Sequence effects on patterns of variation and DNA strand asymmetries observed from whole-genome sequenced UK Biobank participants

David Curtis

UCL Genetics Institute, UCL, Darwin Building, Gower Street, London WC1E 6BT.

Correspondence:

David Curtis *d.curtis@ucl.ac.uk*

07973 906143

Abstract

UK Biobank has released whole genome sequence data for 500,000 participants, including allele counts for hundreds of millions of variants and these were considered in the context of the pentanucleotide background on which they occurred. Frequencies of singleton variants were obtained and compared with frequencies of more common variants. Results were highly correlated across chromosomes, reflecting systematic effects. C>T singleton variants were less frequent in the CG context but the opposite was true for more common variants, suggesting that they are relatively well tolerated and not subject to strong negative selection. The frequencies of singleton variant types were strongly influenced by their trinucleotide context and the total counts of variants in their trinucleotide context could be well approximated by combining five mutational signatures obtained from genomes of cancer cells. For some variant types, there were marked asymmetries in counts between plus and minus DNA strands. The patterns of these asymmetries for singleton variants differed between chromosomes, with five being negatively correlated with the rest. These asymmetries did not appear related to strand-specific gene content. It was noted that there were also strand asymmetries for some pentanucleotide sequences in the reference genome and that these were consistent across chromosomes. The sequence TTCGT is seen 673300 times on the plus strand but only 465807 times on the minus strand. These findings must reflect strand-specific mechanisms affecting mutation and selection which are not currently well understood and which could be investigated further.

This research has been conducted using the UK Biobank Resource.

Introduction

Ultimately, all genome sequences are derived from processes which involve mutation, in which there are imperfections in the molecular processes responsible for copying the genetic code from one cell to another. Investigating patterns of mutation may yield insights into these underlying mechanisms. Additionally, understanding which types of mutation are more or less likely to persist in the population may throw light on the extent to which they are harmful and hence poorly tolerated, which could assist in predicting the pathogenicity of variants (Curtis, 2025). In cancers, mutations occur with increasing frequency as the disease progresses and a fruitful area of research has been to study the mutational signatures seen in different types of cancer (Alexandrov et al., 2020). This involves enumerating the counts for each mutation, for example a single base substitution, specific to the background upon which that mutation occurs, for example the two flanking bases. The mutations observed in a sample can be considered to be produced by number of different drivers of mutation involved in DNA damage, repair and/or replication, possibly involving different endogenous and exogenous exposures, and each driver will produce a characteristic mutational signature. Methods have been devised to disaggregate an observed set of mutations into its component signatures (Díaz-Gay et al., 2023).

Although cancers provide a useful source of data involving large numbers of somatic mutations, information can also be derived from other sources. If parents and children are both sequenced then de novo germline mutations can be observed (Jónsson et al., 2017; McRae et al., 2017; Rahbari et al., 2016). Whole genome sequenced population samples such as 1000 Genomes can be used to reconstruct genealogies and these can then be used to make inferences about mutational processes and their association with different population features (Gao et al., 2023). Sequencing sperm allows for the investigation of mutation within an individual and can demonstrate the relationship between mutation rate and paternal age (Shoag et al., 2025). An alternative approach is to study polymorphisms observed in a large sample of individuals and in this context it has been reported, for example, that C>T substitutions are seen more often in the context of CpG dinucleotides, an observation which has been explained as being due to the spontaneous deamination of 5-methylcytosine (Aggarwala and Voight, 2016; Zhao and Boerwinkle, 2002). Effects on substitution rates were also observed for more distant bases, with the heptanucleotide context explaining >81% of variability in substitution probabilities (Aggarwala and Voight, 2016). Frequencies of all polymorphisms observed in a population reflect not only ongoing mutational processes but also historic processes, including selection. Instead, attention may be confined to variants occurring only as singletons, as these are deemed to be likely to reflect patterns of recent mutation. In a sample of

3560 whole-genome sequences, 36 million singleton variants could be used to estimate relative mutation rates for different mutation types, which were shown to be influenced by nucleotide context and other genomic features (Carlson et al., 2018).

In order to provide some background to the present study, it is relevant to draw attention to two previously reported aspects of the human genome. The first relates to the fact that chromosome 19 is recognised to have unusual sequence characteristics in humans, as well as other primates (Harris et al., 2020). It has the highest GC content of any chromosomes and the largest number of CpG dinucleotides and CpG islands. The second issue of relevance is that substitutional strand asymmetries occur in transcribed regions of the human genome, such that C>T, A>G and G>T mutations occur more frequently on the coding strand, and that there are also compositional strand asymmetries in intergenic regions which are thought to arise from mechanisms related to replication (Mugal et al., 2009; Touchon et al., 2005). Comparing the human genome to other species in order to obtain mutation rates demonstrated that there were mutational strand asymmetries in the region of replication initiation zones (Chen et al., 2011). Analysing whole-genome sequences of 590 tumours revealed asymmetries across mutagenic processes with both transcriptional and replicative asymmetries (Haradhvala et al., 2016).

Whole genome sequence data for 500,000 UK Biobank participants has now been released, including hundreds of millions of variants (The UK Biobank Whole-Genome Sequencing Consortium., 2025). This was used to examine the distribution of singleton variants and the contexts in which they are seen in order to gain further insights into patterns of mutation occurring naturally in humans. It was also used to compare frequencies of singleton variants against those with higher allele counts, since differences might reflect effects of selection. Finally, it was noted that strand asymmetries were present and efforts were made to characterise the sequence patterns related to these asymmetries for variants and in the reference sequence.

Results

Modelling effects of background on variant frequencies

The first column of Table 1 shows the frequencies of singleton variants observed on the background of a C or T nucleotide. (The frequencies of the complementary substitutions were very similar and the issue of strand asymmetry will be addressed below.) It can be seen that the most frequent singleton variant is C>T, with frequency 0.107, and the rarest is T>G, with frequency 0.036. Singleton variant frequencies for backgrounds including different numbers of flanking nucleotides were also calculated. For each partial background sequence, these frequencies could be used to predict the

expected observed frequencies for each singleton variant in the context of the full pentanucleotide background. If just the core nucleotide was used then the predicted frequencies correlated with the observed frequencies with R = 0.71. If a single nucleotide upstream was included the correlation increased to R = 0.81 and including the second upstream nucleotide yielded R = 0.84, while including the two flanking nucleotides yielded R = 0.92. Finally, if two upstream and one downstream nucleotide were included the correlation was R = 0.96. These results suggest that the main driver for the frequency of each variant is the probability of the central base substitution itself but that this probability is influenced to some extent by the neighbouring nucleotides.

Logistic regression modelling was used to better understand which features of the two flanking nucleotides had most effect on the frequency. Figure 1A shows the frequencies of all singleton variants occurring on trinucleotide sequences with either C or T as the central base. Figure 1B shows the odds ratio (OR) for each trinucleotide sequence to produce the target base relative to the odds for just the central base. For example, it can be seen that the largest OR, of 2.14, is for A[T>A]C, meaning that the singleton T>A variants are more than twice as likely to occur in the context of the ATC trinucleotide. It can also be seen that four variants have OR of around 0.5 or less and these are the four X[C>T]G variants, meaning that a singleton C>T variant is less likely to occur in the CG context.

To evaluate the effects of the full pentanucleotide background compared to the trinucleotide background, the modelling procedure was repeated to obtain the OR for each pentanucleotide sequence to produce the target base relative to the odds of the central trinucleotide sequence. Table 2 shows the variants with the ten lowest and highest ORs. This shows that broadly speaking a T or G two bases upstream combined with a T or C two bases downstream reduces the frequency of C[C>G]A variants and that for some backgrounds a T or G two bases upstream reduces the frequency of C[C>A]A variants. The frequency is reduced by about half. Conversely, a C two bases upstream can increase the frequency of C[C>A]C and C[C>A]T variants and a C or G two bases upstream combined with a G two bases downstream increases the frequency of A[T>G]A variants. The frequency is approximately doubled. Results for all pentanucleotide backgrounds are presented in Supplementary Table 1.

For the mutation signature analysis, the counts of standard 96 mutation types, consisting of all trinucleotide singleton variants in which the core base was either C or T, were considered with the counts from the plus and minus strands being combined. As described in the Methods section, 86 reference signatures obtained from the COSMIC database were entered into a multiple linear regression analysis to obtain coefficients for a linear combination of signatures which would provide

the best fit to the observed counts of singleton variants.(Tate et al., 2019) Coefficients were constrained to be positive and the intercept was constrained to zero. When this was done there were 5 signatures for which the ratio of the coefficient to its standard error exceeded 7 and these signatures were retained and the modelling repeated. These five signatures along with their best fitting relative contributions are shown in Table 3. The observed counts of mutation types and the predicted counts which would be obtained from combining these signatures according to the best-fitting weights have correlation coefficient R = 0.82. These observed and predicted counts are shown in Figure 2.

Modelling effects of selection pressures on variant frequencies

The overall ratio of SNP to singleton variants is 0.806 and the second column of Table 1 shows the OR for each core single nucleotide variant relative to this ratio. It can be seen that C>T variants are relatively likely to be retained, with OR = 1.321, while C>A variants are relatively likely to be lost, with OR = 0.781. Similarly to the procedure used for singleton variant frequencies, the logarithm of the OR for the ratio of SNP to singleton variants was calculated for backgrounds including different numbers of flanking nucleotides. For each partial background sequence, these log(OR)s could be used to predict the expected log(OR)s for the variant in the context of the full pentanucleotide background. If just the core SBS was used then the predicted log(OR)s correlated with the observed log(OR)s with R = 0.46. If a single nucleotide upstream was included the correlation increased to R = 0.72 and including the second upstream nucleotide yielded R = 0.73, while including the two flanking nucleotides yielded R = 0.94. Finally, if two upstream and one downstream nucleotide were included the correlation was R = 0.96. Thus, including the two flanking nucleotides strongly predicted the ratio of SNP to singleton variants.

Detailed examination of the results revealed that these effects were largely driven by the fact that C[G>A] variants had OR = 18.6 and that all the trinucleotide variants including this pattern, consisting of C[G>A]X and the complementary X[C>T]G variants, had OR exceeding 14. Thus, C>T variants are very much more likely to be observed as SNPs than singletons if they occur on a CG background. That said, the additional flanking SNP did also have some impact, in that the OR for T[C>T]G was 14.9 whereas that for A[C>T]G was 28.6.

The ORs for the ratio of SNPs to singleton variants for all trinucleotide variants with C or T as the central base and excluding those with the X[C>T]G motif are shown in Figure 3. Figure 3A shows the ORs for all trinucleotide variants and Figure 3B shows the ORs for the full trinucleotide variants relative to the ORs of the single nucleotide variants. It can be seen that the effects are fairly modest, with all ORs in Figure 3A falling within the range 0.5 – 1.3.

The three variants with the lowest ORs are C[C>A]A (OR = 0.540), C[C>A]T (OR = 0.586) and C[C>A]C (OR = 0.647) and the C[C>A]G variant has OR = 0.727, suggesting that C[C>A]X variants are less likely to be maintained within the population. The highest value is for A[T>C]A, with OR = 1.23, suggesting that it may be relatively well tolerated, although this value is still much smaller than the ORs for the X[C>T]G variants.

The modifying effects due to the flanking bases relative to the central SBS, excluding those with the X[C>T]G motif, are shown in Figure 3B. The ORs all fall within the range 0.6 – 1.5. The lowest values were for C[C>T]A (OR = 0.606), T[C>T]T (OR = 0.646), T[C>T]A (OR = 0.648) and T[C>T]C (OR = 0.668), suggesting that these flanking bases all act to reduce the probability of a C>T variant being retained. The two highest values were for A[C>A]G (OR = 1.473) and G[C>A]G (OR = 1.466), suggesting that these motifs render the C>A variant more likely to be retained, though again with a much smaller effect than for the X[C>T]G motifs on C>T variants.

When considering each variant in the context of the full pentanucleotide background, there was considerable variation in the ORs for the ratio of SNPs to singleton variants, with the lowest value being 0.237 for TA[G>T]GG and the highest being 52.7 for TA[C>T]GC. To evaluate the effects of the full pentanucleotide background compared to the trinucleotide background on the ratio of SNP to singleton variants, the modelling procedure was repeated to obtain the OR for each pentanucleotide sequence relative to the central trinucleotide sequence. Table 4 shows the variants with the ten lowest and highest ORs. This shows that a C or G two bases upstream and downstream can reduce the probability of some C>T and C>A variants being retained by about half. Conversely, a number of full pentanucleotide sequences are associated with ORs over 2 and the OR is over 3 for CG[T>A]TG (OR = 3.771) and for three different C>A variants. The results for all pentanucleotide versus trinucleotide effects on the ratios of SNP to singleton counts are presented in Supplementary Table 2.

Modelling strand bias of variant frequencies

Strand bias in mutational processes was studied by taking the log of the ratio of the singleton frequencies on the plus and minus strand for each variant having C or T as the central base. These logs were obtained separately for each chromosome and pairwise correlations were performed across all chromosomes. A heatmap illustrating these correlations is shown in Figure 4, with positive correlations in black and negative correlations in white. It can be seen that the patterns of strand asymmetry are positively correlated for most chromosomes but there are five chromosomes for which this is not the case and the correlation is absent or negative. This second group of

chromosomes tend to be positively correlated with each other. Visually, this means that most of the grid is filled with black circles but there are five rows and columns with white circles, though with black circles where these rows and columns intercept. The five chromosomes forming the second group consist of chromosomes 10, 14, 19, 21 and 22. Chromosomes 5 and 17 show only weak correlations, though are somewhat positively correlated with this second group.

The asymmetry in gene burden for each chromosome was characterised as the logarithm of the ratio of gene counts on the plus versus minus strand and the logarithm of the ratio of total transcript lengths (taking the first listed transcript for each gene). Principal components analysis was carried out for the logarithms of frequencies of each singleton variant on the plus versus minus strand for each chromosome. The first principal component accounted for 0.450 of the total variance and the second principal component accounted for 0.114. There were no significant correlations across chromosomes between either of the measures of gene burden asymmetry and the either of the first two principal components for variant frequency asymmetry. Thus, the discrepancy of strand asymmetries for singleton variant frequencies between chromosomes is not obviously driven by differences in strand asymmetries for gene content.

Chromosomes 1 and 10 were used for separately for more detailed analyses of strand asymmetries because they were the largest chromosomes in each of the two discrepant groups. The third and fourth columns of Table 1 show the ratios for the core SBS variants for chromosomes 1 and 10. It can be seen that the ratios are all very close to 1 and for some the difference from 1 is not statistically significant. However for some variants considered in the context of the full pentanucleotide background there were marked strand asymmetries. The ratios of frequencies of singleton variants on the plus versus minus strand varied on chromosome 1 from 0.647 for GT[T>A]TC to 2.009 for CT[T>A]CG and on chromosome 10 from 0.553 for CG[T>G]AT to 1.728 for TA[C>T]GA. For each partial background sequence, the observed log ratio of frequencies could be used to predict the log ratio for each variant in the context of the full pentanucleotide background. On chromosome 1, when just the core single nucleotide variant was used then the predicted log ratios correlated with the observed log ratios with R = 0.07. If a single nucleotide upstream was included the correlation increased to R = 0.15 and including the second upstream nucleotide yielded R = 0.23, while including the two flanking nucleotides yielded R = 0.25. Finally, if two upstream and one downstream nucleotide were included the correlation was R = 0.43. These correlations are much smaller than had been obtained for partial sequences used to predict singleton variant frequencies and ratios of SNPs to singleton variants. Similar results were obtained for chromosome 10. These results were interpreted as showing that strand asymmetries in singleton variant frequencies are not very

strongly driven by the partial sequence background but that factors relating to the full pentanucleotide background are important.

Because chromosome 10 is smaller than chromosome 1, the absolute counts of variants were smaller and some differences in frequencies of pentanucleotide variants would not be statistically significant after correction for multiple testing. Therefore, further analyses were carried out only on chromosome 1. Table 5 shows the variants which have the highest and lowest strand singleton frequency ratios. This shows that while there are variants with marked asymmetries, only four have a ratio less than 0.8 and only five have a ratio greater than 1.2. It is difficult to discern clear patterns. It may be noteworthy that six out of the ten with lowest ratios as well as six out of ten with the highest all have the T>A substitution, as if this is particularly prone to strand bias effects in one direction or the other. One can also observe that all three CT[T>X]CG variants are seen among the ten with highest ratios, suggesting that the CTTCG background is more likely to result in a mutation on the plus than on the minus strand.

Strand bias for selection effects

Strand bias for selection effects was quantified using the OR for the ratio of SNP to singleton counts on the plus versus minus strand. The variants with the ten lowest and ten highest ORs are shown in Table 6. It can be seen that TC[C>A]AC, TC[C>A]CG and TC[C>A]AA are in the lowest four while TC[C>G]AC and TC[C>G]AA are in the lowest ten, indicating that when these variants occur they are less likely to be retained if they happen on the plus strand than the minus strand, by a factor of around one third. Conversely, AA[T>G]GC and AA[T>A]GC are both among the highest ten, indicating that they are more likely to be retained on the plus strand, by a factor of about three.

Background sequence strand asymmetry

Asymmetries for pentanucleotide sequences in the reference genome were quantified as being the ratio of counts for the sequence with C or T as the central base on the plus compared to minus strand. Table 7 shows the sequences with the lowest and highest ratios. It can be seen that the lowest ten include both AACGA and AACGC as well as TCCAA and TCCAC. Likewise, the ten sequences with highest ratios include TTCGT and TGCGT as well as CTTCG and TTTCG. The ratios most different from 1 have narrow confidence intervals and the results could not have arisen by chance. Presumably some features of these sequences influence the probability for them to eventually be included in the reference sequence differentially on the plus and minus strands.

Discussion

These results seem to demonstrate intriguing relationships between brief DNA sequences and frequencies of singleton and more common variants as well as strand asymmetries for both variants and the reference sequence. There are two potential sources of error which could mean that that these results should not immediately be accepted at face value.

The first potential issue to is that some results might be influenced by features of the technology used to call the variants. If there were genotyping errors then these would be more likely to be relevant to singleton variant calls rather than more common variants or aspects of the reference sequence. One genotyping error might result in calling a singleton variant which was not in fact present. Alternatively, aspects of the calling algorithm might reduce the probability of calling a variant seen only in a single participant rather than one which was more widely present. If any such errors were also affected by the background context in which the call was made or not then this might account for some of results reported for singleton variants, both in terms of their relative probabilities of occurring and in terms of the ratio of their frequencies to more common variants.

A second issue worth commenting on is that it is expected that rare variants present only in groups with minority ancestries are more likely to appear as singleton variants than rare variants in the majority, white European ancestry group. Thus, among the observed singleton variants in the whole sample, it is expected that a larger proportion will consist of those that reflect recent mutational events among white Europeans than groups with other ancestries. Conversely, a larger proportion will reflect rare polymorphisms present in the population in minority ancestries than in white Europeans. However this phenomenon would only bias the results obtained if there were also ancestry-related differences in probabilities for mutations to occur or be retained conditional on DNA sequence and this seems to be *a priori* unlikely.

If we take the results at face value then they demonstrate some effects for which there are no immediately obvious explanations and which would warrant further investigation. One avenue of investigation would be to explore how the observed effects relate to other genomic features, such as proximity to genes, presence within promoter regions or relation to centromeres and telomeres. Another would be to carry out experimental work to better understand the molecular basis of DNA replication in order to characterise aspects which might explain differential probabilities of mutation related to different background sequences.

The observations relating to singleton variants may reflect sequence-based effects on mutation probabilities. Studying the rates of different mutation types in single spermatozoa provides information mainly concerning errors occurring during meiosis (Shoag et al., 2025). Studying cancer

cells provides information about somatic mutations which have arisen either during mitosis or at some other point in the cell cycle but which have occurred in the context of disease, when the normal mechanisms for maintaining and accurately copying DNA have been disrupted (Alexandrov et al., 2020). By contrast, singleton variants observed in a large population sample are expected to reflect mutational events which have occurred fairly recently in essentially healthy cells. They may have arisen during meiosis or as somatic mutations in primordial germ cells. Thus, studying the relative frequency of different types of singleton variant may throw light on the molecular mechanisms used by healthy cells to prevention mutations from occurring and this may lead to better insights into susceptibility to neoplastic disease. Indeed, it is possible that an individual's risk of developing cancer might be reflected in the number and pattern of recent mutations which they carry and this will be the subject of further investigations.

Perhaps the most striking finding is that, among variants with allele count higher than two, C>T substitutions are far more frequently seen in the CG context than other variants or C>T substitutions not in the CG context. Conversely, among singleton variants C>T substitutions are somewhat less frequent in the CG context than in other contexts. Taken at face value, this would suggest that cellular mechanisms have evolved to protect C>T substitutions from occurring in the CG context. However, if any do occur then an unusually high proportion are not under strong negative selection and end up being retained in the population. One avenue to explore would be the extent to which this process is more or less likely to occur for variants within CpG islands involved in the regulation of transcription.

Another finding of note is that there appears to be strand asymmetry for both mutational processes, as evidenced by singleton variants, and processes related to selection, as evidenced by SNP to singleton ratios and to asymmetries in the reference sequence itself. The drivers of asymmetry for mutational processes appear to operate differentially on different chromosomes, with five chromosomes being negatively correlated with the others. It is not obvious why this should be the case but it does not appear to relate directly to strand specific gene burden, as measured by either gene count or total transcript length. It may be that a more detailed analysis of the genomic distribution of these asymmetries might provide additional clues as to possible drivers of this process. Perhaps even more challenging is the observation that there appears to be strand asymmetry for selection processes, in terms of which variants are retained as SNPs and which sequences are seen in the reference genome. In contrast to the mutational asymmetries, these drivers seem to be consistent across chromosomes. It is difficult to think what could account for

these observations but one potentially fruitful approach would be to compare reference sequences across species in order to assess the extent to which asymmetries are shared.

To conclude, this large dataset provides opportunities to investigate associations between short DNA sequences and variant frequencies, reflecting how these sequences may impact probabilities of mutations occurring and of subsequently being retained. The molecular mechanisms underlying these associations are not yet elucidated but the fact that these associations exist should act to encourage further research in this area.

Materials and Methods

The UK Biobank Research Analysis Platform was used to access the whole genome sequence data which had been obtained using the methods previously described (The UK Biobank Whole-Genome Sequencing Consortium., 2025). UK Biobank had obtained ethics approval from the North West Multi-centre Research Ethics Committee which covers the UK (approval number: 11/NW/0382) and had obtained written informed consent from all participants. The UK Biobank approved an application for use of the data (ID 51119) and ethics approval for the analyses was obtained from the UCL Research Ethics Committee (11527/001). For each chromosome, the *pvar* file was downloaded for the DRAGEN population WGS variants in PLINK format. The *pvar* file contains the coordinate, DNA change and allele count for each variant but no information about individual participants. Attention was restricted to single base substitution (SBS) variants.

In the interests of clarity, accounts of variants and their context will be described in terms of the reference strand, also referred to as the plus strand or Watson strand. Each SBS was considered in the context of the two bases 5' and two bases 3', so that the fundamental unit of analysis consisted of a pentanucleotide sequence with the variant consisting of a change from reference to alternate for the central base. A variant type is then defined by six bases, the pentanucleotide background along with the alternate central base, for example AA[G>C]TG. The allele count (AC) field for each *pvar* file was used to categorise the variants into singletons, doubletons and others. Variants with allele counts higher than 2 will here be referred to as SNPs, although many were still very rare, and it is assumed that a larger proportion of SNPs will represent standing variation within the population rather than resulting from recent mutations. There were a total of 510 million singletons, 240 million doubletons and 411 million SNPs. For each six-base variant type and for each chromosome the number of each type of variant was counted. In order to investigate mutation rates conditional on the frequency of the pentanucleotide backgrounds, a sliding window of five bases was moved one base at a time across the GRCh38 reference genome and counts for each pentanucleotide

background sequence were obtained for each chromosome, meaning that in total 2.9 billion pentanucleotide backgrounds were considered.

For some analyses, these background counts were then used to obtain frequencies for the variants, with the frequency defined as the count for each variant divided by the count of the background sequence on which it occurred. For example, the sequence TTTTT occurs 1,526,553 times in the chromosome 1 reference while the sequence CGACG occurs only 6,228 times. Considering singleton variants producing a C as the ALT allele from these backgrounds, TT[T>C]TT occurs 88,620 times while CG[A>C]CG occurs 225 times, so the unadjusted ratio for these variant counts is around 400. However if we divide the variant counts by the background counts then for TT[T>C]TT this yields a frequency of 88,620/1,526,553 = 0.0580 and for CG[A>C]CG a frequency of 225/6,228 = 0.0361, so the ratio of frequencies is 1.6. This ratio is expected to better reflect the relative probabilities for the mutational events producing these variants conditional on the background sequence. Using the variant frequencies provides a way to investigate active mutational processes independent of pre-existing differences between counts of background sequences.

Correlation analyses of pentanucleotide background counts were carried out between all pairs of chromosomes. This revealed that in general the correlations were very high, with R values greater than 0.99 for most chromosome pairs and with an average value of over 0.93 for each chromosome compared with all other chromosomes. However for chromosome 22 the correlations were slightly lower with R values mostly between 0.8 and 0.9 and an average of 0.88 and for chromosome 19 the correlations were lower still (although still high in absolute terms), with an average R of 0.84. To investigate this, the frequencies of background counts between chromosome 19 and the other chromosomes combined were compared in order to see which background sequences were most discrepant. When this was done it was clear that, as previously noted, sequences with high GC content were relatively frequent on chromosome 19 and this phenomenon appeared to be responsible for the somewhat reduced correlation of background counts with other chromosomes.

Correlations of variant counts and variant frequencies adjusted for background were carried out between chromosome pairs. In general these were high and in particular the correlations for frequencies of singleton variants were uniformly high, averaging 0.97 or above for each chromosome, including chromosome 19. This suggested that there were no chromosome-specific differences in mutational processes conditional on background and meant that it seemed legitimate to consider the combined frequencies across all chromosomes for subsequent analyses.

Modelling effects of background sequence on variant frequencies

The general approach to analysing the variants consisted of studying the frequency of observing a certain alternate allele conditional on the reference sequence background. For example, if the T nucleotide occurs 7.78e8 times in reference genome and we observe 6.38e7 singleton T>C variants then we say that the frequency of singleton T>C variants occurring on the background of the T nucleotide is 6.38e7/7.78e8 = 0.0876. Likewise, if the trinucleotide sequence ATG occurs 4.55e7 times in the reference and there are 7.07e6 singleton A[T>C]G variants then the frequency of singleton A[T>C]G variants on the ATG background is 4.55e6 / 7.07e7 = 0.155. Finally, if the pentanucleotide CATGC sequence occurs 3.89e6 times in the reference and there are 6.12e5 CA[T>C]GC singleton variants then the frequency of CA[T>C]GC singleton variants on the CATGC is 6.12e5 / 3.89e6 = 0.159. Using this definition of variant frequency, we can consider the frequency as being the probability that we will observe a variant with a particular alternate base given a particular reference sequence. Each reference sequence will then have the possibility to yield one of three alternate bases so that the singleton variant frequencies for T>C, T>G and T>A conditional on a T nucleotide in the reference sequence might be 0.0876, 0.0370 and 0.0290 respectively. Conversely, a particular alternate base may have different probabilities of being observed conditional on the background sequence, so that the frequencies of singleton T>C, G>C and A>C variants conditional on a T, G or A reference allele might be 0.0876, 0.0515 and 0.0371 respectively.

Using this approach, it is possible to implement a logistic regression framework whereby the outcome to be predicted is a variant with a specific alternate base, referred to as the target, and this outcome is modelled as being influenced by the background reference sequence. Where background sequences are mutually exclusive, for example when considering all trinucleotides consisting of a central base and the flanking base on each side, one can simply total up counts across matching pentanucleotides to obtain the frequency for each background to produce a singleton variant yielding a particular target base. This process was carried out for single nucleotides, dinucleotides including one upstream base, trinucleotides including two upstream or two flanking bases and tetranucleotides including two upstream and one downstream base.

In order to explore the effect on mutational probabilities of including the two flanking nucleotides additional to the core nucleotide, singleton variant frequencies were obtained for background core nucleotides and then each flanking trinucleotide sequence was included in turn in a logistic regression model to estimate its effect on the odds of producing the target base. Similar analyses were done to measure the effects of models with the full pentanucleotide sequences compared to the flanking trinucleotide sequences.

Previous investigations of mutational signatures in cancer have used the raw counts of variants, rather than adjusting them for the frequency of the background sequence, and so for mutational signature analysis of the singleton variants observed in the UK Biobank sample the raw counts were used (Alexandrov et al., 2020). Each SBS was considered along with the single base on either side, so to carry out this analysis the counts for singleton variants were combined across all variant types sharing the same central trinucleotide sequence and alternate base. These analyses typically assume strand symmetry and assume that the central base of the background is either C or T. In order to check for strand bias, the singleton variant frequencies for 96 trinucleotide motifs based on either the plus or minus strand were calculated separately but in fact they were very highly correlated (R=0.9998) so a single analysis was carried out using the counts for the plus strand where the central base was C or T or else for the complementary sequence (i.e. for the same sequence on the minus strand). Sets of mutation-type specific reference signatures have been developed and deposited in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (Alexandrov et al., 2020; Tate et al., 2019). Code was written to combine these signatures linearly to best fit the observed counts, with coefficients constrained to be positive (by fitting the square roots of the coefficients). Those signatures for which the estimate of beta/SE exceeded 7 were retained in the model which was then refitted, to produce a parsimonious model to predict the observed singleton counts from a small number of reference signatures.

Modelling effects of selection pressures on variant frequencies

The relative frequencies of singleton variants within a large sample may be taken to reflect the relative probabilities of different types of mutation occurring, since many singleton variants will be the result of fairly recent mutations. However different variant types may have different probabilities of persisting in the population and, in particular, variant types which tend to be deleterious may be under negative selection pressure and hence less likely to persist than other variants which have a neutral or even benign effect. This phenomenon can be investigated by comparing the frequencies of singleton variants with the frequencies of variants observed in three or more people, which we refer to as SNPs. For example, on chromosome 1 the ratio of the total numbers of SNPs to singletons is 33,786,519 / 41,338,613 = 0.82 but for the variant CA[G>T]GG the corresponding ratio is 5,916 / 33,237 = 0.18. This suggests that variants of this type may tend to have deleterious effects and hence be relatively unlikely to persist in the population. By contrast, the variant GC[G>A]TA occurs only 369 times as a singleton but 15,067 as an SNP so the corresponding ratio is 15,067 / 369 = 40.83, suggesting that this mutational event rarely happens but that when it does the variant is likely to be retained.

To quantify the extent to which variant types were preferentially retained, the logarithms of the ratio of SNP count to singleton count were calculated across all variants for each chromosome. Correlation analyses were then performed across all pairs of chromosomes. These revealed uniformly high positive correlations, with the average correlations for each chromosome with the others ranging from 0.83 to 0.95. These positive correlations were taken to indicate that whatever mechanisms contributed to the preferential retention of some variant types over others were broadly shared across chromosomes. Therefore, subsequent analyses were based on total counts across all chromosomes.

The approach to analysing the effects of different properties of each variant on its probability of being retained was similar to the methods described above to compare frequencies of the singleton variants, except that the outcome of interest was taken to be the ratio of SNP to singleton counts. The background odds, consisting of the ratio of all SNP to singleton counts, was 4.11e8 / 5.10e8 = 0.806. A logistic regression framework was used to find the OR associated with different features of the variant which would produce the best fit to the observed ratio of counts. As previously, this process was carried out for single nucleotide variants, dinucleotides including one upstream base, trinucleotides including two upstream or two flanking bases and tetranucleotides including two upstream and one downstream base. Additionally, modelling was performed to measure the effects of including two flanking nucleotides relative to only the single core nucleotide and the effects of the full pentanucleotide sequence relative to the central trinucleotide sequence.

Modelling effects of background on strand bias of variant frequencies

In the course of these analyses it became apparent that there was not complete strand symmetry between complementary bases or complementary sequences. In the reference genome, the C nucleotide occurs on the plus strand 593,752,912 times while the G nucleotide occurs 596,226,039. Although the difference is small relatively, testing for whether this inequality arises purely by chance yields a chi-squared statistic of 2,570, with one degree of freedom, indicating that the difference is non-random and that there must be some systematic drivers for it. There is a similar small but definite difference between the counts for A and T nucleotides, which are 859,262,665 and 861,982,382 respectively. These strand asymmetries become relatively more pronounced when certain longer sequences are considered. For example, the sequence TTCGT occurs 673,300 times on the plus strand but the complementary sequence ACGAA occurs only 465,807 times. Singleton variants occurring on these backgrounds also differ in frequency, with TT[C>T]GT occurring 32,065 times and AC[G>A]AA occurring 25,476, meaning that the complementary variants have frequencies of 0.048 and 0.055 respectively.

Strand asymmetries in singleton variant frequencies may be taken to reflect asymmetries in mutational processes. In order to explore these asymmetries systematically, complementary pairs of variants were considered together, with the index member of the pair being denoted as the one which had either C or T as its central base. The measure of asymmetry was taken to be the logarithm of the ratio of the frequencies of each member of the pair. Equivalently, this can be considered as the ratio of frequencies for the index variant to be observed on the plus or minus strand.

As an initial step, the logarithms of these strand ratios for singleton variant frequencies were obtained for all variants for each chromosome. Correlation analysis was carried out between chromosomes. In contrast to the results previously obtained for the singleton frequencies and the ratios of SNPs to singletons, the logarithms of strand ratios were not uniformly positively correlated between chromosomes. Instead, for the majority of chromosomes the logarithms were positively correlated with each other but there were five chromosomes for which the logarithms were negatively correlated with the first group but positively correlated with each other, this minority group consisting of chromosomes 10, 14, 19, 21 and 22. The magnitude of the correlation coefficients was lower than those which had been obtained for the correlations for singleton frequencies and for SNP to singleton ratios and tended to range between 0.3 and 0.7. Nevertheless, because there are 1536 complementary pairs of variants any value of R with magnitude greater than 0.2 is highly statistically significant. Because of the fact that there were both positive and negative correlations between chromosome pairs exceeding this value, we can conclude that there are systematic factors which influence strand asymmetry in mutational processes but that these factors operate in opposing directions on different chromosomes. These correlation analyses will be presented in more detail in the Results section but the implication was that it would not be appropriate to combine data across chromosomes.

In order to explore whether strand asymmetries in gene content might contribute to differences between chromosomes, for each chromosome the numbers of genes on the plus and minus strand were obtained and the lengths of the first listed transcripts for every gene were totalled separately for the plus and minus strand. As a measure of gene burden asymmetry, these observations were converted to a log ratio of counts and log ratio of total transcript lengths for the plus versus minus strand for each chromosome. Principal components analysis of the log ratios of singleton variant frequencies was carried out. Then correlation analyses across chromosomes were performed between the log ratio of gene counts or log ratio of total transcript lengths and the first and second principal components of the log ratios of singleton variant frequencies.

As it seemed inappropriate to combine data across chromosomes, further analyses of the sequences associated with strand asymmetry of singleton variant frequencies were carried out separately for chromosomes 1 and 10, selected for being the largest chromosomes in each of the two groups. These analyses were similar to those performed to investigate drivers of mutation and selection, in that the effects of different numbers of nucleotides were modelled up to the full pentanucleotide sequence.

Strand bias for selection effects

As described above, the ratio of SNP to singleton variant counts can be taken as an index of the extent to which a mutation is likely to be retained in the population once it has occurred. Strand bias in this process can be quantified for each variant by taking the logarithm of this ratio for the plus strand versus the minus strand. This was done for each chromosome separately and then correlation analyses were performed of these logarithms between chromosomes. Most pairs of chromosomes were positively correlated with each other, with average correlations generally in the range of 0.40 – 0.45. However chromosomes 10, 14, 19, 21 and 22 had generally low or weakly negative correlations with other chromosomes, with R values between -0.14 and -0.01. Since there were no strongly negative correlations, the counts across all chromosomes were totalled and these total counts were used to identify variants for which there was marked asymmetry between strands for the ratio of SNP to singleton counts.

Background sequence strand asymmetry

As noted above, there was strand asymmetry for the background pentanucleotide counts. For each pentanucleotide having C or T as the central base, this was quantified by taking the logarithm of the ratio of the counts on the plus or minus strand for each chromosome. Correlation analysis revealed these logarithms were positively correlated that all pairs of chromosomes, with average R ranging from 0.59 to 0.68. Given that this indicated a consistent pattern of strand bias across all chromosomes, the counts were totalled across chromosomes and these total counts were used to identify pentanucleotide sequences in the reference genome demonstrating marked strand asymmetry.

Data analysis

Code to count variants and background sequences as well as to carry out modelling was written in C++ and is available at: https://github.com/davenomiddlenamecurtis/countVariantsInContext. Principal components analysis was performed using R v4.3.2 (R Core Team, 2014). This repository

also contains raw counts and calculated frequencies of the variants analysed. Because of the large number of observations utilised, most of the results obtained were very highly statistically significant, sometimes with the estimate exceeding its standard error by orders of magnitude. Thus, p values will not be presented unless they are larger than 1.0e-9 and confidence intervals on estimates will not be presented unless they would affect the second significant figure.

Statements

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 51119. This work uses data provided by patients and collected by NHS England as part of their care and support. This research also used data assets made available by National Safe Haven as part of the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (grants MC_PC_20029 and MC_PC_20058). The author wishes to acknowledge the staff supporting the High Performance Computing Cluster, Computer Science Department, University College London. The author wishes to thank the participants who volunteered for the UK Biobank project.

Statements of Ethics

UK Biobank had obtained ethics approval from the North West Multi-centre Research Ethics Committee which covers the UK (approval number: 11/NW/0382) and had obtained written informed consent from all participants. The UK Biobank approved an application for use of the data (ID 51119) and ethics approval for the analyses was obtained from the UCL Research Ethics Committee (11527/001).

Conflict of Interest Statement

The author declares he has no competing interests.

Funding Sources

No external funding was received for this work.

Data Availability Statement

The raw data is available on application to UK Biobank at https://ams.ukbiobank.ac.uk/ams/. Any legitimate researcher can apply for access to the UK Biobank dataset by filling in the form at this URL. The counts of variants and of pentanucleotide sequences in the reference genome are provided at https://github.com/davenomiddlenamecurtis/countVariantsInContext.

Code Availability

Code and scripts used to carry out the analyses are provided at https://github.com/davenomiddlenamecurtis/countVariantsInContext.

References

Aggarwala, V., Voight, B.F. (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat. Genet. 48, 349.

Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., Ashiqul Islam, S.M., Lopez-Bigas, N., Klimczak, L.J., McPherson, J.R., Morganella, S., Sabarinathan, R., Wheeler, D.A., Mustonen, V., Getz, G., Rozen, S.G., Stratton, M.R., Boutros, P., Chan, K., Fujimoto, A., Kazanov, M., Lawrence, M., Martincorena, I., Nakagawa, H., Polak, P., Prokopec, S., Roberts, S.A., Saini, N., Shibata, T., Shiraishi, Y., Teh, B.T., Vázquez-García, I., Yousif, F., Yu, W. (2020) The repertoire of mutational signatures in human cancer. Nature 2020 578:7793 578, 94–101.

Carlson, J., Locke, A.E., Flickinger, M., Zawistowski, M., Levy, S., Myers, R.M., Boehnke, M., Kang, H.M., Scott, L.J., Li, J.Z., Zöllner, S., Absher, D., Akil, H., Breen, G., Burmeister, M., Cohen-Woods, S., Iacono, W.G., Knowles, J.A., Legrand, L., Lu, Q., McGue, M., McInnis, M.G., Pato, C.N., Pato, M.T., Rivera, M., Sobell, J.L., Vincent, J.B., Watson, S.J. (2018) Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nat. Commun. 9.

Chen, C.L., Duquenne, L., Audit, B., Guilbaud, G., Rappailles, A., Baker, A., Huvet, M., D'Aubenton-Carafa, Y., Hyrien, O., Arneodo, A., Thermes, C. (2011) Replication-Associated Mutational Asymmetry in the Human Genome. Mol. Biol. Evol. 28, 2327–2337.

Curtis, D. (2025) Assessment of ability of a DNA language model to predict pathogenicity of rare coding variants. J. Hum. Genet.

Díaz-Gay, M., Vangara, R., Barnes, M., Wang, X., Islam, S.M.A., Vermes, I., Duke, S., Narasimman, N.B., Yang, T., Jiang, Z., Moody, S., Senkin, S., Brennan, P., Stratton, M.R., Alexandrov, L.B. (2023) Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. Bioinformatics 39.

Gao, Z., Zhang, Y., Cramer, N., Przeworski, M., Moorjani, P. (2023) Limited role of generation time changes in driving the evolution of the mutation spectrum in humans. Elife 12.

Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E., Hess, J.M., Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, L.Z., Kamburov, A., Hanawalt, P.C., Wheeler, D.A., Koren, A., Lawrence, M.S., Getz, G. (2016) Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell 164, 538–549.

Harris, R.A., Raveendran, M., Worley, K.C., Rogers, J. (2020) Unusual sequence characteristics of human chromosome 19 are conserved across 11 nonhuman primates. BMC Evolutionary Biology 2020 20:1 20, 33-.

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A., Ward, L.D., Arnadottir, G.A., Helgason, E.A., Helgason, H., Gylfason, A., Jonasdottir, Adalbjorg, Jonasdottir, Aslaug, Rafnar, T., Frigge, M., Stacey, S.N., Th. Magnusson, O., Thorsteinsdottir, U., Masson, G., Kong, A., Halldorsson, B. V., Helgason, A., Gudbjartsson, D.F., Stefansson, K. (2017) Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature 549, 519–522.

McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., Ambridge, K., Barrett, D.M., Bayzetinova, T., Jones, P., Jones, W.D., King, D., Krishnappa, N., Mason, L.E., Singh, T., Tivey, A.R., Ahmed, M., Anjum, U., Archer, H., Armstrong, R., Awada, J., Balasubramanian, M., Banka, S., Baralle, D., Barnicoat, A., Batstone, P., Baty, D., Bennett, C., Berg, J., Bernhard, B., Bevan, A.P., Bitner-Glindzicz, M., Blair, E., Blyth, M., Bohanna, D., Bourdon, L., Bourn, D., Bradley, L., Brady, A., Brent, S., Brewer, C., Brunstrom, K., Bunyan, D.J., Burn, J., Canham, N., Castle, B., Chandler, K., Chatzimichali, E., Cilliers, D., Clarke, A., Clasper, S., Clayton-Smith, J., Clowes, V., Coates, A., Cole, T., Colgiu, I., Collins, A., Collinson, M.N., Connell, F., Cooper, N., Cox, H., Cresswell, L., Cross, G., Crow, Y., D'Alessandro, M., Dabir, T., Davidson, R., Davies, S., De Vries, D., Dean, J., Deshpande, C., Devlin, G., Dixit, A., Dobbie, A., Donaldson, A., Donnai, D., Donnelly, D., Donnelly, C., Douglas, A., Douzgou, S., Duncan, A., Eason, J., Ellard, S., Ellis, I., Elmslie, F., Evans, K., Everest, S., Fendick, T., Fisher, R., Flinter, F., Foulds, N., Fry, A., Fryer, A., Gardiner, C., Gaunt, L., Ghali, N., Gibbons, R., Gill, H., Goodship, J., Goudie, D., Gray, E., Green, A., Greene, P., Greenhalgh, L., Gribble, S., Harrison, R., Harrison, L., Harrison, V., Hawkins, R., He, L., Hellens, S., Henderson, A., Hewitt, S., Hildyard, L., Hobson, E., Holden, S., Holder, M., Holder, S., Hollingsworth, G., Homfray, T., Humphreys, M., Hurst, J., Hutton, B., Ingram, S., Irving, M., Islam, L., Jackson, A., Jarvis, J., Jenkins, L., Johnson, D., Jones, E., Josifova, D., Joss, S., Kaemba, B., Kazembe, S., Kelsell, R., Kerr, B., Kingston, H., Kini, U., Kinning, E., Kirby, G., Kirk, C., Kivuva, E., Kraus, A., Kumar, D., Ajith Kumar, V.K., Lachlan, K., Lam, W., Lampe, A., Langman, C., Lees, M., Lim, D., Longman, C., Lowther, G., Lynch, S.A., Magee, A., Maher, E., Male, A., Mansour, S., Marks, K., Martin, K., Maye, U., McCann, E., McConnell, V., McEntagart, M., McGowan, R., McKay, K., McKee, S., McMullan, D.J., McNerlan, S., McWilliam, C., Mehta, S., Metcalfe, K., Middleton, A., Miedzybrodzka, Z., Miles, E., Mohammed, S., Montgomery, T., Moore, D., Morgan, S., Morton, J., Mugalaasi, H., Murday, V., Murphy, H., Naik, S., Nemeth, A., Nevitt, L., Newbury-Ecob, R., Norman, A., O'Shea, R., Ogilvie, C., Ong, K.R., Park, S.M., Parker, M.J., Patel, C., Paterson, J., Payne, S., Perrett, D., Phipps, J., Pilz, D.T., Pollard, M., Pottinger, C., Poulton, J., Pratt, N., Prescott, K., Price, S., Pridham, A., Procter, A., Purnell, H., Quarrell, O., Ragge, N., Rahbari, R., Randall, J., Rankin, J., Raymond, L., Rice, D., Robert, L., Roberts, E., Roberts, J., Roberts, P., Roberts, G., Ross, A., Rosser, E., Saggar, A., Samant, S., Sampson, J., Sandford, R., Sarkar, A., Schweiger, S., Scott, R., Scurr, I., Selby, A., Seller, A., Sequeira, C., Shannon, N., Sharif, S., Shaw-Smith, C., Shearing, E., Shears, D., Sheridan, E., Simonic, I., Singzon, R., Skitt, Z., Smith, A., Smith, K., Smithson, S., Sneddon, L., Splitt, M., Squires, M., Stewart, F., Stewart, H., Straub, V., Suri, M., Sutton, V., Swaminathan, G.J., Sweeney, E., Tatton-Brown, K., Taylor, C., Taylor, R., Tein, M., Temple, I.K., Thomson, J., Tischkowitz, M., Tomkins, S., Torokwa, A., Treacy, B., Turner, C., Turnpenny, P., Tysoe, C., Vandersteen, A., Varghese, V., Vasudevan, P., Vijayarangakannan, P., Vogt, J., Wakeling, E., Wallwark, S., Waters, J., Weber, A., Wellesley, D., Whiteford, M., Widaa, S., Wilcox, S., Wilkinson, E., Williams, D., Williams, N., Wilson, L., Woods, G., Wragg, C., Wright, M., Yates, L., Yau, M., Nellåker, C., Firth, H. V., Wright, C.F., FitzPatrick, D.R., Barrett, J.C., Hurles, M.E. (2017) Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433–438.

Mugal, C.F., Von Grünberg, H.H., Peifer, M. (2009) Transcription-induced mutational strand bias and its effect on substitution rates in human genes. Mol. Biol. Evol. 26, 131–142.

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Al Turki, S., Dominiczak, A., Morris, A., Porteous, D., Smith, B., Stratton, M.R., UK10K Consortium, Hurles, M.E. (2016) Timing, rates and spectra of human germline mutation. Nat. Genet. 48, 126–33.

Shoag, J.E., Srinivasa, A., Loh, C.A., Liu, M.H., Lassen, E., Melanaphy, S., Costa, B.M., Grońska-Pęski, M., Jabara, N.T., Picciotto, S., Choi, U., Bohorquez, A.D., Barbieri, C.E., Callum, P., Skytte, A.B., Evrony, G.D. (2025) Direct measurement of the male germline mutation rate in individuals using sequential sperm samples. Nature Communications 2025 16:1 16, 2546-.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S.C., Kok, C.Y., Noble, K., Ponting, L., Ramshaw, C.C., Rye, C.E., Speedy, H.E., Stefancsik, R., Thompson, S.L., Wang, S., Ward, S., Campbell, P.J., Forbes, S.A. (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 47, D941–D947.

The UK Biobank Whole-Genome Sequencing Consortium. (2025) Whole-genome sequencing of 490,640 UK Biobank participants. Nature 645, 692–701.

Touchon, M., Nicolay, S., Audit, B., Brodie Of Brodie, E.B., D'Aubenton-Carafa, Y., Arneodo, A., Thermes, C. (2005) Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins. Proc. Natl. Acad. Sci. U. S. A. 102, 9836–9841.

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, 1679–1686.

Table 1

The table shows different quantitative measures for the six different single nucleotide variants having C or T as the background (reference) base. The first column shows the frequency of singleton variants relative to the counts of the reference base. The second column shows the odds ratios for counts of SNPs (seen in three or more participants) relative to singleton variants. The third and fourth columns show the ratios of the frequencies of singleton variants on the plus versus minus strand for chromosomes 1 and 10.

Substitution	Frequencies of singleton variants	ORs for SNP versus singleton variant counts	Ratios of frequencies of singleton variants on plus versus minus strands (0.95 confidence interval)	
			Chromosome 1	Chromosome 10
C>A	0.062	0.781	0.994 (0.992 - 0.996)	1.014 (1.008 - 1.020)
C>G	0.049	0.864	1.001 (0.999 - 1.003)	0.986 (0.980 - 0.992)
C>T	0.107	1.321	1.002 (1.001 - 1.003)	1.007 (1.002 - 1.011)
T>A	0.028	0.934	0.998 (0.996 - 1.001)	1.006 (0.999 - 1.014)
T>C	0.081	0.986	0.998 (0.997 - 1.000)	1.022 (1.017 - 1.026)
T>G	0.036	0.807	0.997 (0.995 - 0.998)	0.989 (0.982 - 0.995)

Table 2

Table showing the effect of the full pentanucleotide sequence versus the central trinucleotide sequence on the frequencies of singleton variants. Results are shown for the ten variants for which the frequency is most strongly reduced or increased.

Variant	OR	(0.95 CI)
TC[C>A]AG	0.462	(0.459 - 0.464)
TC[C>G]AT	0.475	(0.472 - 0.479)
TC[C>G]AC	0.477	(0.472 - 0.481)
CG[C>A]AC	0.494	(0.486 - 0.503)
GC[C>A]TG	0.498	(0.495 - 0.501)
GA[C>G]AT	0.511	(0.507 - 0.515)
GC[C>A]AG	0.511	(0.508 - 0.514)
TC[C>A]AC	0.519	(0.515 - 0.522)
TC[C>A]TG	0.535	(0.533 - 0.538)
GC[C>G]AC	0.537	(0.532 - 0.542)
CA[T>G]AG	1.867	(1.856 - 1.879)
CC[C>A]CT	1.898	(1.891 - 1.905)
CC[C>A]TG	1.910	(1.903 - 1.917)
AT[T>G]AG	1.919	(1.909 - 1.928)
CC[C>A]CA	1.923	(1.917 - 1.930)
CC[C>A]TC	1.979	(1.971 - 1.986)
AC[C>G]AA	2.023	(2.013 - 2.033)
CC[C>A]TA	2.272	(2.262 - 2.283)
GA[T>G]AG	2.273	(2.259 - 2.287)
CC[C>A]TT	2.493	(2.484 - 2.502)

Table 3

Mutational signatures producing the best fit to observed counts of trinucleotide singleton variants. The mutational signatures are obtained from somatic mutations observed in cancer cells and their proposed aetiologies are as stated for the COSMIC database at https://cancer.sanger.ac.uk/signatures/sbs/.

Signature	Relative contribution	Proposed aetiology
SBS12	0.461	Unknown.
SBS19	0.102	Unknown.
SBS30	0.164	Deficiency in base excision repair due to inactivating mutations in NTHL1.
SBS32	0.197	Prior treatment with azathioprine to induce immunosuppression.
SBS54	0.076	Possible sequencing artefact. Possible contamination with germline variants.

Table 4

Table showing the effect of the full pentanucleotide sequence versus the central trinucleotide sequence on the ratio of SNP to singleton variant counts. Results are shown for the ten variants for which the ratio is most strongly reduced or increased.

Variant	OR	(0.95 CI)
CG[C>T]GG	0.400	(0.394 - 0.406)
CC[C>A]TG	0.406	(0.403 - 0.410)
CC[C>A]TA	0.448	(0.444 - 0.453)
CG[C>T]GC	0.526	(0.518 - 0.534)
GC[C>T]GC	0.550	(0.544 - 0.557)
GT[C>A]AG	0.559	(0.553 - 0.565)
CC[T>A]TA	0.566	(0.557 - 0.575)
GC[T>A]TA	0.568	(0.558 - 0.578)
AA[T>G]AC	0.572	(0.565 - 0.580)
CT[C>A]TA	0.575	(0.570 - 0.581)
TT[T>A]CG	2.360	(2.316 - 2.406)
CT[T>A]CG	2.511	(2.459 - 2.564)
CG[T>C]TG	2.730	(2.700 - 2.760)
CG[C>G]TT	2.747	(2.700 - 2.794)
CG[T>G]TG	2.824	(2.778 - 2.871)
AA[C>G]GG	2.978	(2.937 - 3.019)
TT[C>A]GT	3.038	(3.003 - 3.074)
AA[C>A]GG	3.227	(3.188 - 3.267)
CG[C>A]TT	3.757	(3.694 - 3.821)
CG[T>A]TG	3.771	(3.705 - 3.837)

Table 5

Table showing chromosome 1 variants with the lowest and highest ratios of singleton frequencies on the plus versus minus strand.

Variant	Ratio	CI
GT[T>A]TC	0.646	(0.621 - 0.673)
CG[T>A]TT	0.764	(0.710 - 0.823)
TC[T>A]AG	0.769	(0.733 - 0.805)
TC[C>G]AC	0.780	(0.747 - 0.814)
CG[C>G]AC	0.808	(0.734 - 0.889)
TC[T>A]GT	0.836	(0.812 - 0.859)
CT[T>A]GC	0.838	(0.801 - 0.876)
GT[T>G]AC	0.843	(0.799 - 0.889)
AA[C>A]GG	0.849	(0.806 - 0.894)
TT[T>A]CC	0.853	(0.828 - 0.879)
CT[T>A]TG	1.154	(1.113 - 1.197)
CG[T>A]CG	1.170	(0.964 - 1.420)
CT[T>C]CG	1.176	(1.101 - 1.257)
GT[T>A]GG	1.190	(1.134 - 1.248)
CG[T>G]TG	1.192	(1.110 - 1.280)
CG[T>G]AT	1.218	(1.092 - 1.358)
CG[T>A]AT	1.220	(1.105 - 1.347)
CG[T>A]TG	1.368	(1.252 - 1.496)
CT[T>G]CG	1.592	(1.473 - 1.721)
CT[T>A]CG	2.009	(1.828 - 2.208)

Table 6

Table showing variants with the lowest and highest odds ratios for SNP to singleton variant counts on the plus versus minus strand.

Variant	OR (0.95 CI)
TC[C>A]AC	0.301 (0.297 - 0.305)
TC[C>A]CG	0.332 (0.325 - 0.339)
TC[C>G]AC	0.345 (0.339 - 0.352)
TC[C>A]AA	0.359 (0.355 - 0.363)
CG[C>G]AC	0.384 (0.369 - 0.401)
AA[C>A]GA	0.399 (0.392 - 0.405)
TC[C>G]AA	0.404 (0.400 - 0.409)
CC[C>A]GT	0.406 (0.400 - 0.412)
GT[T>G]TC	0.417 (0.412 - 0.421)
GT[C>A]TA	0.419 (0.412 - 0.425)
CT[T>A]GC	2.298 (2.270 - 2.326)
GT[T>A]TC	2.523 (2.466 - 2.580)
AA[C>G]GA	2.566 (2.473 - 2.662)
CG[C>A]AC	2.581 (2.524 - 2.638)
TC[C>G]CG	2.788 (2.708 - 2.870)
CG[C>A]TC	2.907 (2.852 - 2.963)
AA[T>A]GC	3.017 (2.960 - 3.076)
TA[T>G]CC	3.355 (3.252 - 3.461)
CA[C>G]AA	3.429 (3.284 - 3.581)
AA[T>G]GC	4.051 (3.914 - 4.193)

Table 7

Table showing pentanucleotide sequences in the reference genome with the lowest and highest ratios of counts on the plus versus minus strand

Pentanucleotide	Plus strand count	Minus strand count	Ratio	(0.95 CI)
AACGA	470030	611654	0.768	(0.767 - 0.770)
CCCGT	542317	680590	0.797	(0.795 - 0.798)
CGTTT	686619	825273	0.832	(0.831 - 0.833)
TCCCG	742920	865677	0.858	(0.857 - 0.860)
GTTTC	3585077	4152246	0.863	(0.863 - 0.864)
TACGA	289993	332914	0.871	(0.869 - 0.873)
TCCAA	3397443	3833783	0.886	(0.886 - 0.887)
TCCAC	2781088	3135301	0.887	(0.886 - 0.888)
GTCTA	1687459	1895823	0.890	(0.889 - 0.891)
AACGC	322055	359574	0.896	(0.893 - 0.898)
CGTTC	421273	389799	1.081	(1.078 - 1.083)
CTTTG	5114710	4719443	1.084	(1.083 - 1.084)
CGTAT	440600	403921	1.091	(1.088 - 1.093)
TGCGT	481833	441485	1.091	(1.089 - 1.094)
CATTC	3936190	3565845	1.104	(1.103 - 1.105)
TTTCG	599225	498471	1.202	(1.200 - 1.204)
CTTCG	522495	410566	1.273	(1.270 - 1.275)
CGTTG	524759	377494	1.390	(1.387 - 1.393)
AACGG	530951	380526	1.395	(1.392 - 1.398)
TTCGT	673300	465807	1.445	(1.443 - 1.448)

Figure 1

Effects of different trinucleotide sequences on the frequencies of different singleton variants. Results are shown for all trinucleotides having either C or T as the central base.

Figure 1A

Frequencies of all singleton variants occurring in the context of different trinucleotide backgrounds. The frequency is defined as the count of the number of times each singleton variant is observed divided by the count of the number of times the background is seen in the reference genome.

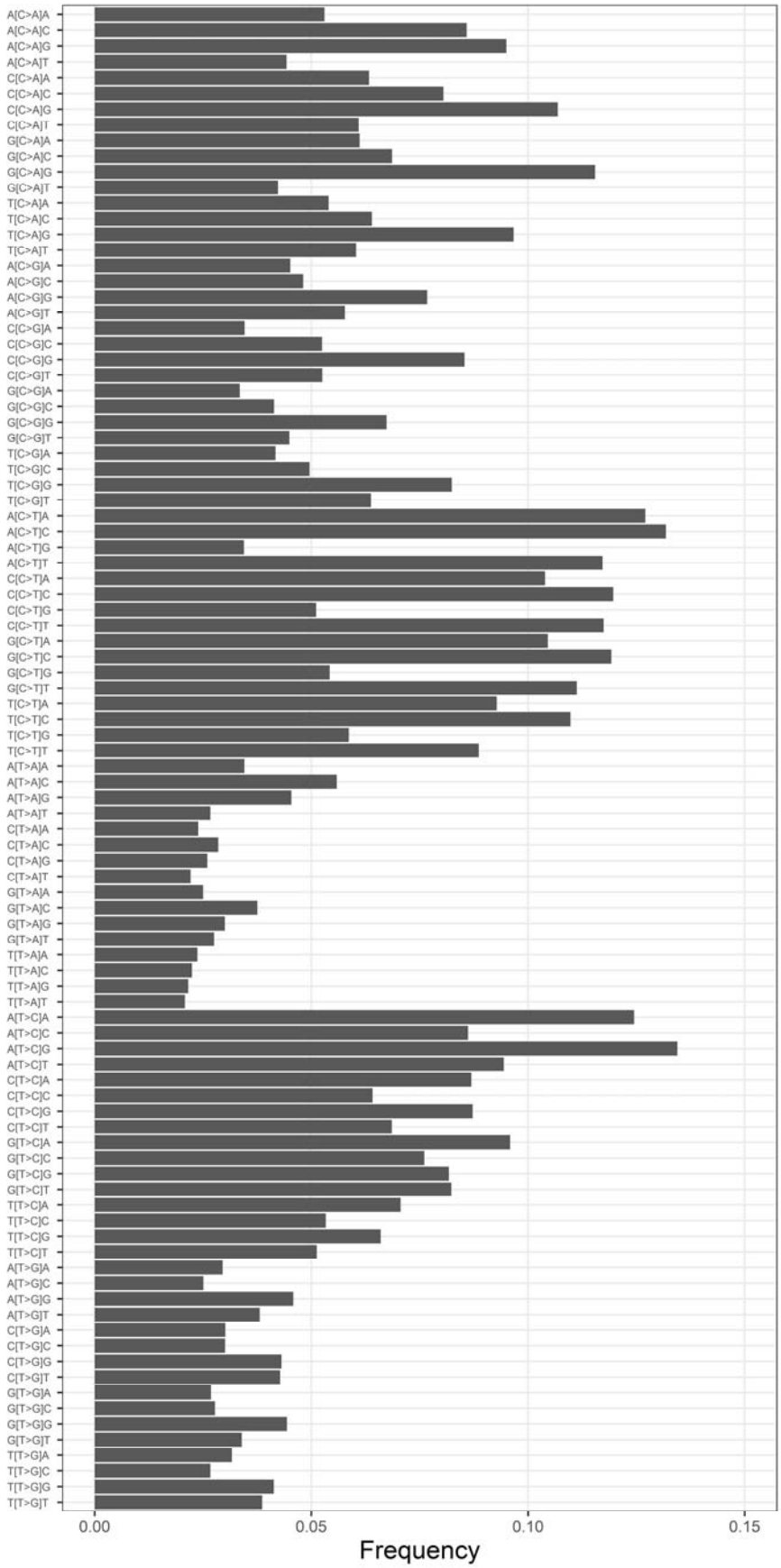

A[C>A]A
A[C>A]C
A[C>A]G
A[C>A]T
C[C>A]A
C[C>A]C
C[C>A]G
C[C>A]T
G[C>A]A
G[C>A]C
G[C>A]G
G[C>A]T
T[C>A]A
T[C>A]C
T[C>A]G
T[C>A]T
A[C>G]A
A[C>G]C
A[C>G]G
A[C>G]T
C[C>G]A
C[C>G]C
C[C>G]G
C[C>G]T
G[C>G]A
G[C>G]C
G[C>G]G
G[C>G]T
T[C>G]A
T[C>G]C
T[C>G]G
T[C>G]T
A[C>T]A
A[C>T]C
A[C>T]G
A[C>T]T
C[C>T]A
C[C>T]C
C[C>T]G
C[C>T]T
G[C>T]A
G[C>T]C
G[C>T]G
G[C>T]T
T[C>T]A
T[C>T]C
T[C>T]G
T[C>T]T
A[T>A]A
A[T>A]C
A[T>A]G
A[T>A]T
C[T>A]A
C[T>A]C
C[T>A]G
C[T>A]T
G[T>A]A
G[T>A]C
G[T>A]G
G[T>A]T
T[T>A]A
T[T>A]C
T[T>A]G
T[T>A]T
A[T>C]A
A[T>C]C
A[T>C]G
A[T>C]T
C[T>C]A
C[T>C]C
C[T>C]G
C[T>C]T
G[T>C]A
G[T>C]C
G[T>C]G
G[T>C]T
T[T>C]A
T[T>C]C
T[T>C]G
T[T>C]T
A[T>G]A
A[T>G]C
A[T>G]G
A[T>G]T
C[T>G]A
C[T>G]C
C[T>G]G
C[T>G]T
G[T>G]A
G[T>G]C
G[T>G]G
G[T>G]T
T[T>G]A
T[T>G]C
T[T>G]G
T[T>G]T
0.00
0.05
0.10
0.15
Frequency

Figure 1B

Odds ratio for each singleton variant associated with its trinucleotide background, defined as the frequency with which the variant occurs on this background compared with the overall frequency of singleton single base substitutions of this type (e.g. frequency of A[T>A]C versus frequency of T>A).

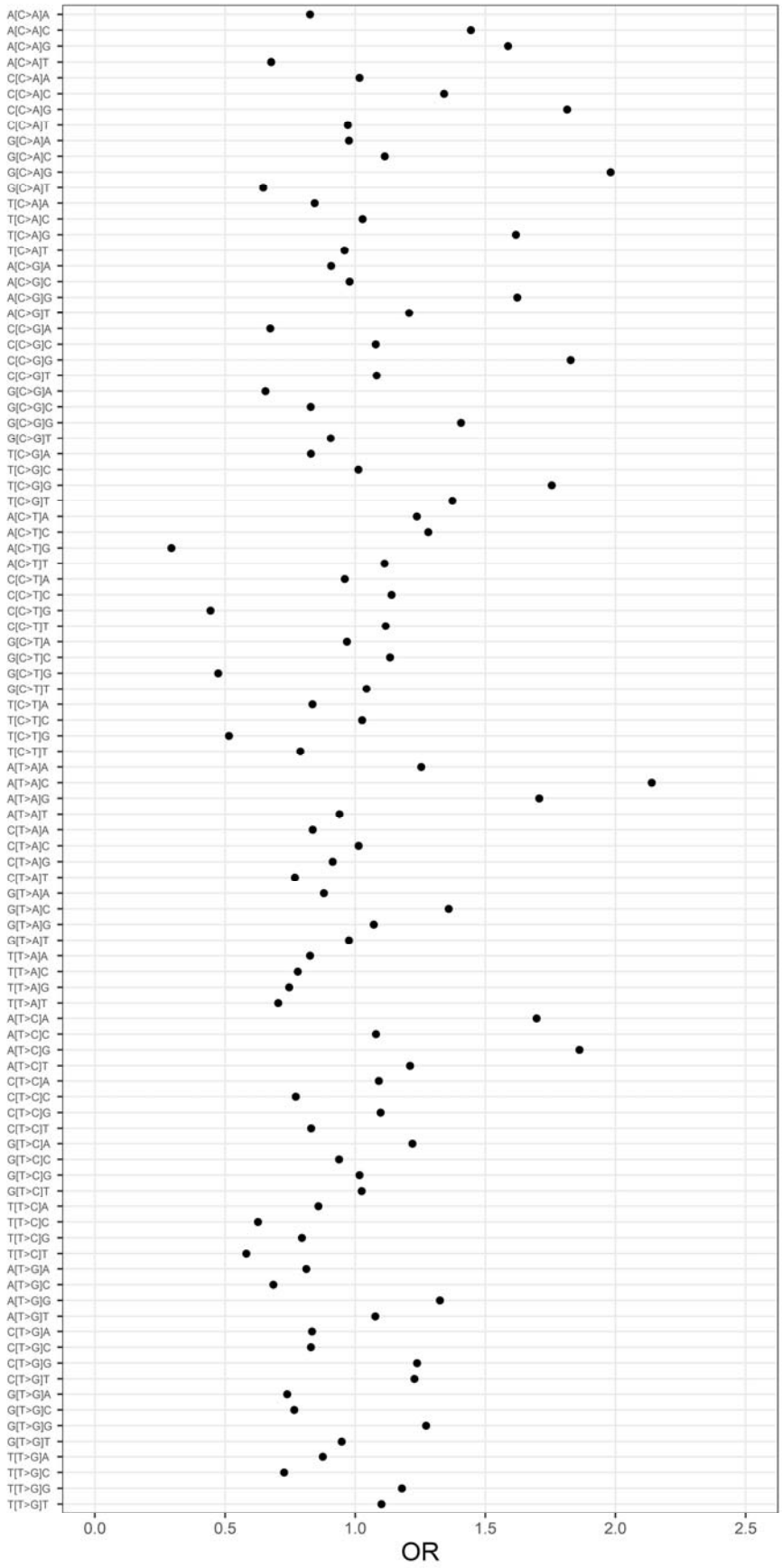

A[C>A]A
A[C>A]C
A[C>A]G
A[C>A]T
C[C>A]A
C[C>A]C
C[C>A]G
C[C>A]T
G[C>A]A
G[C>A]C
G[C>A]G
G[C>A]T
T[C>A]A
T[C>A]C
T[C>A]G
T[C>A]T
A[C>G]A
A[C>G]C
A[C>G]G
A[C>G]T
C[C>G]A
C[C>G]C
C[C>G]G
C[C>G]T
G[C>G]A
G[C>G]C
G[C>G]G
G[C>G]T
T[C>G]A
T[C>G]C
T[C>G]G
T[C>G]T
A[C>T]A
A[C>T]C
A[C>T]G
A[C>T]T
C[C>T]A
C[C>T]C
C[C>T]G
C[C>T]T
G[C>T]A
G[C>T]C
G[C>T]G
G[C>T]T
T[C>T]A
T[C>T]C
T[C>T]G
T[C>T]T
A[T>A]A
A[T>A]C
A[T>A]G
A[T>A]T
C[T>A]A
C[T>A]C
C[T>A]G
C[T>A]T
G[T>A]A
G[T>A]C
G[T>A]G
G[T>A]T
T[T>A]A
T[T>A]C
T[T>A]G
T[T>A]T
A[T>C]A
A[T>C]C
A[T>C]G
A[T>C]T
C[T>C]A
C[T>C]C
C[T>C]G
C[T>C]T
G[T>C]A
G[T>C]C
G[T>C]G
G[T>C]T
T[T>C]A
T[T>C]C
T[T>C]G
T[T>C]T
A[T>G]A
A[T>G]C
A[T>G]G
A[T>G]T
C[T>G]A
C[T>G]C
C[T>G]G
C[T>G]T
G[T>G]A
G[T>G]C
G[T>G]G
G[T>G]T
T[T>G]A
T[T>G]C
T[T>G]G
T[T>G]T
0.0
0.5
1.0
1.5
2.0
2.5
OR

Figure 2

Results of modelling singleton SBS counts using reference mutational signatures obtained from studies of somatic mutations observed in cancers.

Figure 2B

Observed counts of singleton SBSs on trinucleotide backgrounds.

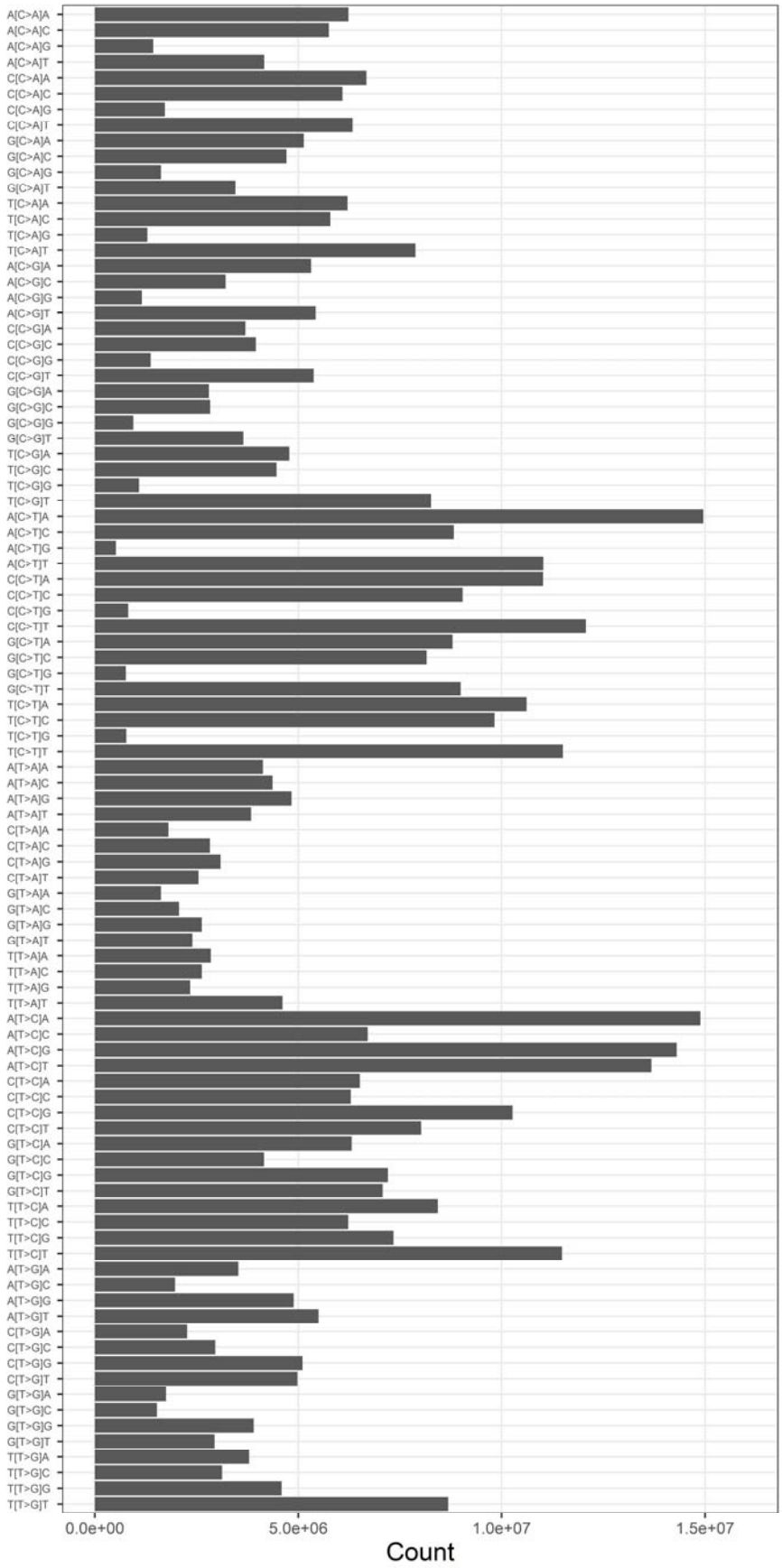

A[C>A]A
A[C>A]C
A[C>A]G
A[C>A]T
C[C>A]A
C[C>A]C
C[C>A]G
C[C>A]T
G[C>A]A
G[C>A]C
G[C>A]G
G[C>A]T
T[C>A]A
T[C>A]C
T[C>A]G
T[C>A]T
A[C>G]A
A[C>G]C
A[C>G]G
A[C>G]T
C[C>G]A
C[C>G]C
C[C>G]G
C[C>G]T
G[C>G]A
G[C>G]C
G[C>G]G
G[C>G]T
T[C>G]A
T[C>G]C
T[C>G]G
T[C>G]T
A[C>T]A
A[C>T]C
A[C>T]G
A[C>T]T
C[C>T]A
C[C>T]C
C[C>T]G
C[C>T]T
G[C>T]A
G[C>T]C
G[C>T]G
G[C>T]T
T[C>T]A
T[C>T]C
T[C>T]G
T[C>T]T
A[T>A]A
A[T>A]C
A[T>A]G
A[T>A]T
C[T>A]A
C[T>A]C
C[T>A]G
C[T>A]T
G[T>A]A
G[T>A]C
G[T>A]G
G[T>A]T
T[T>A]A
T[T>A]C
T[T>A]G
T[T>A]T
A[T>C]A
A[T>C]C
A[T>C]G
A[T>C]T
C[T>C]A
C[T>C]C
C[T>C]G
C[T>C]T
G[T>C]A
G[T>C]C
G[T>C]G
G[T>C]T
T[T>C]A
T[T>C]C
T[T>C]G
T[T>C]T
A[T>G]A
A[T>G]C
A[T>G]G
A[T>G]T
C[T>G]A
C[T>G]C
C[T>G]G
C[T>G]T
G[T>G]A
G[T>G]C
G[T>G]G
G[T>G]T
T[T>G]A
T[T>G]C
T[T>G]G
T[T>G]T
0.0e+00
5.0e+06
1.0e+07
1.5e+07
Count

Figure 2B

Predicted singleton SBS counts using best-fitting combination of five reference mutational signatures.

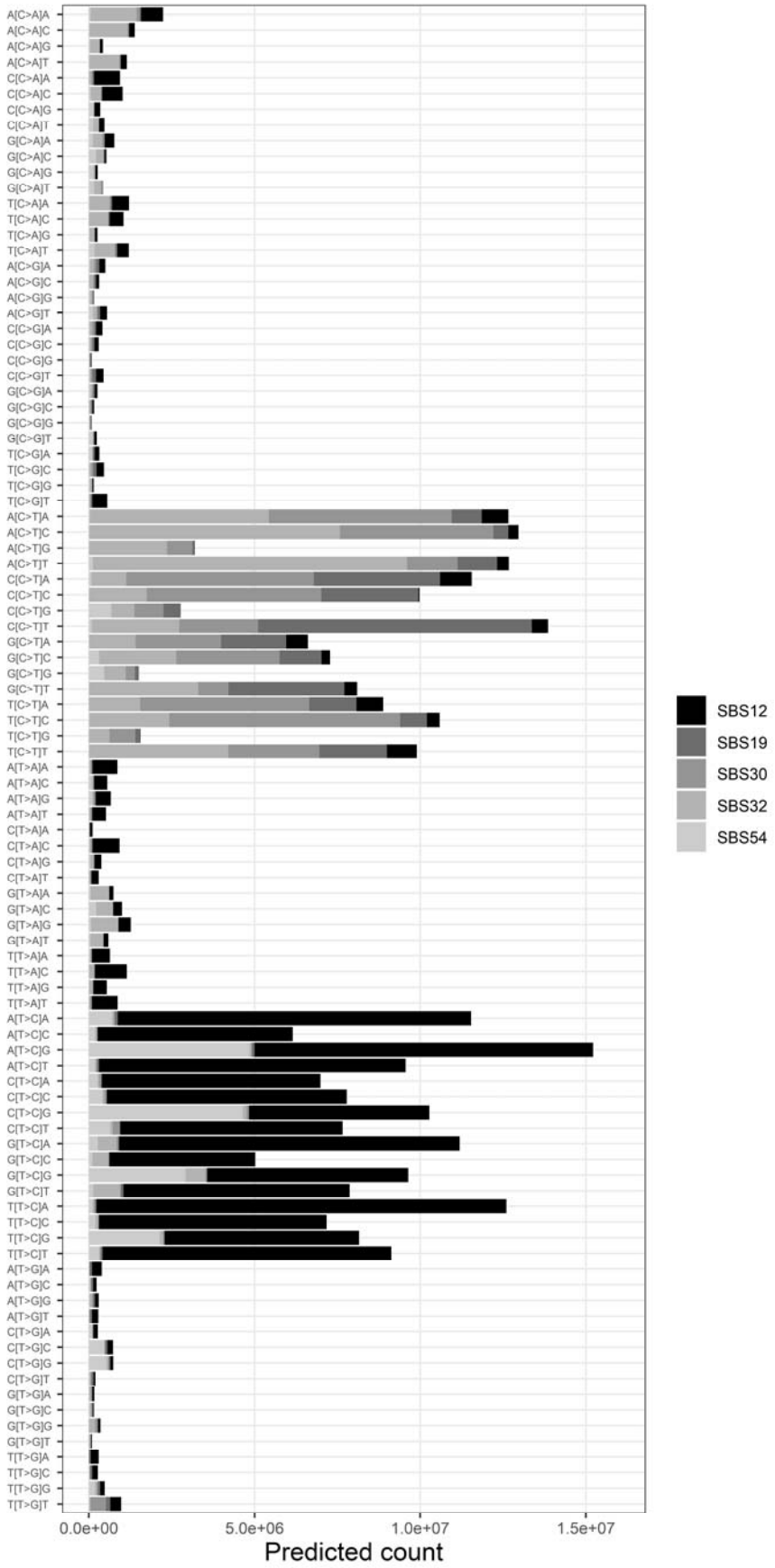

A[C>A]A
A[C>A]C
A[C>A]G
A[C>A]T
C[C>A]A
C[C>A]C
C[C>A]G
C[C>A]T
G[C>A]A
G[C>A]C
G[C>A]G
G[C>A]T
T[C>A]A
T[C>A]C
T[C>A]G
T[C>A]T
A[C>G]A
A[C>G]C
A[C>G]G
A[C>G]T
C[C>G]A
C[C>G]C
C[C>G]G
C[C>G]T
G[C>G]A
G[C>G]C
G[C>G]G
G[C>G]T
T[C>G]A
T[C>G]C
T[C>G]G
T[C>G]T
A[C>T]A
A[C>T]C
A[C>T]G
A[C>T]T
C[C>T]A
C[C>T]C
C[C>T]G
C[C>T]T
G[C>T]A
G[C>T]C
G[C>T]G
G[C>T]T
T[C>T]A
T[C>T]C
T[C>T]G
T[C>T]T
A[T>A]A
A[T>A]C
A[T>A]G
A[T>A]T
C[T>A]A
C[T>A]C
C[T>A]G
C[T>A]T
G[T>A]A
G[T>A]C
G[T>A]G
G[T>A]T
T[T>A]A
T[T>A]C
T[T>A]G
T[T>A]T
A[T>C]A
A[T>C]C
A[T>C]G
A[T>C]T
C[T>C]A
C[T>C]C
C[T>C]G
C[T>C]T
G[T>C]A
G[T>C]C
G[T>C]G
G[T>C]T
T[T>C]A
T[T>C]C
T[T>C]G
T[T>C]T
A[T>G]A
A[T>G]C
A[T>G]G
A[T>G]T
C[T>G]A
C[T>G]C
C[T>G]G
C[T>G]T
G[T>G]A
G[T>G]C
G[T>G]G
G[T>G]T
T[T>G]A
T[T>G]C
T[T>G]G
T[T>G]T
0.0e+00
5.0e+06
1.0e+07
1.5e+07
Predicted count
SBS12
SBS19
SBS30
SBS32
SBS54

Figure 3

Odds ratios for SNP versus singleton variant counts for each trinucleotide variant with C or T as the central base, excluding variants with the X[C>T]G motif (which all had OR > 14).

Figure 3A

Odds ratio for each trinucleotide variant relative to ratio of counts for all SNP and singleton variants.

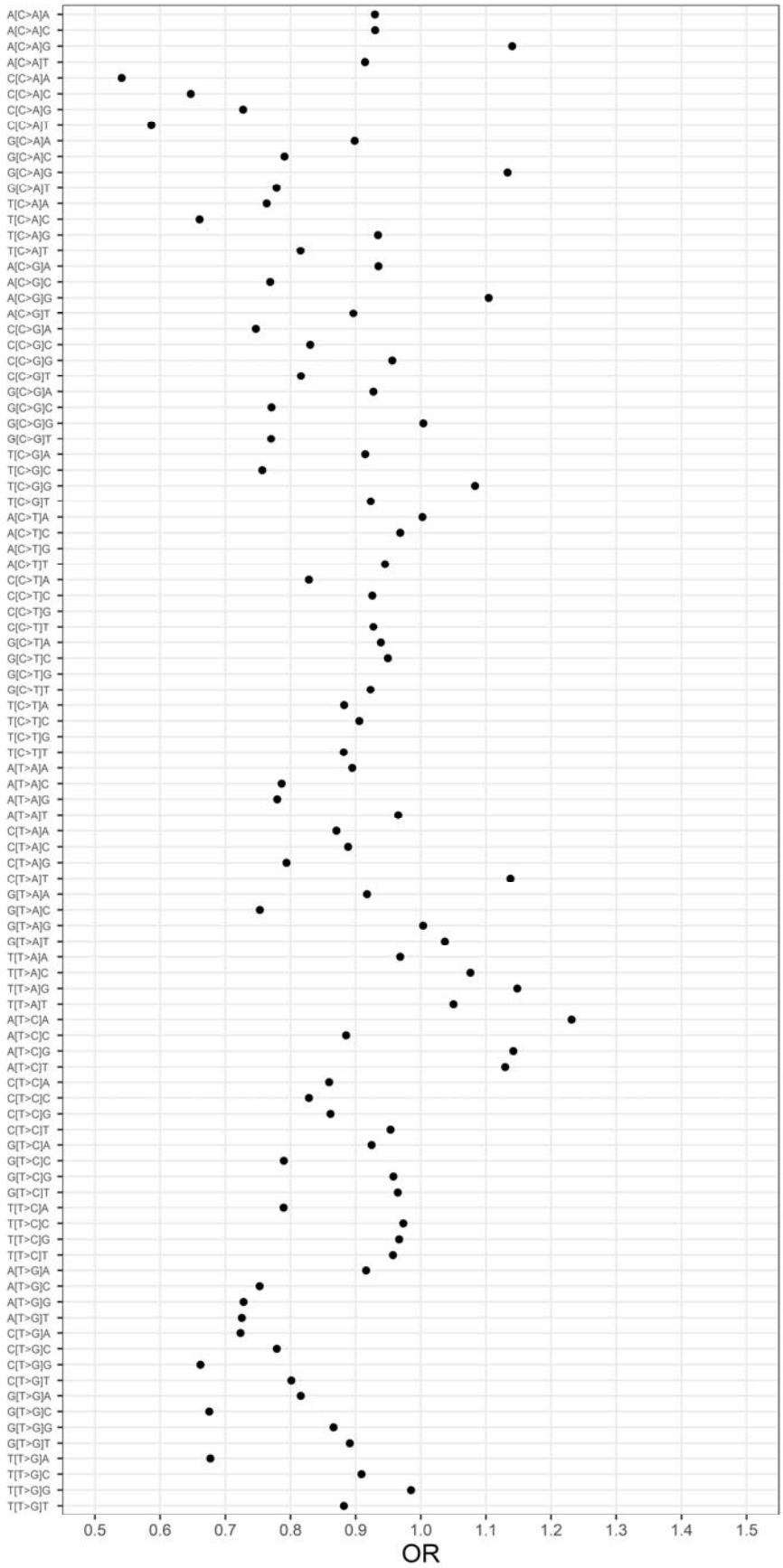

A[C>A]A
A[C>A]C
A[C>A]G
A[C>A]T
C[C>A]A
C[C>A]C
C[C>A]G
C[C>A]T
G[C>A]A
G[C>A]C
G[C>A]G
G[C>A]T
T[C>A]A
T[C>A]C
T[C>A]G
T[C>A]T
A[C>G]A
A[C>G]C
A[C>G]G
A[C>G]T
C[C>G]A
C[C>G]C
C[C>G]G
C[C>G]T
G[C>G]A
G[C>G]C
G[C>G]G
G[C>G]T
T[C>G]A
T[C>G]C
T[C>G]G
T[C>G]T
A[C>T]A
A[C>T]C
A[C>T]G
A[C>T]T
C[C>T]A
C[C>T]C
C[C>T]G
C[C>T]T
G[C>T]A
G[C>T]C
G[C>T]G
G[C>T]T
T[C>T]A
T[C>T]C
T[C>T]G
T[C>T]T
A[T>A]A
A[T>A]C
A[T>A]G
A[T>A]T
C[T>A]A
C[T>A]C
C[T>A]G
C[T>A]T
G[T>A]A
G[T>A]C
G[T>A]G
G[T>A]T
T[T>A]A
T[T>A]C
T[T>A]G
T[T>A]T
A[T>C]A
A[T>C]C
A[T>C]G
A[T>C]T
C[T>C]A
C[T>C]C
C[T>C]G
C[T>C]T
G[T>C]A
G[T>C]C
G[T>C]G
G[T>C]T
T[T>C]A
T[T>C]C
T[T>C]G
T[T>C]T
A[T>G]A
A[T>G]C
A[T>G]G
A[T>G]T
C[T>G]A
C[T>G]C
C[T>G]G
C[T>G]T
G[T>G]A
G[T>G]C
G[T>G]G
G[T>G]T
T[T>G]A
T[T>G]C
T[T>G]G
T[T>G]T
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
OR

Figure 3B

Odds ratio for each trinucleotide variant relative to the odds for the central SBS (e.g. A[T>C]G versus T>C).

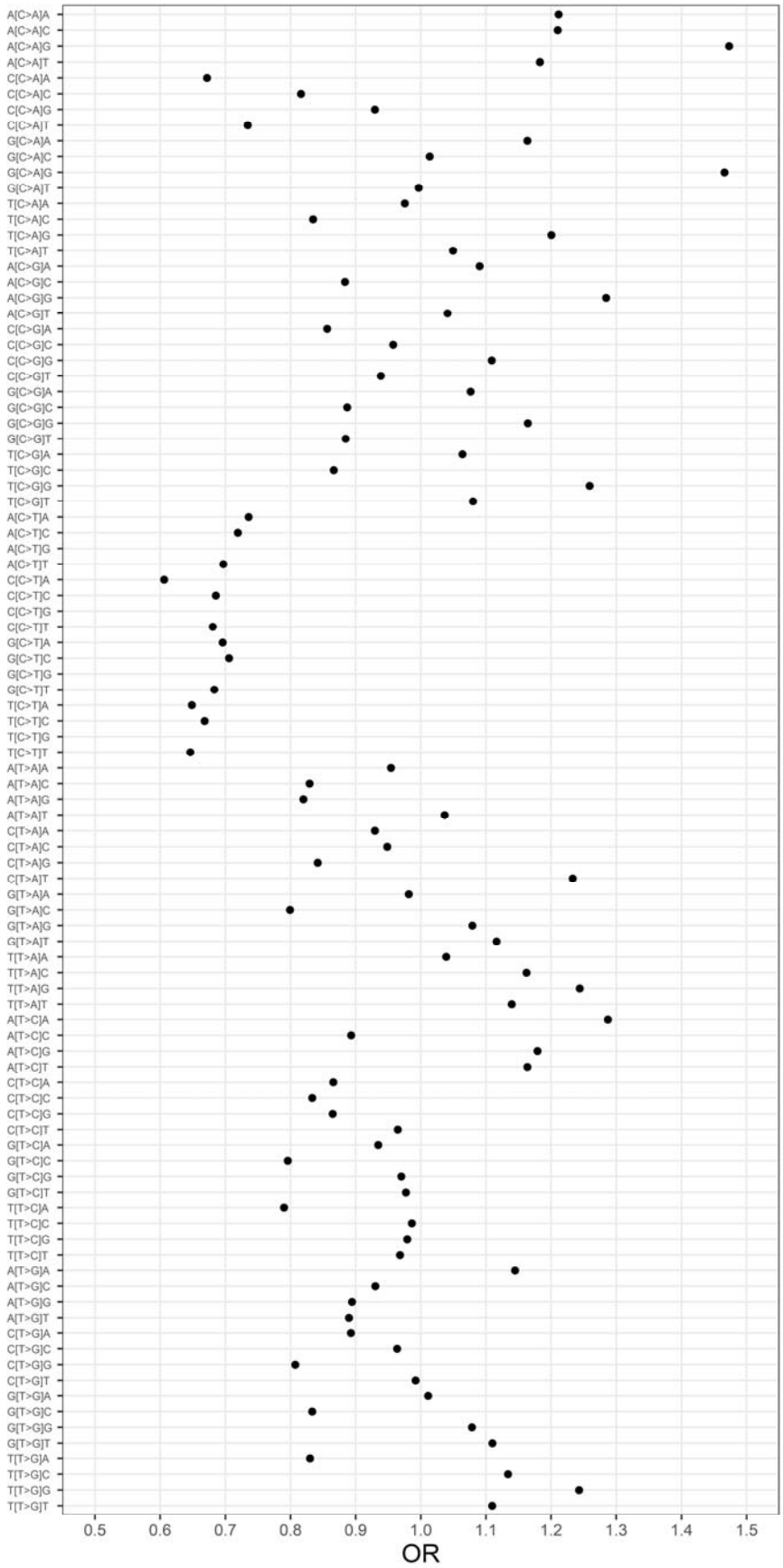
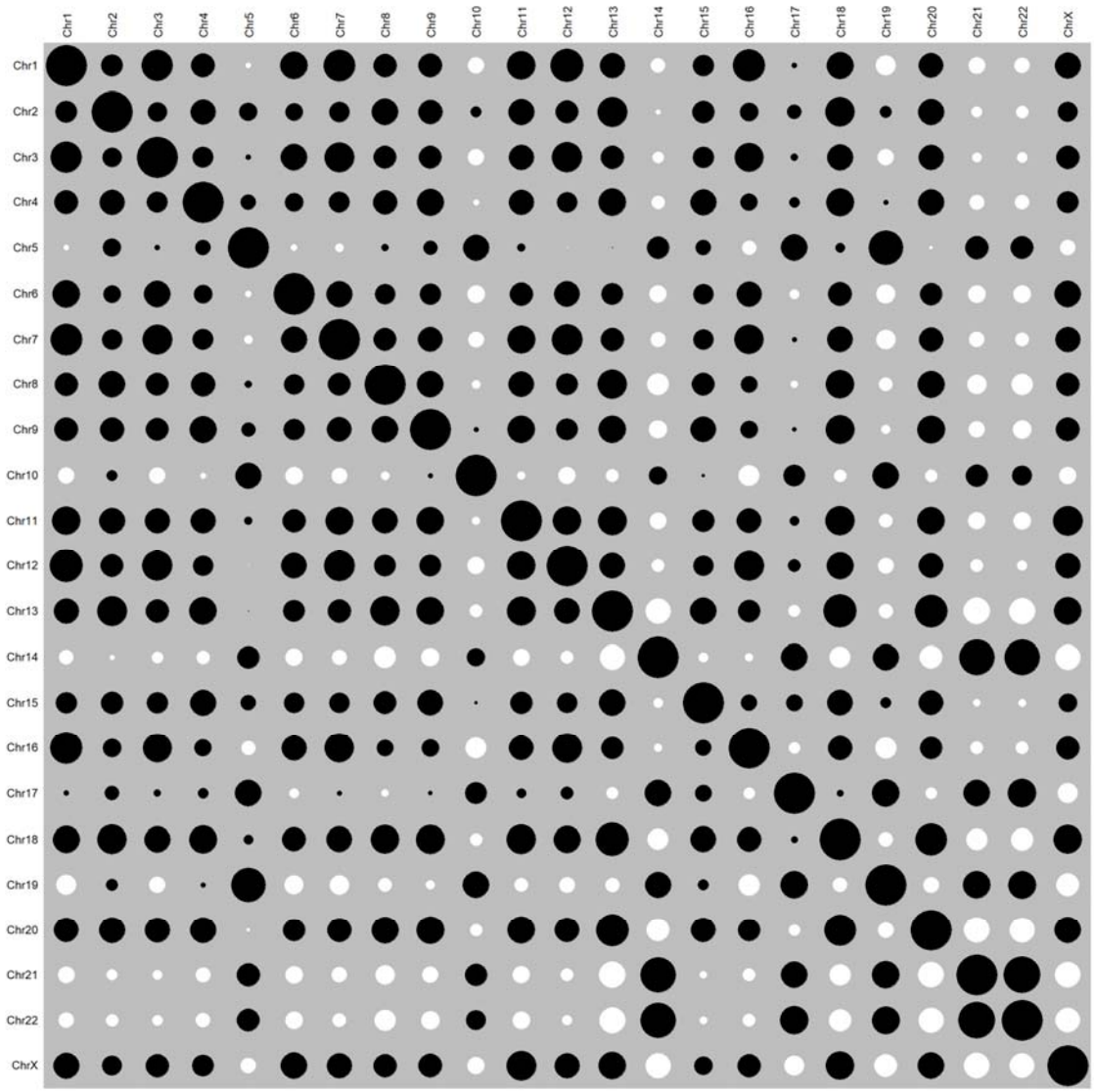

A[C>A]A
A[C>A]C
A[C>A]G
A[C>A]T
C[C>A]A
C[C>A]C
C[C>A]G
C[C>A]T
G[C>A]A
G[C>A]C
G[C>A]G
G[C>A]T
T[C>A]A
T[C>A]C
T[C>A]G
T[C>A]T
A[C>G]A
A[C>G]C
A[C>G]G
A[C>G]T
C[C>G]A
C[C>G]C
C[C>G]G
C[C>G]T
G[C>G]A
G[C>G]C
G[C>G]G
G[C>G]T
T[C>G]A
T[C>G]C
T[C>G]G
T[C>G]T
A[C>T]A
A[C>T]C
A[C>T]G
A[C>T]T
C[C>T]A
C[C>T]C
C[C>T]G
C[C>T]T
G[C>T]A
G[C>T]C
G[C>T]G
G[C>T]T
T[C>T]A
T[C>T]C
T[C>T]G
T[C>T]T
A[T>A]A
A[T>A]C
A[T>A]G
A[T>A]T
C[T>A]A
C[T>A]C
C[T>A]G
C[T>A]T
G[T>A]A
G[T>A]C
G[T>A]G
G[T>A]T
T[T>A]A
T[T>A]C
T[T>A]G
T[T>A]T
A[T>C]A
A[T>C]C
A[T>C]G
A[T>C]T
C[T>C]A
C[T>C]C
C[T>C]G
C[T>C]T
G[T>C]A
G[T>C]C
G[T>C]G
G[T>C]T
T[T>C]A
T[T>C]C
T[T>C]G
T[T>C]T
A[T>G]A
A[T>G]C
A[T>G]G
A[T>G]T
C[T>G]A
C[T>G]C
C[T>G]G
C[T>G]T
G[T>G]A
G[T>G]C
G[T>G]G
G[T>G]T
T[T>G]A
T[T>G]C
T[T>G]G
T[T>G]T
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
OR

Figure 4

Heatmap showing correlations between patterns of strand asymmetry across variant types between pairs of chromosomes. Positive correlations are shown by black circles and negative correlations by white circles. The size of the circle indicates the magnitude of the correlation, with circles having size R = 1 on the leading diagonal.