AAA>ATA and the second change ATA>ATC, $P(Change2|Change1)$ would correspond to the number of ATA>ATC intergenic SNPs divided by the sum of all ATA>ATB and ATA>BTA changes (where B can be any nucleotide except A). This therefore accounts for the fact that the second change in an MNP had to occur at a base neighbouring, but not at, the location of the first, and involve the triplet the first change had created.

Statistical analyses

To test whether the impact of CpG dynamics on MNP numbers depended on the form of the original base change, we fit the negative binomial generalized linear model specified in Eq (2).

$$\begin{aligned} obsMNPCount_i = \mu + expMNPCount_i\beta_l + firstBaseChange_i\beta_m + \\ cpgChange_i\beta_n + firstBaseChange_i cpgChange_i\beta_{mn} + \epsilon_i \end{aligned} \quad (2)$$

Where $obsMNPCount_i$ corresponds to the observed number of MNPs of type i (e.g. AAA>ATA>ATT), $expMNPCount_i$ is the expected number given background SNP frequencies (see Eq 1) and $firstBaseChange_i$ and $cpgChange_i$ are the first base change and impact of both changes on any CpG sites respectively. The significance of the interaction term was assessed using ANOVA.

Significance testing in Fig 4 and S8 Fig was carried out as in [15] i.e. we used their iterative approach of undertaking conditionally independent chi-square tests to try and minimise false positive significant results. To test for the enrichment of specific codon changes among coding MNPs, having accounted for background rates of coding SNP and intergenic MNP changes, we used multiple linear regression as specified in Eq 3 to Eq 6.

$$\begin{aligned} firstCodingMNPCount_i = \mu + firstIntergMNPCount_i\beta_l + funcClass_i\beta_m + \\ cpgChange_i\beta_n + funcClass_i cpgChange_i\beta_{mn} + \epsilon_i \end{aligned} \quad (3)$$

$$\begin{aligned} firstCodingMNPCount_i = \mu + codingSNPCount_i\beta_l + funcClass_i\beta_m + \\ cpgChange_i\beta_n + funcClass_i cpgChange_i\beta_{mn} + \epsilon_i \end{aligned} \quad (4)$$

$$secondCodingMNPCount_i = \mu + secondIntergMNPCount_i\beta_l + funcClass_i\beta_m + cpgChange_i\beta_n + funcClass_i cpgChange_i\beta_{mn} + \epsilon_i \quad (5)$$

$$secondCodingMNPCount_i = \mu + codingSNPCount_i\beta_l + funcClass_i\beta_m + cpgChange_i\beta_n + funcClass_i cpgChange_i\beta_{mn} + \epsilon_i \quad (6)$$

Where $firstCodingMNPCount_i$ corresponds to the number of distinct first changes among coding MNPs that match change i, where each i is one of the 576 possible single base difference between two codons. $secondCodingMNPCount_i$ is the corresponding count among the second changes of coding MNPs and $firstIntergMNPCount_i$ and $secondIntergMNPCount_i$ are the corresponding counts among the same nucleotide triplets at intergenic MNPs. $codingSNPCount_i$ is the count of the same change observed among coding SNPs, $funcClass_i$ is the functional impact of the corresponding change (missense, stop gained etc) and $cpgChange_i$ is the impact on any CpG sites (lost or created). Any interaction effect between the functional impact of the given codon change and its impact on CpG sites is represented by β_{mn}. The results of these four models are shown in S11 Fig and S12 Fig.

Multiple linear regression was also used in the test for epistatic selection among coding MNPs as specified in Eq 7.

$$codingMNPCount_i = \mu + intergMNPCount_i\beta_l + funcClass2_i\beta_m + codonCount_i\beta_n + cpgChange1_i\beta_o + cpgChange2_i\beta_p + \epsilon_i \quad (7)$$

Where $codingMNPCount_i$ is the number of each MNP, i, in the genome where i is restricted to the 1628 MNPs where the first change is missense. $intergMNPCount_i$ is the count of the same MNP, i, in intergenic regions. $funcClass2_i$ is the functional impact of the second change in MNP i, and $cpgChange1_i$ and $cpgChange2_i$ are the impact on CpG sites of the first and second changes in the MNP respectively. $codonCount_i$ is the number of codons that encode the final amino acid created by the MNP, to account for the fact that amino acids with only one codon are generally favoured by MNPs. A goodness-of-fit test confirmed that the Poisson model suitably fit the data (Chi-squared p=0.997).

Table 1. Counts and proportions (by ancestral triplet) of MNPs passing through the intermediate triplet TTA.

Ancestral triplet	Inter-mediate triplet	Derived triplet	Intergenic MNPs	Intergenic propor-tion by start triplet	Intronic MNPs	Intronic propor-tion by start triplet	Coding SNP count (first change)	Coding SNP count (second change)
ATA	TTA	TAA	38	0.34	32	0.42	314	68
ATA	TTA	TCA	56	0.50	36	0.47	314	558
ATA	TTA	TGA	17	0.15	9	0.12	314	91
CTA	TTA	TAA	36	0.21	20	0.19	1207	68
CTA	TTA	TCA	104	0.61	69	0.64	1207	558
CTA	TTA	TGA	30	0.18	18	0.17	1207	91
GTA	TTA	TAA	21	0.25	19	0.30	403	68
GTA	TTA	TCA	54	0.64	34	0.53	403	558
GTA	TTA	TGA	9	0.11	11	0.17	403	91
TTC	TTA	TAA	62	0.44	55	0.49	656	68
TTC	TTA	TCA	54	0.38	43	0.38	656	558
TTC	TTA	TGA	25	0.18	15	0.13	656	91
TTG	TTA	TAA	44	0.17	38	0.20	1865	68
TTG	TTA	TCA	162	0.64	121	0.64	1865	558
TTG	TTA	TGA	49	0.19	31	0.16	1865	91
TTT	TTA	TAA	1817	0.94	1696	0.95	257	68
TTT	TTA	TCA	81	0.04	55	0.03	257	558
TTT	TTA	TGA	28	0.01	27	0.02	257	91
TAA	TTA	ATA	61	0.03	52	0.03	15	168
TAA	TTA	CTA	46	0.02	44	0.03	15	606
TAA	TTA	GTA	27	0.01	19	0.01	15	244
TAA	TTA	TTC	23	0.01	18	0.01	15	182
TAA	TTA	TTG	76	0.04	47	0.03	15	1046
TAA	TTA	TTT	1611	0.87	1489	0.89	15	209
TCA	TTA	ATA	80	0.15	55	0.14	1503	168
TCA	TTA	CTA	130	0.25	90	0.23	1503	606
TCA	TTA	GTA	58	0.11	43	0.11	1503	244
TCA	TTA	TTC	35	0.07	34	0.09	1503	182
TCA	TTA	TTG	160	0.31	127	0.33	1503	1046
TCA	TTA	TTT	60	0.11	39	0.10	1503	209
TGA	TTA	ATA	32	0.14	11	0.06	24	168
TGA	TTA	CTA	47	0.21	29	0.17	24	606
TGA	TTA	GTA	19	0.09	16	0.09	24	244
TGA	TTA	TTC	25	0.11	18	0.10	24	182
TGA	TTA	TTG	54	0.24	43	0.25	24	1046
TGA	TTA	TTT	45	0.20	55	0.32	24	209

Figures

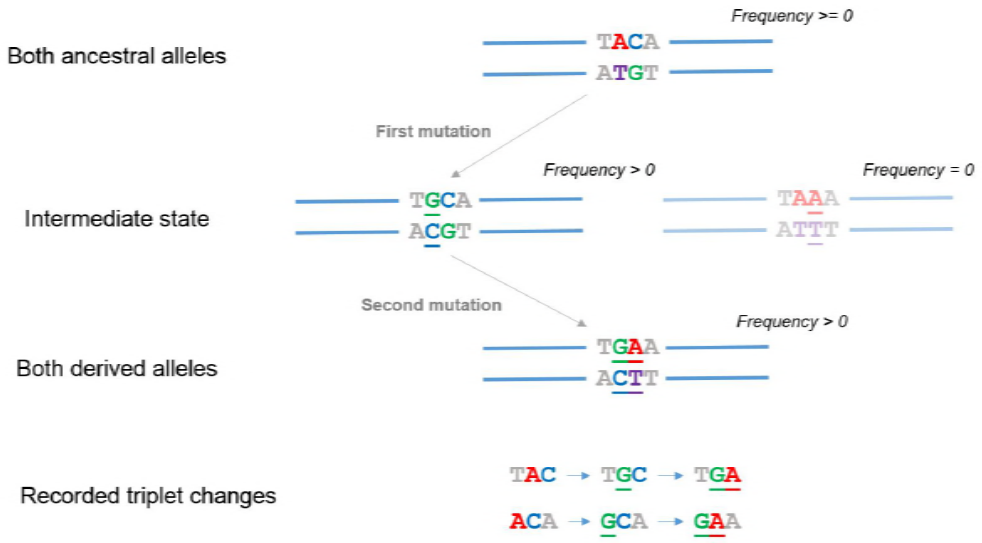


Fig 1. Defining the MNPs studied in this analysis. In this study we focused on MNPs involving neighbouring nucleotides that could not be readily explained by a single mutational event. Haplotypes containing just one of the two changes that make up the MNP had to be observed among the individuals (we assumed that reverting mutations and recombination events between neighbouring bases were rare). Using information on which alleles were ancestral we then inferred the order of changes, with each MNP recorded twice, according to their immediate 5' and 3' nucleotides. This allowed downstream analysis of their impact on codons and comparisons to SNPs in the same triplet contexts.

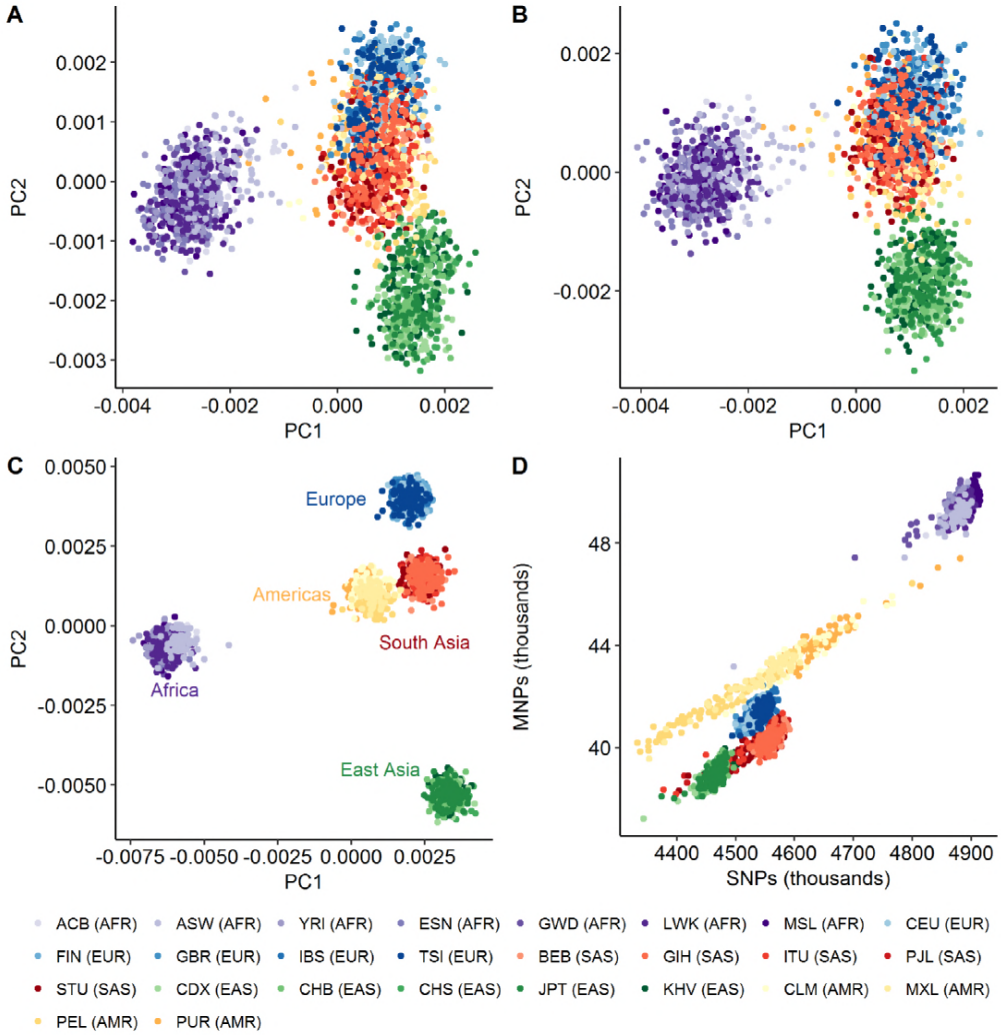


Fig 2. Divergence in MNP frequencies between human populations. A) Principal component analysis of individuals according to the frequency of each first change in the MNPs they carry, defined by their triplet context. Points are coloured according to the continental group to which they belong and the corresponding 1000 genomes consortium three letter population codes are shown below. Although restricted to SNPs forming part of MNPs only, this plot largely mirrors that derived using all SNPs in the same populations presented in [15] at `https://doi.org/10.7554/eLife.24284.006`. B) The same as A but for the second change in MNPs. C) Principal component analysis of individuals according to the frequency of the MNP changes they carry defined by triplet context. D) Number of SNPs and MNPs each individual carries.

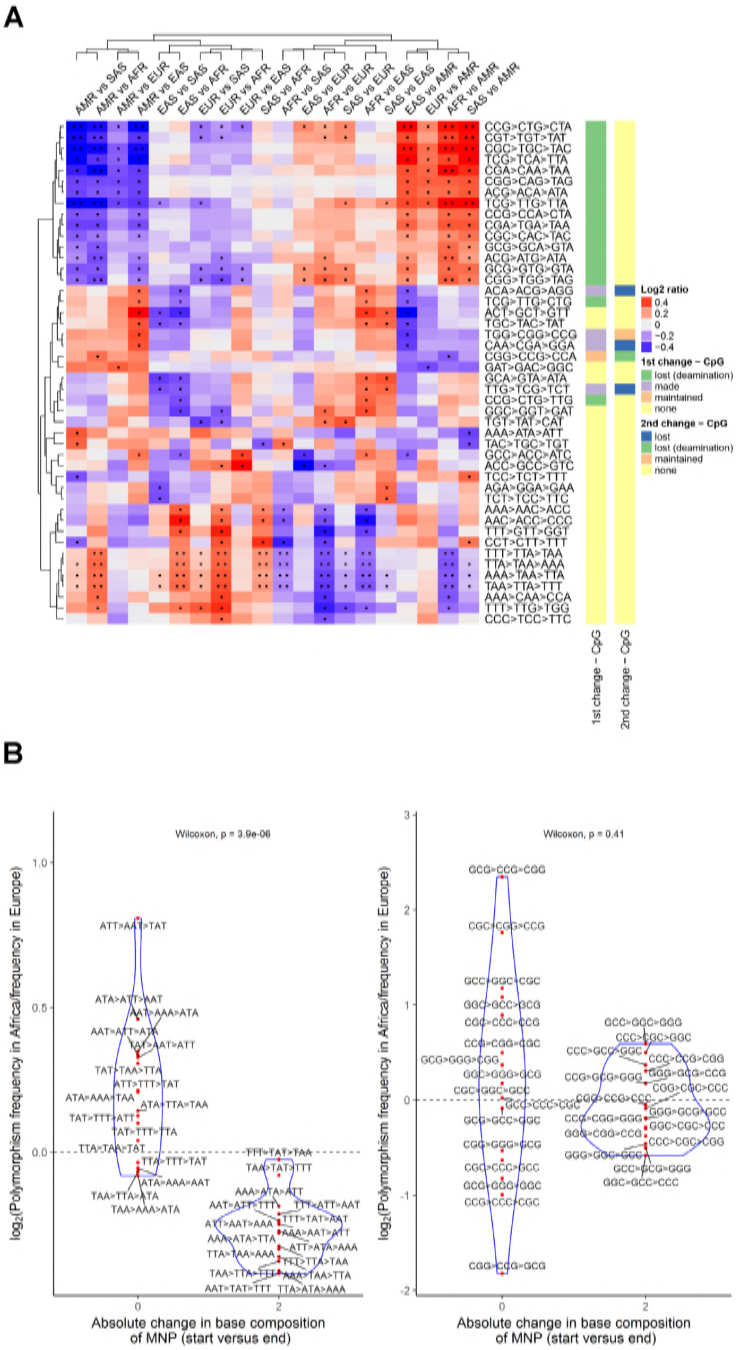


Fig 3. MNPs that differ between population groups. A) Log_2 ratios of the frequencies with which selected MNPs occur in different continental groups. Only MNPs found at at least 150 sites in each population and with a log_2 ratio >= 0.3 and a P <0.05 in at least one population comparison are shown in this plot. One dot indicates the corresponding comparison was associated with an uncorrected Fisher's exact P <0.05, two dots that the false discovery rate (q value) was <0.05. Red cells indicate the corresponding MNP is relatively enriched in the first named continental group, blue that the change is enriched in the second population. B) The relative enrichment of selected MNPs in African versus European populations. (Left) MNPs that exclusively involve weak base pairs (adenine and thymine) broken down by their net impact on the base composition of each strand. (Right) MNPs that exclusively involve strong base pairs (guanine and cytosine).

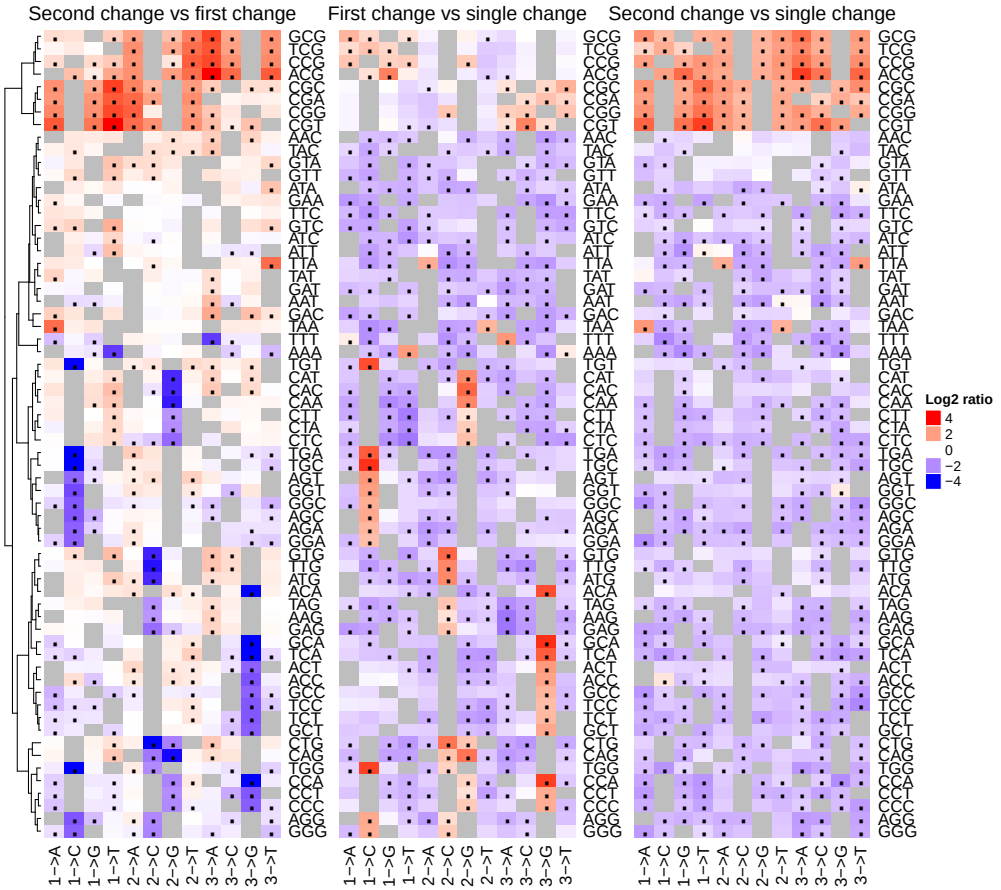


Fig 4. The frequencies of different forms of first and second changes among MNPs. (Left) The ratio of the frequencies of the first and second changes by triplet context in the 1000 genomes cohort. Each cell corresponds to a particular change defined by the original triplet (labelled on the right of the plot) and the observed change (labelled at the bottom). The numbers in the bottom row indicate which base in the triplet is polymorphic followed by the mutation that occurred i.e. 1->A in the bottom GGG row indicates the respective cell corresponds to GGG->AGG changes. A dot in the cell indicates that the frequency of the corresponding change is significantly different between the first and second changes of MNPs (Bonferroni corrected chi-squared P <0.05, see methods for more details). (Middle and right) The ratio of the frequencies of the first (middle) and second changes (right) in MNPs versus the observed frequencies of the same change at SNPs. In all three heatmaps red cells indicate the change is enriched among the first named change, and blue that it is enriched among the second named change.

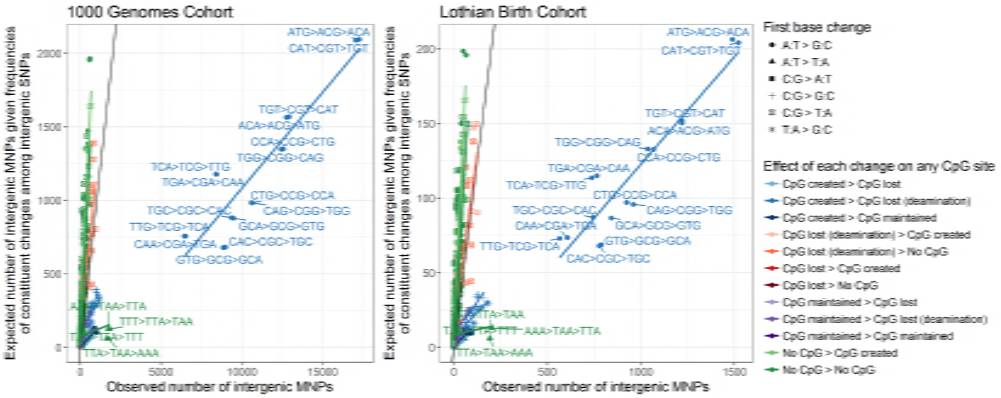


Fig 5. The observed number of intergenic MNPs against the number expected given the frequencies with which the constituent changes occur among SNPs. MNPs are broken down by the original base change and the impact on any CpG sites of each constituent change (impact of first change >impact of second change). The line of parity is shown in black. Together these three factors (the expected number of MNPs from background SNP rates, the first base change and impact of the changes on CpG sites) explain the majority of the variation in MNP frequencies (McFadden's pseudo-R^2: 0.84 (Lothian Birth Cohort), 0.85 (1000 Genomes Cohort)) .

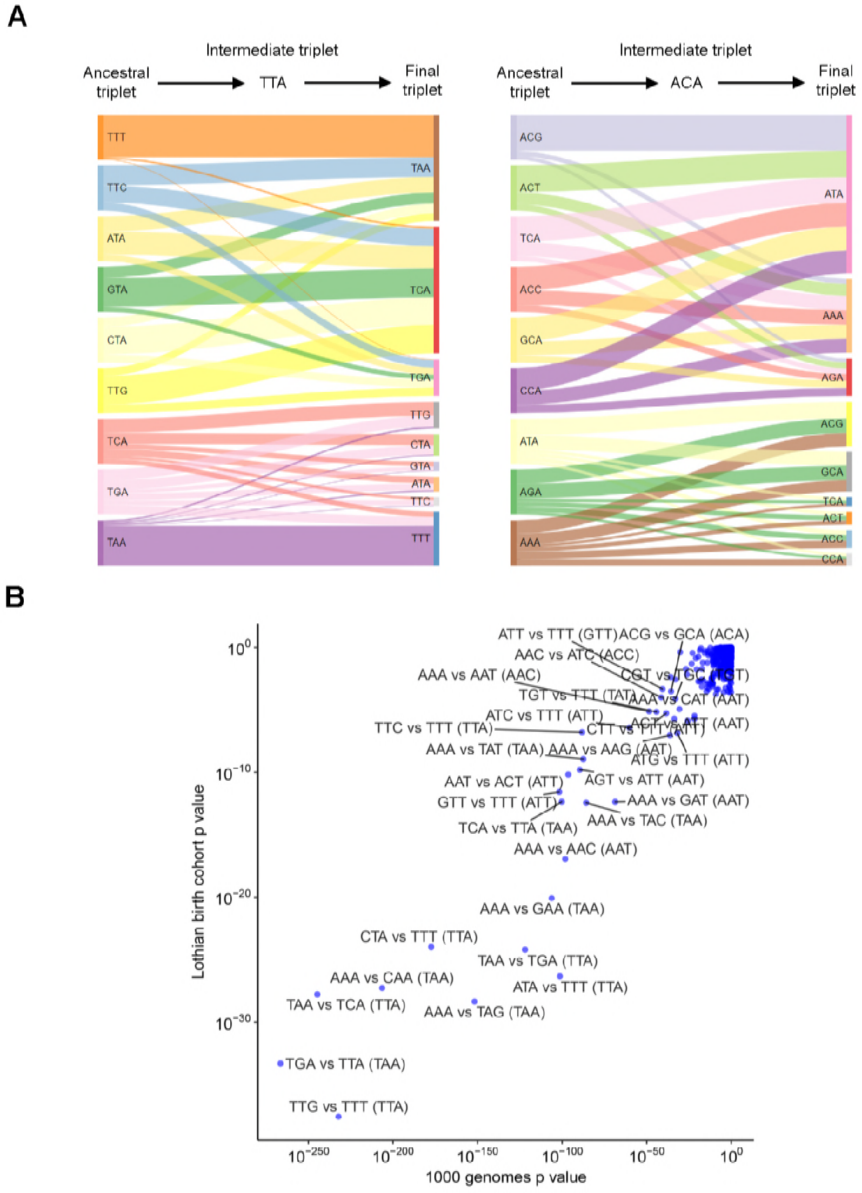


Fig 6. The form of the second change is related to the ancestral state of MNPs. A) Sankey plots representing the changes observed among intergenic MNPs sharing the same intermediate state. The width of each bar represents the proportion of MNPs sharing the corresponding ancestral state that ended at the indicated final triplet. B) P values indicating pairs of ancestral states where intergenic MNPs show different biases for final triplets despite passing through the same intermediate state (indicated in brackets). Chi-squared test p values for both the 1000 genomes and Lothian Birth Cohorts are shown.

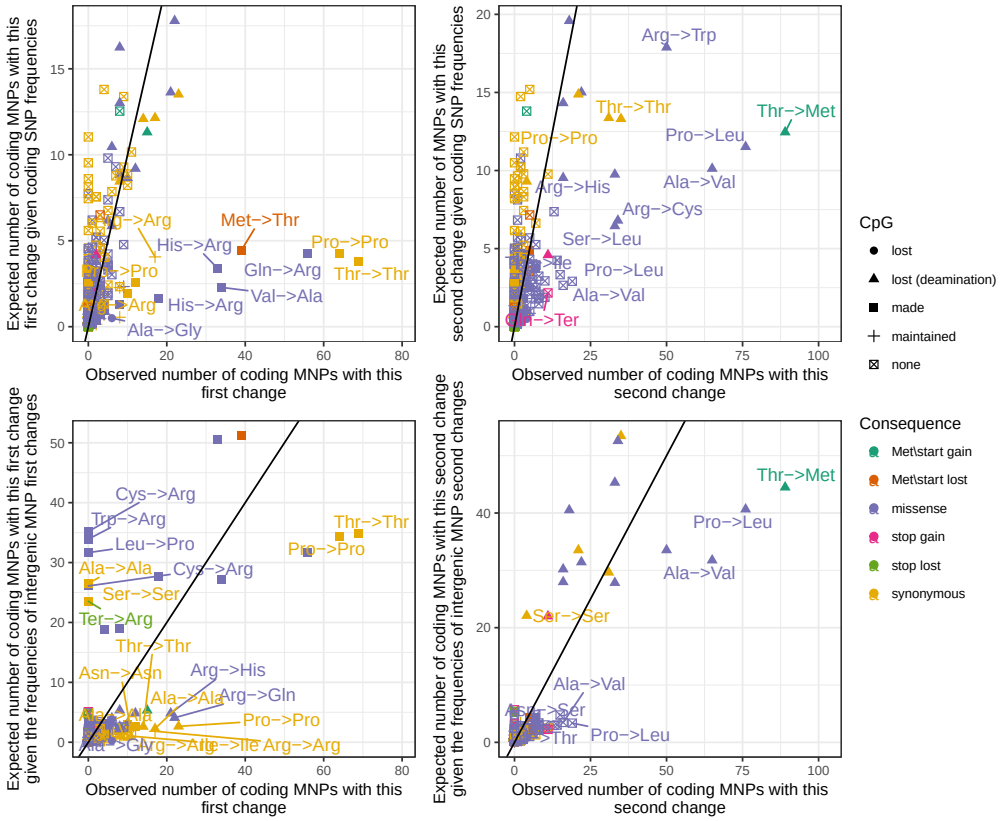


Fig 7. The observed and expected number of coding MNPs in the 1000 genomes cohort. The x axes in the left and right columns correspond to the observed number of first and second changes in coding MNPs respectively. Each point represents a distinct single base change between two codons and points are coloured by the impact of the change on the corresponding protein. The impact of the change on any CpG sites is also indicated by the shape of the point. The y axis in the top row corresponds to the number of the changes expected given their frequency among coding SNPs. The y axis in the bottom row gives the expected numbers given the frequencies of the changes amongst the first and second changes of intergenic MNPs. Points showing a significant difference between the observed and expected counts after correcting for multiple testing (Chi-squared $P<2.4e^{-5}$) are labelled with the corresponding amino acid change. The line of equality is also indicated.

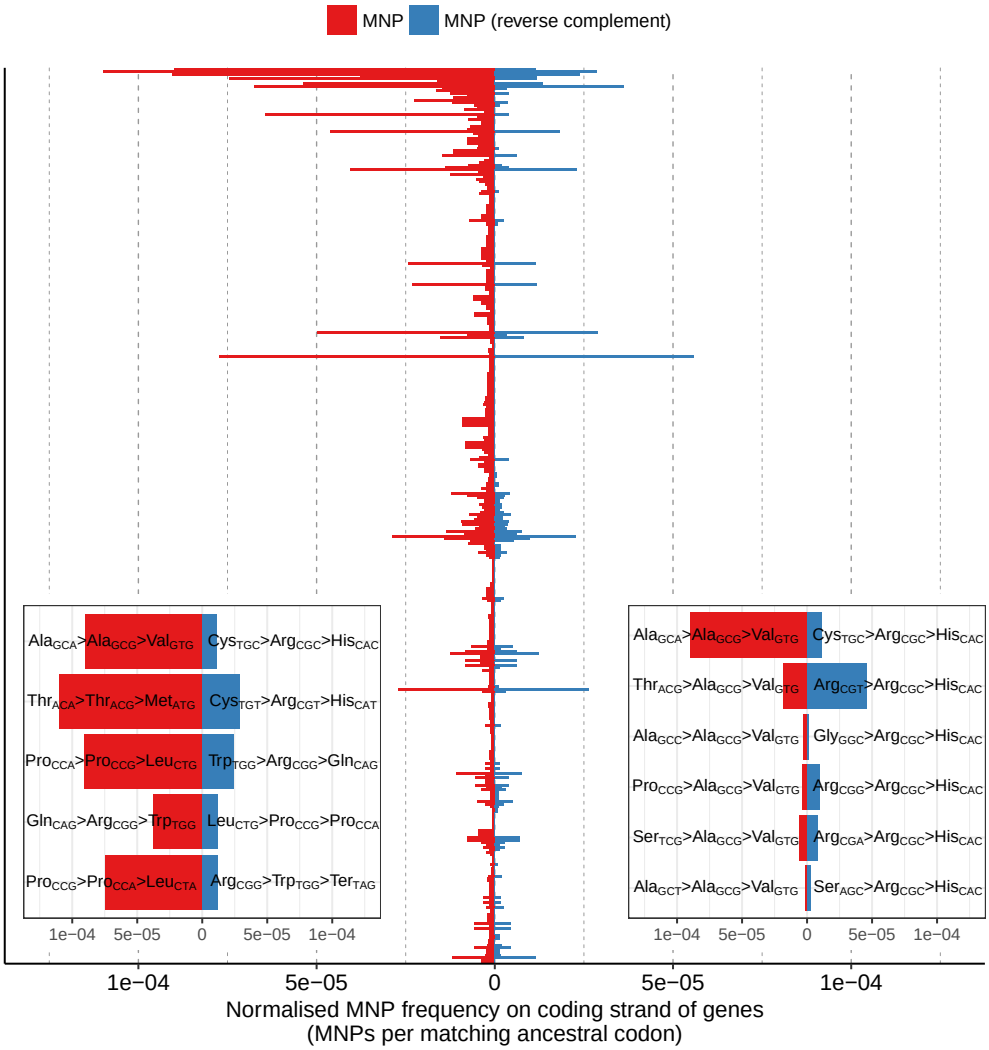
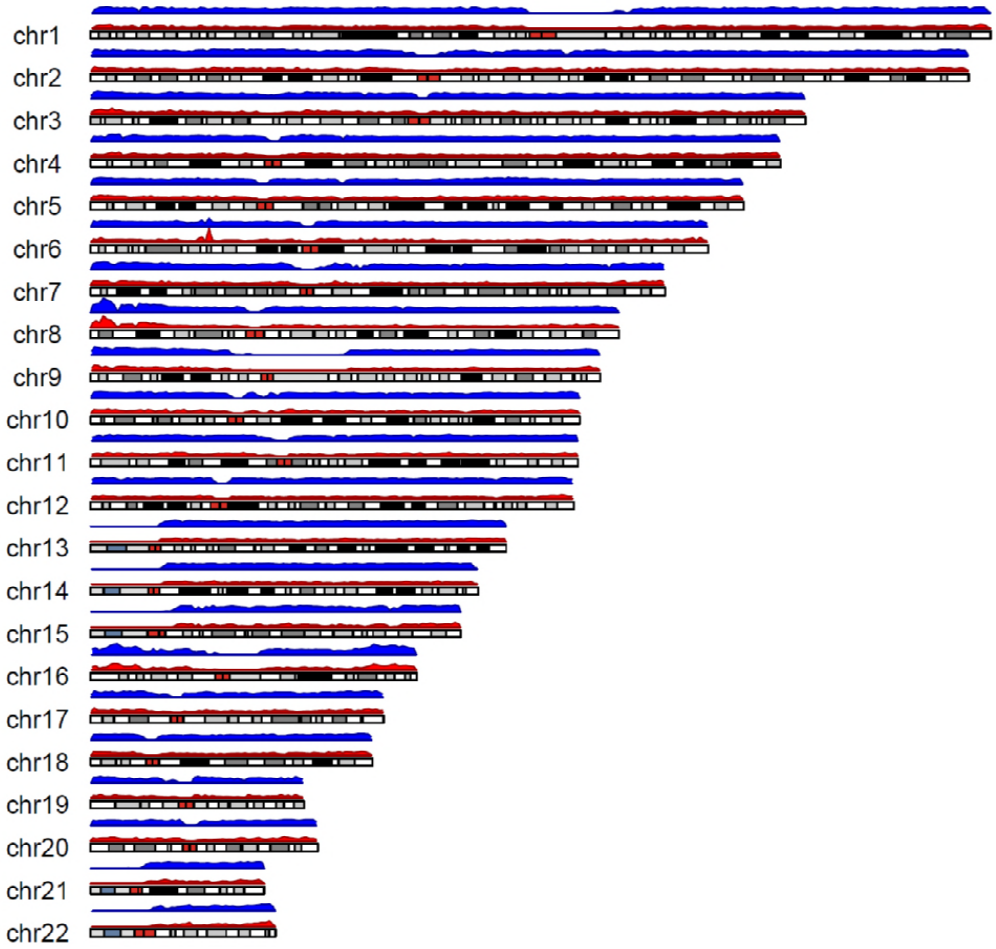
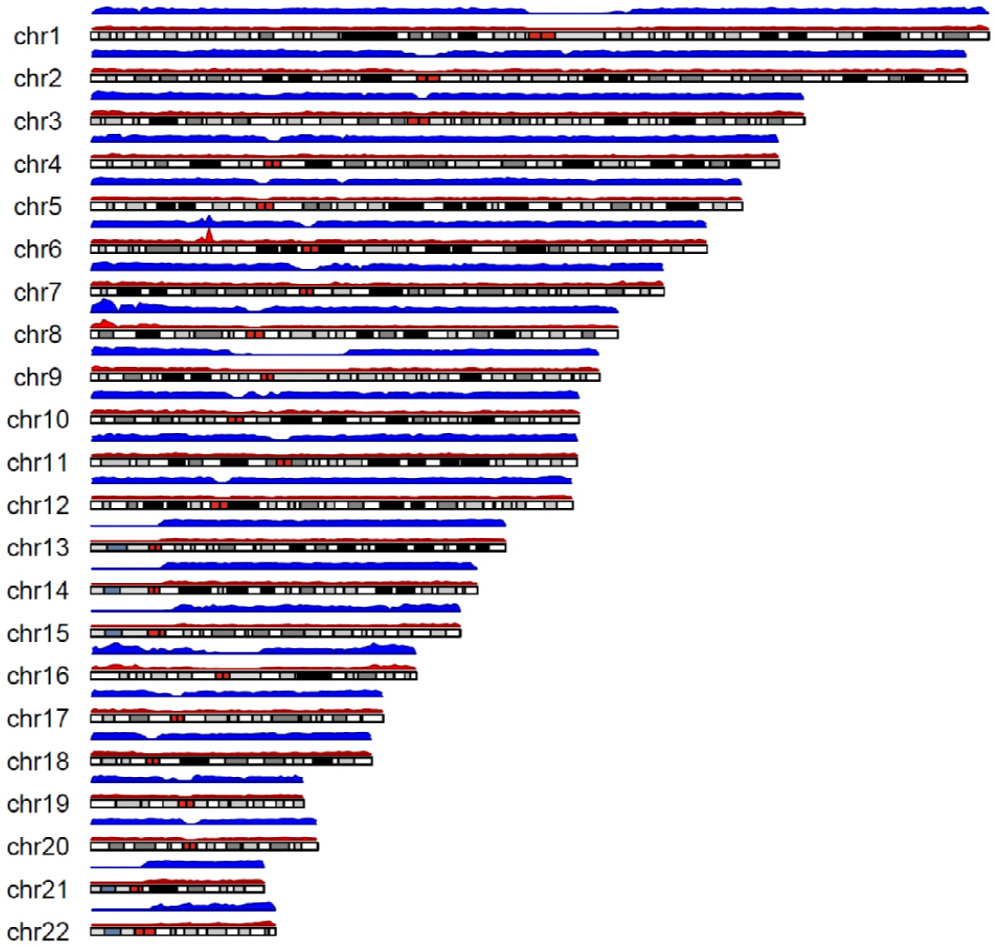


Fig 8. Asymmetry in the occurrence of changes on the coding and non-coding strand of genes. The observed count of each MNP change on the coding strand of genes, normalised by the frequency with which the ancestral codon occurs in coding regions, is shown. Each change is paired with its reverse complement change, with the count of the change displaying a relative enrichment on the coding strand shown in red. Where both changes are equally common on coding strands, one was randomly chosen to be shown in red. To test for asymmetry in the frequencies of MNP pairs, a 2x2 contingency table of (a) the count of each MNP in coding regions and (b) the count of the matching ancestral codons across all genes was constructed, and differences in the proportion of ancestral codons carrying the respective MNPs was tested using the Fishers exact test. MNP pairs are ranked by p value (smallest at the top). The five most significant pairs (P $<5x10^{-5}$) are shown in the bottom-left inset and the six MNPs involving a final $Ala_{GCG} \rightarrow Val_{GTG}$ change are shown in the bottom right.

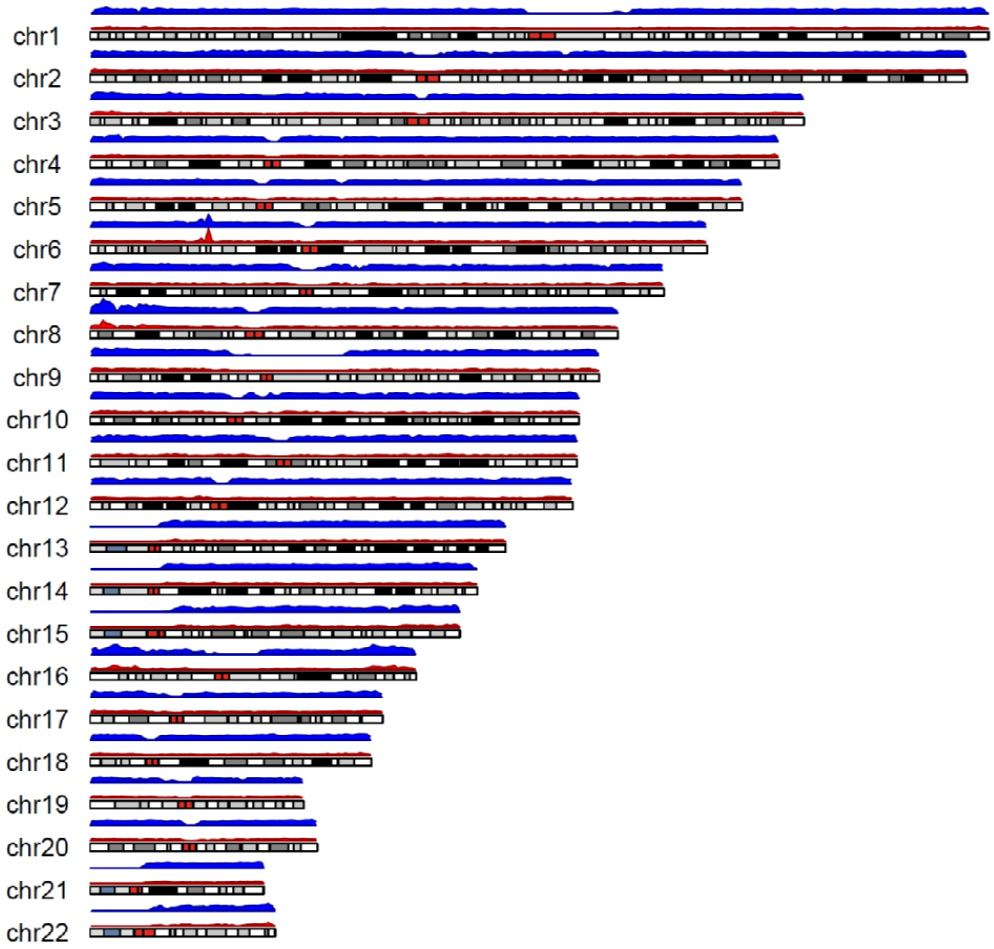
Supporting information



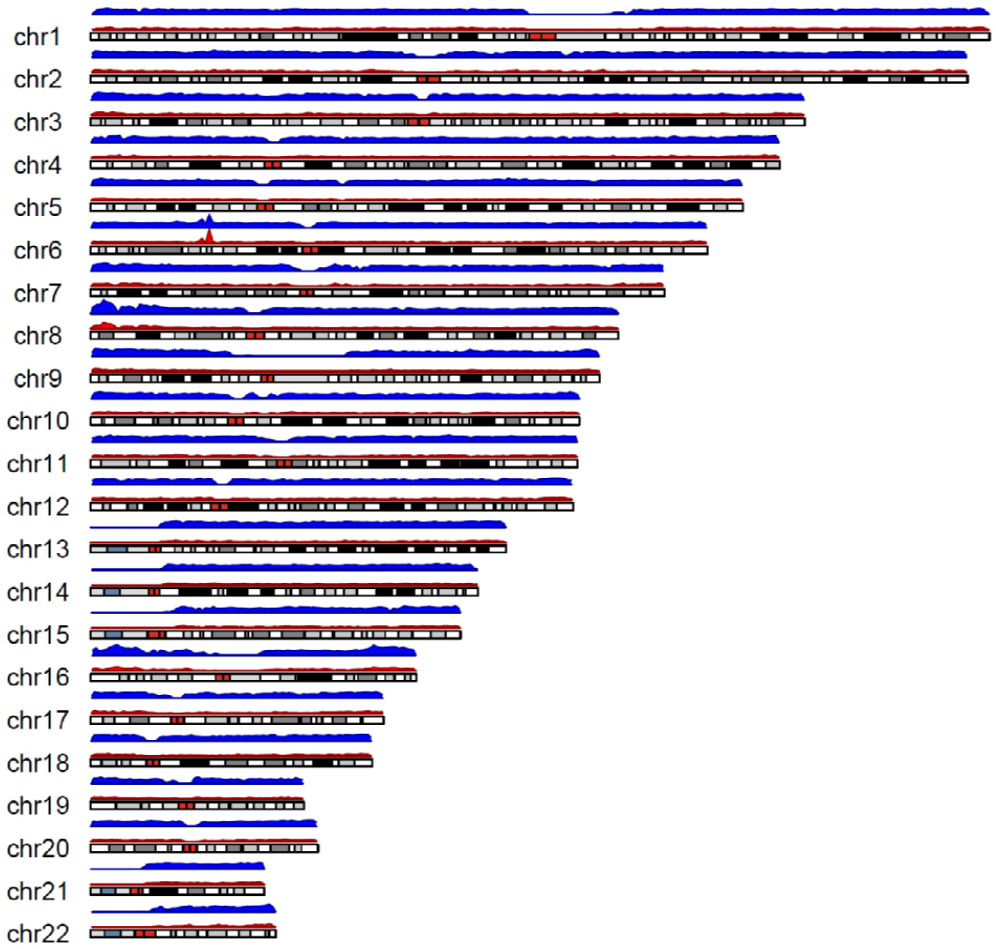
S1 Fig. African population MNP (red) and SNP (blue) densities in 1Mb windows across the genome



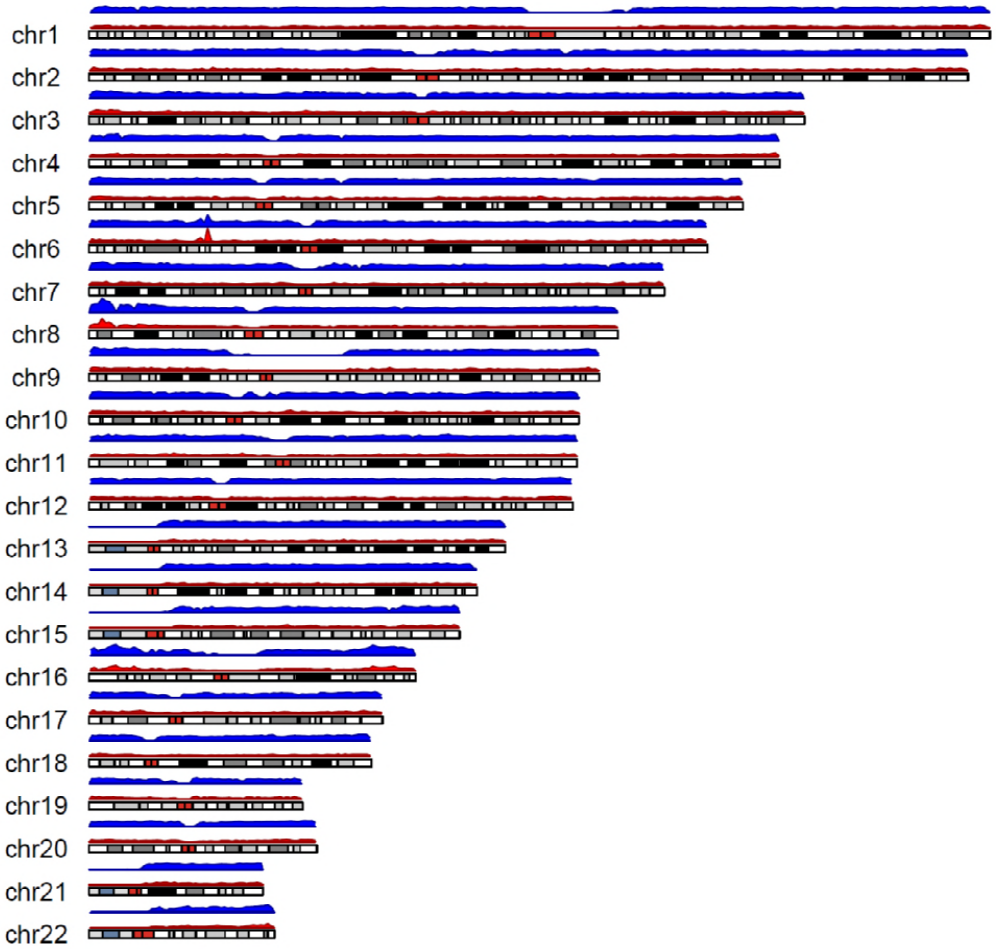
S2 Fig. American population MNP (red) and SNP (blue) densities in 1Mb windows across the genome



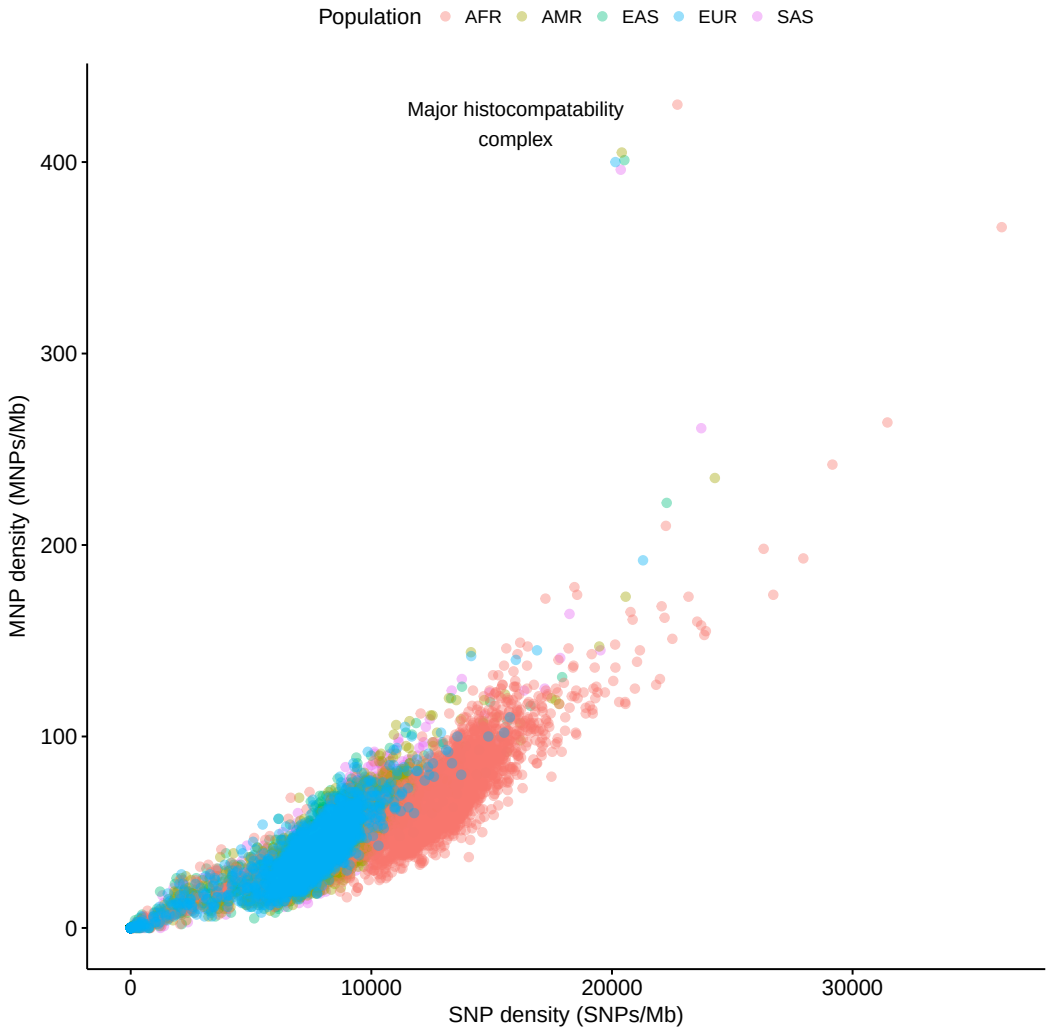
S3 Fig. East Asian population MNP (red) and SNP (blue) densities in 1Mb windows across the genome



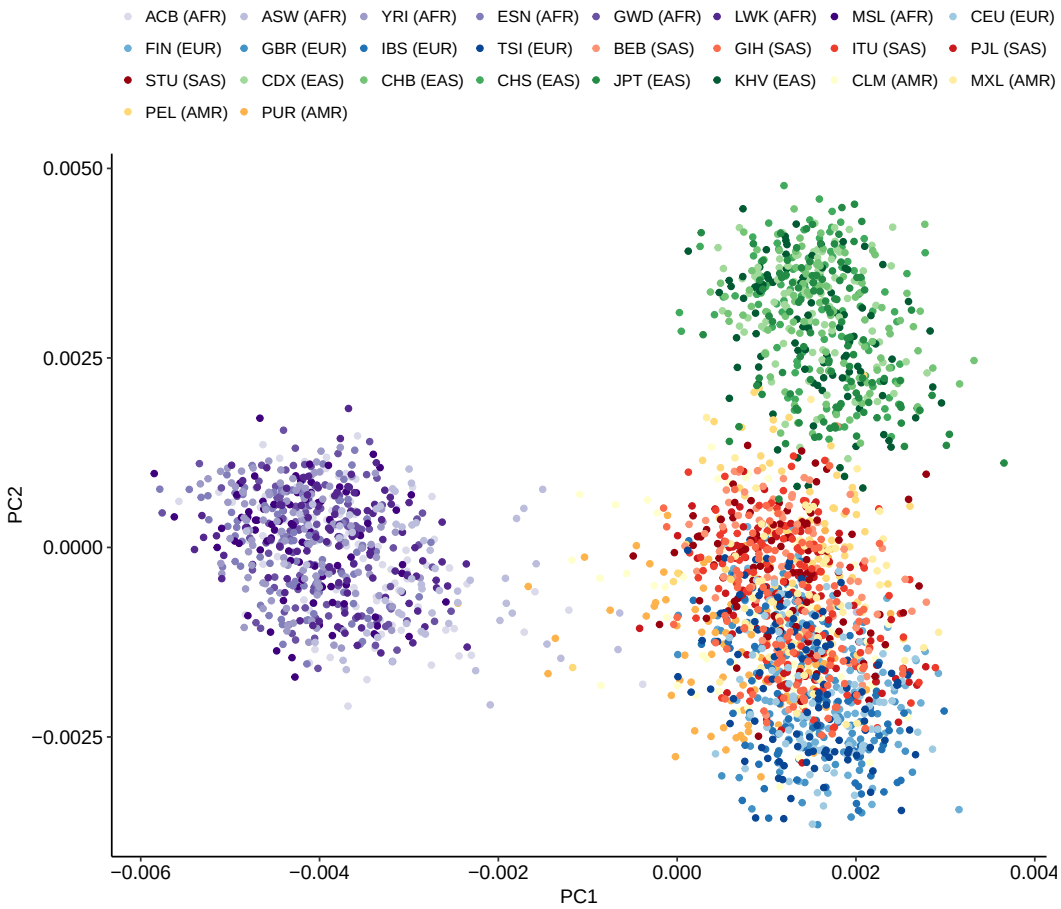
S4 Fig. European population MNP (red) and SNP (blue) densities in 1Mb windows across the genome



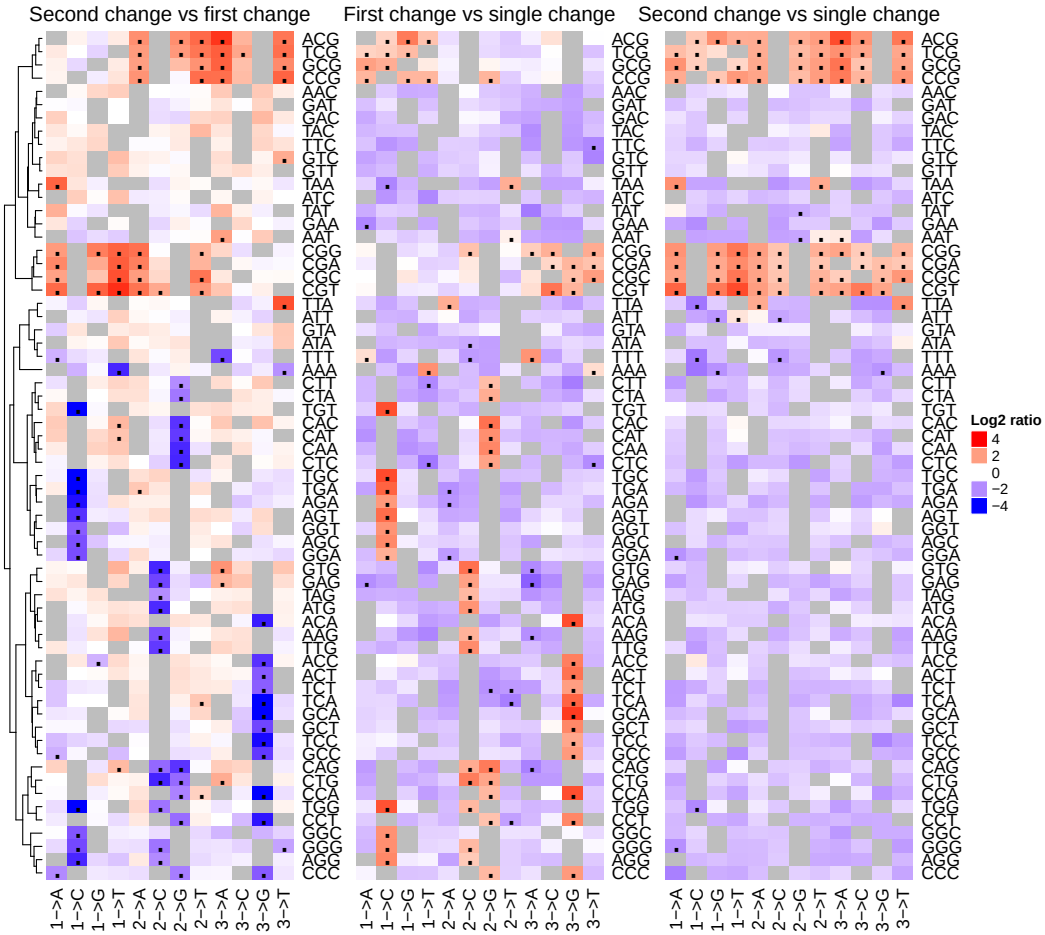
S5 Fig. South Asian population MNP (red) and SNP (blue) densities in 1Mb windows across the genome



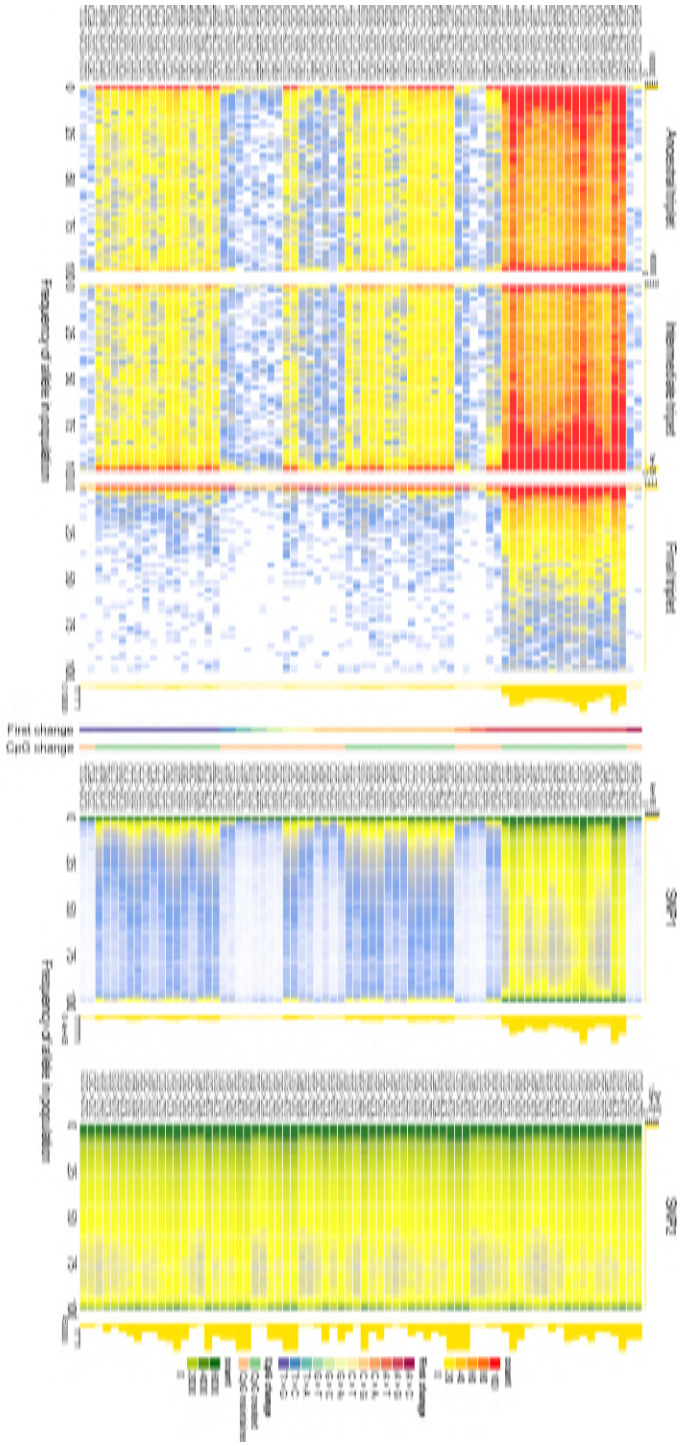
S6 Fig. MNP versus SNP densities in 1Mb windows along the genome. Densities in different population groups are shown and the outlier windows corresponding to the MHC region on chromosome 6 are indicated.



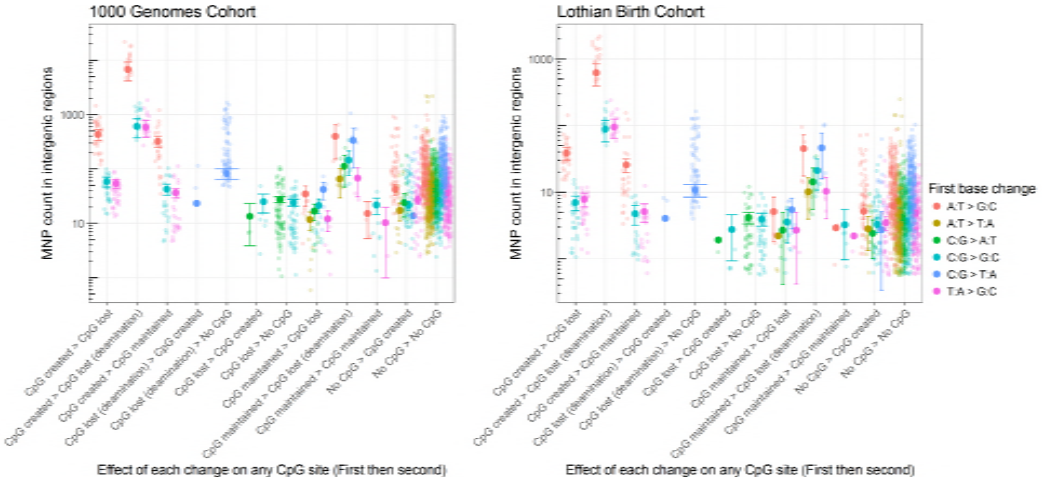
S7 Fig. Principal component analysis of individuals according to the difference in frequencies of first and second changes among MNPs. The frequencies of each second change in an MNP defined by their base change and triplet nucleotide context was subtracted from the corresponding frequency of first changes. The resulting differences were then used as input to the PCA. The continental groups still separate in this analysis suggesting the first and second changes display different biases between populations.



S8 Fig. Same as Fig 4 but for the Lothian Birth Cohort genomes.



S9 Fig. Frequencies and counts of a subset of intergenic MNPs and SNPs across the 1000 genomes cohort. The three leftmost heatmaps indicate allele frequencies of the intergenic MNPs where the second change is $CpG \rightarrow TpG$, i.e. the potential deamination of a CpG site. The numbers of intergenic MNPs of the corresponding allele frequency are indicated by the intensity of colour. The derived allele frequencies and numbers of intergenic SNPs matching each constitutive change are indicated in the final two heatmaps. The original base change in the MNP and its impact on any CpG sites is also indicated (the base change being orientated with respect to the $CpG \rightarrow TpG$ change). The yellow histograms represent the total counts of each corresponding intergenic MNP or SNP.



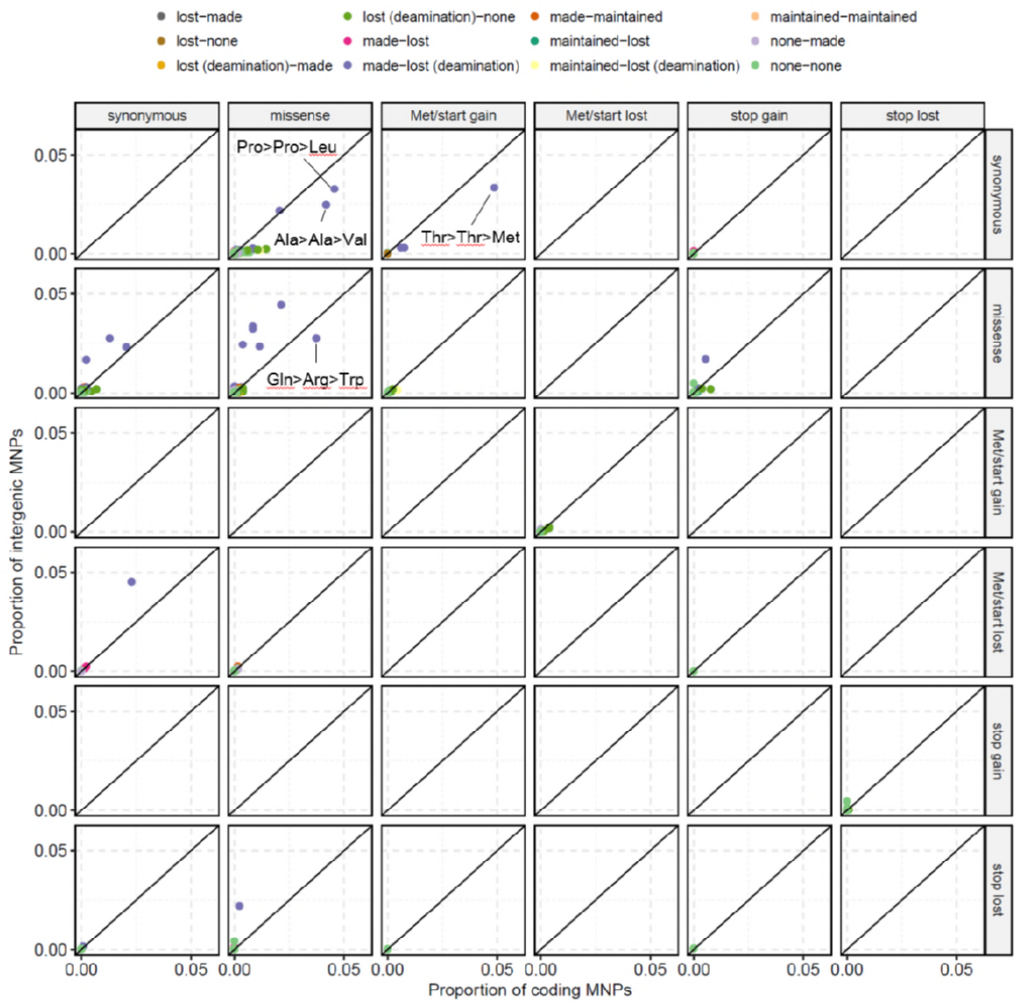
S10 Fig. The importance of CpG dynamics in shaping MNP numbers depends on the form of the original base change. Each dot represents the observed count of a particular form of MNP. Filled circles and error bars represent the expected number of MNPs of the given category having controlled for background rates of change at SNPs. These expected numbers were generated from prediction outputs of Eq 2 when the expected number of MNPs variable was fixed at its median observed value. The interaction term in Eq 2, assessed using ANOVA, was significant ($p < 2.2x10^{-16}$) in both cohorts.

	Dependent variable:	
	first variant in coding MNP	second variant in coding MNP
single coding variant in SNP	**0.001*** **	**0.0004*** **
	(0.0001)	**(0.0001)**
Consequence (reference level: synonymous)		
missense	0.79	1.01**
	(0.58)	(0.43)
Met/start gain	0.33	0.12
	(1.76)	(1.30)
Met/start lost	0.26	0.93
	(1.76)	(1.30)
stop gain	0.96	0.76
	(1.18)	(0.87)
stop lost	1.24	0.37
	(1.19)	(0.88)
CpG change (reference level: no CpG)		
lost	1.10	1.05
	(1.60)	(1.18)
lost (deamination)	6.27**	**18.59*** **
	(2.60)	**(1.92)**
made	**10.69*** **	0.03
	(1.37)	(1.01)
maintained	3.79**	0.87
	(1.47)	(1.09)
Interactions (consequence and CpG change)		
missense:lost	-0.58	-1.37
	(1.95)	(1.44)
stop gain:lost	-0.93	-1.48
	(5.15)	(3.80)
missense:lost (deamination)	-5.18*	**12.76*** **
	(2.88)	**(2.13)**
Met/start gain:lost (deamination)	-0.61	**66.23*** **
	(5.63)	**(4.15)**
stop gain:lost (deamination)	-7.76	-9.61**
	(5.49)	(4.05)
missense:made	**-5.13*** **	0.23
	(1.66)	(1.22)
Met/start lost:made	**25.28*** **	2.68
	(5.27)	(3.89)
stop lost:made	-10.73**	-0.045
	(3.79)	(2.80)
missense:maintained	-2.54	-0.72
	(1.97)	(1.45)
Constant	-1.27**	-0.39
	(0.57)	(0.42)
Observations	576	576
R^2	0.36	0.76
Adjusted R^2	0.34	0.75
Residual Std. Error (df = 556)	4.78	3.53
F Statistic (df = 19; 556)	16.42***	93.52***
Note:		*p<0.1; **p<0.05; ***p<0.0026

S11 Fig. The enrichment and depletion of changes in coding MNPs having accounted for the corresponding frequency of changes observed among SNPs. TThe observed number of each codon change among the first and second change of coding MNPs was fitted against its consequence, effect on CpG sites and observed frequency among SNPs. See Eq 4 and Eq 6 in the methods section for more details.

	Dependent variable:	
	first variant in coding MNP	second variant in coding MNP
first variant in intergenic MNP	**0.002*** **	
	(0.0001)	
second variant in intergenic MNP		**0.002*** **
		(0.0002)
Consequence (reference level: synonymous)		
missense	-0.72	0.27
	(0.48)	(0.39)
Met/start gain	-0.34	-0.20
	(1.52)	(1.22)
Met/start lost	-0.74	0.22
	(1.52)	(1.23)
stop gain	-1.46	-0.39
	(1.00)	(0.80)
stop lost	-1.30	-0.89
	(1.00)	(0.80)
CpG change (reference level: no CpG)		
lost	-0.34	-0.24
	(1.37)	(1.11)
lost (deamination)	**12.89*** **	-6.00*
	(2.11)	(3.23)
made	2.73**	-0.12
	(1.25)	(0.95)
maintained	2.10*	-0.13
	(1.27)	(1.02)
Interactions (consequence and CpG change)		
missense:lost	0.59	-0.84
	(1.69)	(1.36)
stop gain:lost	0.83	-0.69
	(4.45)	(3.59)
missense:lost (deamination)	-5.19**	**12.04*** **
	(2.50)	**(2.01)**
Met/start gain:lost (deamination)	-2.40	**58.35*** **
	(4.87)	**(4.00)**
stop gain:lost (deamination)	-11.79**	-0.89
	(4.73)	(3.96)
missense:made	**-5.77*** **	0.29
	(1.43)	(1.15)
Met/start lost:made	-2.38	1.19
	(4.90)	(3.67)
stop lost:made	**-11.74*** **	0.43
	(3.28)	(2.64)
missense:maintained	-1.35	-0.13
	(1.70)	(1.37)
Constant	0.83	0.11
	(0.42)	(0.34)
Observations	576	576
R^2	0.52	0.79
Adjusted R^2	0.50	0.78
Residual Std. Error (df = 556)	4.14	3.33
F Statistic (df = 19; 556)	31.77***	108.48***
Note:		*p<0.1; **p<0.05; ***p<0.0026

S12 Fig. The enrichment and depletion of changes in coding MNPs having accounted for the corresponding frequency of changes observed among intergenic MNPs. The observed number of each codon change among the first and second change of coding MNPs was fitted against its consequence, effect on CpG sites and observed frequency among intergenic MNPs. See Eq 3 and Eq 5 in the methods section for more details.



S13 Fig. The relative frequencies of coding and intergenic MNPs depends on the consequences of the two individual changes they contain. Row labels indicate the consequence of the MNP's first change (when found in coding regions) and the column labels reflect the consequence of the second change. Point colours indicate the impact on CpG sites of the two changes in the MNP (first then second). The line of equality is shown. Only MNPs where the first change was synonymous show a general enrichment in coding regions. The one outlier below the line of equality labelled in the missense-missense plot corresponds to the $Gln_{CAG} > Arg_{CGG} > Trp_{TGG}$ MNP that creates a new tryptophan codon.

	Dependent variable:
	number of MNPs in coding regions (only MNPs where first change is missense)
number of MNPs in intergenic regions	**0.0002*** **
	(0.00001)
Consequence of first change (reference level: synonymous)	
missense	-0.29**
	(0.10)
Met/start gain	**-1.054*** **
	(0.30)
stop gain	-0.27
	(0.29)
reversion	-0.14
	(451)
Number of codons for final amino acid (reference level: 1	
2	**-1.14*** **
	(0.16)
3	-0.87**
	(0.27)
4	**-0.96*** **
	(0.18)
6	**-1.26*** **
	(0.19)
CpG impact of first change (reference level: no CpG)	
lost	-0.60**
	(0.30)
lost (deamination)	**2.22*** **
	(0.12)
made	0.12
	(0.26)
maintained	-0.39
	(0.35)
CpG impact of second change (reference level: no CpG)	
lost	-0.25
	(0.32)
lost (deamination)	**1.84*** **
	(0.28)
made	-0.46*
	(0.25)
Constant	-0.22
	(0.20)
Observations	1628
Log Likelihood	-1104.33
Akaike Inf. Crit.	2,242.67
Note:	*p<0.1; **p<0.05; ***p<0.0033

S14 Fig. Testing for epistatic selection across neighbouring, coding nucleotide changes. Restricting the analysis to MNPs where the first change leads to a missense change in coding regions we tested whether the frequency of these MNPs was related to the consequence of the second change. We controlled for the number of codons that encode the final amino acid, the impact of the first and second changes in the MNP on CpG sites and the corresponding frequency of the same MNPs in intergenic regions in this analysis. See Eq 7 in the methods section for more details.

Acknowledgments

JGDP is supported by the Biotechnology and Biological Sciences Research Council (BBSRC, Grant No. BBS/E/D/10002071) and the whole genome sequencing of the Lothian Birth Cohorts was funded through an institutional award to the Roslin Institute from the BBSRC. The Lothian Birth Cohorts are supported by Age UK (Disconnected Mind programme). IJD is supported by the Centre for Cognitive Ageing and Cognitive Epidemiology, which is funded by the Medical Research Council and the Biotechnology and Biological Sciences Research Council (Grant No. MR/K026992/1). We thank the Lothian Birth Cohorts' participants and research team for their help.

References

1. Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. Human Molecular Genetics. 2000;9(16):2403–2408. doi:10.1093/hmg/9.16.2403.

2. Brookes AJ. The essence of SNPs. Gene. 1999;234(2):177–186. doi:10.1016/S0378-1119(99)00219-X.

3. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Computational Biology. 2012;8(12). doi:10.1371/journal.pcbi.1002822.

4. Morin PA, Luikart G, Wayne RK, the SNP workshop group. SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution. 2004;19(4):208–216. doi:10.1016/j.tree.2004.01.009.

5. Schrider DR, Hourmozdi JN, Hahn MW. Pervasive Multinucleotide Mutational Events in Eukaryotes. Current biology : CB. 2011;21(12):1051–1054. doi:10.1016/j.cub.2011.05.013.

6. Amos W. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? Proceedings of the Royal Society B: Biological Sciences. 2010;277(1686):1443–1449. doi:10.1098/rspb.2009.1757.

7. Hodgkinson A, Eyre-Walker A. Human triallelic sites: evidence for a new mutational mechanism? Genetics. 2010;184(1):233–241. doi:10.1534/genetics.109.110510.

8. Bazykin GA, Kondrashov FA, Ogurtsov AY, Sunyaev S, Kondrashov AS. Positive selection at sites of multiple amino acid replacements since rat–mouse divergence. Nature. 2004;429(6991):558–562. doi:10.1038/nature02601.

9. Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, et al. Multi-nucleotide de novo Mutations in Humans. PLoS genetics. 2016;12(11):e1006315. doi:10.1371/journal.pgen.1006315.

10. Harris K, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Research. 2014;24(9):1445–1454. doi:10.1101/gr.170696.113.

11. Davis BH, Poon AFY, Whitlock MC. Compensatory mutations are repeatable and clustered within proteins. Proceedings of the Royal Society B: Biological Sciences. 2009;276(1663):1823–1827. doi:10.1098/rspb.2008.1846.

12. Lappalainen T, Montgomery S, Nica A, Dermitzakis E. Epistatic Selection between Coding and Regulatory Variation in Human Evolution and Disease. The American Journal of Human Genetics. 2011;89(3):459–463. doi:10.1016/j.ajhg.2011.08.004.

13. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution. Nature. 2012;490(7421):535–538. doi:10.1038/nature11510.

14. Rosenfeld JA, Malhotra AK, Lencz T. Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. Nucleic Acids Research. 2010;38(18):6102–6111. doi:10.1093/nar/gkq408.

15. Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. eLife. 2017;6:e24284. doi:10.7554/eLife.24284.

16. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. doi:10.1038/nature15393.

17. Bernatchez L, Landry C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? Journal of Evolutionary Biology. 2003;16(3):363–377.

18. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. International Journal of Epidemiology. 2012;41(6):1576–1584. doi:10.1093/ije/dyr197.

19. Taylor AM, Pattie A, Deary IJ. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. International Journal of Epidemiology. 2018;doi:10.1093/ije/dyy022.

20. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father/'s age to disease risk. Nature. 2012;488(7412):471–475. doi:10.1038/nature11396.

21. Koonin EV, Novozhilov AS. Origin and evolution of the genetic code: the universal enigma. Iubmb Life. 2009;61(2):99–111. doi:10.1002/iub.146.

22. Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nature Genetics. 2015;47(2):126–131. doi:10.1038/ng.3186.

23. Wei WH, Hemani G, Haley CS. Detecting epistasis in human complex traits. Nat Rev Genet. 2014;advance online publication.

24. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779.

25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

26. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–498. doi:10.1038/ng.806.

27. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics (Oxford, England). 2014;30(7):1006–1007. doi:10.1093/bioinformatics/btt730.

28. Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Research. 2008;18(11):1829–1843. doi:10.1101/gr.076521.108.

29. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016;17:122. doi:10.1186/s13059-016-0974-4.