Sequence effects on mutation frequencies investigated in 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*

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. Variants with an allele count of 25 were found to closely mirror previously reported de novo mutations (DNMs) in terms of the frequencies of variant types. Therefore these variants, referred to as AC25 variants, were used to investigate factors relevant to mutation frequency. The counts of AC25 variant types in a trinucleotide context could be well approximated by combining seven mutational signatures previously obtained from studies of cancer cells. Frequencies of variants were strongly influenced by context. C>T variants occurred more often in the CpG context but other features of the trinucleotide context also had marked effects. Although the trinucleotide context was a very strong predictor of the variant frequencies for the full pentanucleotide context, there were examples where the more distant nucleotides did have marked effects. Here, the configuration of bases two upstream or downstream could have a more than two-fold effect on variant frequencies, presumed to reflect mutation rates. Investigating the molecular mechanisms whereby these contexts moderate mutation frequencies could lead to a better understanding of how DNA fidelity is conserved in normal tissue.

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. 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 mutations (DNMs) can be observed (Jónsson et al., 2017; McRae et al., 2017; Rahbari et al., 2016). A sample of 2,976 whole genome-sequenced trios was used to identify 200,435 DNMs and these were used to assess the mutation probabilities for different trinucleotides as well as how these were impacted by crossover events (Halldorsson et al., 2019).

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, on the assumption that these are likely to reflect patterns of recent mutation. In a

sample of 3,560 whole-genome sequences for the Bipolar Research in Deep Genome and Epigenome Sequencing (BRIDGES) study, 36 million singleton variants were 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).

More recent work has thrown into question the appropriateness of using singleton variants to estimate mutation rates. With a constant population size and a neutral mutation rate u, the allele frequency spectrum is such that the expected number of variants to have exactly i copies of the alternate allele in a sample is proportional to $\underline{u}$ / i under the "infinite sites" assumption that there are no recurrent mutations (Kimura, 1969). However for sites with high mutation rates at which recurrent mutations can occur there will be a reduction in low frequency variants and an excess at higher frequencies (Wakeley et al., 2023). This means that the singleton count is low relative the mutation rate for variant types whose mutation rate is high and so overall the singleton counts across different variant types will not correlate well with their mutation rates. A follow-up study of the BRIDGES samples confirmed that the allele frequency spectrum differed between variant types, with the singleton to doubleton ratio being lower for variant types with higher mutation rates, and showed that this heterogeneity could impact inferences about population genetic history (Liao et al., 2023).

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 large sample was used to obtain estimates of relative mutation rates for different types of variant according to their pentanucleotide context.

Methods

Obtaining variant counts

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. All analyses were confined to non-centromeric regions because the centromeres include repetitive DNA sequences which are highly homogeneous and hence their sequence content is not representative of the rest of the genome (Altemose et al., 2022). The telomeric regions were also excluded. In the included regions there were a total of 1.06e9 SBS variants, of which 4.05e6 were singletons and 1.48e6 were doubletons.

Each SBS variant 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, meaning that there was a total of 4^5*3 = 3,072 types. The allele count (AC) field for each *pvar* file was used to obtain total counts for each variant type binned according to the AC field, consisting of singletons, doubletons, AC = 3, etc. up to AC=200.

A list of DNMs obtained from 2,786 whole genome sequenced trios described above was downloaded from https://www.science.org/doi/suppl/10.1126/science.aau1043/suppl_file/aau1043_datas5_revision1.tsv (Halldorsson et al., 2019). Counts of variant types among these DNMs, consisting of the variant in the context of its pentanucleotide background, were enumerated. In the regions excluding the centromeres and telomeres, there were a total of 178,610 SBS variants.

For each AC bin, the frequency of each variant type was obtained, consisting of the count divided by the number of times the background pentanucleotide occurred in the reference genome, and the frequencies for the DNMs were obtained in the same way. As previously reported, the ratio of singletons to doubletons was markedly reduced for variant types with higher mutation rates, consisting of those including the [C>T]G motif, and across the 3,072 types the correlation of singleton frequencies with the DNM frequencies was only 0.043. This correlation increased to 0.442 for doubletons and to 0.704 for tripletons and continued increasing with allele count until it reached 0.990 for variants with AC=25. Thus these variants, each of which occurs exactly 25 times in the sample and now referred to as AC25 variants, almost exactly predict the relative mutation frequencies observed directly in DNMs. There were in total 2,654,130 AC25 variants so this much larger sample size was taken to mean that these variants could be used to obtain more accurate estimates of effects of sequence on mutation probability than had been possible with the DNMs.

Mutational signature analysis

Previous investigations of mutational signatures in cancer have used counts of variants occurring on a trinucleotide background sequence and use the 96 variant types having C or T as the central base, using the sequence on the minus strand when the central base on the plus strand is G or A (Alexandrov et al., 2020). The AC25 counts for the variant types based on pentanucleotide backgrounds were pooled according to this scheme. 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 AC25 counts for the 96 types, 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 AC25 counts from a small number of reference signatures.

Modelling effects of background on variant frequencies

The general approach to analysing sequence effects on mutation probabilities involved 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 the reference genome and we observe 3.23e5 AC25 T>C variants then we say that the frequency of AC25 T>C variants occurring on the background of the T nucleotide is 3.23e5/7.78e8 = 0.000415. Likewise, if the trinucleotide sequence ATG occurs 4.55e7 times in the reference and there are 4.00e4 AC25 A[T>C]G variants then the frequency of AC25 A[T>C]G variants on the ATG background is 4.00e4 / 7.07e7 = 0.000880. Finally, if the pentanucleotide CATGC sequence occurs 3.89e6 times in the reference and there are 2,680 CA[T>C]GC AC25 variants then the frequency of CA[T>C]GC AC25 variants on the CATGC background is 2,680 / 3.89e6 = 0.000689.

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 AC25 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, AC25 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 additional 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.

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/countDNAVariants. This repository also contains raw counts and calculated frequencies of the variants analysed. Additional data management and figure preparation was performed using R v4.3.2 (R Core Team, 2014).

Results

Mutational signature analysis

For the mutation signature analysis, the counts of standard 96 mutation types, consisting of all trinucleotide AC25 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 AC25 variants.(Tate et al., 2019). Coefficients were constrained to be positive and the intercept was constrained to zero. When this was done there were seven signatures for which the ratio of the coefficient to its standard error exceeded 7 and these signatures were retained and the modelling repeated. These seven signatures along with their best fitting relative contributions are shown in Table 1. 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 = 0.993. These observed and predicted counts are shown in Figure 1.

Modelling effects of background on variant frequencies

Table 2 shows the frequencies of AC25 variants observed on the background of a C or T nucleotide. It can be seen that the most frequent AC25 variant is C>T, with frequency 0.000922, and the rarest is T>A, with frequency 0.000106. AC25 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 AC25 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 = 0.507. If a single nucleotide upstream was included the correlation increased to R = 0.748 and including the second upstream nucleotide yielded R = 0.749, while including the two flanking nucleotides yielded R = 0.992. Finally, if two upstream and one downstream nucleotide were included the correlation was R = 0.996. These results suggest that a major driver for the frequency of each variant is the probability of the central base substitution itself but that this probability is strongly influenced 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 2A shows the frequencies of all AC25 variants occurring on trinucleotide sequences with either C or T as the central base. Figure 2B shows the odds ratio (OR) for each trinucleotide sequence to produce the target base relative to the odds for just the central base. The graph does not show the results for variant types having the pattern X[C>T]G because these all had very large ORs compared to the others, with ORs of 8.66, 9.16, 9.97 and 11.15 for T[C>T]G, G[C>T]G, C[C>T]G and A[C>T]G respectively. Among the other variant types, the two for which the flanking nucleotides have the strongest positive effect are A[T>C]A and A[T>C]G with ORs of 2.32 and 2.29 respectively. Conversely T[C>T]T and T[C>T]A have ORs of 0.408 and 0.439, indicating that these flanking nucleotides act to reduce the probability of the variant occurring.

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 variant types with the ten lowest and highest ORs. These results indicate that the more distant nucleotides can have major effects. For example, if there is a T upstream and a C or T downstream then the frequency of C[C>G]A variants is reduced by a factor of more than 2. Similar results are seen for G[C>A]A variants if there is a C upstream and a C or G downstream. Conversely, TT[T>A]AA and CA[T>C]TG variants have a four-fold higher frequency than the mean for T[T>A]A and A[T>C]T variants. For A[T>C]T variants, having a G downstream and a G or T upstream more than doubles the frequency. Results for all pentanucleotide backgrounds are presented in Supplementary Table 1.

Discussion

Unless they occur as recurrent mutations, DNMs in occurring in research participants within a sample will appear as singleton variants and it might seem tempting to use singleton variants to gain information about mutation rates. However in this large cohort each participant carries around 900 singleton variants and is expected to have only 60 or so DNMs, so the vast majority of singleton variants consist of rare variants segregating in the population. Moreover, as discussed previously, singleton counts can be reduced for variant types with high mutation rates so that overall the counts of singleton variants do not accurately reflect mutation probabilities. Instead, we show that if one considers variants each of which has an allele count of 25, rather than 1, then one finds that the frequencies of the AC25 variant types are very highly correlated with the frequencies of DNMs directly observed in a previously reported, smaller sample.

The large sample size allows for the study of sequence-based effects on mutation probabilities in a normal population. Studying the rates of different mutation types in spermatozoa provides information mainly concerning errors occurring during spermatogenesis (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, AC25 variants observed in a large population sample are expected to reflect mutational events which have occurred 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 AC25 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.

We show that the counts of these variants, which we take to reflect mutation probabilities, when considered in a trinucleotide context can be very accurately modelled using a small number of mutational signatures obtained from analysing somatic mutations in cancer tissue. The signature with the largest contribution, SBS5, has been described as clock-like, in that number of these mutations increases with the age of the individual, and may be due to DNA damage to adenine (Alexandrov et al., 2020). Likewise the signature with the second largest contribution, SBS1, is also clock-like and may reflect deamination of methylcytosine. Comparing these variant frequencies with those observed in cancer will hopefully facilitate a fuller understanding of pathways to carcinogenesis.

We replicate the previously reported finding that the frequency of C>T variants is much increased in the CpG context but we also demonstrate how other aspects of the context can have quite marked

effects on variant frequency and hence, we assume, mutation probability (Zhao and Boerwinkle, 2002). Although the trinucleotide context largely predicts mutation probability, we demonstrate that in some situations the fully pentanucleotide context has relevance and we see examples where the more distant bases have two-fold or higher impacts. Investigations into why these specific bases have the effects that they do could lead to a better understanding of the molecular mechanisms involved in generating these mutations.

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.

Altemose, N., Logsdon, G.A., Bzikadze, A. V., Sidhwani, P., Langley, S.A., Caldas, G. V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J., Sauria, M.E.G., Borchers, M., Gershman, A., Mikheenko, A., Shepelev, V.A., Dvorkina, T., Kunyavskaya, O., Vollger, M.R., Rhie, A., McCartney, A.M., Asri, M., Lorig-Roach, R., Shafin, K., Lucas, J.K., Aganezov, S., Olson, D., de Lima, L.G., Potapova, T., Hartley, G.A., Haukness, M., Kerpedjiev, P., Gusev, F., Tigyi, K., Brooks, S., Young, A., Nurk, S., Koren, S., Salama, S.R., Paten, B., Rogaev, E.I., Streets, A., Karpen, G.H., Dernburg, A.F., Sullivan, B.A., Straight, A.F., Wheeler, T.J., Gerton, J.L., Eichler, E.E., Phillippy, A.M., Timp, W., Dennis, M.Y., O'Neill, R.J., Zook, J.M., Schatz, M.C., Pevzner, P.A., Diekhans, M., Langley, C.H., Alexandrov, I.A., Miga, K.H. (2022) Complete genomic and epigenetic maps of human centromeres. Science (1979). 376.

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.

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.

Halldorsson, B. V., Palsson, G., Stefansson, O.A., Jonsson, H., Hardarson, M.T., Eggertsson, H.P., Gunnarsson, B., Oddsson, A., Halldorsson, G.H., Zink, F., Gudjonsson, S.A., Frigge, M.L., Thorleifsson, G., Sigurdsson, A., Stacey, S.N., Sulem, P., Masson, G., Helgason, A., Gudbjartsson, D.F., Thorsteinsdottir, U., Stefansson, K. (2019) Human genetics: Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science (1979). 363.

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.

Kimura, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61, 893–903.

Liao, K., Carlson, J., Zöllner, S. (2023) The effect of mutation subtypes on the allele frequency spectrum and population genetics inference. G3 Genes|Genomes|Genetics 13, 35.

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.

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.

Wakeley, J., Fan, W.T.L., Koch, E., Sunyaev, S. (2023) Recurrent mutation in the ancestry of a rare variant. Genetics 224, iyad049.

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

Mutational signatures producing the best fit to observed counts of trinucleotide AC25 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
SBS1	0.178	Spontaneous or enzymatic deamination of 5-methylcytosine (clock-like signature)
SBS5	0.486	Unknown (clock-like signature)
SBS6	0.061	Defective DNA mismatch repair
SBS10b	0.061	Polymerase epsilon exonuclease domain mutations
SBS54	0.050	Possible sequencing artefact
SBS87	0.107	Thiopurine chemotherapy treatment
SBS97	0.057	Unknown

Table 2

The table shows frequencies for the six different single nucleotide variants having C or T as the background (reference) base relative to the counts of the reference base.

Substitution	Frequencies of AC25 variants
C>A	0.000235
C>G	0.000202
C>T	0.000922
T>A	0.000106
T>C	0.000415
T>G	0.000125

Table 3

Table showing the effect of the full pentanucleotide sequence versus the central trinucleotide sequence on the frequencies of AC25 variants. Results are shown for the ten variants with T or C as the central base for which the frequency is most strongly reduced or increased.

Variant	OR	(0.95 CI)
TC[C>G]AC	0.335	(0.280 - 0.400)
CT[T>A]AG	0.368	(0.359 - 0.377)
TC[C>G]AT	0.386	(0.335 - 0.445)
CG[C>A]AC	0.395	(0.289 - 0.540)
GC[C>A]CG	0.420	(0.339 - 0.519)
CG[C>A]AG	0.424	(0.326 - 0.552)
TC[C>A]CA	0.429	(0.397 - 0.463)
CT[C>A]TG	0.431	(0.392 - 0.474)
GC[C>G]AC	0.435	(0.373 - 0.508)
TC[C>G]GC	0.437	(0.359 - 0.532)
CC[C>A]TC	2.437	(2.281 - 2.603)
GA[T>G]AG	2.486	(2.269 - 2.724)
CC[T>G]TC	2.496	(2.338 - 2.665)
CG[T>C]TG	2.531	(2.282 - 2.808)
TA[T>C]TG	2.586	(2.508 - 2.667)
AA[T>C]TG	2.795	(2.715 - 2.877)
GA[T>C]TG	2.954	(2.852 - 3.060)
CC[C>A]CC	3.267	(3.100 - 3.443)
CA[T>C]TG	4.000	(3.893 - 4.110)
TT[T>A]AA	4.517	(4.303 - 4.741)

Figure 1

Results of modelling AC25 SBS counts using reference mutational signatures obtained from studies of somatic mutations observed in cancers.

Figure 1A

Observed counts of AC25 SBSs on trinucleotide backgrounds.

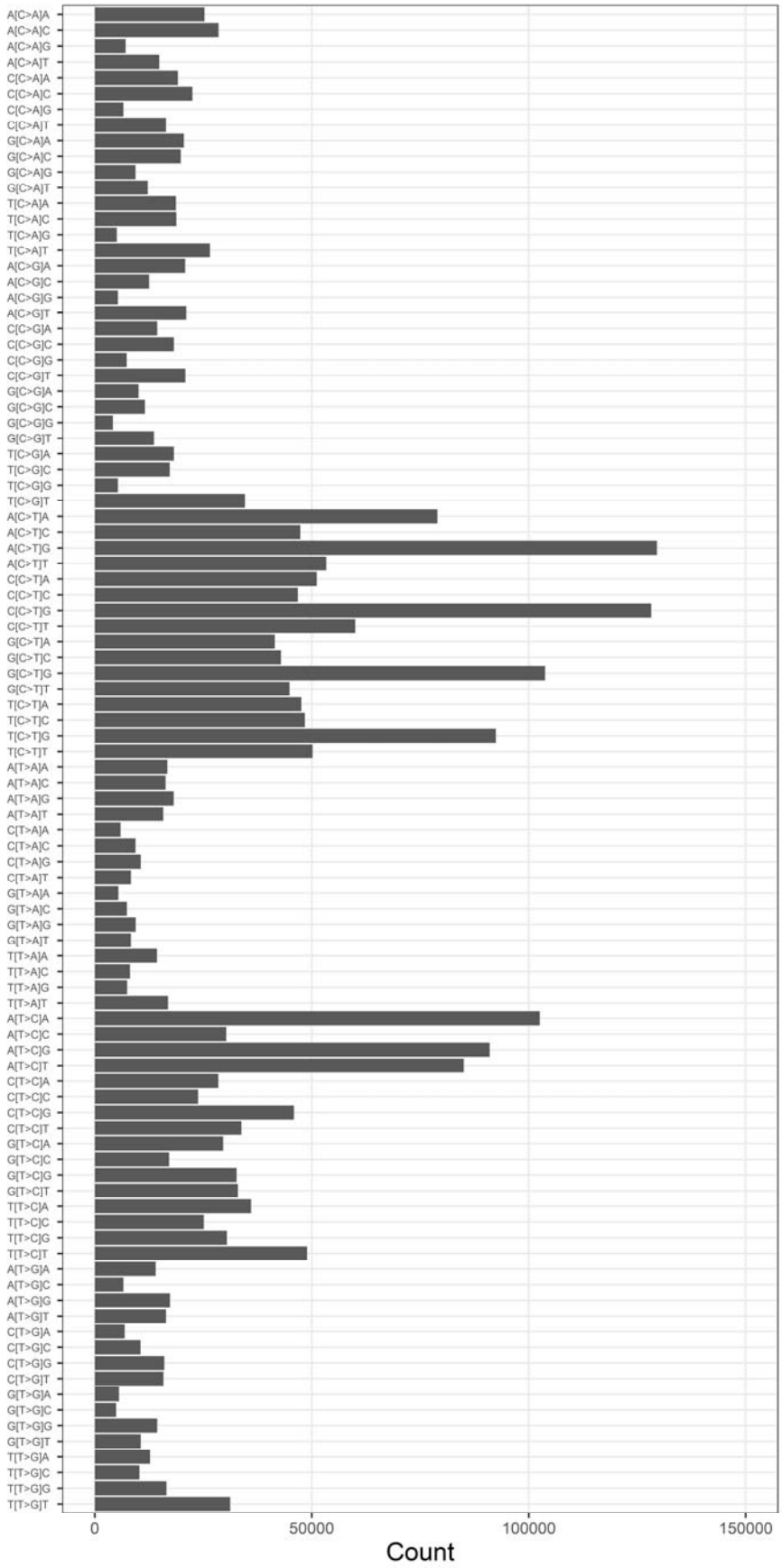

A[C>A]A
A[C>A]C
A[C>A]G
A[C>A]T
C[C>A]A
C[C>A]C
C[C>A]G
C[C>A]T
G[C>A]A
G[C>A]C
G[C>A]G
G[C>A]T
T[C>A]A
T[C>A]C
T[C>A]G
T[C>A]T
A[C>G]A
A[C>G]C
A[C>G]G
A[C>G]T
C[C>G]A
C[C>G]C
C[C>G]G
C[C>G]T
G[C>G]A
G[C>G]C
G[C>G]G
G[C>G]T
T[C>G]A
T[C>G]C
T[C>G]G
T[C>G]T
A[C>T]A
A[C>T]C
A[C>T]G
A[C>T]T
C[C>T]A
C[C>T]C
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
50000
100000
150000
Count

Figure 1B

Predicted AC25 SBS counts using best-fitting combination of seven reference mutational signatures.

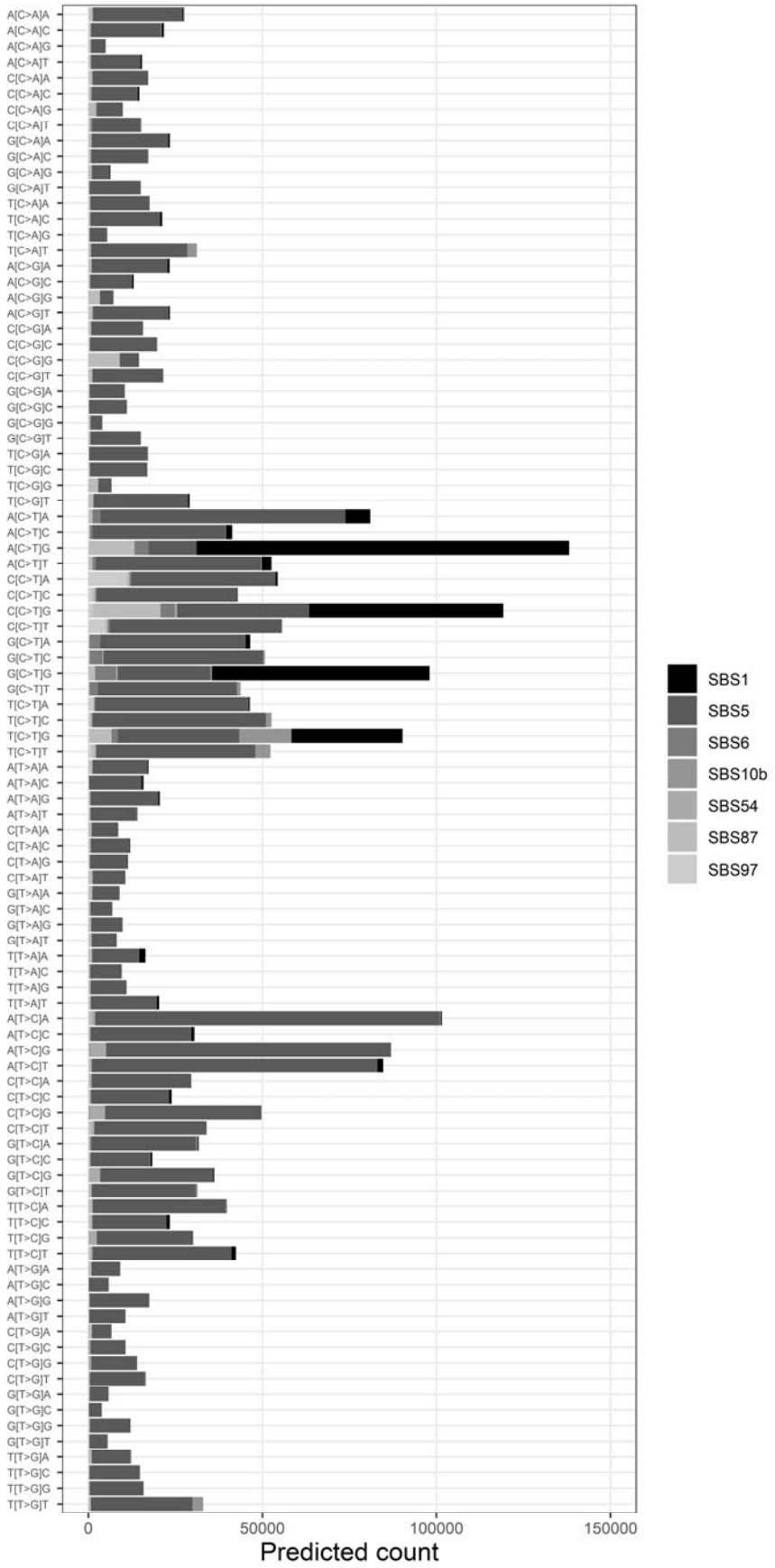

A[C>A]A
A[C>A]C
A[C>A]G
A[C>A]T
C[C>A]A
C[C>A]C
C[C>A]G
C[C>A]T
G[C>A]A
G[C>A]C
G[C>A]G
G[C>A]T
T[C>A]A
T[C>A]C
T[C>A]G
T[C>A]T
A[C>G]A
A[C>G]C
A[C>G]G
A[C>G]T
C[C>G]A
C[C>G]C
C[C>G]G
C[C>G]T
G[C>G]A
G[C>G]C
G[C>G]G
G[C>G]T
T[C>G]A
T[C>G]C
T[C>G]G
T[C>G]T
A[C>T]A
A[C>T]C
A[C>T]G
A[C>T]T
C[C>T]A
C[C>T]C
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
50000
100000
150000
Predicted count
SBS1
SBS5
SBS6
SBS10b
SBS54
SBS87
SBS97

Figure 2

Effects of different trinucleotide sequences on the frequencies of different AC25 variants. Results are shown for all trinucleotides having either C or T as the central base.

Figure 2A

Frequencies of all AC25 variants occurring in the context of different trinucleotide backgrounds. The frequency is defined as the count of the number of times each AC25 variant is observed divided by the count of the number of times the background is seen in the reference genome.

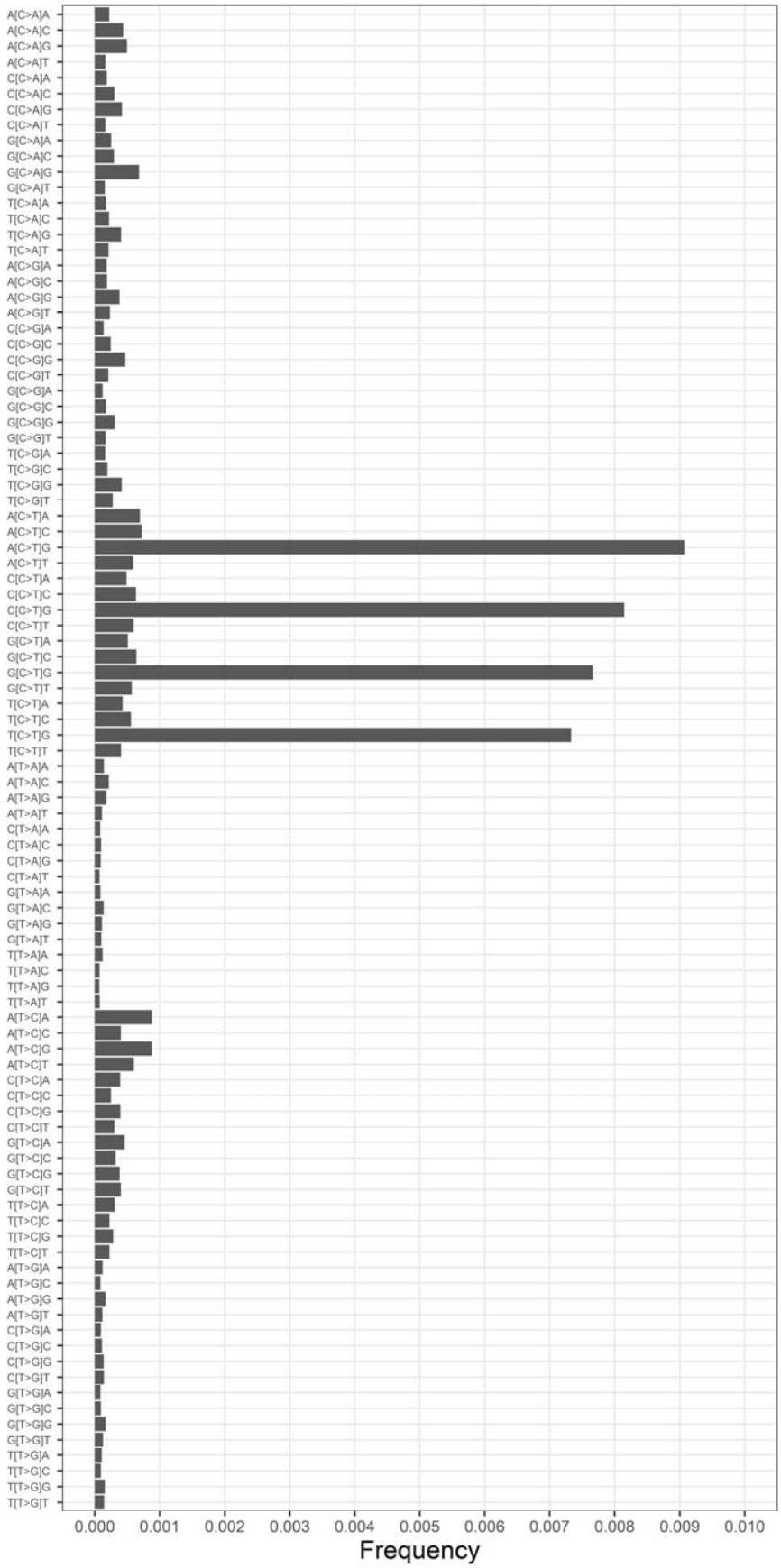
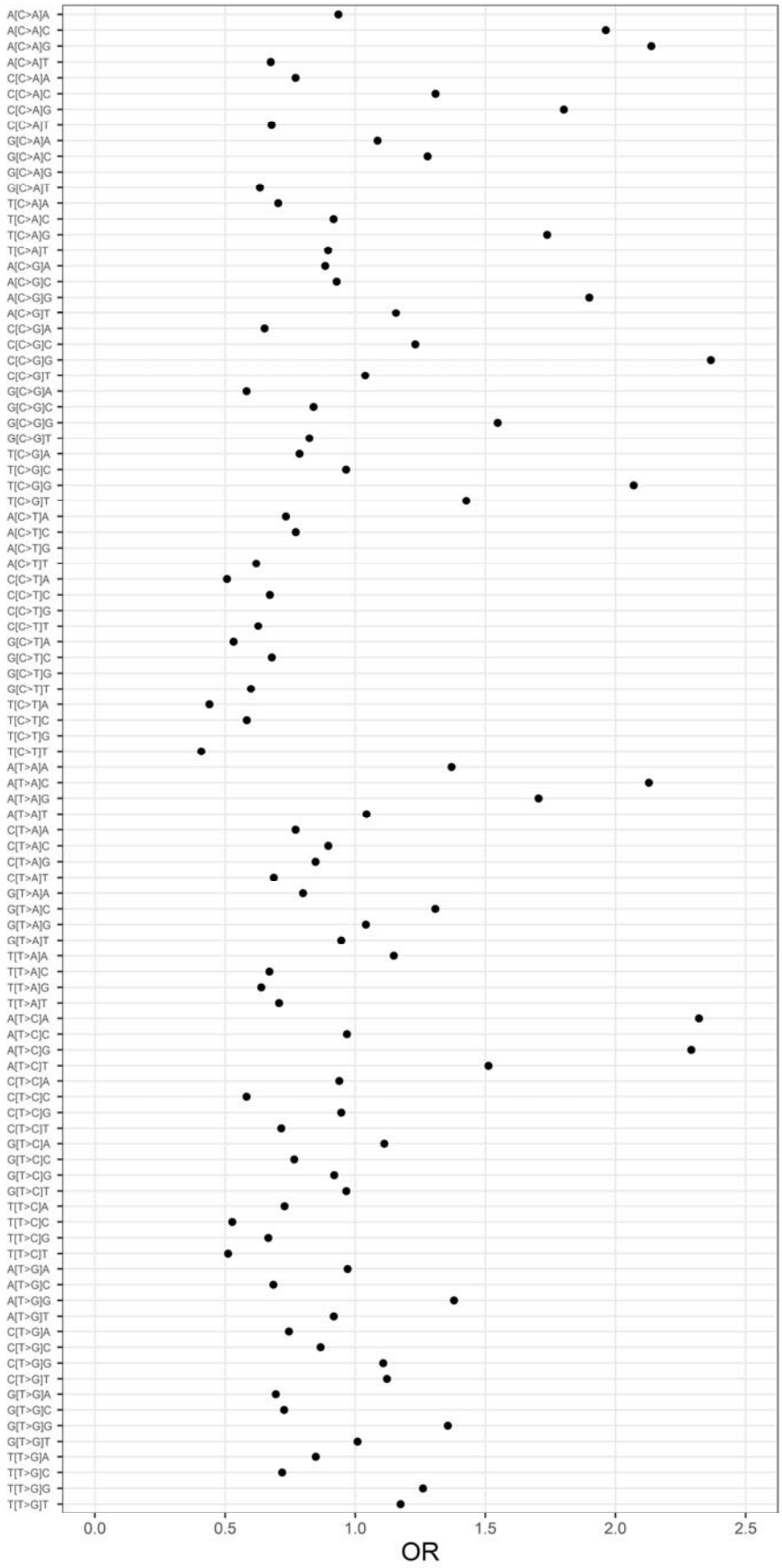

A[C>A]A
A[C>A]C
A[C>A]G
A[C>A]T
C[C>A]A
C[C>A]C
C[C>A]G
C[C>A]T
G[C>A]A
G[C>A]C
G[C>A]G
G[C>A]T
T[C>A]A
T[C>A]C
T[C>A]G
T[C>A]T
A[C>G]A
A[C>G]C
A[C>G]G
A[C>G]T
C[C>G]A
C[C>G]C
C[C>G]G
C[C>G]T
G[C>G]A
G[C>G]C
G[C>G]G
G[C>G]T
T[C>G]A
T[C>G]C
T[C>G]G
T[C>G]T
A[C>T]A
A[C>T]C
A[C>T]G
A[C>T]T
C[C>T]A
C[C>T]C
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
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
0.010
Frequency

Figure 2B

Odds ratio (OR) for each AC25 variant associated with its trinucleotide background, defined as the frequency with which the variant occurs on this background compared with the overall frequency of AC25 single base substitutions of this type (e.g. frequency of A[T>A]C versus frequency of T>A). The results for X[C>T]G variants are not shown because the ORs are too high.


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