Evolutionary Bioinformatics

Libertas Academica
FREEDOM TO RESEARCH

**OPEN ACCESS**
Full open access to this and thousands of other papers at http://www.la-press.com.

ORIGINAL RESEARCH

# Conserved Nonsense-Prone CpG Sites in Apoptosis-Regulatory Genes: Conditional Stop Signs on the Road to Cell Death

Yongzhong Zhao[1] and Richard J. Epstein[2]

[1]Department of Genetics, Mount Sinai School of Medicine, New York, USA. [2]Laboratory of Genome Evolution and Informatics, The Kinghorn Cancer Centre, St. Vincent's Hospital, University of New South Wales, Sydney, Australia. Corresponding author email: repstein@stvincents.com.au

**Abstract:** Methylation-prone CpG dinucleotides are strongly conserved in the germline, yet are also predisposed to somatic mutation. Here we quantify the relationship between germline codon mutability and somatic carcinogenesis by comparing usage of the nonsense-prone CGA (→TGA) codons in gene groups that differ in apoptotic function; to this end, suppressor genes were subclassified as either apoptotic (gatekeepers) or repair (caretakers). Mutations affecting CGA codons in sporadic tumors proved to be highly asymmetric. Moreover, nonsense mutations were 3-fold more likely to affect gatekeepers than caretakers. In addition, intragenic CGA clustering nonrandomly affected functionally critical regions of gatekeepers. We conclude that human gatekeeper suppressor genes are enriched for nonsense-prone codons, and submit that this germline vulnerability to tumors could reflect in utero selection for a methylation-dependent capability to short-circuit environmental insults that otherwise trigger apoptosis and fetal loss.

**Keywords:** nonsense mutations, stop codons, molecular evolution, apoptotic resistance, teratogenesis, carcinogenesis

*Evolutionary Bioinformatics* 2013:9 275–283

doi: 10.4137/EBO.S11759

This article is available from http://www.la-press.com.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article published under the Creative Commons CC-BY-NC 3.0 license.

## Introduction

The 'fifth base' of DNA, 5-methylcytosine, functions as an endogenous mutagen, increasing mutation frequency (C→T, and cross-strand G→A) more than 10-fold.[1] The asymmetry of such mutations in human tumors[2] is not attributable to transcription-coupled repair, translational efficiency, or the Hill-Robertson effect, suggesting that the high frequency of methyl-CpG mutation in cancer-causing genes[3] reflects selection. The existence of such tumorigenic mutational hotspots raises the question as to why such CpG-containing codons are not evolutionarily purged.[4] One illustration of CpG non-suppression relates to codons for arginine, which are encoded either by methylation-prone CG*N* trinucleotides or by more stable AGG/AGA codons. The most drastic CG*N* mutation is the creation of a nonsense codon via single-step deamination of methyl-CGA to TGA[5] (see Fig. 1). Hence, the distribution of CGA codons—identified by Cusack et al as a 'fragile' (nonsense-prone) codon uncommon in single-exon genes[6]—could help to explain why mutable CpG sites are conserved in the germline.

We previously reported that the 2 main subclasses of tumor suppressor genes—DNA repair-type 'caretaker' genes, and pro-apoptotic 'gatekeeper' genes—differ in their phylogenetic behavior: caretakers evolve faster and are more CpG-suppressed than gatekeepers, implying that methylation-dependent mutability is evolutionarily advantageous for repair genes exposed to damage in the male germline.[7] A similar defensive role has been proposed for the evolution of DNA methylation.[8] Although germline mutation is less well tolerated for gatekeepers than for caretakers, mutation during somatic tumorigenesis is more frequent for gatekeepers,[7] with ~50% such mutations arising from methyl-CpG mutation.[9] Furthermore, many carcinogenetic errors in gatekeeping genes like *APC* are nonsense mutations,[10,11] consistent with a crucial role for loss of apoptosis in tumors.

Apoptosis also underlies the pattern-forming activities of embryogenesis, however. Environmental threats to the fetus include teratogenic exposures such as hyperthermia, xenobiotics or oxidative damage,[12,13] any of which may drive apoptosis[14,15] and thus cause birth defects such as limb truncations or microphthalmia.[16] Such teratogen-induced apoptosis is mediated by gatekeeper genes like *TP*53,[17] and may be prevented by DNA methylation.[18] Low-level exposure to pro-apoptotic teratogens could trigger a negative-feedback inhibition of embryonic gatekeeper gene function, whether via promoter methylation, nonsense-mediated mRNA decay, or methylation-dependent point mutation, limiting teratogenesis.[15] Here we examine the relation between germline CpG retention and somatic mutation by assessing the conservation of CGA

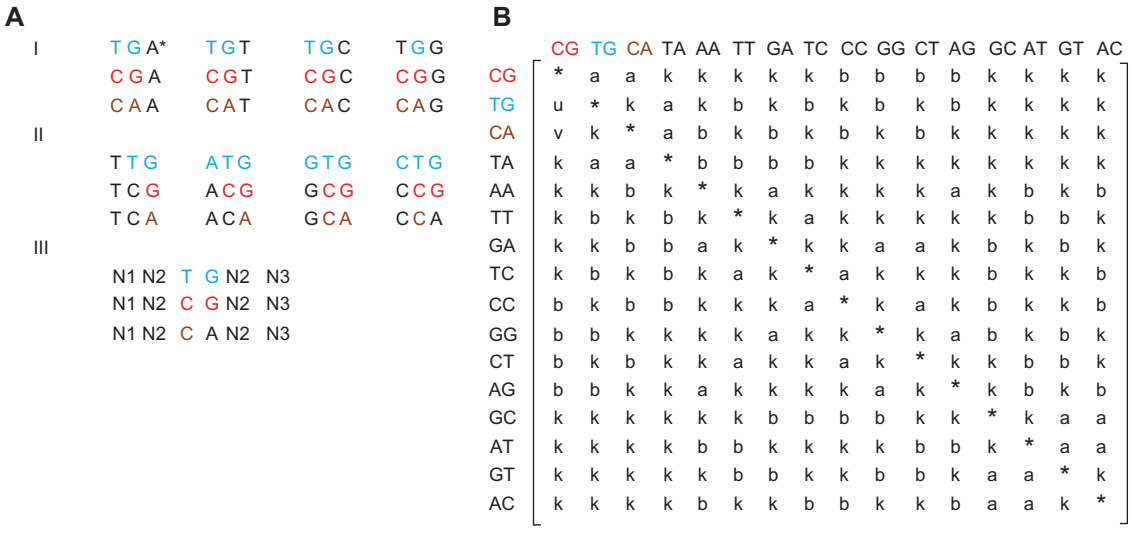


**Figure 1.** Strand- and frame-specific dinucleotide mutation model. (**A**) Open reading frame and strand specific cataloging of CpG-dependent deamination possibilities: I. frame 1, 2, including arginine encoding codons; II, frame 2,3; III, frame 3, 1. *Indicates nonsense mutation, ie, untranscribed strand CGA to TGA mutation. (**B**) Markov transition matrix with 5 parameters for each frame (total 15 parameters), including transition rate *a*, transversion rate *b*, untranscribed strand CpG deamination rate *u*, transcribed CpG deamination rate *v*, and dinucleotide substitution rate *k*. We define the asymmetry parameter *A* in terms of strand-specific methylation/deamination, $A = u/v$.

codon patterns in gatekeepers and non-apoptotic genes.

## Materials and Methods

### Biostatistical database analyses

Listings of human cancer-related genes were created using classifications of viral oncogenes and familial cancer genes[7] (Supplemental Table S1). Databases were compared in terms of nucleotide composition (GC%), intragenic CpG sites, and stop codon frequencies using ClustalW for alignment of human-mouse orthologs and CodonW for codon pattern analysis. Scripts were written in PERL 5.8.6. Human and mouse reference sequences were downloaded from NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/Entrez/Gene), and mutation data from the Human Gene Mutation Database. A variety of packages from R 2.14.1 (http://www.r-project.org) were used for statistical analysis, including coin, biomaRt, GeneR and nlmc. For analysis of multiple splicing forms, the longest coding sequence was used; mono- and dinucleotide composition was assessed using Perl scripts and/or the GeneR package in R2.14.1. Comparison of mutations in tumors was based on the Cancer Genome Anatomy Project Cancer Gene Census (http://www.sanger.ac.uk/genetics/CGP/Census).[19,20] Frame-dependent dinucleotide composition and asymmetries were analyzed using GeneR. For the analyses of 5′ and 3′ untranslated regions (UTR), reference sequences were downloaded from ENSEMBL (Release 52) using R package Biomart (http://www.r-project.org and http://bioconductor.org).

### Biomathematical calculations

We derived a model to quantify the asymmetry of DNA mutations between 2 DNA strands. The model is a binomial distribution; ie, for the total of $n$ mutations at the same double-stranded nucleotide site—which by definition will have a probability of 0.5 if symmetric—if we observe $x$ mutations of $n$ total mutations in 1 strand, then:

$$P = 1 - \sum_{i=0}^{x} \binom{n}{x_i} 0.5^n$$

However, for tests of $m$ codons, we need a Dunn-Šidák correction, such that:

$$\alpha_{ds} = 1 - (1 - \alpha)^m$$

Therefore:

$$\alpha = 1 - (1 - \alpha_{ds})^{1/m}$$

So the critical value $Y_c$ of our test is:

$$\sum_{i=0}^{Y_c - 1} \binom{n}{Y_c - 1} 0.5^n = 1 - (1 - \alpha_{ds})^{1/m}$$

The statistical power is:

$$P_{power} = 1 - \beta = \sum_{i=0}^{Y_c - 1} \binom{n}{Y_c - 1} 0.5^n$$

Our analysis also sought to quantify the extent to which nonsense-prone codons are localized towards the 5′ or 3′ sense strand, corresponding to the N-terminus or C-terminus of the peptide encoded, implying greater or lesser phenotypic effects, respectively, in the event of a nonsense mutation.

For the frequency of a selected codon $f$, for observed first position $w$, the first codon generally is fixed, such as ATG or GTG, such that we have a geometric distribution,

$$P_{N-bias} = 1 - \sum_{i=2}^{w} f\ (1 - f)^{w_i - 2}$$

Similarly, for tests of $u$ codons, we have a Dunn-Šidák correction,

$$\alpha_{ds} = 1 - (1 - \alpha)^u$$

Therefore,

$$\alpha = 1 - (1 - \alpha_{ds})^{1/u}$$

Such that the critical value of our test is,

$$\sum_{i=2}^{Y_c - 1} f\ (1 - f)^{Y_i - 2} = 1 - (1 - \alpha_{ds})^{1/u}$$

In turn, the statistical power is,

$$P_{power} = 1 - \beta = \sum_{i=2}^{Y_c - 1} f(1-f)^{Y_i - 2}$$

## Codon cluster analysis

For codon cluster analysis, we calculated a negative binomial distribution cumulative mass function:

$$X \sim BN(2, k/K)$$

where $K$ is the number of total codons-2; $k$, total number of given codons; $X$, the distance of adjacent given codons in codon number. The positions of selected codons were computed sequentially, with a critical value set as defined above, where $u$ = the total selected codons in the gene of interest, such that for distance $d$, $d < Y_c$, a cluster is defined. This can be completed recursively, so that we generated pseudo-codes according to the following steps: (1) compute selected codon frequency, $f = k/K$; (2) compute codon positions; (3) set critical distance value; (4) recursively compute cluster with this critical value; (5) plot such clusters.

## Nonsense mutations in cancer gene census, and phylogenomics of nonsense-prone codons

We downloaded the cancer gene census (http://www.sanger.ac.uk/genetics/CGP/Census/, dataset version updated on March 15 2012) and the cancer encyclopedia (947 cell lines).[21] From the public database of whole-genome sequencing, and open reading frames therein, we computed codon usage for the nonsense-prone codons using the Biomart database with R statistical computing (http://www.r-project.org and http://bioconductor.org), based on the most up-to-date data of 29 mammals.[22] We aligned these nonsense-prone codons in the above dataset, and computed the cluster pattern. For loss of function analysis in mouse genes, we mined data from the mouse genome informatics database.

# Results

## Asymmetric pattern of codons predisposing to nonsense mutation in a single step

Next-generation sequencing technologies have enabled population genetic information to be available at the whole-genome level, making it possible to visualize nonsense mutation patterns using the depicted graph model of the full repertoire (Fig. 2) based on 1000-genome data.[23] Among the total of 559 nonsense mutations, 206 (36.85%) mutants were G to A; eg, TGG to either TGA or TAG; it is thus possible to quantify the asymmetry of TGG-associated nonsense mutations using this approach. Hence, of the 64 codons in the human genes, 18 can mutate to nonsense mutations with in a single step;[6,24] there are 23 nonsense trajectories so defined, including 7 codons mutable to TAA (the ancestral stop codon) and 8 codons each for TGA or TAG. The asymmetry of this layer derives from the dual trajectories of synonymous stop codon mutations to TAA (ie, TGA to TAA, and TAG to TAA), as compared to only 1 path towards either TGA or TAG (Fig. 2, dotted lines). Moreover, of the 18 nonsense prone codons, 9 (with 10 paths) arise from the first codon position, including 3 trajectories with methylation-related codons (CGA, CAG and CAA); 5 codons with 6 paths arise from the second codon position, including 1 methylation-related codon path (TGG to TAG); while 5 codons with 7 paths arise from the third codon position, including one methylation related codon (TGG to TGA). Interestingly, the codon TGG, which encodes tryptophan, is susceptible to both second- and third-codon position nonsense mutations; in addition, if the broader 'methylation' view is considered (ie, rather than CpG only), we note that the antisense strand of the TGG codon is also predisposed to methylation-related nonsense mutation, whereas CGA, CAG, and CAA are nonsense-prone only on the sense strand. Accordingly, we submit that this asymmetry of nonsense-propensity could link codon methylation and transcription.

## Phylogenetic correlations between stop codon and nonsense-prone codon frequency

A positive correlation exists between species-specific genome GC content and TGA stop codon frequency, as well as a negative correlation with TAA stop codon frequency (Supplemental Fig. 1). There is a similarly strong correlation between species-specific CGA and TGA codon contents (Fig. 3; $P = 0.013$). These findings support the view that TGA stop codons arise by single-stranded methyl-CGA deamination events (ie, predominating in lightly-methylated genomes

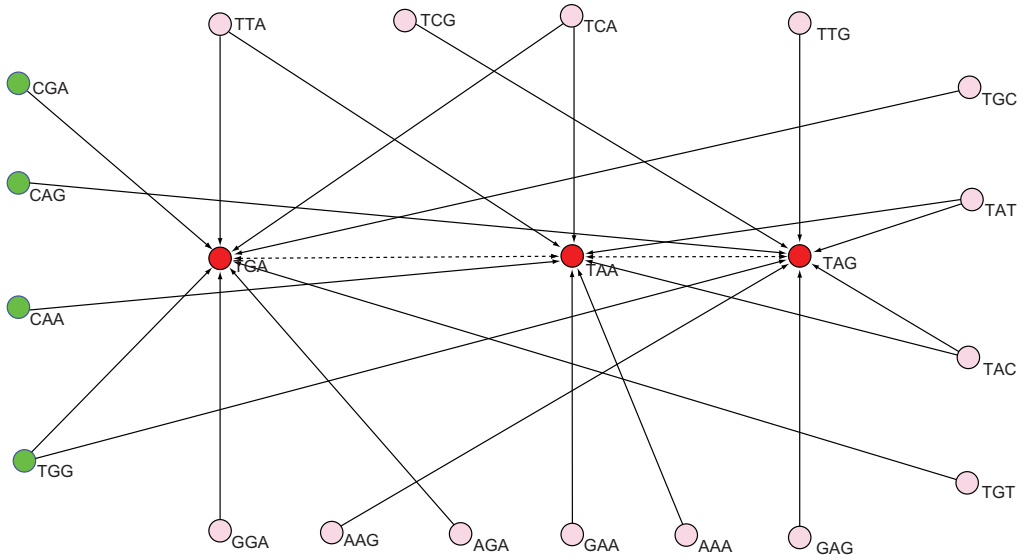


**Figure 2.** 1-step pathways to nonsense mutations, highlighting the C-to-T deamination trajectories. Vertices representing stop codons are labeled red, whereas the 4 trajectories of synonymous stop codon interchange are represented as dotted lines. 23 1-step pathways to stop codons for amino acid encoding codons were labeled with solid line; the vertices representing the 4 codons predisposed to C to deamination which is enhanced when undergoing DNA methylation are labeled with green color. Codons predisposed to nonsense mutation at the first, second, or the third codon positions are depicted at the bottom, the upper, or the left side successively. The graph was drawn using Pajek software (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

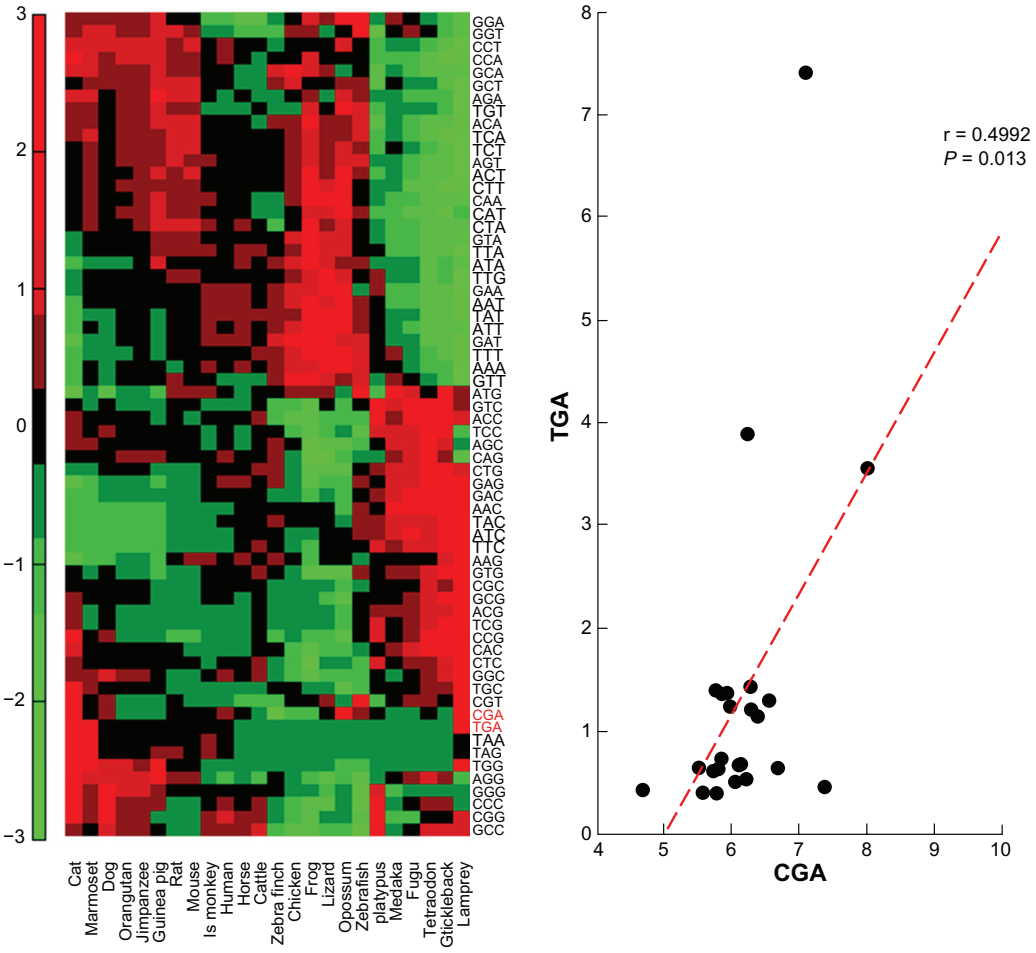


**Figure 3.** Heat-map graphical analysis of relationship between CGA and TGA codon content. Sequences of 24 species from UCSC genome site were analyzed.

with higher residual GC content), whereas TAA stop codons tend to arise from double-stranded methyl-CGA mutations in AT-rich genomes. Moreover, of 328 tumorigenic (somatic) CGA mutations in human tumor suppressor genes, 321 involved formation of a stop (TGA) codon rather than a missense mutation (CAA; Table 1), confirming a selectable advantage for loss of function in tumors.

## Predilection of nonsense-prone codons for gatekeeper over caretaker suppressor genes

Of 129 instances of methylation-dependent CGA mutation affecting gatekeeper genes in tumors, 119 were nonsense mutations (CGA→TGA) versus 10 missense (CGA→CAA; $P = 3.94 \times 10^{-25}$; Table 2). Comparing the frequency of CGA→TGA mutations affecting the 2 main classes of tumor suppressor genes, the pro-repair caretakers (141 CGA codons of 19 genes) and the pro-apoptotic gatekeepers (181 CGA codons of 35 genes), greater selection pressure for nonsense mutations is evident for gatekeepers (119 mutated versus 52 non-mutated) relative to caretakers (57 mutated versus 79 non-mutated; $\chi$-square = 23.7, $P = 1.6 \times 10^{-6}$). This represents a 3-fold higher risk of such mutations in gatekeeper than in caretaker genes (OR = 3.172, 95% CI 1.98–5.082; $P < 0.0003$, using Pearson's $\chi$-square = 13.35, with Yates continuity correction).

## Nonrandom spatial intragenic clustering of nonsense-prone codons in gatekeeper genes

The canonical gatekeeper suppressor gene *TP53* exhibits an inverse relationship between the amino acid site-specificity of sporadic carcinogenic mutations and evolutionary rate (Supplemental Fig. 2). As shown in Supplemental Table 2, *TP53* also contains 4 CGA sites at positions 196, 213, 306, 342, the *P* value of the 196/213 cluster being 0.0175; whereas for the 306/342 cluster, $P = 0.0575$. Moreover, for all 4 CGA codons, the calculated probability is still significant, $P = 0.0699$; similarly, for all 3 CGG sites, $P = 0.0101$. In contrast, for the 3 AGA sites, $P = 0.2361$, while for the 2 closest sites, $P = 0.1104$. Again, for the 3 AGG sites, no pair of sites reached significant levels of clustering ($P = 0.1849$; $P = 0.1204$, respectively). These results confirm that CGN clustering, unlike arginine clustering per se, has a nonrandom (selectable) significance.

We also note that CpG-containing arginine codons (CGN) tend to be localized to the central or 3′ end of the *TP*53 gene (cf. NCG codons, situated mainly in the 5′ region). CGG codons, which typically give rise to missense mutations, cluster in the 3′ end of the DNA-binding domain where they are bounded by the 2 CGA clusters. This CpG microanatomy suggests 3 broad categories of pre-programmed methylation-dependent mutation: C-terminal deletions affecting the oligomerization/RUNT domains, 3′ DBD missense mutations, or more drastic 5′ DBD deletions (this result agrees with that of Yang et al, who reported that missense mutations are more common within essential tumor suppressor gene domains, whereas nonsense mutations cluster in linked regions).[25] Our statistical analysis confirms that these nonsense-prone codons correlate to calpain or caspase cleavage sites; hence, as an extension of the Anfinsen dogma, we infer that protein folding and degradation information are primarily encoded in the codon (ie, nucleotide) sequence

**Table 1.** Asymmetric pattern of CG*N* codon germline mutation in tumor suppressor genes.

| Gene group | Gene | CGA | | CGT | | CGG | |
|---|---|---|---|---|---|---|---|
| | | TGA | CAA | TGT | CAT | TGG | CAG |
| Gate-keepers | *TP53* | 10** | 0 | 12 | 10 | 16 | 17 |
| | *RB1* | 126** | 0 | 0 | 0 | 0 | 0 |
| | *APC* | 125** | 0 | 0 | 0 | 3 | 0 |
| | *VHL* | 18* | 7 | 0 | 0 | 24 | 28 |
| Care-takers | *ATM* | 4 | 0 | 0 | 0 | 0 | 2 |
| | *MSH2* | 6* | 0 | 0 | 0 | 0 | 0 |
| | *MLH1* | 26** | 0 | 0 | 0 | 0 | 0 |
| Total | | 321** | 7 | 12 | 10 | 43 | 47 |

**Notes:** $^{**}P < 0.001$; $^{*}P < 0.05$, based on binomial distribution.

**Table 2.** Comparison of asymmetry in CGA mutations of CTs versus GKs.

| Mutation at CpG sites | Caretakers | Gatekeepers |
|---|---|---|
| Total CGA sites | 141 | 181 |
| Unmutated | 79 | 52 |
| CGA→CAA | 3 | 10 |
| CGA→TGA | 57 | 119 |
| Asymmetry *P* value | $3.13 \times 10^{-14}$ | $3.94 \times 10^{-25}$ |
| Missense mutations (total) | 716 | 973 |
| Nonsense mutations (total) | 451 | 813 |

rather than amino acid sequence; the N-end (protein cleavage and degradation) rule thus reflects a direct link between genetic and epigenetic information.

The nonrandom arrangement of CGA codon clusters in a further sample of gatekeeper genes (*RB1, NF1*, and *HPRT2*) are illustrated in Supplemental Figure 3. The difference between the intragenic topography of these CG*N* codons and their AGA/AGG equivalents is detailed for the *RB1* gene in Supplemental Table 3. Both the frequency of clustering and the statistical significance of the clustering is greater for CG*N* than for AG*N* codons. The frequency of clustered codons is 13/16 (81%) for CGA, 2/4 (50%) for CGG, 6/13 (46%) for AGA, and 2/6 (33%) for AGG; of these clusters, 100% were significant for CGA and CGG, but only 33% for AGA.

## Discussion

Our study shows that pro-apoptotic gatekeepers are more often mutated in adult-onset tumors than are repair-style caretakers, and that this somatic mutability correlates with an abundance of hypermutable CpG-containing codons that selectively predispose to protein-truncating nonsense mutations. Why should such an apparently maladaptive vulnerability remain conserved within pro-apoptotic genes despite availability of more stable codons? Teratogenic drugs like thalidomide, retinoids and methotrexate all have established anti-cancer utility, reflecting their ability to enhance apoptosis, whereas apoptosis is reduced via epigenetic repression of pro-apoptotic tumor suppressor genes[26] by carcinogens (eg, from smoking) as well as by DNA-damaging heavy metal poisoning[27] or oxygen radicals.[28] Teratogens like diethylstilbestrol – a tumorilytic (pro-apoptotic) drug which, like decitabine and retinoids,[29] triggers initial genomic hypomethylation – could thus select for an abundance of methylation-induced gatekeeper (epi) mutations in utero[30] with the long-term result of adult tumors like vaginal clear cell carcinoma supervening.[31]

Stress-induced mutagenesis is an incompletely understood evolutionary process that benefits fitness.[32,33] Our study supports the latter view by suggesting that CpG sites may act as methylation-sensitive 'sensors' of microenvironmental threats in utero, while also acting as effectors of transcriptional repression (in the case of CpG island methylation and/or nonsense-mediated mRNA decay). Given that germline gatekeepers are highly conserved relative to caretakers, it is surprising to note that somatic nonsense mutations occur more often than missense mutations in gatekeepers. These results suggest that CGA is conserved in gatekeepers as a 'conditional stop' signal that protects developing embryos from excessive apoptosis and miscarriage, while simultaneously predisposing ageing adults to cancer. Since gatekeeper promoter methylation is a common response to DNA damage in adult tissues[26] that increases mutability of intragenic methyl-CpG sites by reducing transcription and repair, we submit that noxious insults in utero could select for such methylation-dependent mutability.

This study confirms for the first time that epigenetic modification potential—whether germline or somatic—is encoded within the germline DNA sequence, thus raising central questions as to mechanisms of synonymous germline codon sequence conservation. To this end we have initiated new work using synonymous CpG-variable codon constructs in vivo to test the somatic and carcinogenetic consequences predicted by our findings here.

## Acknowledgements

We thank Professor Allan Spigelman and the St. Vincent's Clinic Foundation for support.

## Author Contributions

Conceived and designed the experiments: RJE. Analyzed the data: YZ, RJE. Wrote the first draft of the manuscript: RJE. Contributed to the writing of the manuscript: YZ, RJE. Agree with manuscript results and conclusions: YZ, RJE. Jointly developed the structure and arguments for the paper: YZ, RJE. Made critical revisions and approved final version: YZ, RJE. All authors reviewed and approved of the final manuscript.

## Funding
Author(s) disclose no funding sources.

## Competing Interests
Author(s) disclose no potential conflicts of interest.

## Disclosures and Ethics
As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

## References
1. Rideout WM, Coetzee GA, Olumi AF, Jones PA. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science*. 1990;249(4974):1288–90.
2. Rodin SN, Rodin AS. Strand asymmetry of CpG transitions as indicator of G1 phase-dependent origin of multiple tumorigenic p53 mutations in stem cells. *Proc Natl Acad Sci U S A*. 1998;95(20):11927–32.
3. Soussi T, Béroud C. Significance of TP53 mutations in human cancer: a critical analysis of mutations at CpG dinucleotides. *Hum Mutat*. 2003;21(3): 192–200.
4. Cooper DN, Gerber-Huber S. DNA methylation and CpG suppression. *Cell Differ*. 1985;17:199–205.
5. Mort M, Ivanov D, Cooper DN, Chuzhanova NA. A meta-analysis of nonsense mutations causing human genetic disease. *Hum Mutat*. 2008;29(8):1037–47.
6. Cusack BP, Arndt PF, Duret L, Roest Crollius H. Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. *PLoS Genet*. 2011;7(10):e1002276.
7. Zhao Y, Epstein RJ. Programmed genetic instability: a tumor-permissive mechanism for maintaining the evolvability of higher species through methylation-dependent mutation of DNA repair genes in the male germ line. *Mol Biol Evol*. 2008;25(8):1737–49.
8. Aravin AA, Sachidanandam R, Bourc'his D, et al. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell*. 2008;31(6):785–99.
9. Krawczak M, Smith-Sorensen B, Schmidtke J, Kakkar VV, Cooper DN, Hovig E. Somatic spectrum of cancer-associated single basepair substitutions in the TP53 gene is determined mainly by endogenous mechanisms of mutation and by selection. *Hum Mutat*. 1995;5(1):48–57.
10. Floquet C, Rousset JP, Bidou L. Readthrough of premature termination codons in the adenomatous polyposis coli gene restores its biological activity in human cancer cells. *PLoS ONE*. 2011;6(8):e24125.
11. Zilberberg A, Lahav L, Rosin-Arbesfeld R. Restoration of APC gene function in colorectal cancer cells by aminoglycoside- and macrolide-induced read-through of premature termination codons. *Gut*. 2010;59(4):496–507.
12. Knobloch J, Schmitz I, Götz K, Schulze-Osthoff K, Rüther U. Thalidomide induces limb anomalies by PTEN stabilization, Akt suppression, and stimulation of caspase-dependent cell death. *Mol Cell Biol*. 2008;28(2): 529–38.
13. Kochhar DM, Jiang H, Harnish DC, Soprano DR. Evidence that retinoic acid-induced apoptosis in the mouse limb bud core mesenchymal cells is gene-mediated. *Prog Clin Biol Res*. 1993;383B:815–25.
14. Toder V, Carp H, Fein A, Torchinsky A. The role of pro- and anti-apoptotic molecular interactions in embryonic maldevelopment. *Am J Reprod Immunol*. 2002;48(4):235–44.
15. Torchinsky A, Fein A, Toder V. Teratogen-induced apoptotic cell death: does the apoptotic machinery act as a protector of embryos exposed to teratogens? *Birth Defects Res C Embryo Today*. 2005;75(4):353–61.
16. Knobloch J, Rüther U. Shedding light on an old mystery: thalidomide suppresses survival pathways to induce limb defects. *Cell Cycle*. 2008;7(9):1121–7.
17. Hosako H, Little SA, Barrier M, Mirkes PE. Teratogen-induced activation of p53 in early postimplantation mouse embryos. *Toxicol Sci*. 2007;95(1):257–69.
18. Carambula SF, Oliveira LJ, Hansen PJ. Repression of induced apoptosis in the 2-cell bovine embryo involves DNA methylation and histone deacetylation. *Biochem Biophys Res Commun*. 2009;388(2):418–21.
19. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer*. 2004;4(3):177–83.
20. Vogelstein B, Kinzler KW. *The Genetic Basis of Human Cancer*, 2*nd Ed*. 2002; New York: McGraw-Hill.
21. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603–7.
22. Lindblad-Toh K, Garber M, Zuk O, et al. Broad Institute Sequencing Platform and Whole Genome Assembly Team; Baylor College of Medicine Human Genome Sequencing Center Sequencing Team; Genome Institute at Washington University. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*. 2011;478(7370):476–82.
23. MacArthur DG, Balasubramanian S, Frankish A, et al. 1000 Genomes Project Consortium. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 2012;335(6070):823–8.
24. Modiano G, Battistuzzi G, Motulsky AG. Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? *Proc Natl Acad Sci U S A*. 1981;78(2):1110–4.
25. Yang Z, Ro S, Rannala B. Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics*. 2003;165(2):695–705.
26. Toyooka S, Toyooka KO, Miyajima K, et al. Epigenetic down-regulation of death-associated protein kinase in lung cancers. *Clin Cancer Res*. 2003;9(8):3034–41.
27. Salnikow K, Zhitkovich A. Genetic and epigenetic mechanisms in metal carcinogenesis and cocarcinogenesis: nickel, arsenic, and chromium. *Chem Res Toxicol*. 2008;21(1):28–44.
28. Franco R, Schoneveld O, Georgakilas AG, Panayiotidis MI. Oxidative stress, DNA methylation and carcinogenesis. *Cancer Lett*. 2008;266:6–11.
29. Kuriyama M, Udagawa A, Yoshimoto S, et al. DNA methylation changes during cleft palate formation induced by retinoic acid in mice. *Cleft Palate Craniofac J*. 2008;45(5):545–51.
30. Sato K, Fukata H, Kogo Y, Ohgane J, Shiota K, Mori C. Neonatal exposure to diethylstilbestrol alters expression of DNA methyltransferases and methylation of genomic DNA in the mouse uterus. *Endocr J*. 2009;56(1):131–9.
31. Manning FC, Patierno SR. Apoptosis: inhibitor or instigator of carcinogenesis? *Cancer Invest*. 1996;14(5):455–65.
32. Koonin EV, Wolf YI. Is evolution Darwinian or/and Lamarckian? *Biol Direct*. 2009;4:42.
33. Rosenberg SM. Mutation for survival. *Curr Opin Genet Dev*. 1997;7: 829–34.

## Supplementary materials

Supplementary Table 1. List of genes analyzed.

Supplementary Table 2. Spatial distribution of hypermutable CGA(CGN) codons in *TP53* gene. The cluster model is calculated as a negative binomial distribution of codons.

Supplementary Table 3. Significantly clustered distribution of CGA (13/16) > AGA (2/13) codons in the *RB1* gene.

Supplementary Figure 1. Phylogenetic relationship between genomic GC content and the frequency of either stop codons or CGA codons.

Supplementary Figure 2. Inverse relationship between *TP*53 CpG somatic mutation rates (upper diagram, blue) and germline conservation (lower diagram, Ka/Ks, red).

Supplementary Figure 3. Mapping of CGA positions in gatekeeper genes, showing non-random clustering.