Mutation patterns in cancer genomes

Alan F. Rubin[1] and Phil Green[1]

Department of Genome Sciences, University of Washington, Foege Building S-250, Box 355065, 1705 NE Pacific Street, Seattle, WA 98195-5065

Contributed by Phil Green, October 29, 2009 (sent for review September 27, 2009)

Recent large-scale cancer sequencing studies have focused primarily on identifying cancer-associated genes, but as an important byproduct provide ''passenger mutation'' data that can potentially illuminate the mutational mechanisms at work in cancer cells. Here, we explore patterns of nucleotide substitution in several cancer types using published data. We first show that selection (negative or positive) has affected only a small fraction of mutations, allowing us to attribute observed trends to underlying mutational processes rather than selection. We then show that the increased CpG mutation frequency observed in some cancers primarily occurs outside of CpG islands and CpG island shores, thus rejecting the hypothesis that the increase is a byproduct of island or shore methylation followed by deamination. We observe an A→G vs. T→C mutational asymmetry in some cancers similar to one that has been observed in germline mutations in transcribed regions, suggesting that the mutation process may be influenced by gene expression. We also demonstrate that the relative frequency of mutations at dinucleotide ''hotspots'' can be used as a tool to detect likely technical artifacts in large-scale studies.

CpG island | dinucleotide hotspot | mutation and selection | mutational asymmetry |

Large-scale sequencing of cancer genomes is beginning to have a major impact on cancer research (1–5). The primary target of such studies is ''driver'' mutations, i.e., those that play a role in cancer initiation or progression or provide a cell growth or survival advantage. However, the majority of mutations actually identified are presumed to be ''passenger'' mutations that are not advantageous to the cancer cell (2). From one point of view, passenger mutations are an annoying ''haystack'' complicating the search for causal mutations. However, they are also a potentially rich source of information about the specific mutational mechanisms at work in somatic cells and cancer, an aspect we pursue in this paper.

Analyses of neutrally evolving genomic sequences in a variety of organisms have revealed several patterns that are presumed to reflect the nature of the mutational processes in germline cells. Transition mutations occur at a significantly higher rate than transversion mutations (6–8). Substitution rates depend on flanking nucleotides, a notable example being cytosine in CpG dinucleotides which, in mammals, is usually methylated at the 5-carbon and undergoes hydrolytic deamination to thymine at a relatively high rate (8–10). The germline mutation rate at CpGs is much lower within CpG islands (regions enriched in CpGs surrounding or near the transcription start sites of many genes) reflecting, at least in part, the fact that most islands are likely unmethylated in the germline (10). Substitution rates are asymmetric in transcribed regions of the genome, with A→G transitions occurring at a higher rate than T→C transitions on the coding strand (11), the magnitude of the effect correlating with gene expression level (12, 13). While mechanisms for transcriptional mutagenesis have been described in model systems (14, 15), the mechanism responsible for this asymmetry is currently unknown.

Cancer sequencing studies have provided preliminary evidence for several mutational rate trends, some, but not all, of which are similar to those seen in germline mutations. Transitions are more frequent than transversions, and CpG dinucleotides are a mutation hotspot, particularly in colorectal cancer (1–5). In contrast, TpC dinucleotides are a mutation hotspot in breast cancers (16), but not in the germline (17). Analysis of mutations in *TP53* suggests that diverse tumor types are affected by different mutation processes (18). Note that such trends could reflect mutations arising before carcinogenesis that are carried along by subsequent clonal expansion, in addition to mutations that occur within the cancer cells themselves.

Here, we use published data from several studies to further explore patterns of nucleotide substitution in cancer cells. We first examine the overall impact of selection on the mutation spectra by comparing synonymous and nonsynonymous substitution frequencies (mutations per site sequenced) in pancreatic cancer and glioblastoma multiforme (4, 5) and by examining the nature of amino acid changes in breast and colorectal cancers (1). Our results suggest that most coding sequence mutations in cancer are neutral with respect to cancer growth. This finding justifies our assumption that mutation patterns in the data are more likely to reflect the mutation process than selection.

Since methylation of CpG islands and of regions within 2 kb of islands, called CpG island shores, are known features of some cancers (19–22), it has been hypothesized that the elevated CpG mutation frequency is due to increased DNA methylation of islands and shores, followed by deamination of the methylated CpGs (23). We reject this hypothesis by showing that the elevated frequency primarily reflects increased mutation of CpGs outside of islands and shores. We also find an A→G vs. T→C mutational asymmetry similar to that previously observed in the germline (11), which suggests that the mutation pattern within a gene is influenced by its expression. Finally, we show that the fraction of mutations occurring at dinucleotide hotspots can be a useful metric for identifying technical artifacts in cancer sequencing studies, by detecting an inconsistency between the discovery and validation screens in one study that is likely due to error-prone sample amplification.

Results and Discussion

Strength of Selection on Coding Sequence Mutations. The majority of mutations observed in cancer sequencing studies are believed to be ''passenger'' mutations having little impact on the cancer cell (2). However, it remains possible that many mutations occurring during cancer growth are deleterious to the cell and consequently eliminated by selection. If so, the set of passenger mutations would not faithfully reflect the underlying mutation process. Since selection on germline mutations in coding sequences acts mainly at the amino acid level (24), we assume that this is also true of somatic mutations and that we can therefore explore the effects of selection by comparing frequencies of nonsynonymous and synonymous substitutions. Using data from two studies, where both types of substitution were catalogued (4, 5), we find the overall nonsynonymous/synonymous frequency

Author contributions: A.F.R. and P.G. designed research; A.F.R. performed research; A.F.R. analyzed data; and A.F.R. and P.G. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. E-mail: afrubin@u.washington.edu or phg@u.washington.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0912499106/DCSupplemental.

ratios for pancreatic cancer to be 0.95, and for glioblastoma multiforme to be 1.10, neither of which is significantly different from 1 (P = 0.57 and 0.43, respectively). This indicates that (in contrast to germline mutations) the set of nonsynonymous mutations in cancer is not strongly biased by selection. We also calculated separate ratios for dinucleotide contexts that are known to have elevated mutation frequencies in cancer (1, 16) and again found that the nonsynonymous/synonymous substitution frequency ratios are not significantly different from 1 (see Fig. S1 in *SI Appendix*).

It has been claimed that hundreds of genes are subject to significant positive selection in cancer (1), although this has been disputed (25–27). Analyzing data in ref. 1, we find that the amino acid substitutions in known cancer genes are significantly different from substitutions in other putative (according to ref. 1) cancer-associated genes (P = 1.61e-4), which in turn are not significantly different from genes that were not cancer-associated in ref. 1 (P = 0.133) (see Fig. S2 in *SI Appendix*). This suggests that the mutations in the putative cancer-associated genes in ref. 1 are primarily passenger mutations, and the set of mutations is therefore unlikely to be strongly biased by positive selection. This is not to say that there is no such selection, and indeed a more sensitive test can detect evidence for positive selection in cancer (2). However, our results suggest the proportion of selected mutations is small. In combination, these results suggest that, while selection certainly acts on some mutations, the set of mutations in cancer cells is not significantly biased by negative or positive selection, and we therefore assume in the following that the sets of nonsynonymous nucleotide substitutions reported by cancer sequencing studies mainly reflect the underlying mutation process that generated them.

CpG Mutations. CpG dinucleotides are known to be mutation hotspots in some cancer types (1, 3–5). Thirty-two percent of the CpGs found in the coding sequence of genes we analyzed (302,780 of 940,319) are in CpG islands, and another 11% (106,691 of 940,319) are found within 2 kb of CpG islands, in regions called CpG island shores (22, 28). Since methylation of CpG islands and/or shores is known to occur in some genes in cancer (19–22), and the higher germline mutation rate at CpGs appears confined to methylated CpGs (29), it has been hypothesized that the elevated CpG mutation rate in some cancers could result primarily from mutations in methylated CpG islands or shores (23). To test this, we classified the CpG sites in the studied genes as island, shore, or other, and calculated mutation frequencies for each class using published data (3–5). In three of the four cancer types analyzed (colorectal, pancreatic, and glioblastoma), CpG frequencies are significantly different among classes (Fig. 1), but in all cases the island frequency is lowest, and the "other" frequency is highest. Moreover, the dramatic increase in overall CpG frequency in colorectal cancer relative to the other cancer types is mostly confined to the non-island, non-shore CpGs. The elevated overall mutation frequency at CpG dinucleotides in the cancers is therefore not due primarily to deamination of cytosines in methylated CpG islands or shores. A similar analysis using 500-bp shores [which have a higher density of CpGs and a stronger tendency to cancer-specific hypermethylation than the rest of the 2 kb shore (22)] gives similar results (see Fig. S3 in *SI Appendix*).

We also recalculated CpG mutation frequencies for protein kinase genes in various cancers sequenced in another study (see Fig. S4*A* in *SI Appendix*) (2). Although the smaller size of this dataset prevents robust conclusions, in colorectal, lung, and ovarian cancers, the CpG mutation frequency outside of islands is again higher, suggesting this may be a general trend.

Excision of thymine in a deaminated CpG is performed by MBD4 and TDG DNA glycosylases (30). Neither enzyme has coding sequence mutations in the samples used in our analysis,

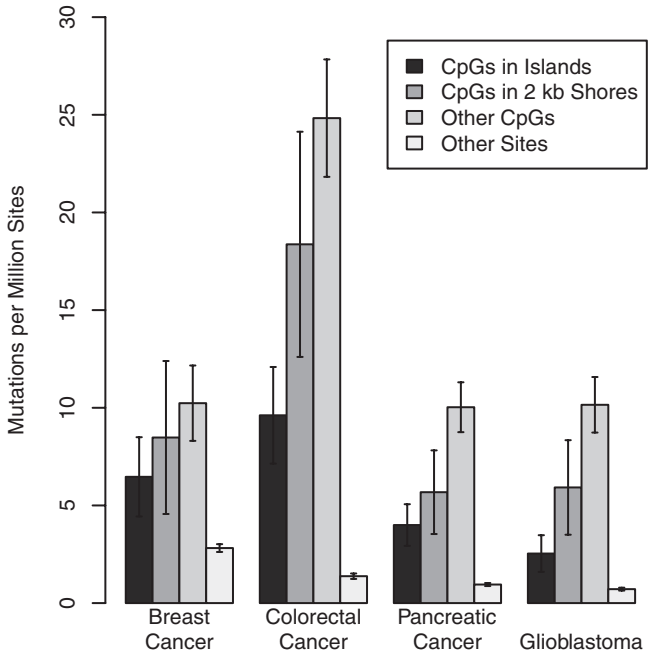

Fig. 1. CpG mutation frequencies in CpG islands, 2 kb CpG island shores, and remainder of gene. Frequencies were computed from discovery screen data in refs. 3–5 by dividing the observed number of nonsynonymous changes at the site by the number of sequenced sites of that type. See Table S3 in *SI Appendix* for details. Error bars indicate 95% confidence intervals.

but there could be noncoding changes affecting expression or splicing, or mutations in interaction partner genes. Other possibilities are that cellular changes increasing methylation or deamination rates, or shortening cell division times could reduce the probability of repair occurring before DNA replication, in tumor or precursor tissue.

A→G/T→C Mutational Asymmetry. An elevated rate of germline A→G mutations compared to the complementary rate of T→C mutations on the coding strand has been associated with transcribed regions in mammals (11). We tested whether this asymmetry is detectable in cancer mutations, using data from three studies (Fig. 2) (3–5). All cancer types appear to have some excess of A→G mutations, and the asymmetry is statistically significant in breast cancer (P = 2.52e-3). To examine whether this asymmetry is associated with expression, we calculated separate mutation frequencies for genes with higher or lower than median expression, using expression data from 10 breast cancer samples to classify the genes (31). The asymmetry is more pronounced in genes with higher than median expression, suggestive of an expression-related effect, although the difference does not reach statistical significance (see Fig. S5*A* in *SI Appendix*). We also performed this test using expression data from normal breast tissue (31), hypothesizing that the effect might be stronger if most of the mutations had occurred before carcinogenesis. However, the results were essentially identical (see Fig. S5*B* in *SI Appendix*), presumably reflecting the fact that a relatively small number of genes (<200) are differentially expressed between normal and cancer tissue in this dataset (31). The fact that the same asymmetry is associated with germline transcription (11) could point to a common (currently unknown) mechanism, for example involving transcription-coupled DNA repair (32) as hypothesized in ref. 11.

Dinucleotide Hotspots as a Signature. TpC (and the complementary GpA) dinucleotides are mutation hotspots for C→G transver-

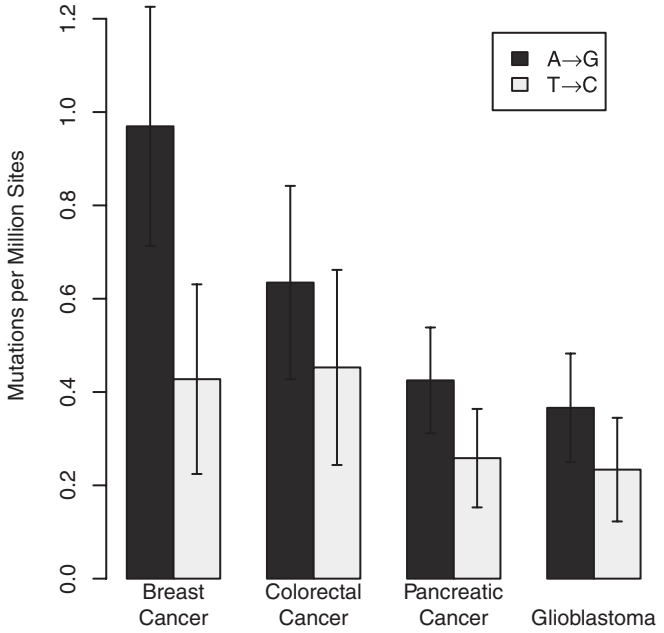

Fig. 2. A→G/T→C mutational asymmetry. Frequencies were computed as for Fig. 1. See Table S5*A* in *SI Appendix* for counts. Error bars indicate 95% confidence intervals.

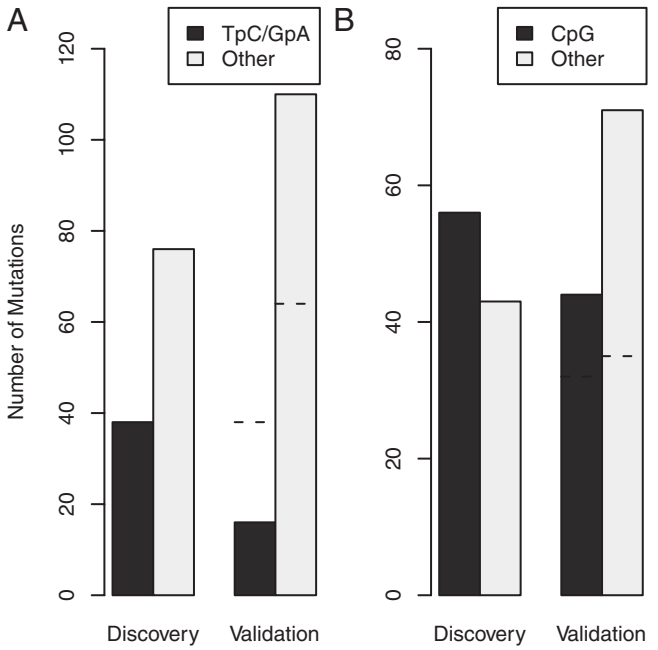

Fig. 3. Screen differences in dinucleotide hotspot mutation proportion. (*A*) breast cancer and (*B*) colorectal cancer. We used mutation data from ref. 1, excluding known cancer genes (breast cancer: TP53; colorectal cancer: APC, KRAS, TP53, SMAD4, FBXW7), genes that were unmutated in one or both screens, and mutations in breast validation tumor BB23, because it was excluded from analysis in ref. 3. Dashed lines indicate expected number of mutations in the validation screen based on discovery screen frequencies, calculated by multiplying the number of bases of each dinucleotide context (CpG in island, CpG in 2 kb shore, other CpG, TpC/GpA, or non-hotspot) sequenced in the validation screen by the appropriate discovery screen mutation rate. Only the C:G base pair in TpC/GpA dinucleotides is considered to be the hotspot. TpCpG and CpGpA trinucleotides were counted as CpGs. See Table S6*A* in *SI Appendix* for counts.

sions at the C:G base pair in breast cancer (1, 16). In addition to breast cancers, three other cancer types sequenced in ref. 2 have a significantly elevated mutation frequency at TpC/GpA dinucleotides: lung cancer, melanoma, and ovarian cancer (see Fig. S4*B* in *SI Appendix*). These results indicate that the TpC/GpA dinucleotide is a mutation hotspot in a subset of cancer types. As discussed above, CpG dinucleotides are a mutation hotspot in breast cancer, colorectal cancer, pancreatic cancer, and glioblastoma (1, 3–5). The relative fractions of mutations occurring in TpC and CpG hotspots vs. other sites may thus be viewed as a "signature" that presumably reflects the nature of the mutational mechanisms in different cancers or their precursor somatic tissues. As such, it may provide a useful tool for monitoring data quality in large sequencing studies.

We compared the proportion of mutations at TpC and CpG sites in the discovery and validation screens for ref. 1 in the genes that were mutated in both screens and found a statistically significant difference between the screens, for each tissue type (Fig. 3) (breast, P = 1.73e-4; colorectal, P = 9.06e-3), with the validation screen samples having a higher proportion of non-hotspot mutations. This difference suggests differing mutational mechanisms underlying the data from the two screens. It is not due to enrichment for cancer-associated genes in the validation screen, because our calculations use the same gene set for both screens. One possible explanation is biological differences in the tissue samples: the breast cancer discovery screen used cell lines, whereas the breast cancer validation screen used primary tumor tissue, and the colorectal cancer validation screen had a higher proportion of xenografts (as opposed to cell lines) than the colorectal cancer discovery screen (1). An alternative possibility is that the difference is due to the whole genome amplification protocol, which was applied to the validation screen samples, but not the discovery screen samples. To discriminate between these possibilities, we examined a subsequent study by the same investigators in ref. 3 that analyzed additional genes in the same samples as in ref. 1 (excluding one validation screen breast tumor) using generally similar methods, but excluding mutations detected in amplified samples that could not be verified by sequencing of the unamplified sample. We analyzed hotspot proportions in the genes sequenced in ref. 3 and found no significant difference between screens (see Fig. S6 in *SI Appendix*). Because the same samples were used in both studies, this excludes sample differences as the primary cause for the discrepancy between screens in ref. 1 and points instead to amplification-induced mutations.

The error rate of the ϕ-29 polymerase used in the whole genome amplification reactions in ref. 1 is at least 12.5 mutations per Mb (33), which is significantly higher than the cancer genome mutation rates reported in recent studies (1). Thus, some strategy is needed to eliminate amplification-induced mutations. In ref. 1, five independent amplification reactions were pooled; our results suggest this strategy was not successful, possibly because of variable amplification efficiency among replicates or reproducibility of particular amplification errors. However, combining it with sequencing of unamplified samples to validate putative mutations appears to have successfully eliminated artifacts in ref. 3. Note that the presence of amplification errors may contribute to the higher overall mutation rate observed in the validation screen for ref. 1, an observation originally attributed to enrichment for cancer-associated genes (1).

Conclusions

Our results indicate that only a small subset of nonsynonymous substitutions in cancer are affected by selection, thus making it possible to interpret substitution trends as reflecting underlying mutational processes. Our analyses eliminate CpG island methylation as a major factor in increased CpG mutation frequencies and detect a mutational asymmetry in breast cancers that may be

linked to gene expression. We also demonstrate the utility of mutation patterns for detecting technical artifacts in cancer sequencing studies.

Methods

Sources of Data. Mutation data were obtained from supplementary online material provided with refs. 1 and 3 (breast and colorectal cancers), ref. 5 (glioblastoma multiforme), ref. 4 (pancreatic cancer), and ref. 2 (protein kinases). Gene sequences and CpG island annotations were downloaded from the University of California, Santa Cruz (UCSC) Genome Browser database, release hg18, except for Fig. 3, which used hg17 (34). Where necessary, we used the UCSC liftOver tool to convert mutation coordinates from human genome build 35 to build 36. For type-specific analyses, tumors were classified by tissue of origin. Normalized gene expression data were obtained from GEO (35) series accession number GSE5764 (31).

Data Filtering. We obtained the list of genes sequenced in ref. 3 and their RefSeq (28) identifiers from the table of PCR primers in supplemental online material (3). We discarded genes that are no longer in RefSeq (release 35) and 68 genes containing a total of 112 nonsynonymous mutations that could not be reconciled with sequences obtained from the UCSC Genome Browser database (34). For gene models that overlap (share a portion of a coding exon), the gene model reported to contain a mutation was selected for analysis, or if no mutation was found, the gene model with the longest coding sequence was selected. Our final gene set contained 17,180 non-overlapping genes present in RefSeq release 35 (March 2009). Genomic coordinates in some FASTA headers were corrected to reflect the fact that short intergenic sequences had been included. We removed all mutations from glioblastoma multiforme tumor BR27P, which had a 17-fold excess of mutations apparently resulting from chemotherapy (5). The majority of mutation data for breast and colorectal cancer is from ref. 1, which did not report synonymous substitutions. For compatibility across studies, synonymous substitutions were excluded from our analyses of data from refs. 3–5, except when analyzing synonymous versus nonsynonymous substitution frequencies. Analyses of data from refs. 3–5 included only those genes analyzed in ref. 3. We also excluded mutations in intronic splice sites. Analysis of mutation data from ref. 1 for Fig. 3 used version 1 of the consensus coding sequence (CCDS) gene set (36) instead of RefSeq.

RefSeq (28) identifiers and sequences were obtained from the UCSC genome browser (release hg18) for 498 of 518 genes analyzed in ref. 2. Point mutation positions reported in ref. 2 were manually curated to resolve inconsistencies between RefSeq sequences and sequences downloaded from the Cancer Genome Project (CGP) database (http://www.sanger.ac.uk/genetics/CGP/Studies/Kinases/). For five genes, we manually corrected apparent frameshift errors in the RefSeq sequences. Genes containing mutations that were inconsistent with CGP sequences were not included. Only coding sequence point mutations identified as part of the main screen were considered. We excluded data from gliomas and acute lymphocytic leukemias (because only the kinase domains were sequenced in these samples), from MMR-deficient tumors, and from cancer types with <20 point mutations after filtering, which left breast, colorectal, gastric, lung, melanoma, ovarian, and renal cancers. We also discarded 47 mutations for which the mutated nucleotide and two flanking nucleotides on each side did not all match in the RefSeq and CGP sequences or which had inconsistent codon positions in the two sequences. Of 1,007 total mutations reported for the main screen in ref. 2, 610 passed these filters.

Analysis. Custom software was written to determine nucleotide context and CpG island membership of bases in sequenced genes. Statistical analysis was performed using the R statistical package (37).

Nonsynonymous frequencies were calculated by dividing the number of mutations at second codon positions (at which substitutions are always nonsynonymous) by the number of codons sequenced. Synonymous mutation frequencies were calculated by dividing the number of mutations at 4-fold degenerate sites (at which substitutions are always synonymous) by the number of 4-fold degenerate sites sequenced. We exclude sites at which both synonymous and nonsynonymous substitutions are possible to avoid issues relating to rate differences between substitution types. Nonsynonymous and synonymous frequencies were calculated separately for each cancer type and dinucleotide context (CpG in island, CpG in 2 kb shore, other CpG, TpC/GpA, or other). Overall nonsynonymous/synonymous frequency ratios were computed by taking a weighted average of frequency ratios for each nucleotide context, using as weights the inverse of the variance of each frequency ratio, calculated as described in section 3.1 of ref. 38. *P* values for overall frequency ratios were calculated using a two-tailed Z test.

Mutation frequencies for data from refs. 1 and 3–5 were calculated by dividing the observed number of nonsynonymous mutations of a given type (e.g. CpG in island) by the number of nonsynonymous sequenced sites of that type using discovery screen data. Mutation frequencies for data from ref. 2 were calculated by dividing the observed number of synonymous and nonsynonymous mutations of a given type by the number of sequenced sites of that type using main screen data. For frequency comparisons, 2×2 contingency tables were constructed with entries equal to the number of mutated and unmutated sequenced bases for each of the two frequencies being compared. *P* values were calculated using Fisher's exact test. Confidence intervals were calculated using the normal approximation to the binomial.

Gene expression data were processed using custom scripts. Single expression values for each RefSeq gene and condition were calculated by taking the mean of probe sets mapping to that gene, according to Affymetrix annotation Build 28. Genes were then ranked by their median expression across all cancer or normal tissue replicates and split into two equal-sized categories (high- and low-expression). Software is available from the authors by request.

ACKNOWLEDGMENTS. We thank Steve Henikoff, Larry Loeb, Ray Monnat, Jesse Salk, Bernard Strauss, Thomas Kunkel, Darryl Shibata, and Raju Kucherlapati for their critical reading of the manuscript and Graham McVicker for helpful discussion. This work was supported by the National Institutes of Health and the Howard Hughes Medical Institute.

1. Sjöblom T, et al. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–274.
2. Greenman C, et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature* 446:153–158.
3. Wood LD, et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318:1108–1113.
4. Jones S, et al. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321:1801–1806.
5. Parsons DW, et al. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* 321:1807–1812.
6. Gojobori T, Li WH, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. *J Mol Evol* 18:360–369.
7. Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol* 21:58–71.
8. Blake RD, Hess ST, Nicholson-Tuell J (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J Mol Evol* 34:189–200.
9. Hess ST, Blake JD, Blake RD (1994) Wide variations in neighbor-dependent substitution rates. *J Mol Biol* 236:1022–1033.
10. Antequera F (2003) Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci* 60:1647–1658.
11. Green P, Ewing B, Miller W, Thomas PJ, NISC Comparative Sequencing Program, Green ED (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* 33:514–517.
12. Majewski J (2003) Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet* 73:688–692.
13. Comeron JM (2004) Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. *Genetics* 167:1293–1304.
14. Frank AC, Lobry JR (1999) Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. *Gene* 238:65–77.
15. Francino MP, Ochman H (2001) Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. *Mol Biol Evol* 18:1147–1150.
16. Stephens P, et al. (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat Genet* 37:590–592.
17. Hwang DG, Green P (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA* 101:13994–14001.
18. Strauss BS (2000) Role in tumorigenesis of silent mutations in the TP53 gene. *Mutat Res* 457:93–104.
19. Costello JF, et al. (2000) Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat Genet* 24:132–138.
20. Esteller M (2002) CpG island hypermethylation and tumor suppressor genes: A booming present, a brighter future. *Oncogene* 21:5427–5440.
21. Shen L, et al. (2007) Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci USA* 104:18654–18659.
22. Irizarry RA, et al. (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 41:178–186.
23. Gonzalgo ML, Jones PA (1997) Mutagenic and epigenetic effects of DNA methylation. *Mutat Res* 386:107–118.
24. The Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

25. Forrest WF, Cavet G (2007) Comment on ''The consensus coding sequences of human breast and colorectal cancers.'' *Science* 317:1500a.
26. Getz G, et al. (2007) Comment on ''The consensus coding sequences of human breast and colorectal cancers.'' *Science* 317:1500b.
27. Rubin AF, Green P (2007) Comment on ''The consensus coding sequences of human breast and colorectal cancers.'' *Science* 317:1500c.
28. Pruitt KD, Tatusova T, Maglott DR (2007) National Center for Biotechnology Information reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35:D61–D65.
29. Bird AP (1986) CpG-rich islands and the function of DNA methylation. *Nature* 321:209–213.
30. Barnes DE, Lindahl T (2004) Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu Rev Genet* 38:445–476.
31. Turashvili G, et al. (2007) Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* 7:55.
32. Hanawalt PC, Spivak G (2008) Transcription-coupled DNA repair: Two decades of progress and surprises. *Nat Rev Mol Cell Biol* 9:958–970.
33. Paez JG, et al. (2004) Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res* 32:e71.
34. Karolchik D, et al. (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* 36:D773–D779.
35. Barrett T, et al. (2007) National Center for Biotechnology Information GEO: Mining tens of millions of expression profiles—Database and tools update. *Nucleic Acids Res* 35:D760–D765.
36. Pruitt KD, et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* 19:1316–1323.
37. R Development Core Team (2008) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).
38. Gart JJ, Nam J (1988) Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* 44:323–338.