# An analysis of intrinsic base substitution patterns in diverse species reveals connections to cancer and metabolism

Kin Chan*[1], Kassidy N. Doherty[2,3], and Salma Alasmar[2,4]

[1] Department of Biochemistry, Microbiology and Immunology, and Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Canada

[2] Biopharmaceutical Sciences Undergraduate Program, University of Ottawa

Current affiliations:

[3] Common Law Section, University of Ottawa, Ottawa, Canada

[4] Department of Chemistry and Biomolecular Sciences, University of Ottawa

* Corresponding Author: kin.chan@uottawa.ca

**Abstract**

**Mutations in genomic DNA are the starting material for evolution by natural selection. Analyses of large-scale sequencing data have revealed that mutagenic processes often create distinctive patterns of base substitutions, called mutational signatures. In this work, we analyzed the base substitution patterns within mutational data from a large number of strains or individuals among seven model species (totalling >784 million mutations), as well as single nucleotide polymorphisms (SNPs) in 41 species from the NCBI SNP database (>603 million SNPs). We found that the intrinsic base substitution pattern for most of these species closely matches mutational signature 5 from the Catalog of Somatic Mutations in Cancer (abbreviated as COSMIC). Signature 5 is ubiquitous in cancers and normal human cells, suggesting that the similar patterns of mutation across many species are likely due to conserved biochemical processes. Finally, we show that a similar pattern of base substitutions can be obtained using a yeast model system that allows controlled generation of genomic single-stranded DNA. Taken together, we propose that intrinsic biochemical processes in cells are coupled to the continuous generation of mutations, which in turn, are acted upon by natural selection to drive the evolution of species.**

Mutagenesis is often thought of as a stochastic process, wherein the temporal and positional occurrence of each individual mutational event is not predictable (*1*). And yet, sets of mutations induced by a given DNA damaging agent or process do create reproducible, characteristic patterns, called mutational signatures (*2*). Mutational signatures derived from cancer genomics data have been published, which have yielded considerable insight into specific sources of DNA damage in humans (*3*). These signatures provide a composite picture of the recurrent patterns of base substitutions found in cancers, and in some cases, normal tissues as well. The observations in humans led us to investigate whether similar patterns of base substitutions are found when comparing genetic variation within and between other species.

We analyzed mutational data (784,102,480 base substitutions) from seven model species, comprising: 1,011 varieties of budding yeast (*Saccharomyces cerevisiae*) (*4*); 161 varieties of fission yeast (*Schizosaccharomyces pombe*) (*5*); 152 isolates of roundworm (*Caenorhabditis elegans*) (*6*); 85 isolates of fruit fly (*Drosophila melanogaster*) (*7*); 36 strains of mouse (*Mus musculus*) (*8*); 42 strains of rat (*Rattus norvegicus*) (*9*); and 49 individual dogs (*Canis familiaris*) (*10*). When the data were plotted using the convention for reporting mutational signatures (six substitution types, each subdivided into 16 contexts with one flanking base on either side, 5′ and 3′) (*11*), it became clear that the majority of mutations in all seven species were either C→T or T→C transitions (see Figure 1A). To assess the similarity of these base substitution patterns to known

mutational signatures, we computed the cosine similarity to the set of published COSMIC signatures (see Figure 1B) (*2*). For context, a cosine similarity value of 1 is obtained when comparing identical signatures. The substitution patterns for all seven model species were each most similar to COSMIC signature 5 (referred to hereafter as COSMIC 5), which is one of two 'clock-like' mutational signatures found in human, as the numbers of such mutations essentially always increase monotonically with age (*12*). COSMIC 5 also accounts for ~75% of de novo mutations in healthy human children (*13*). These results show that, just as COSMIC 5 is found ubiquitously in human, intrinsic base substitution patterns similar to COSMIC 5 are found in multiple model species.

Given that COSMIC 5-like base substitution patterns are intrinsic to the model species, we then investigated if similar patterns are observed among SNPs in a larger, more diverse set of species. We downloaded sets of available VCF (variant call format) files from NCBI SNP for 44 species (*14*). Two data sets were obviously invalid, as they completely lacked two substitution types (C→G and T→A, likely due to a data processing error), while another data set was dropped from analysis because of probable sequencing artefacts (see Supplementary Figure 1). Among the remaining 41 species (yielding 603,722,210 SNPs, see Supplementary Figure 2), the large majority showed the highest cosine similarity to COSMIC 5 (see Figure 2). A particularly interesting exception is the mosquito (*Anopheles gambiae*), with the highest similarity to COSMIC signature 3, which is attributed to deficiency in homologous recombination DNA repair

(*3*). In contrast to COSMIC 5, the mosquito base substitution pattern is relatively 'flat' across substitutions and trinucleotide contexts (see Supplementary Figure 2). The bee (*Apis mellifera*) and the grape (*Vitis vinifera*) are also notable outliers whose intrinsic base substitution patterns do not particularly resemble any COSMIC signature. Such exceptions notwithstanding, intrinsic base substitution patterns that resemble COSMIC 5 do tend to be shared among collections of SNPs in many other organisms. Taken altogether, the observations from prior studies in human and our results here suggest the likelihood that conserved biochemical processes in many species are responsible for the set of COSMIC 5-like intrinsic base substitution patterns.

To quantify and visualize the extent of similarity or divergence among base substitution patterns, we generated profiles showing means and 95% confidence intervals of relative frequency for each mutation type (see Figure 3A). For the seven model species, the 95% confidence intervals for most substitution types were small, i.e., indicating low variance (see Supplementary Figure 3A). The most highly variable trinucleotides with respect to differences in frequencies of mutations were TTA and ATT for T→A transversions (see also Figure 1A). Frequencies of C→T at ACG and CCG motifs are also quite varied, in parallel with T→C at ATG and CTG. These motifs were particularly mutable in rat and dog, but less so in the other model species. Despite these instances of (relatively) high variability, the mean values do provide an estimated consensus COSMIC 5-

like pattern among the seven model species. The cosine similarity between this consensus pattern and COSMIC 5 itself is 0.928 (see Supplementary Figure 3B).

Likewise, we applied this analysis to the SNP data for the 41 species from the NCBI database (see Figure 3B). The 95% confidence intervals for the means were more compact, reflecting the larger number of species examined and yielding a lower variance. Again, the substitution types with higher mean relative frequencies showed more variability (see Supplementary Figure 3C). The consensus pattern from SNP data very closely matched the corresponding pattern from mutations among the seven model organisms, with a cosine similarity of 0.981 (see Figure 3C). All four NCG motifs exhibit more common C→T transitions in the SNP data, suggestive of possible contributions from deamination of 5-methylcytosine at those motifs in a subset of species (e.g., goat, horse, macaque, sheep, etc.), similar to COSMIC signature 1 (*3*). Also similar to the model organism data, the consensus SNP pattern has a cosine similarity of 0.931 compared to COSMIC 5 (see Supplementary Figure 3D). Taken together, we propose that for many species, their respective sets of intrinsic base substitutions form a range of related patterns, and COSMIC 5 is simply the corresponding human counterpart in this continuum. Human is a special case in the sense that our species' intrinsic base substitution pattern also includes a ~25% contribution of COSMIC signature 1 (*13*, *15*). If there is a COSMIC 1-like component to the intrinsic base substitution patterns of other species as well, then it is comparatively minor and outweighed by the COSMIC 5-like contribution.

In principle, many species can be used as models to study the molecular mechanisms underlying these COSMIC 5-like mutational patterns, as these biochemical mechanisms are likely to be well conserved. We previously used a budding yeast system which features controlled generation of genomic single-stranded DNA (ssDNA) to elucidate the relative contributions of the APOBEC3A and APOBEC3B cytidine deaminases to cancer mutagenesis (*16*). This system relies on the *cdc13-1* temperature sensitive mutation (*17*), so that shifting to 37°C causes telomere uncapping, extensive end resection, and cell cycle arrest in $G_2$ via the DNA damage checkpoint (*18*). Since the exposed ssDNA is much more prone to damage than double-stranded DNA, this system can acquire mutations for analysis much more readily. When we sequenced the genomes that were allowed to accumulate mutations, we obtained a pattern of base substitutions that resembled the consensus base substitution pattern in the seven model species, the consensus pattern from SNPs of 41 species, and COSMIC 5 (cosine similarities = 0.941, 0.931, and 0.933, respectively, see Figure 4A and Supplementary Figure 4). This further underscores the generality of COSMIC 5-like intrinsic base substitution patterns.

We then carried out genetic experiments to investigate molecular mechanisms responsible for the mutational pattern. Mutagenesis is dependent on the translesion DNA synthesis (TLS) polymerases zeta and Rev1 (*19*), i.e., on the error-prone bypass of damaged bases (*20*) (see Figure 4B). In contrast, deletion of *RAD30* (encoding polymerase eta) had no effect, since Rad30 specializes in

error-free bypass of ultraviolet light-induced cyclobutane dimer photoproducts (*21*) and apparently plays no role in mutagenesis opposite DNA damage that gives rise to the COSMIC 5-like pattern. Deletion of genes important for error-free bypass also had no significant effect (see Supplementary Figure 5). Further, we found that energy metabolism is key for this form of mutagenesis. Cells maintained in glucose (a fermentative, glycolytic energy source (*22*)) acquired mutations most frequently, while cells in acetate (a carbon source that must be utilized by aerobic respiration (*23*)) had the fewest acquired mutations (see Figure 4C). Moreover, increasing the concentration of glucose from 2% (the amount normally used in media) to 8% led to a five-fold increase in the frequency of mutations (see Figure 4D). Finally, blocking the last step of glycolysis using auxin-induced degron alleles of *CDC19* (encoding pyruvate kinase) caused a large decrease in mutagenesis (see Figure 4E), further corroborating the central role of energy metabolism to generate the DNA damage which produces a COSMIC 5-like base substitution pattern.

Taken altogether, our results suggest that the intrinsic base substitution patterns of many organisms are linked to energy metabolism. This would provide a plausible explanation for why such similar patterns are found across many species, including human, since the main pathways of energy metabolism are well conserved (*24*). This model would also imply that processes of energy metabolism will inevitably produce DNA damage, which leads to mutations that fit a similar pattern across many species. In turn, these mutations are the starting

material for evolution by natural selection. Since energy generation is a prerequisite for living, the continuous basal level of mutagenesis which results would ensure that natural selection will have an ever-renewing repertoire of genetic variants to act upon. We propose that the DNA damage itself is likely very similar between species, but the resulting patterns will be more or less divergent depending on the similarity of the respective sets of error-prone translesion synthesis polymerases when comparing between species. Species with polymerases that are more closely related would have more similar patterns than those from less related species. Finally, a coupling between metabolism and COSMIC 5-like mutational patterns could help explain why the extent of that type of mutagenesis can be highly variable in cancers, when comparing between different patients (*12*) or temporally within the same patient (*25*): namely, such variability could be linked to differences in energy metabolism between individuals or over the course of disease progression in particular individuals.

## Materials and Methods

### Data Acquisition

VCF files were downloaded from the following sources: for *Saccharomyces cerevisiae*, http://1002genomes.u-strasbg.fr/files/; for *Schizosaccharomyces pombe*, ftp://ftp.pombase.org/pombe; for *Caenorhabditis elegans*, https://www.elegansvariation.org/data/release/20160408; for *Drosophila melanogaster*, https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP050151; for *Mus musculus*, ftp://ftp-mouse.sanger.ac.uk/; for *Rattus norvegicus*, ftp://ftp.rgd.mcw.edu/pub/strain_specific_variants/Hermsen_et_al_40Genomes_Variants/; for *Canis familiaris*, http://www.dog10kgenomes.org/; and for NCBI SNP data, ftp://ftp.ncbi.nih.gov/snp/organisms/archive/.

### Mutational Analyses

Where possible, we used the MutationalPatterns R package (version 1.10.0) to extract mutational and trinucleotide contexts from VCF files, to plot mutational signatures, to calculate cosine similarities, and to plot heatmaps (*11*). We constructed BSgenome (*26*) reference genome packages in order to use MutationalPatterns to process VCF files if none of the stock reference genomes were appropriate. In cases where an organism's reference genome proved to be incompatible with BSgenome (or more precisely, its dependency GenomeInfoDB (*27*)), we used a script that extracted mutational and trinucleotide context data. This script used a combination of Unix shell functions, R (*28*, *29*), and Samtools (*30*). Summed matrices of mutational counts were loaded into

MutationalPatterns for graphics rendering in order to produce a consistent style for figures. Most analyses were run on a PC operating Ubuntu (Linux) version 18.04, with the remaining analyses run on a MacBook Pro operating macOS version 10.14.5 or 10.14.6. R version 3.6.0 was used on both computing platforms.

Reagents

Bacto peptone and yeast extract were purchased from Becton, Dickinson and Company (Franklin Lakes, New Jersey). Auxin and canavanine were purchased from MilliporeSigma (St. Louis, Missouri). Agar, carbon sources (i.e., glucose, galactose, raffinose, and potassium acetate), hygromycin, and G418 were purchased from BioBasic (Markham, Ontario).

Yeast Genetics

All yeast strains were constructed in the ySR127 background, a MAT alpha haploid with the *cdc13-1* temperature sensitive mutation. In addition, there is a cassette of three reporter genes (*CAN1*, *URA3*, and *ADE2*) near the de novo left telomere of chromosome V. Further details on ySR127 have been described previously (*31*).

Gene knockouts were constructed by one step gene replacement using antibiotic resistance marker cassettes (*32*). To construct the CDC19 degron strains, we first introduced the *Oryza sativa* TIR1 gene cassette (OsTIR1) into the HO locus of ySR127, using the protocol described in (*33*). We then inserted the 3mAID degron tag (containing three tandem repeats of the core domain of the IAA17 protein from *Arabidopsis thaliana*), using the protocol described in (*34*).

Mutagenesis experiments were initiated by inoculating single colonies each into 5 mL of rich media (2% Bacto peptone, 1% Bacto yeast extract, 2% carbon source, supplemented with 0.001% adenine hemisulfate) in round bottom glass tubes. Cells were grown at permissive temperature (23°C) for three days. Then, cells were diluted ten-fold into fresh media and shifted to restrictive temperature (37°C). Auxin (iodoacetic acid) was added to 1 mM final concentration for degron experiments. After the temperature shift, cells were collected by light centrifugation, washed in water, and plated onto synthetic complete media to assess survival and onto canavanine-containing media to select for mutants. Further details of this procedure are described in (*35*). Genomic DNA from mutants of interest were extracted and sequenced essentially as described previously (*16*). Illumina sequencing was outsourced to Genome Quebec (McGill University). We used Bowtie 2 (*36*), SAMtools (*30*), and bcftools (*37*) to map reads and call variants.

Figure Legends

Figure 1. (A) 96-bar graphs showing the intrinsic base substitution patterns of seven model species, derived by analysis of data from (*4–10*). The patterns tended to show mostly C→T or T→C substitutions. Interestingly, both invertebrate species and fission yeast had prominent T→A peaks at ATT and TTA motifs. (B) Heatmap showing cosine similarity values to 30 signatures from COSMIC. For all seven species, the highest cosine similarity values suggested closest resemblance to COSMIC signature 5.

Figure 2. Heatmap depicting cosine similarity values when comparing sets of SNPs from 41 species vs. the 30 COSMIC signatures. Most species had the highest cosine similarity values compared to COSMIC signature 5. See Supplementary Figure 2 for the 96-bar substitution patterns of these species.

Figure 3. (A) 96-bar graph showing mean and 95% confidence interval for relative frequencies of each substitution type among the seven model species. (B) 96-bar graph showing mean and 95% confidence interval for relative frequencies of each SNP type among the 41 species from NCBI. (C) The mean values of the two profiles are shown, along with the difference in relative frequencies when one is subtracted from the other. The cosine similarity is quite high (0.981).

Figure 4. (A) 96-bar graph showing the base substitution pattern resulting from damage to exposed single-stranded DNA in yeast. (B) The inactivation frequency of the *CAN1* reporter was dependent on *REV3* (encoding the catalytic subunit of polymerase zeta) and *REV1*, i.e., on error-prone translesion DNA

synthesis. Deletion of *RAD30* (encoding polymerase eta) had no effect. Cells were temperature shifted to 37°C for 6 hours. (C) The frequency of mutagenesis was dependent on carbon source. Acetate (which must be metabolized by aerobic respiration) produced the fewest mutations. Glucose (which is metabolized fermentatively) produced the most. Galactose and raffinose (which have a mixed 'respirofermentative' metabolism were intermediate. (D) Increasing glucose concentration to 8% led to a significant increase in mutagenesis within single-stranded DNA. Controls at 23°C did not show any significant differences throughout the tested range of glucose concentrations. Cells were shifted to 37°C for 24 hours. (E) Mutagenesis was dependent on completion of glycolysis. Inducible degradation of Cdc19 (pyruvate kinase) blocked the last step of the glycolytic pathway and caused over a ten-fold decrease in mutagenesis. Cells were shifted to 37°C for 24 hours in the presence or absence of 1 mM auxin.

Supplementary Figure 1. (A) Analysis of the SNP dataset for apple (*Malus domestica*) revealed an apparent data processing issue that deleted all C→G and T→A substitutions. (B) The SNP dataset for flycatcher (*Ficedula albicollis*) was also affected by this issue. (C) The SNP pattern for *Nomascus leucogenys* (the northern white-cheeked gibbon, but referred to simply as "primates" on NCBI) was also suspiciously unusual with very high peaks at the reciprocal motifs ACA and ATA, as well as ACG and ATG, for both C→T and T→C substitutions. As this was very much an outlier, this dataset was dropped for further consideration due to suspected sequencing artefact.

Supplementary Figure 2. 96-bar profiles showing the intrinsic base substitution patterns inferred from SNPs of 41 species downloaded from NCBI. Most patterns showed, to varying extents, a preponderance of C→T and T→C substitutions where these substitution types were similar in overall proportion, forming a range of COSMIC 5-like patterns. Exceptions to this overarching meta-pattern were mosquito, bee, grape, and corn.

Supplementary Figure 3. (A) Scatter plot of 95% confidence interval sizes vs. their corresponding mean relative frequency for the 96 substitution types shown in Figure 3A (model species). Most of the variance among the patterns for the seven model species was found among substitution types that have over 0.01 relative frequency. (B) Comparison of the mean relative frequencies among the seven model species to COSMIC 5. Root sum of squares (RSS) and cosine similarity values are shown. (C) Scatter plot of 95% confidence interval sizes vs. their corresponding mean relative frequency for the 96 substitution types shown in Figure 3B (NCBI SNP data). Notice the 95% confidence intervals are smaller than in (A) because of the larger number of species among the SNP data analysis. (D) Comparison of the mean relative frequencies among the 41 species with NCBI SNP data to COSMIC 5.

Supplementary Figure 4. Comparison of the base substitution pattern observed in the yeast single-stranded DNA vs.: (A) the consensus pattern from seven model species; (B) the consensus pattern from the 41 species from NCBI

1 SNP; and (C) COSMIC 5. Root sum of squares (RSS) and cosine similarity val-

2 ues are shown as well.

3 Supplementary Figure 5. Deletion of genes involved in homologous re-

4 combination (*RAD51*, *RAD52*) or template switching to bypass base lesions

5 (*RAD5*, *RAD18*, *MMS2*) had no significant effect on mutagenesis after correction

6 for false discovery from multiple testing.

References


1. N. Timoféeff-Ressovsky, K. Zimmer, M. Delbrück, The nature of genetic mutations and structure of the gene. *Nachrichten Aus Biol. Ges. Wiss. Gött.* **1**, 189–245 (1935).

2. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, M. R. Stratton, Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3**, 246–259 (2013).

3. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Borresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjord, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinsk, N. Jager, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. Lopez-Otin, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdes-Mas, M. M. van Buuren, L. van /'t Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P. Andrew Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, Signatures of mutational processes in human cancer. *Nature*. **500**, 415–421 (2013).

4. J. Peter, M. De Chiara, A. Friedrich, J.-X. Yue, D. Pflieger, A. Bergström, A. Sigwalt, B. Barre, K. Freel, A. Llored, C. Cruaud, K. Labadie, J.-M. Aury, B. Istace, K. Lebrigand, P. Barbry, S. Engelen, A. Lemainque, P. Wincker, G. Liti, J. Schacherer, Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature*. **556**, 339–344 (2018).

5. A. Lock, K. Rutherford, M. A. Harris, J. Hayles, S. G. Oliver, J. Bähler, V. Wood, PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Res.* **47**, D821–D827 (2018).

6. D. E. Cook, S. Zdraljevic, J. P. Roberts, E. C. Andersen, CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic Acids Res.* **45**, D650–D657 (2016).

7. J. K. Grenier, J. R. Arguello, M. C. Moreira, S. Gottipati, J. Mohammed, S. R. Hackett, R. Boughton, A. J. Greenberg, A. G. Clark, Global diversity lines - a five-continent reference panel of sequenced Drosophila melanogaster strains. *G3 Bethesda Md*. **5**, 593–603 (2015).

8. T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellåker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assunção, L. R. Donahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Birney, J. Flint, D. J. Adams, Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*. **477**, 289 (2011).

9. R. Hermsen, J. de Ligt, W. Spee, F. Blokzijl, S. Schäfer, E. Adami, S. Boymans, S. Flink, R. van Boxtel, R. H. van der Weide, T. Aitman, N. Hübner, M. Simonis, B. Tabakoff, V. Guryev, E. Cuppen, Genomic landscape of rat strain and substrain variation. *BMC Genomics*. **16**, 357–357 (2015).

10. E. A. Ostrander, G.-D. Wang, G. Larson, B. M. vonHoldt, B. W. Davis, V. Jagannathan, C. Hitte, R. K. Wayne, Y.-P. Zhang, Dog10K Consortium, Dog10K: an international sequencing effort to advance studies of canine domestication, phenotypes and health. *Natl. Sci. Rev.* (2019), doi:10.1093/nsr/nwz049.

11. F. Blokzijl, R. Janssen, R. van Boxtel, E. Cuppen, MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).

12. L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, M. R. Stratton, Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402 (2015).

13. R. Rahbari, A. Wuster, S. J. Lindsay, R. J. Hardwick, L. B. Alexandrov, S. Al Turki, A. Dominiczak, A. Morris, D. Porteous, B. Smith, M. R. Stratton, UK10K Consortium, M. E. Hurles, Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126 (2015).

14. A. Kitts, L. Phan, M. Ward, J. B. Holmes, *The Database of Short Genetic Variation (dbSNP)* (National Center for Biotechnology Information (US), 2014; https://www.ncbi.nlm.nih.gov/books/NBK174586/).

15. Y. S. Ju, I. Martincorena, M. Gerstung, M. Petljak, L. B. Alexandrov, R. Rahbari, D. C. Wedge, H. R. Davies, M. Ramakrishna, A. Fullam, S. Martin, C. Alder, N. Patel, S. Gamble, S. O'Meara, D. D. Giri, T. Sauer, S. E. Pinder, C. A. Purdie, Å. Borg, H. Stunnenberg, M. van de Vijver, B. K. T. Tan, C. Caldas, A. Tutt, N. T. Ueno, L. J. van 't Veer, J. W. M. Martens, C. Sotiriou, S. Knappskog, P. N. Span, S. R. Lakhani, J. E. Eyfjörd, A.-L. Børresen-Dale, A. Richardson, A. M. Thompson, A. Viari, M. E. Hurles, S. Nik-Zainal, P. J.

Campbell, M. R. Stratton, Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*. **543**, 714 (2017).

16. K. Chan, S. A. Roberts, L. J. Klimczak, J. F. Sterling, N. Saini, E. P. Malc, J. Kim, D. J. Kwiatkowski, D. C. Fargo, P. A. Mieczkowski, G. Getz, D. A. Gordenin, An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*. **47**, 1067–1072 (2015).

17. L. H. Hartwell, R. K. Mortimer, J. Culotti, M. Culotti, Genetic Control of the Cell Division Cycle in Yeast: V. Genetic Analysis of cdc Mutants. *Genetics*. **74**, 267–286 (1973).

18. B. Garvik, M. Carson, L. Hartwell, Single-stranded DNA arising at telomeres in cdc13 mutants may constitute a specific signal for the RAD9 checkpoint. *Mol. Cell. Biol.* **15**, 6128–6138 (1995).

19. J. F. Lemontt, Mutants of yeast defective in mutation induced by ultraviolet light. *Genetics*. **68**, 21–33 (1971).

20. S. Prakash, R. E. Johnson, L. Prakash, Eukaryotic translesion synthesis DNA polymerases: specificity of structure and function. *Annu. Rev. Biochem.* **74**, 317–353 (2005).

21. R. E. Johnson, S. Prakash, L. Prakash, Efficient Bypass of a Thymine-Thymine Dimer by Yeast DNA Polymerase, Polη. *Science*. **283**, 1001 (1999).

22. R. H. Hopkins, The selective fermentation of glucose and fructose by brewer's yeast. *Biochem. J.* **22**, 1145–1156 (1928).

23. M. S. Esposito, R. E. Esposito, M. Arnaud, H. O. Halvorson, Acetate utilization and macromolecular synthesis during sporulation of yeast. *J. Bacteriol.* **100**, 180–186 (1969).

24. J. M. Peregrín-Alvarez, C. Sanford, J. Parkinson, The conservation and evolutionary modularity of metabolism. *Genome Biol.* **10**, R63 (2009).

25. S. C. Dentro, I. Leshchiner, K. Haase, M. Tarabichi, J. Wintersinger, A. G. Deshwar, K. Yu, Y. Rubanova, G. Mcintyre, I. Vazquez-Garcia, K. Kleinheinz, D. G. Livitz, S. Malikic, N. Donmez, S. Sengupta, J. Demeulemeester, P. Anur, C. Jolly, M. Cmero, D. Rosebrock, S. Schumacher, Y. Fan, M. Fittall, R. M. Drews, X. Yao, J. Lee, M. Schlesner, H. Zhu, D. J. Adams, G. Getz, P. C. Boutros, M. Imielinski, R. Beroukhim, S. C. Sahinalp, Y. Ji, M. Peifer, I. Martincorena, F. Markowetz, V. Mustonen, K. Yuan, M. Gerstung, P. T. Spellman, W. Wang, Q. D. Morris, D. C. Wedge, P. Van Loo, Portraits of genetic intra-

tumour heterogeneity and subclonal selection across cancer types. *bioRxiv* (2018), doi:10.1101/312041.

26. H. Pagès, *BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs* (https://doi.org/doi:10.18129/B9.bioc.BSgenome), *R package*.

27. S. Arora, M. Morgan, M. Carlson, H. Pagès, *GenomeInfoDb: Utilities for manipulating chromosome names, including modifying them to follow a particular naming style* (https://doi.org/doi:10.18129/B9.bioc.GenomeInfoDb), *R package*.

28. B. J. Knaus, N. J. Grünwald, vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* **17**, 44–53 (2017).

29. J. Kim, K. W. Mouw, P. Polak, L. Z. Braunstein, A. Kamburov, G. Tiao, D. J. Kwiatkowski, J. E. Rosenberg, E. M. Van Allen, A. D. D'Andrea, G. Getz, Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600 (2016).

30. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

31. K. Chan, J. F. Sterling, S. A. Roberts, A. S. Bhagwat, M. A. Resnick, D. A. Gordenin, Base Damage within Single-Strand DNA Underlies In Vivo Hypermutability Induced by a Ubiquitous Environmental Agent. *PLoS Genet*. **8**, e1003149 (2012).

32. R. J. Rothstein, in *Methods in Enzymology* (Academic Press, 1983; http://www.sciencedirect.com/science/article/pii/0076687983010150), vol. 101, pp. 202–211.

33. A. Papagiannakis, B. Niebel, E. C. Wit, M. Heinemann, Autonomous Metabolic Oscillations Robustly Gate the Early and Late Cell Cycle. *Mol. Cell*. **65**, 285–295 (2017).

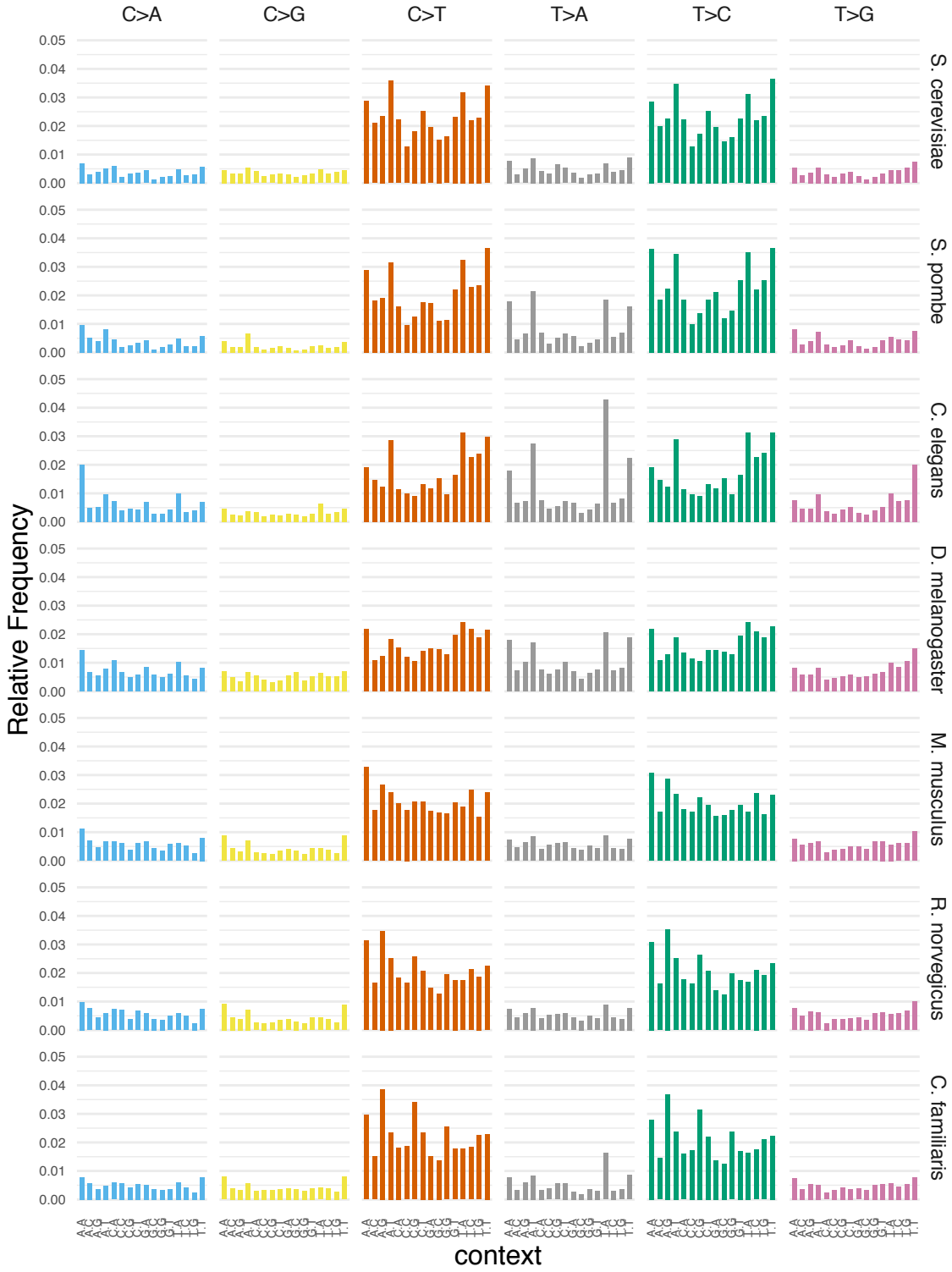
34. Nishimura Kohei, Kanemaki Masato T., *Curr. Protoc. Cell Biol.*, in press, doi:10.1002/0471143030.cb2009s64.

35. K. Chan, in *Genome Instability: Methods and Protocols*, M. Muzi-Falconi, G. W. Brown, Eds. (Springer New York, New York, NY, 2018; https://doi.org/10.1007/978-1-4939-7306-4_4), pp. 33–42.

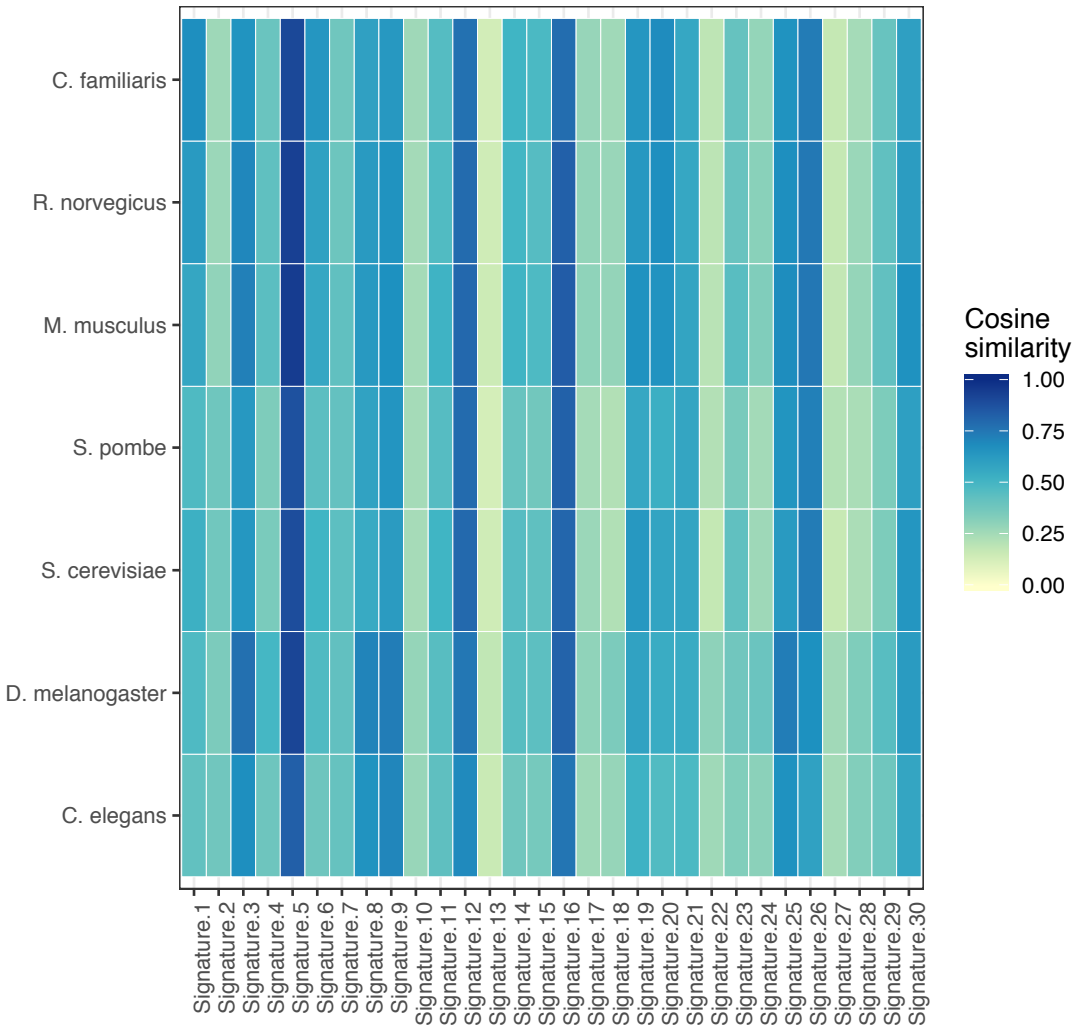
36. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Meth*. **9**, 357–359 (2012).

37. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. **27**, 2987–2993 (2011).

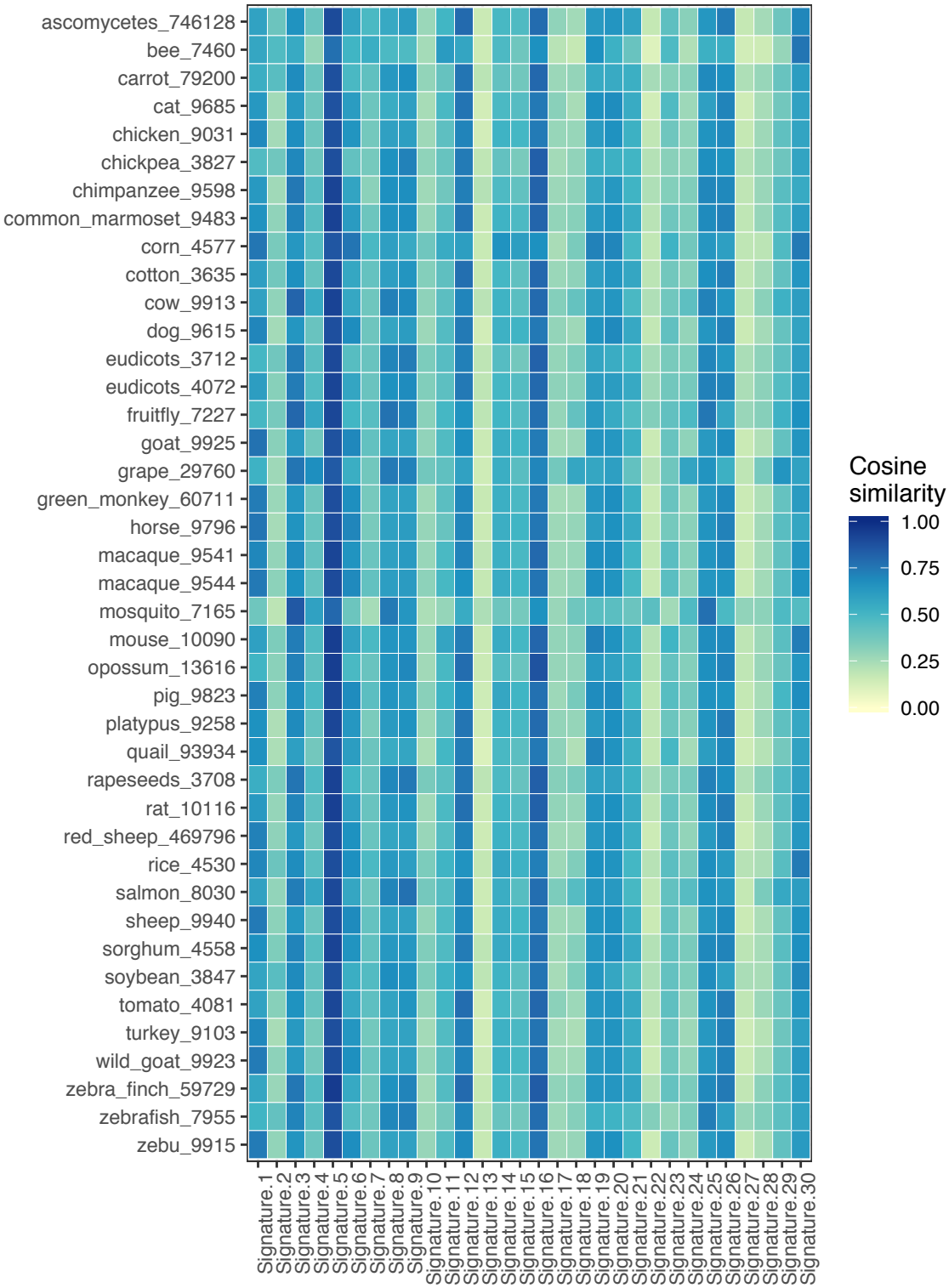
# Figure 1A


C>A
C>G
C>T
T>A
T>C
T>G
S. cerevisiae
S. pombe
C. elegans
D. melanogaster
M. musculus
R. norvegicus
C. familiaris
Relative Frequency
0.05
0.04
0.03
0.02
0.01
0.00
context

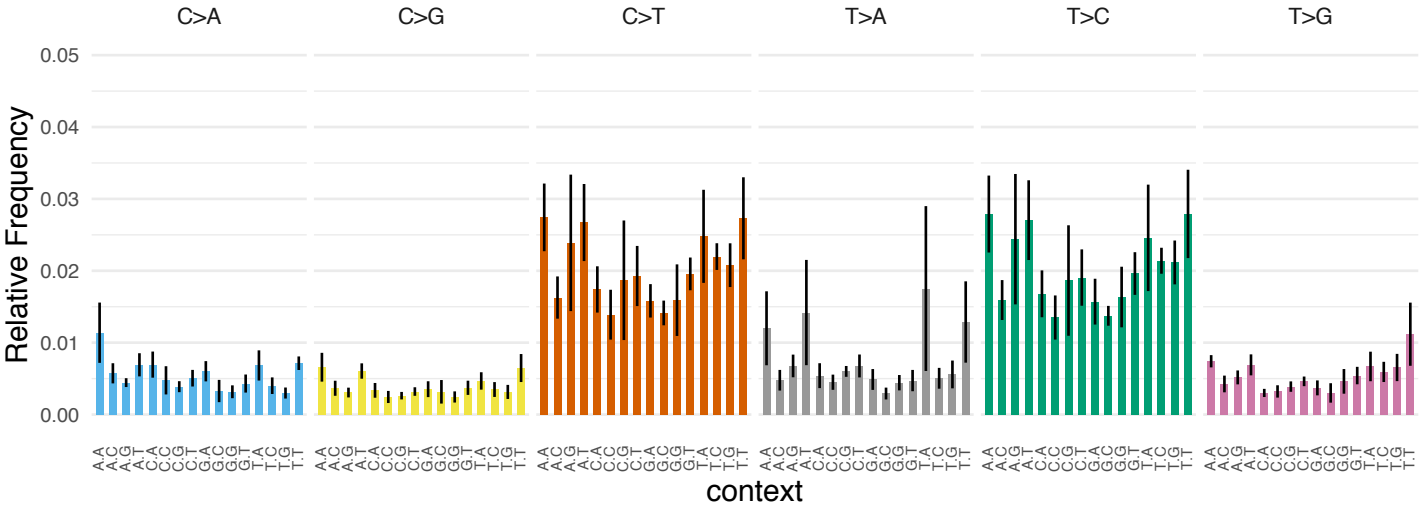
# Figure 1B


C. familiaris
R. norvegicus
M. musculus
S. pombe
S. cerevisiae
D. melanogaster
C. elegans
Signature.1
Signature.2
Signature.3
Signature.4
Signature.5
Signature.6
Signature.7
Signature.8
Signature.9
Signature.10
Signature.11
Signature.12
Signature.13
Signature.14
Signature.15
Signature.16
Signature.17
Signature.18
Signature.19
Signature.20
Signature.21
Signature.22
Signature.23
Signature.24
Signature.25
Signature.26
Signature.27
Signature.28
Signature.29
Signature.30
Cosine similarity
1.00
0.75
0.50
0.25
0.00

# Figure 2


ascomycetes_746128
bee_7460
carrot_79200
cat_9685
chicken_9031
chickpea_3827
chimpanzee_9598
common_marmoset_9483
corn_4577
cotton_3635
cow_9913
dog_9615
eudicots_3712
eudicots_4072
fruitfly_7227
goat_9925
grape_29760
green_monkey_60711
horse_9796
macaque_9541
macaque_9544
mosquito_7165
mouse_10090
opossum_13616
pig_9823
platypus_9258
quail_93934
rapeseeds_3708
rat_10116
red_sheep_469796
rice_4530
salmon_8030
sheep_9940
sorghum_4558
soybean_3847
tomato_4081
turkey_9103
wild_goat_9923
zebra_finch_59729
zebrafish_7955
zebu_9915
Signature.1
Signature.2
Signature.3
Signature.4
Signature.5
Signature.6
Signature.7
Signature.8
Signature.9
Signature.10
Signature.11
Signature.12
Signature.13
Signature.14
Signature.15
Signature.16
Signature.17
Signature.18
Signature.19
Signature.20
Signature.21
Signature.22
Signature.23
Signature.24
Signature.25
Signature.26
Signature.27
Signature.28
Signature.29
Signature.30
Cosine similarity
1.00
0.75
0.50
0.25
0.00

# Figure 3A




# Figure 3B

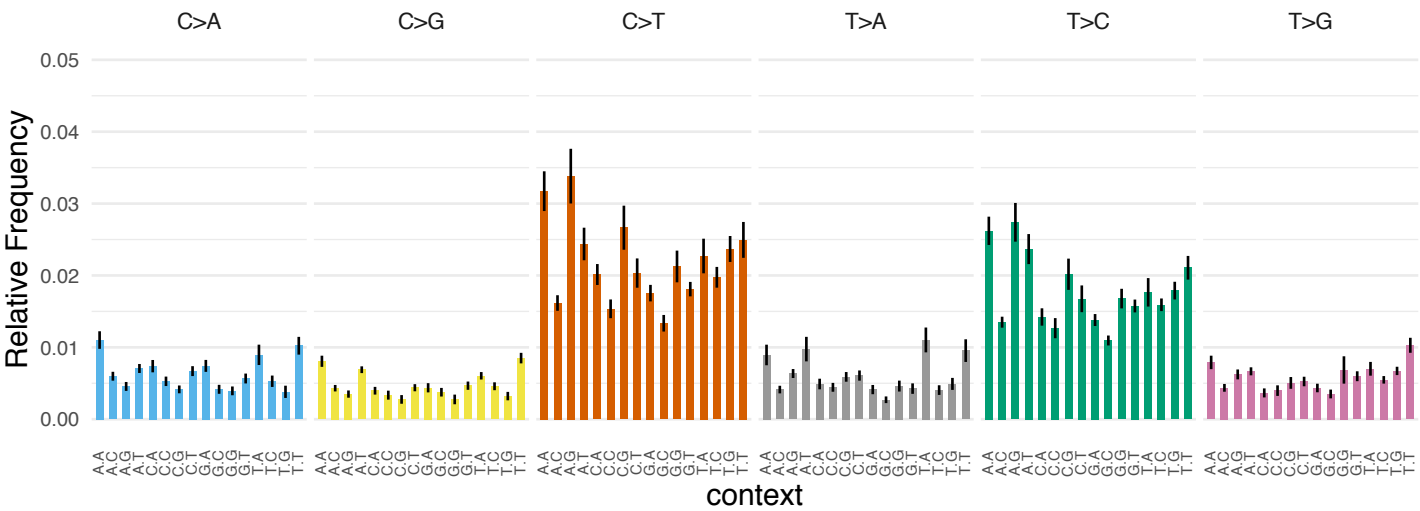


Figure 3C

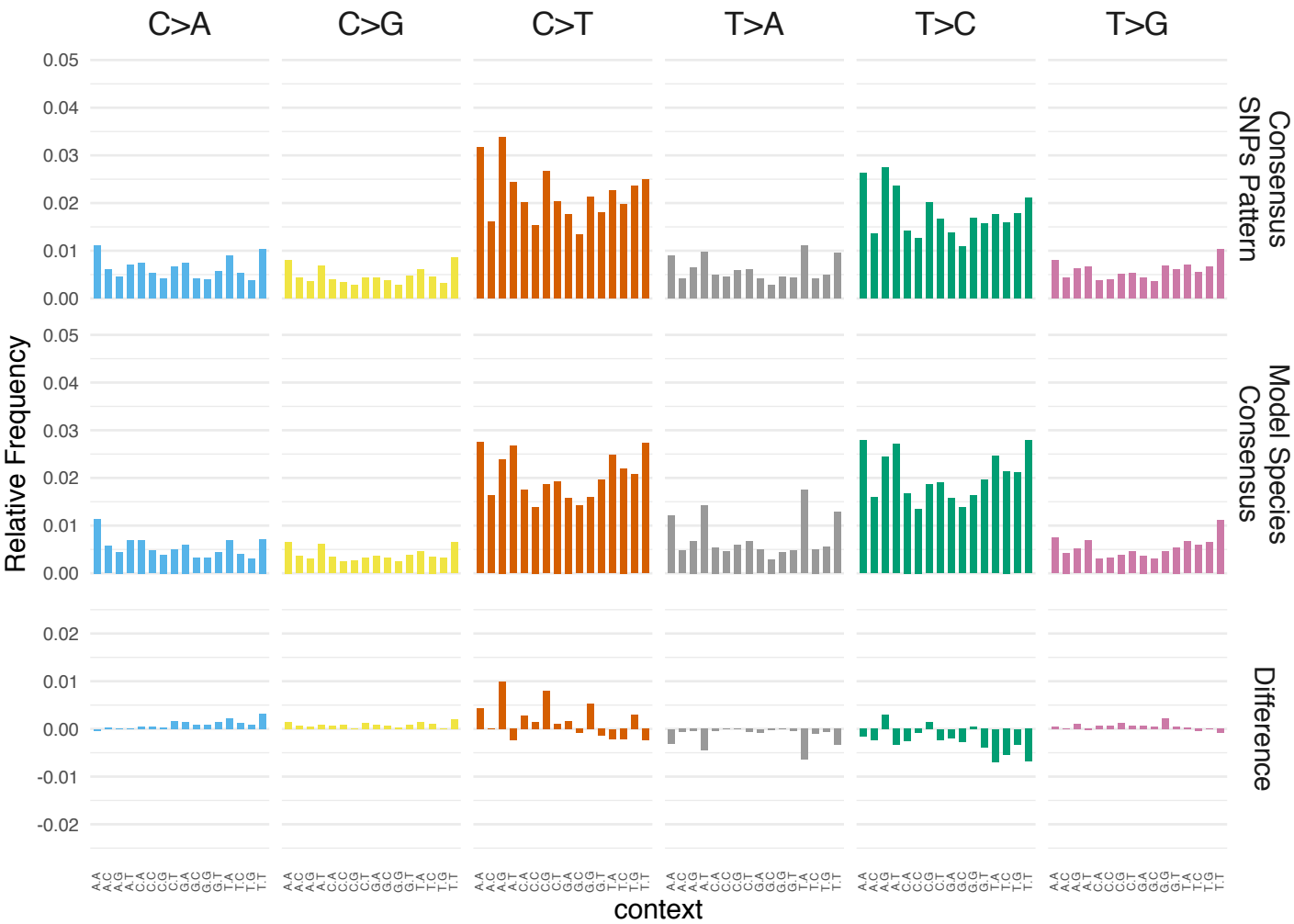

C>A
C>G
C>T
T>A
T>C
T>G
Consensus SNPs Pattern
Model Species Consensus
Difference
Relative Frequency
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
-0.02
context

Figure 4A

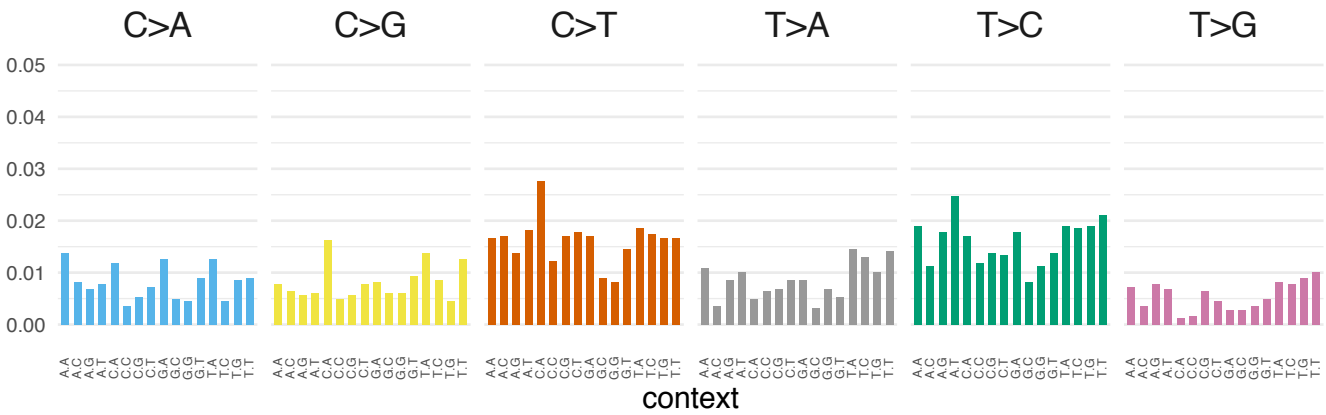



Figure 4B

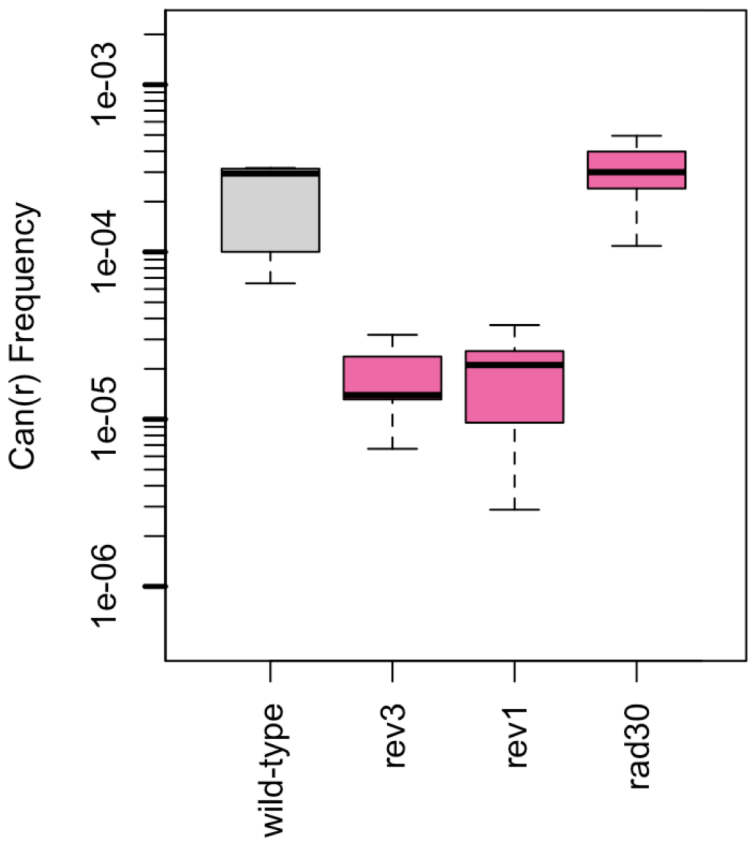


Figure 4C

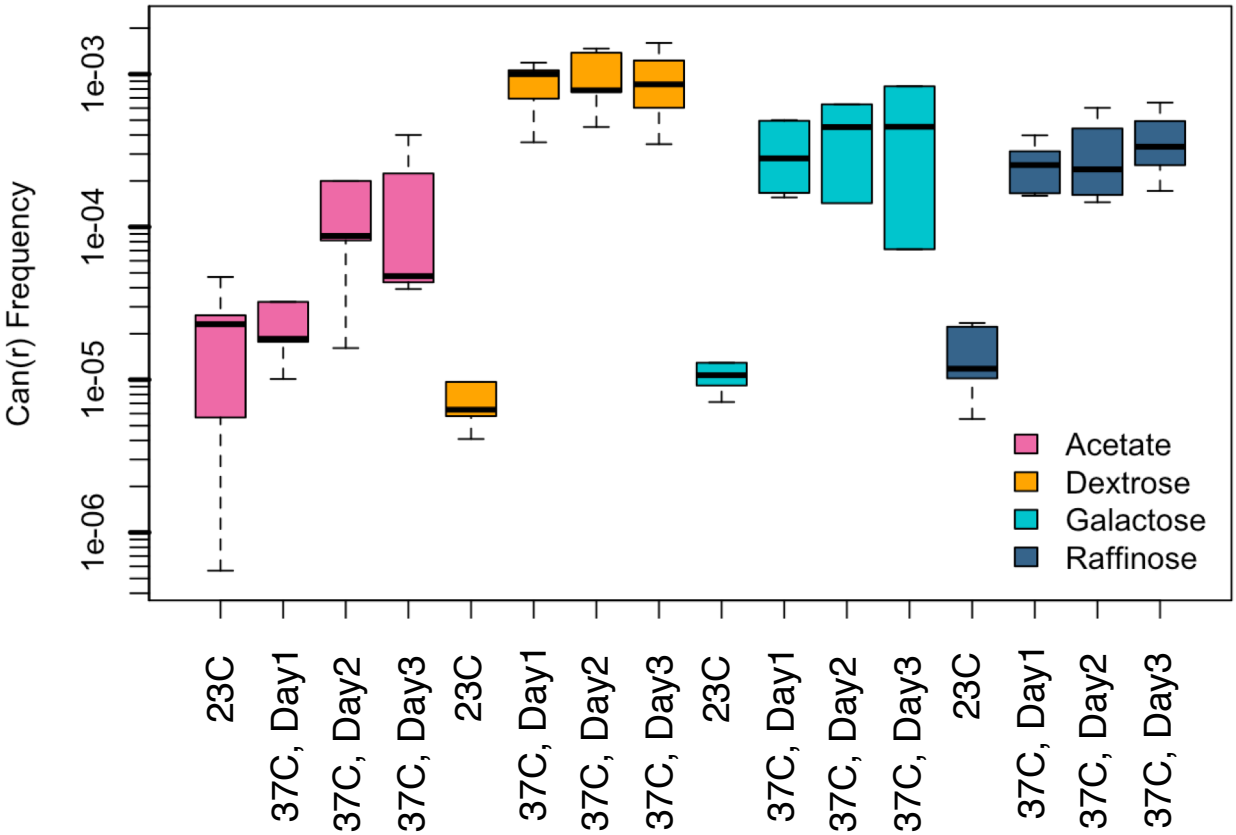

Can(r) Frequency
1e-06
1e-05
1e-04
1e-03
23C
37C, Day1
37C, Day2
37C, Day3
23C
37C, Day1
37C, Day2
37C, Day3
23C
37C, Day1
37C, Day2
37C, Day3
23C
37C, Day1
37C, Day2
37C, Day3
Acetate
Dextrose
Galactose
Raffinose

Figure 4D

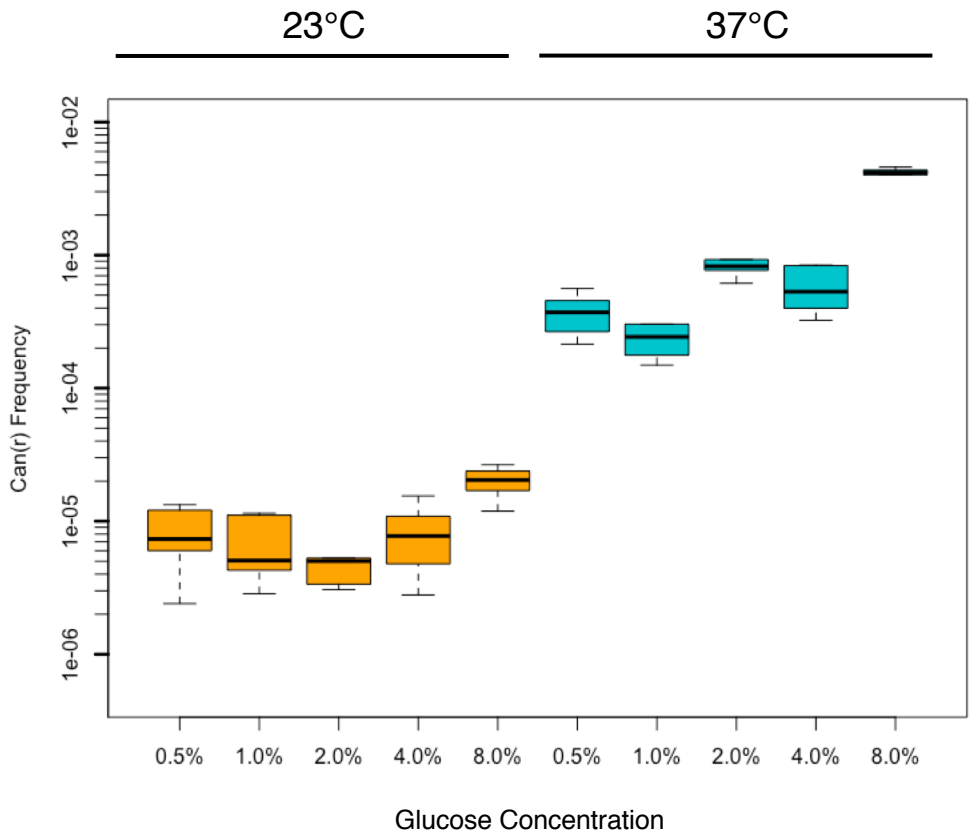
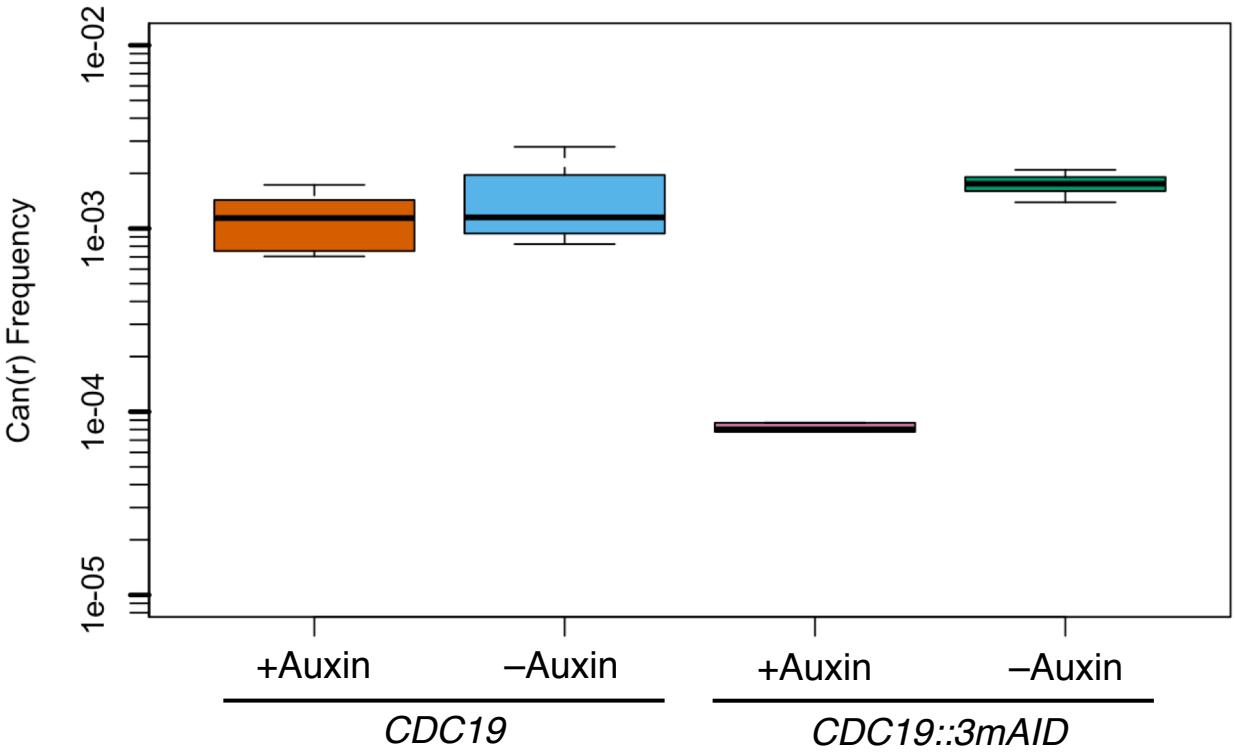
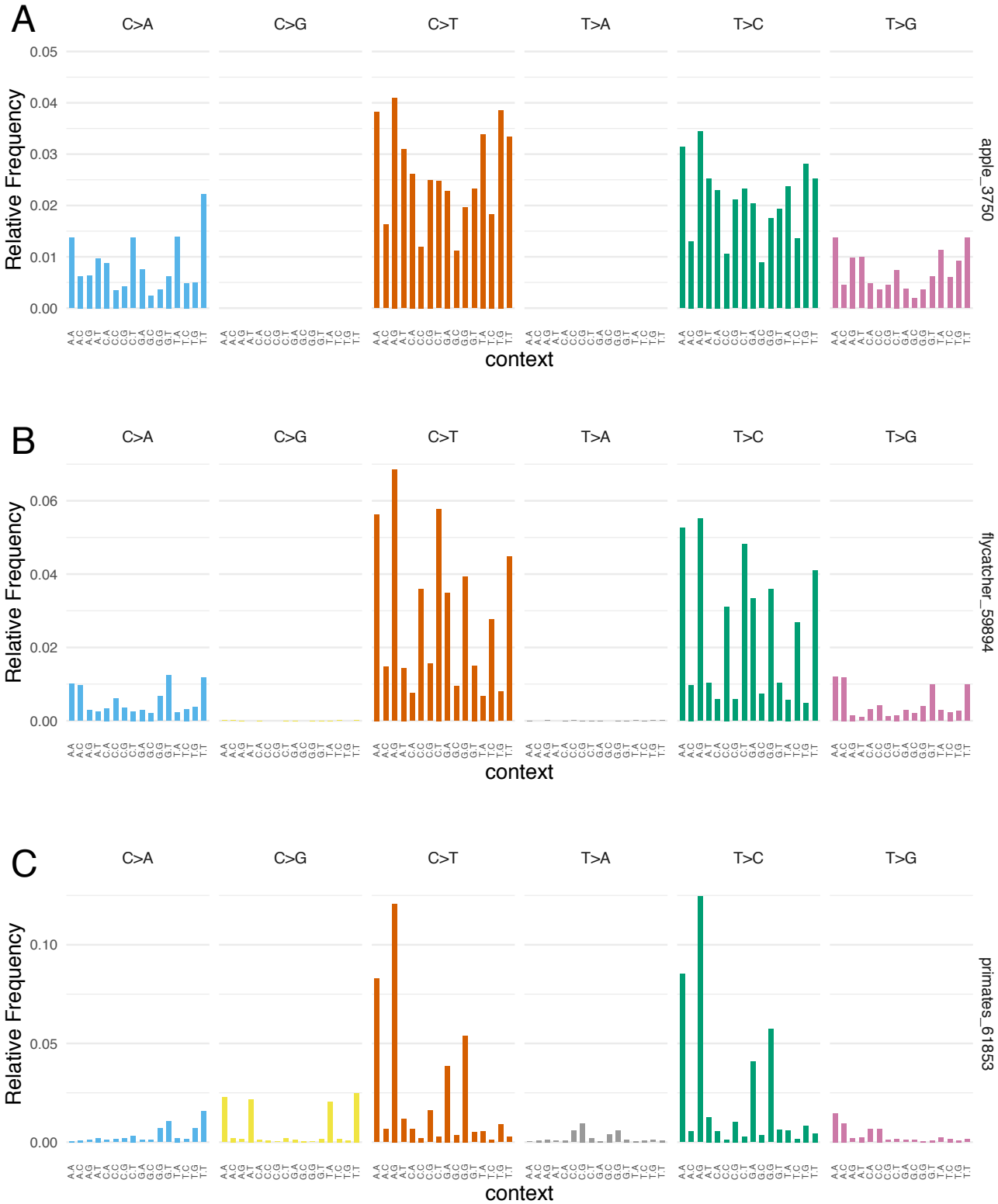

23°C
37°C
Can(r) Frequency
1e-02
1e-03
1e-04
1e-05
1e-06
0.5%
1.0%
2.0%
4.0%
8.0%
0.5%
1.0%
2.0%
4.0%
8.0%
Glucose Concentration

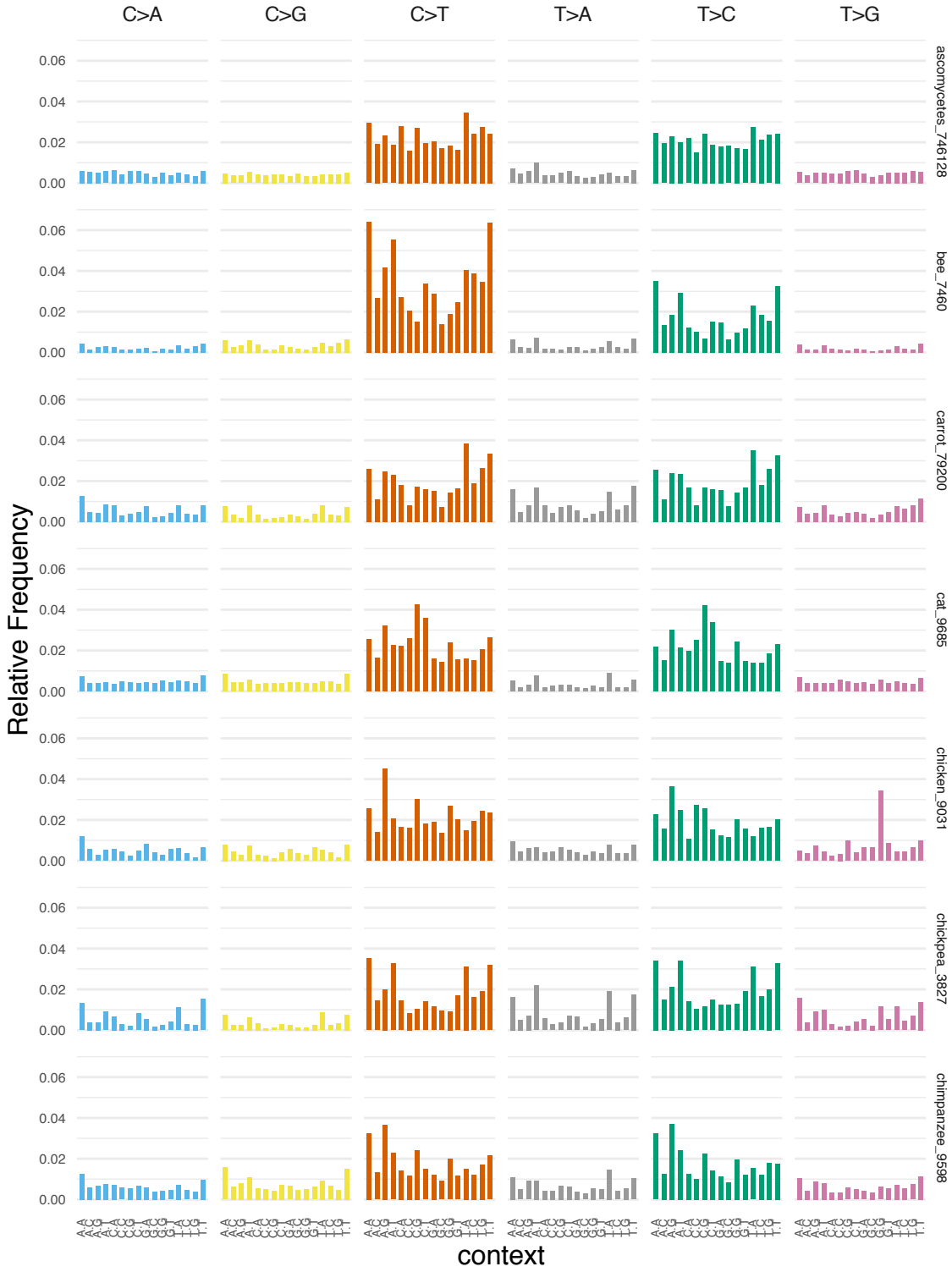
Figure 4E


Can(r) Frequency
1e-02
1e-03
1e-04
1e-05
+Auxin
–Auxin
+Auxin
–Auxin
CDC19
CDC19::3mAID

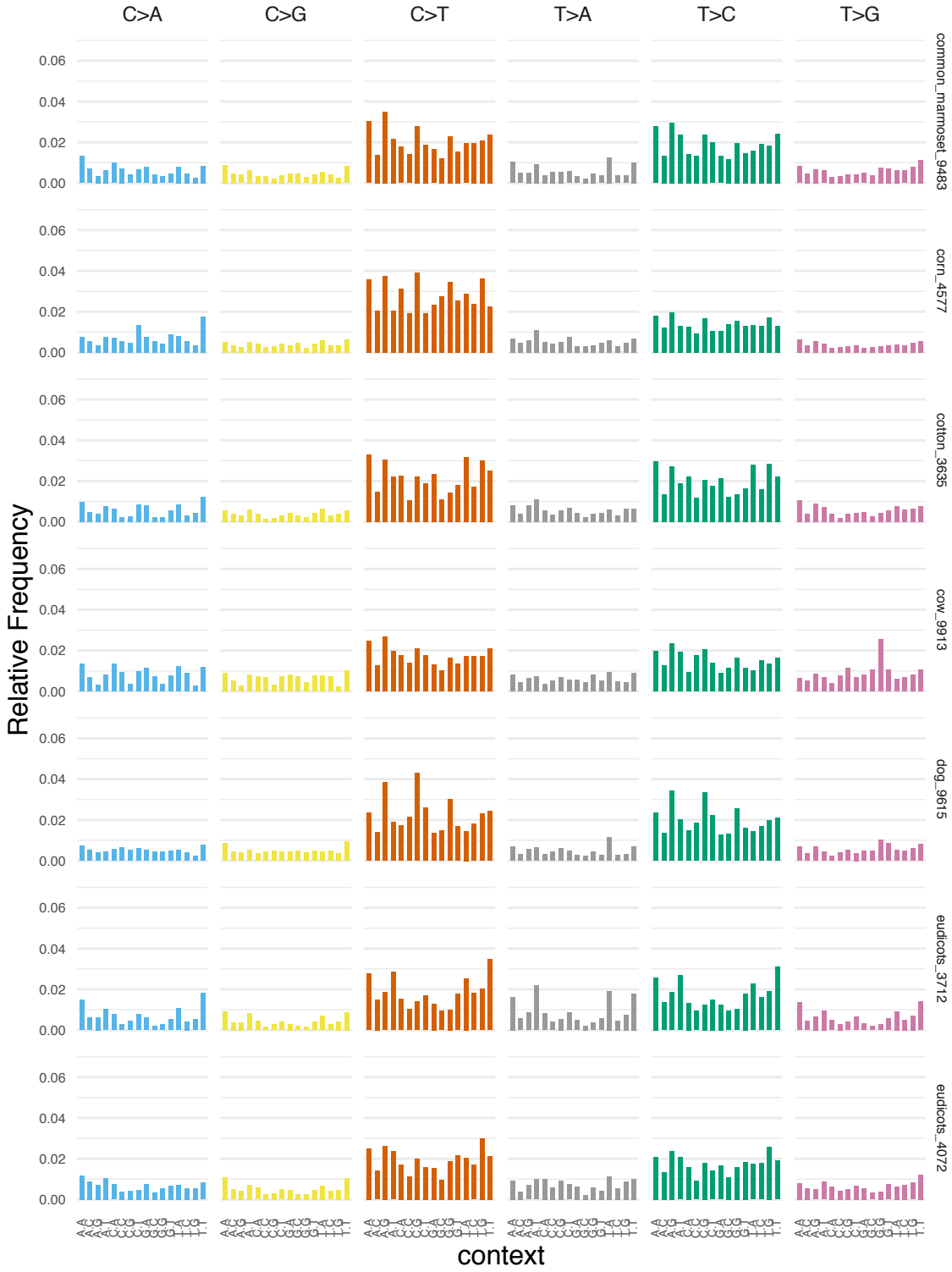
# Supplementary Figure 1



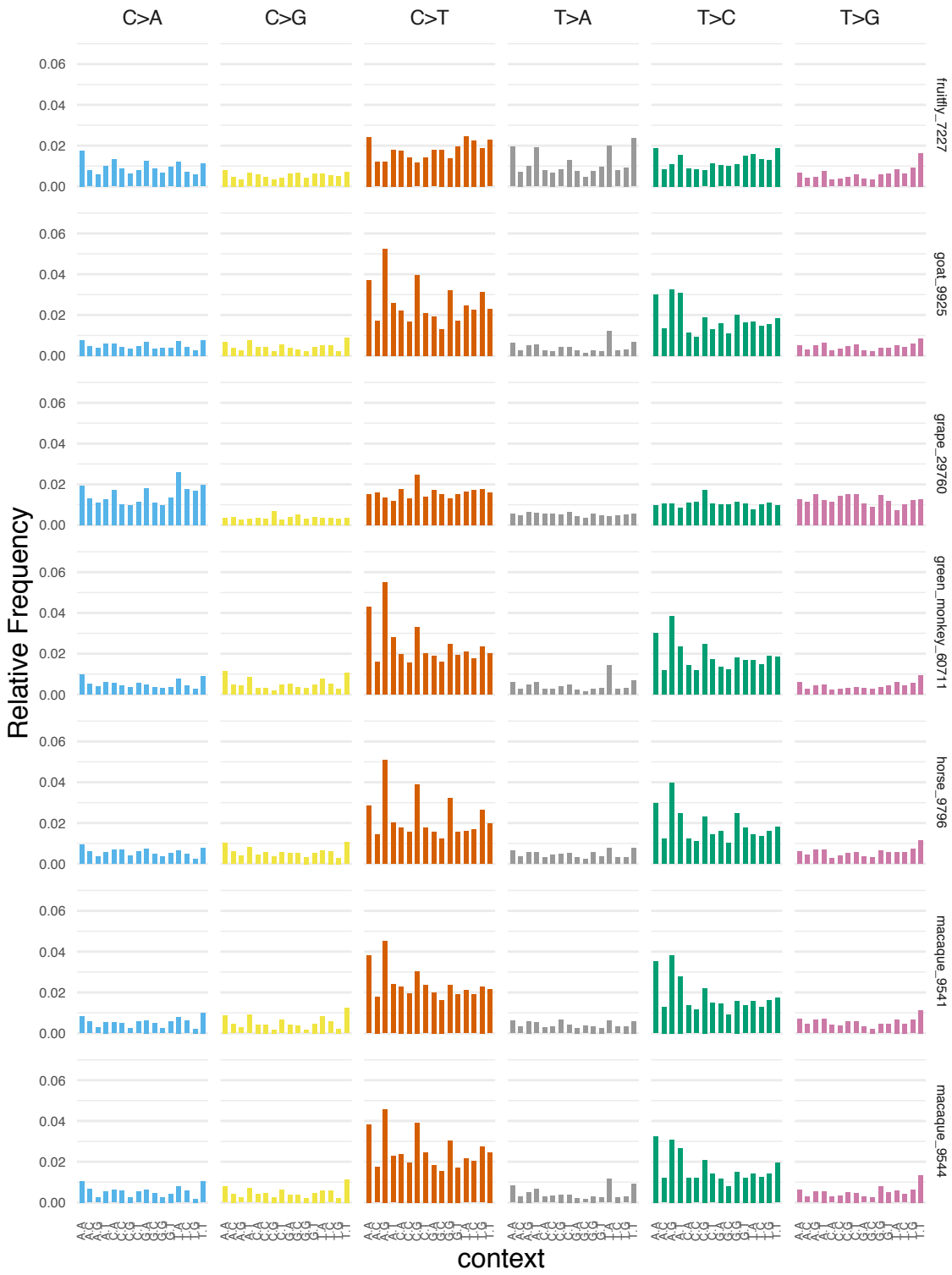
# Supplementary Figure 2



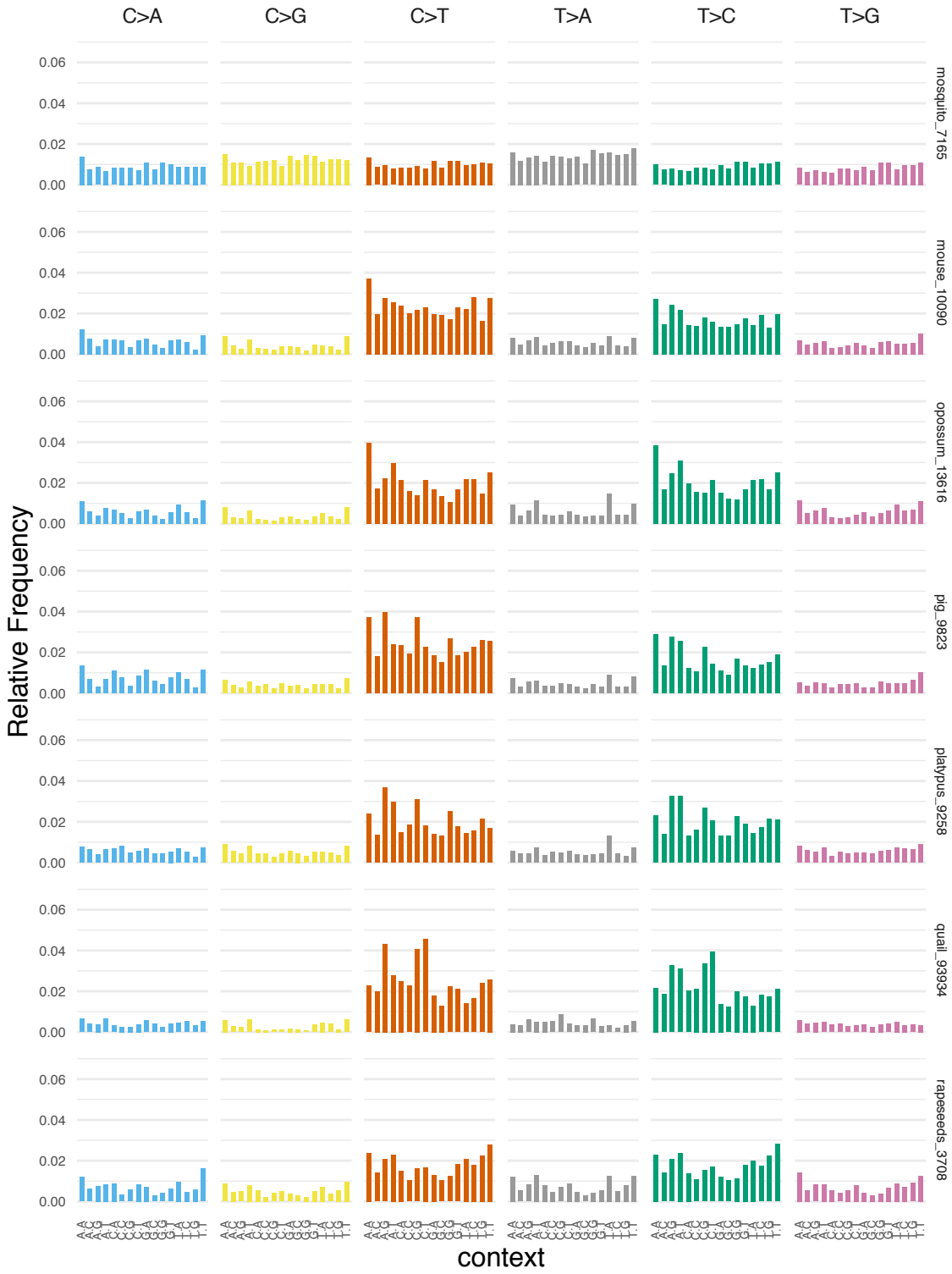
# Supplementary Figure 2, continued



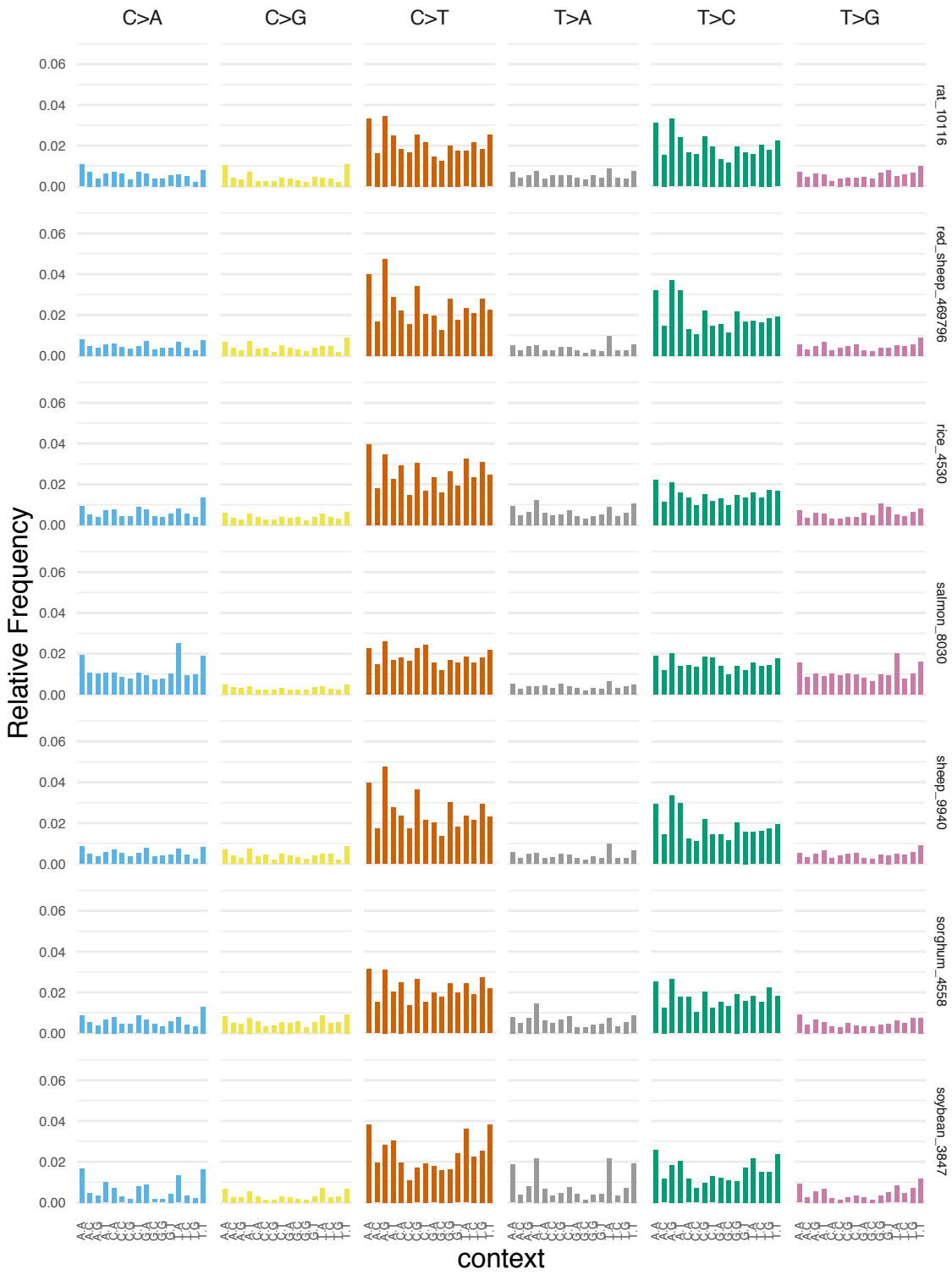
# Supplementary Figure 2, continued



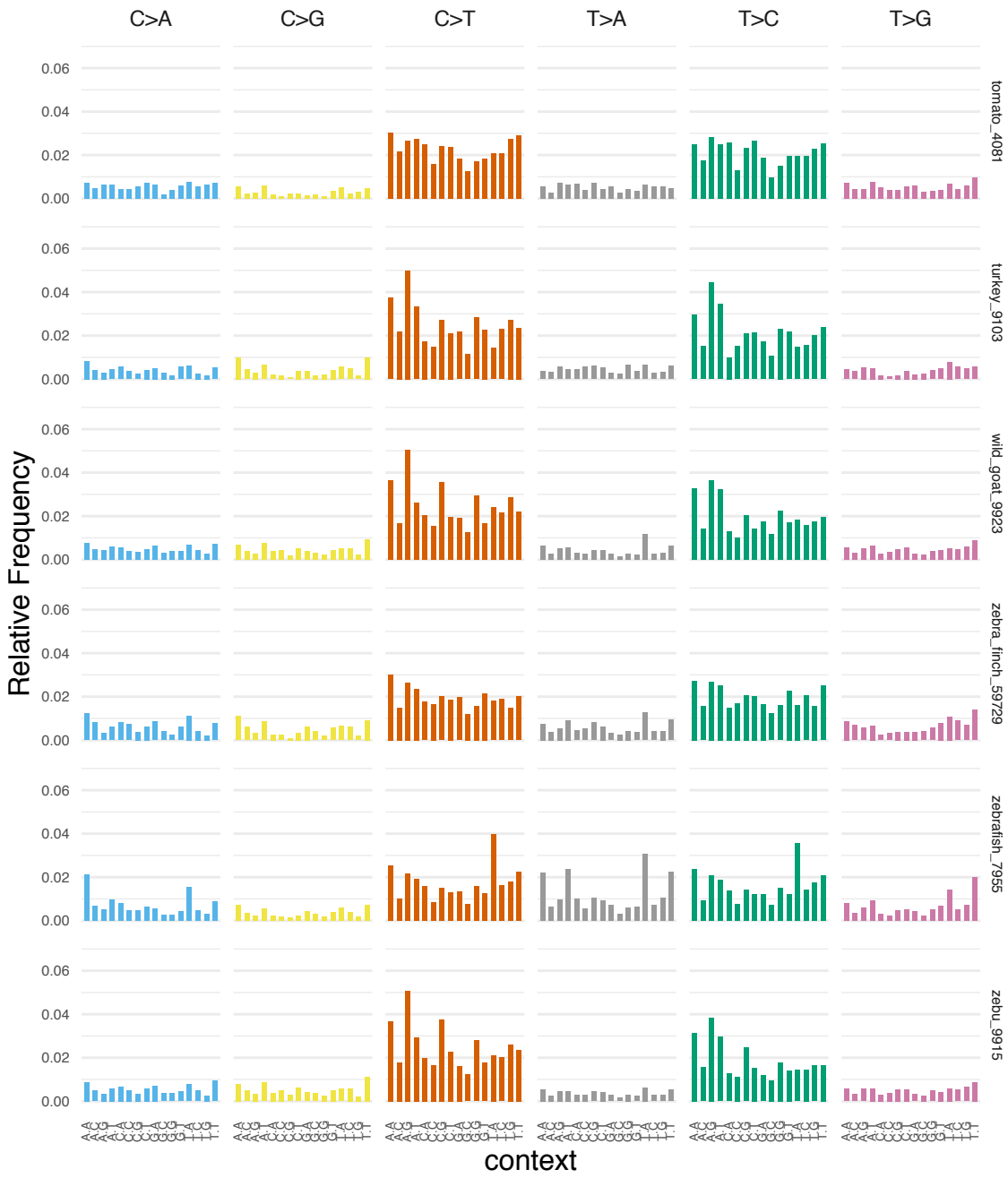
# Supplementary Figure 2, continued



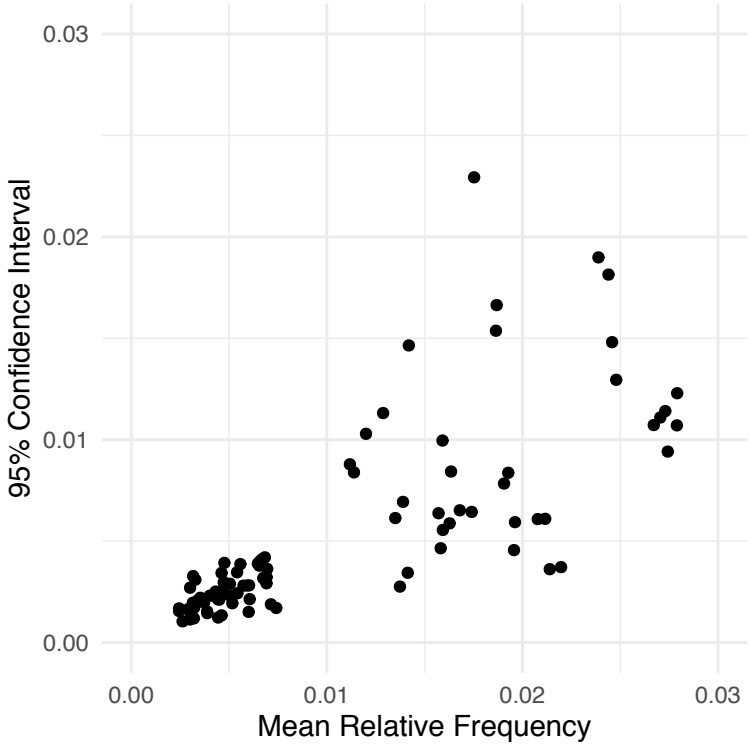
# Supplementary Figure 2, continued



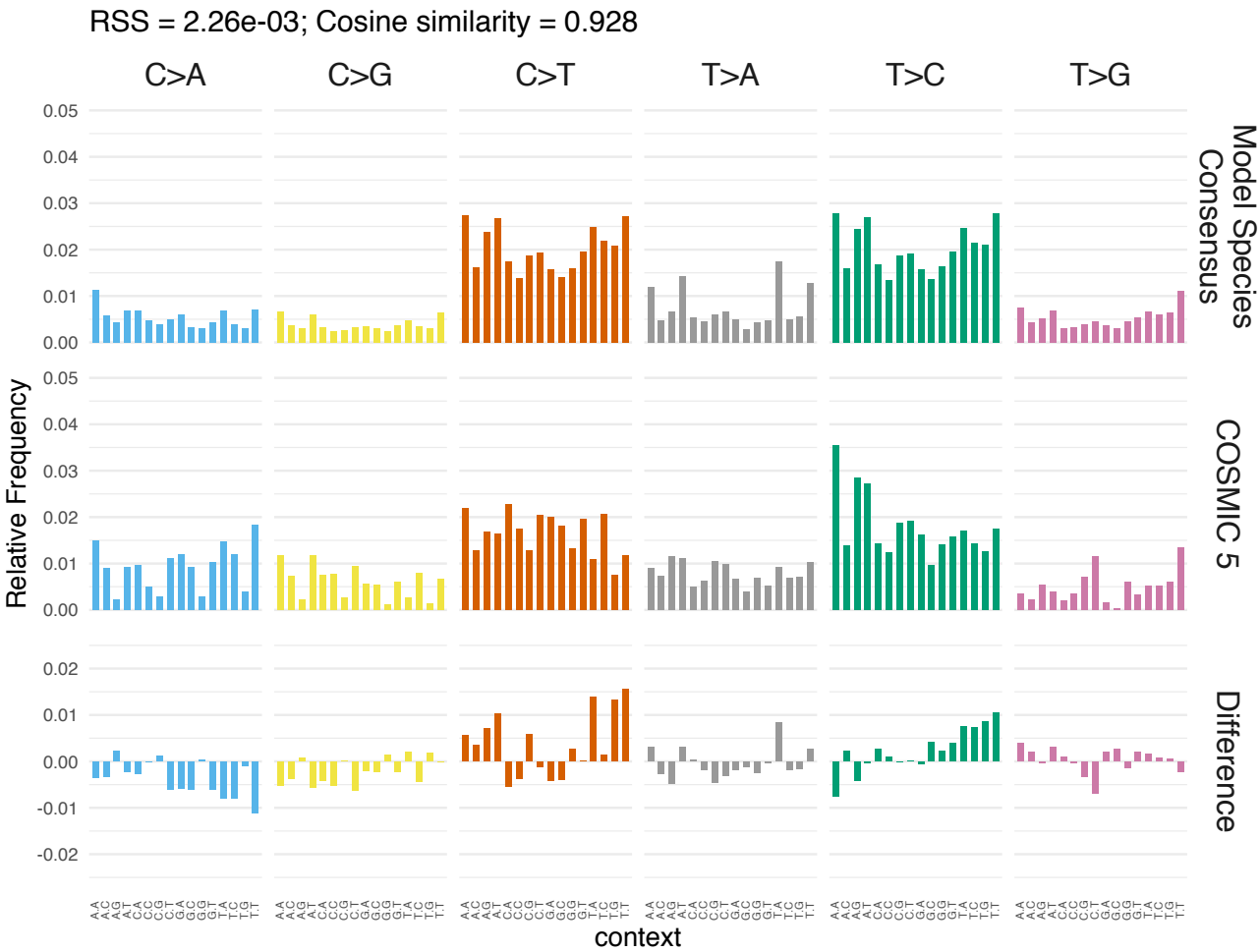
# Supplementary Figure 2, continued



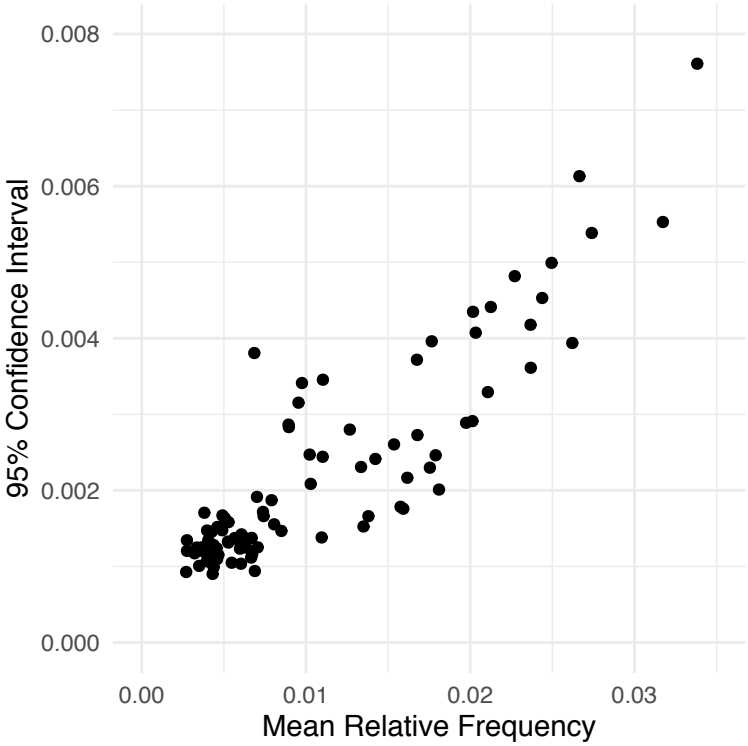
Supplementary Figure 3A



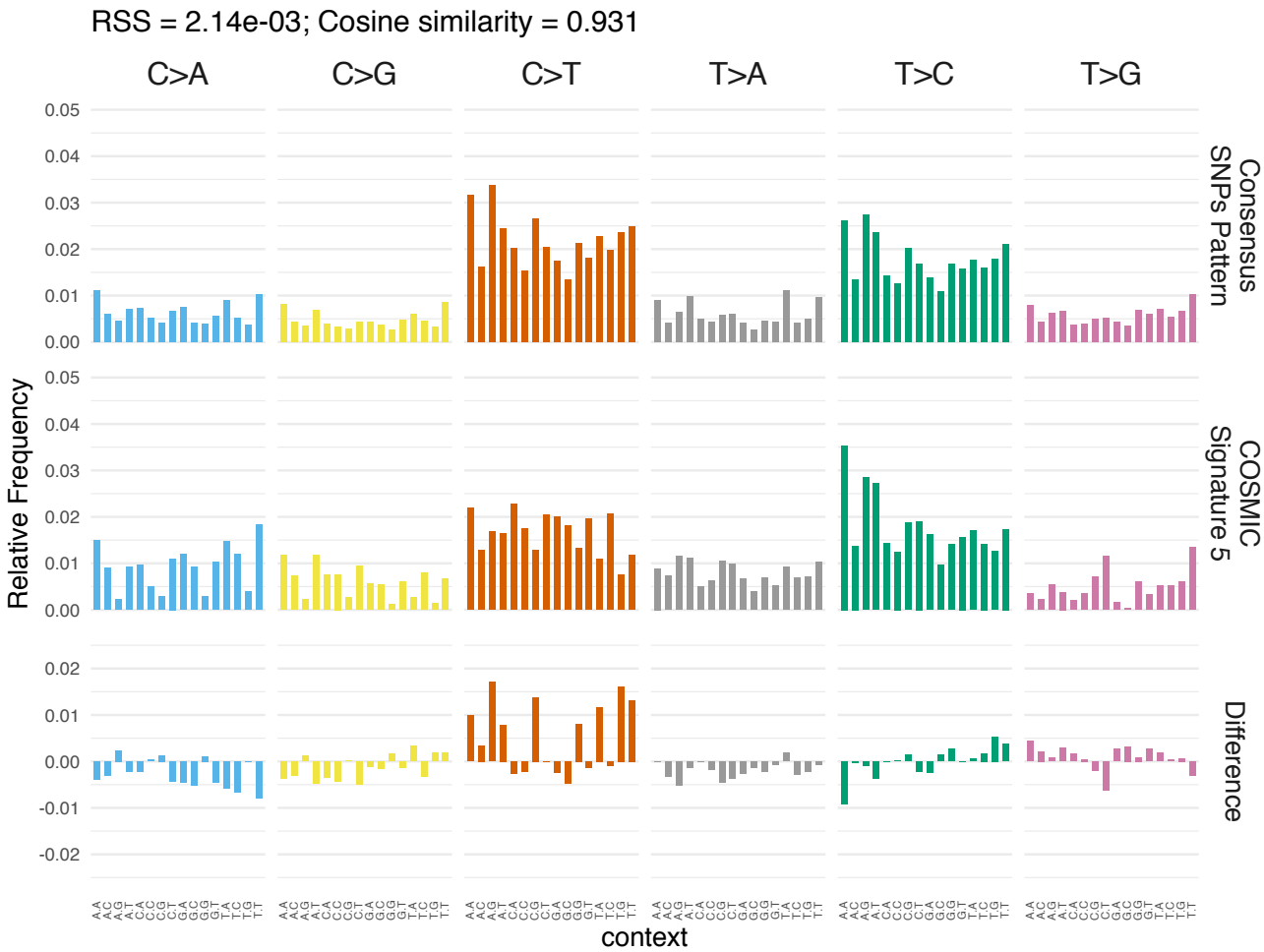
# Supplementary Figure 3B



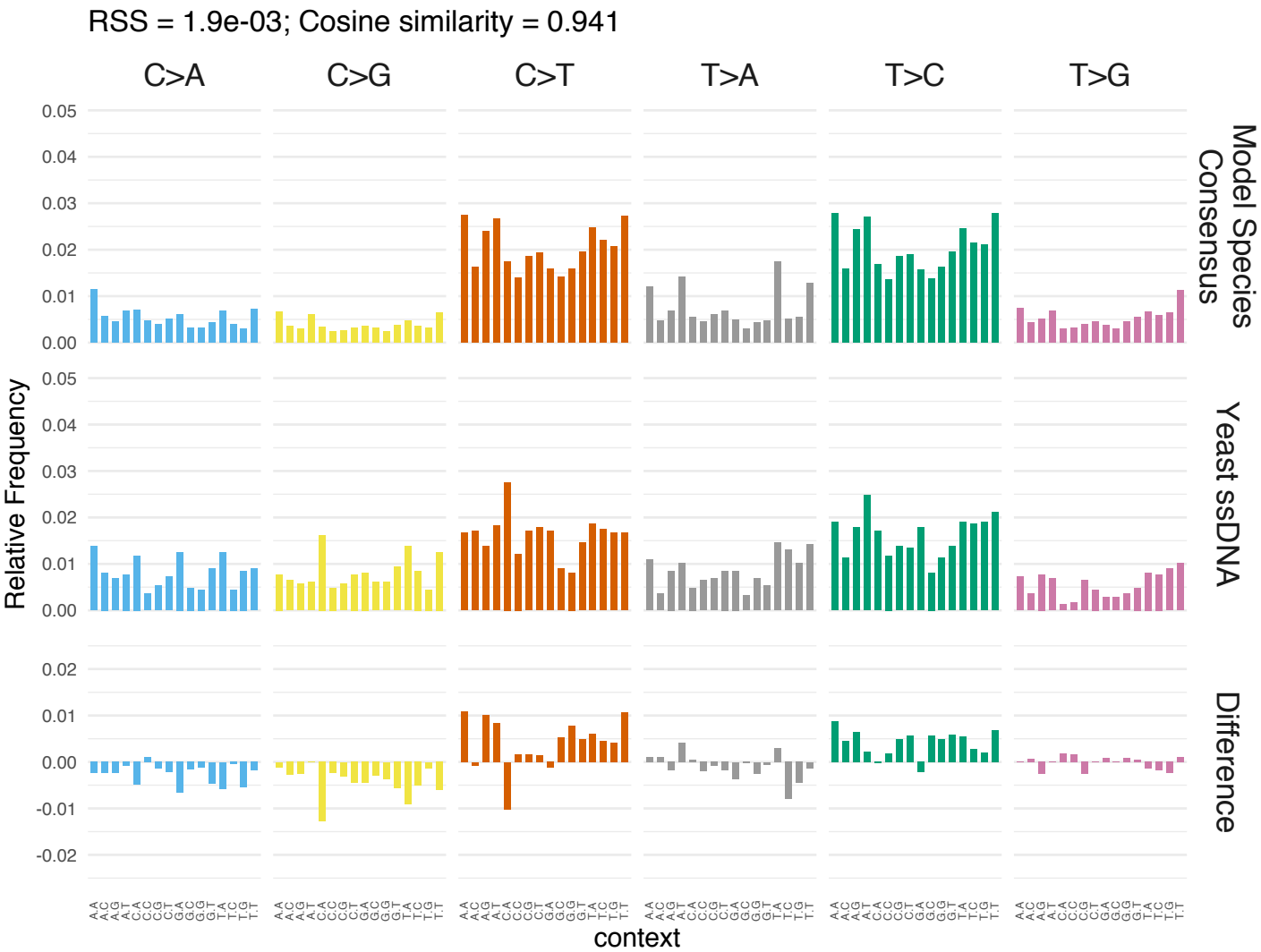
Supplementary Figure 3C



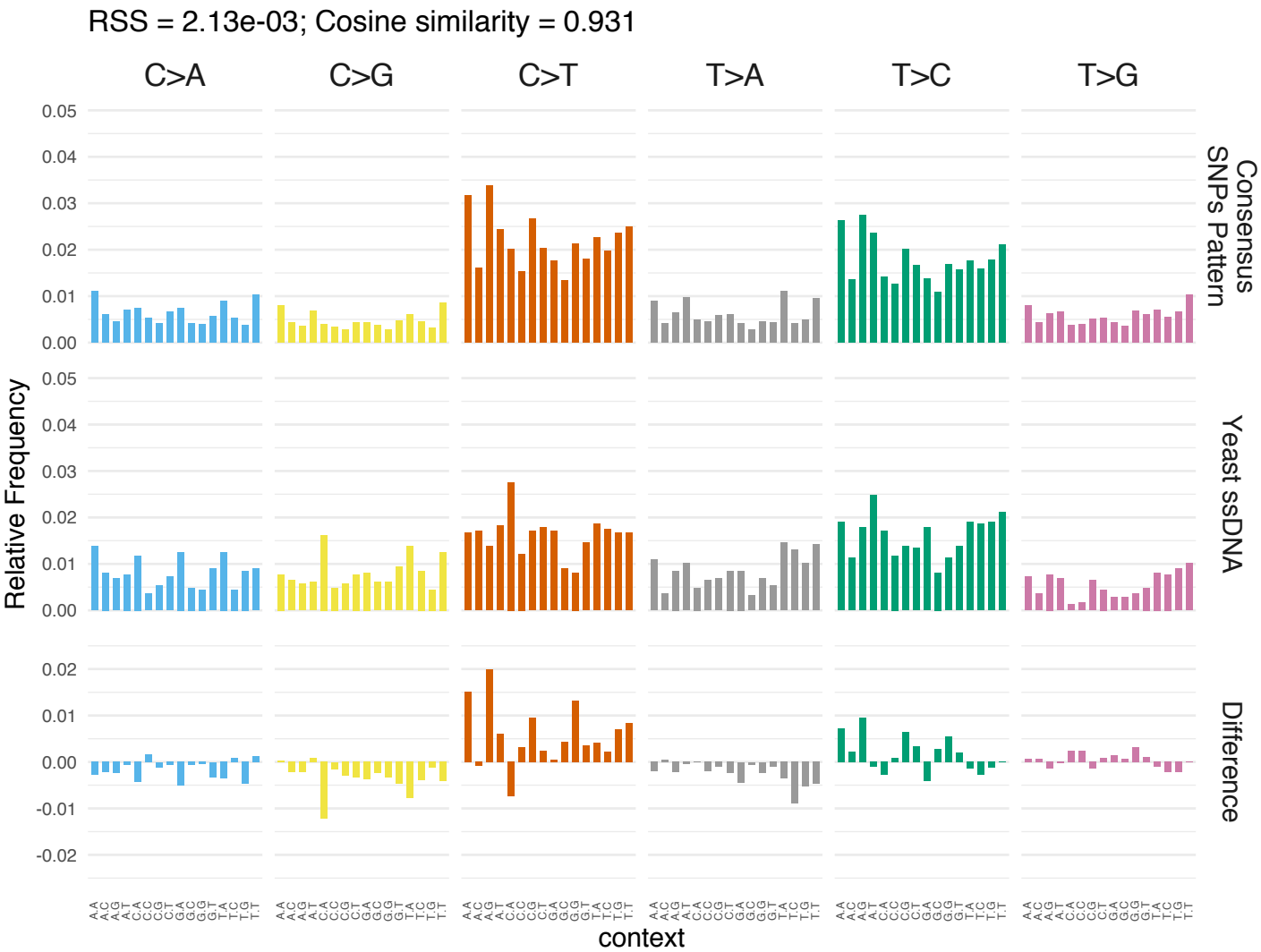
# Supplementary Figure 3D



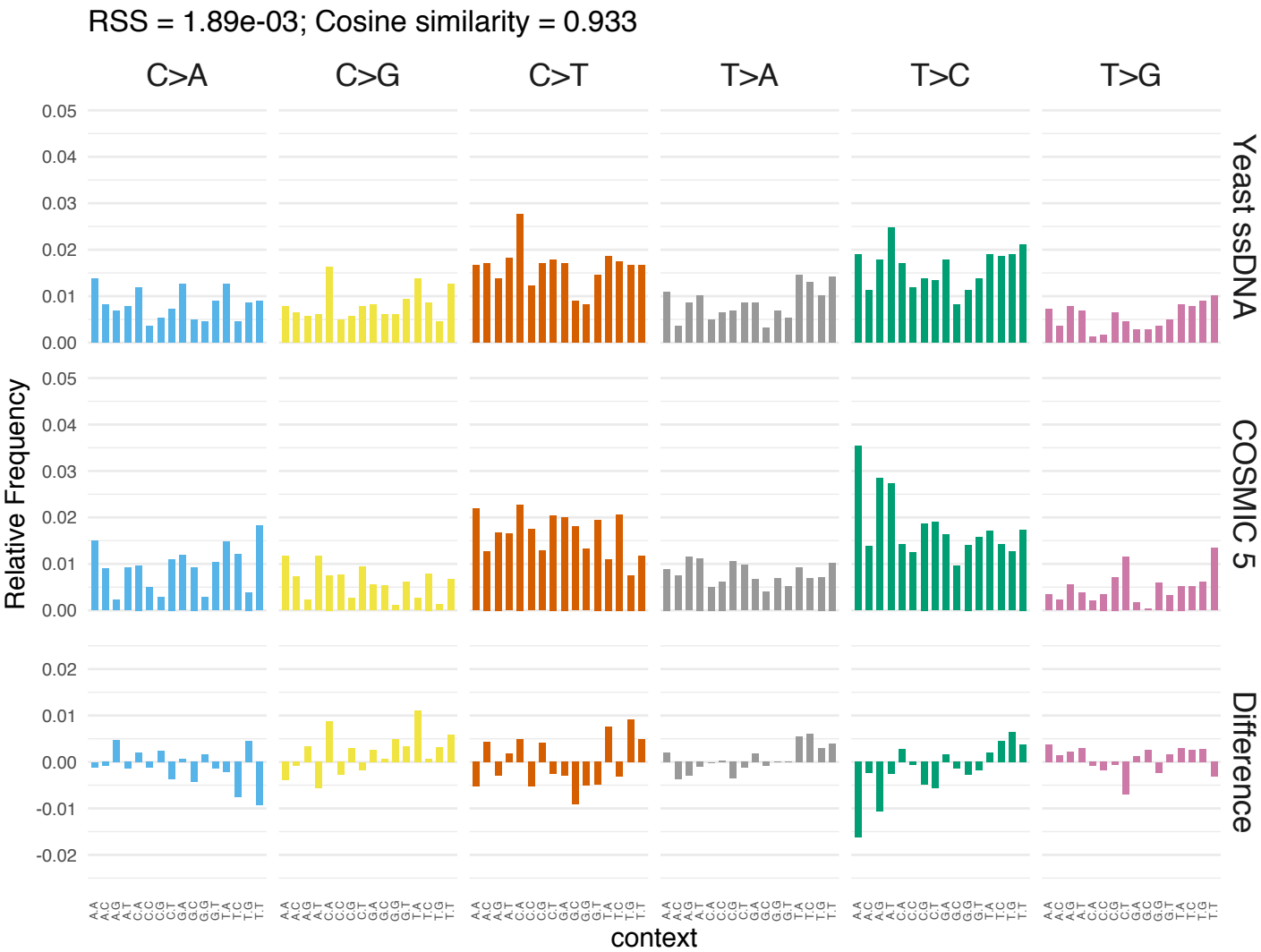
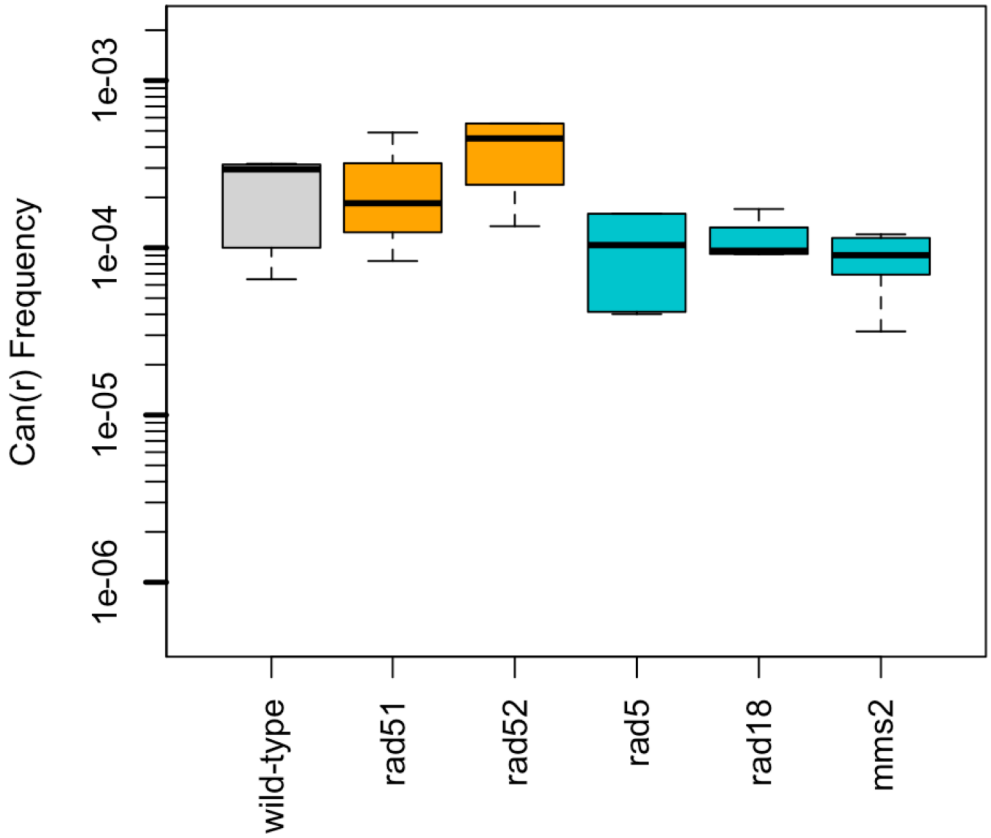
# Supplementary Figure 4A



# Supplementary Figure 4B



# Supplementary Figure 4C



Supplementary Figure 5