Guanine Holes Are Prominent Targets for Mutation in Cancer and Inherited Disease

Albino Bacolla1,2, Nuri A. Temiz2*na, Ming Yi2, Joseph Ivanic2, Regina Z. Cer2, Duncan E. Donohue2, Edward V. Ball1, Uma S. Mudunuri2, Guliang Wang1, Aklank Jain1nb, Natalia Volfovsky2, Brian T. Luke2, Robert M. Stephens2, David N. Cooper3, Jack R. Collins2, Karen M. Vasquez1*

1 Division of Pharmacology and Toxicology, The University of Texas at Austin, Dell Pediatric Research Institute, Austin, Texas, United States of America, 2 Advanced Biomedical Computing Center, SAIC-Frederick, Inc., Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States of America, 3 Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom

Abstract

Single base substitutions constitute the most frequent type of human gene mutation and are a leading cause of cancer and inherited disease. These alterations occur non-randomly in DNA, being strongly influenced by the local nucleotide sequence context. However, the molecular mechanisms underlying such sequence context-dependent mutagenesis are not fully understood. Using bioinformatics, computational and molecular modeling analyses, we have determined the frequencies of mutation at G+C bp in the context of all 64 5′-NGNN-3′ motifs that contain the mutation at the second position. Twenty-four datasets were employed, comprising >530,000 somatic single base substitutions from 21 cancer genomes, >77,000 germline single-base substitutions causing or associated with human inherited disease and 16.7 million benign germline single-nucleotide variants. In several cancer types, the number of mutated motifs correlated both with the free energies of base stacking and the energies required for abstracting an electron from the target guanines (ionization potentials). Similar correlations were also evident for the pathological missense and nonsense germline mutations, but only when the target guanines were located on the non-transcribed DNA strand. Novel candidate driver mutations and tissue-specific mutational patterns were also identified in the cancer datasets. We conclude that electron transfer reactions within the DNA molecule contribute to sequence context-dependent mutagenesis, involving both somatic driver and passenger mutations in cancer, as well as germline alterations causing or associated with inherited disease.

Introduction

At least fifteen cancer genome sequencing projects were reported between 2007 and 2011 [1–15], and this number is now increasing very rapidly. These studies have been critical for addressing mechanisms of somatic mutation, such as those associated with single base substitutions (SBSs), which not only represent the vast majority of lesions in most patients, but also (in the case of some driver mutations) alter gene function, thereby initiating tumor development. Such investigations have demonstrated that SBSs do not occur randomly throughout the genome. Indeed, frequent C→T transitions have been noted at CpG dinucleotides [4,9,16–18], which are attributable to the high rate of spontaneous 5-methylcytosine (5mC) deamination at methylated 5mCpG sites [19,20]. In individuals with a history of exposure to cigarette smoke or radio/chemotherapy, high proportions of G→T transversions, G→A and G→C substitutions at GpA and CpG dinucleotides, and A→T and A→G substitutions at TpA dinucleotides have also been reported [6,8,9,21], suggestive of DNA damage through exogenous mechanisms [22,23]. Likewise, large numbers of C→T transitions at YpC (Y = C/T) dinucleotides in melanoma in sun-exposed areas of the skin [11,21,24] have been attributed to cyclobutane pyrimidine dimer (CPD) formation following UV photoexcitation [25]. For less common types of substitution, such as T→C at ApT dinucleotides in hepatocellular carcinoma [16], underlying mutational mechanisms have yet to be proposed.

Studies aimed at identifying the mechanisms underlying the sequence context dependency of SBSs observed in inherited disease [26], cancer and phylogenetic analyses [20,27–31] are few in number. A recent analysis of breast cancer genomes identified five types of trinucleotide motif enriched in SBSs, all of which contained either a CpG or a GpA motif [18]. Substitutions at CpG were attributed to 5mCpG deamination, whereas mutations at the
Author Summary

A large number of DNA mutations identified in cells from patients with cancer or human inherited disease were analyzed to address a fundamental issue in human pathology, viz, the mutational mechanisms that cause irreversible changes to DNA. By using bioinformatics and computational methods, we found that mutations do not occur randomly, but instead affect specific bases, most often guanines flanked by other guanines or adenines. We attribute this effect to electron transfer, a chemical reaction known to underlie basic biological processes such as cellular respiration and photosynthesis. Certain types of carcinogens, oxidants or radiation can interact with DNA and abstract an electron. Our results imply that the ensuing sites of electron loss can migrate from their original position in the DNA to neighboring guanines where they become trapped, leading to further chemical modifications that may eventually result in mutations. Many of the mutations known to be important for tumor growth (driver mutations), as well as passenger mutations and mutations associated with inherited disease, appear to be caused by electron transfer. Beyond pathological mutations, electron transfer may represent a universal mechanism by which genetic changes occur in all life forms to drive population fitness over evolutionary time.

Results

SBSs Occur Preferentially at G-C Base-Pairs (bp)

We collected the publicly available data from cancer genome studies reported in PubMed between 2007 and 2011, together with the 5 largest datasets from the International Cancer Genome Consortium (ICGC) (Table 1). Twenty-one datasets, 13 from exome-wide (EWS) and 8 from genome-wide (GWS) sequencing scans, together represented cancers from 14 different tissues comprising 1,149 patient samples (1 to 393 samples per dataset) and 2 cell lines, for a total of 533,482 SBSs (Table S1). Additional SBSs from 3 other datasets were included in the analysis: a subset of nonsense and missense germline mutations of pathological significance in the context of inherited disease derived from the Human Gene Mutation Database (HGMD); a subset of splicing mutations from HGMD causing human inherited disease; and the set of SNVs from the “1000 Genomes Project” included in dbSNP build 129 (“rs” set), to allow direct comparison of the cancer-associated somatic mutations with germline mutations and polymorphisms present in the general population. The median fractions of SBSs that occurred at G-C bp were 0.78 (mean ± SD, 0.75 ± 0.11) for the EWS datasets and 0.56 (mean ± SD, 0.60 ± 0.13) for the GWS datasets (Fig. 1A), significantly higher than the average GC-content exome-wide (0.55) [38] and genome-wide (0.41), respectively [39] (both P values were ~2.2 x 10^-16, the smallest computable number by implementation of the binomial exact test in R). Thus, SBSs occurred more frequently at G-C bp, as compared to A-T bp, than expected by chance alone, both in cancer genomes and in the germline as a cause of inherited disease.

SBS Frequencies Are Modulated by Flanking Base Composition

We addressed the sequence-dependent occurrence of SBSs at G-C bp by retrieving the 5’-NGNN-3’ and their complementary 5’-NNCN-3’ sequences (henceforth referred to as NGNN), either genome-wide or exome-wide, and calculating the fractions of mutated motifs, f(NGNN), for each of the 64 sequence combinations. There were two confounding factors in computing f(NGNN): the first related to the fact that only certain portions of the human genome may be effectively mapped by next-generation sequencing; the second originating from the various methodologies used during base-variant mapping and calling (see Materials and Methods). Therefore, we first assessed the relative representation of NGNN sequences in the homologous portions (Segmental Duplications, repetitive elements and simple repeats) as compared with the unique portions of the human genome (Text S1 and Table S1). These analyses indicated that CGNN motifs are significantly overrepresented in Segmental Duplications (Text S1 and Table S2). We then used the genomic mutation sites (Table S3) to compute f(NGNN) according to three methods (Duke35, CRG50 and T_hgt9) for the GWS datasets and three methods (AgilentV2, CGR50_exons and T_exons) for the EWS datasets (Text S1), and used these fractions to assess the extent to which base composition at positions 1, 3 and 4 (P1, P3 and P4) would influence G-C mutations at position 2 (P2) for different classes of NGNN sequences. Mutations were observed at each NGNN sequence combination with the exception of the two small colorectal cancer datasets (Table 1; Table S4, Panel A). Thus, for these two datasets, z-tests, rather than t-tests, were used to assess statistical significance. Irrespective of the mapping method used, f(CGNN) mean values were significantly greater than f(DGNN) (D = A/G/T) mean values in all cancer (2–11 fold, depending on cancer type with gastric and colorectal cancers...
## Table 1. List of datasets and SBSs at G:C base pairs.

| Dataset                          | # Exome or genome wide seq. | Number of mutations | Fraction of SBSs at G:C bp | \(-\log P\) CGNN vs. DGNN | \(-\log P\) DGRN vs. DGYN | Number of samples | Reference |
|----------------------------------|-------------------------------|---------------------|-----------------------------|---------------------------|---------------------------|-------------------|-----------|
| Acute myeloid leukemia           | EWS                           | 1834                | 0.51                        | 12                        | ns                        | 1                 | [1]       |
| Breast                           | EWS                           | 1344                | 0.80                        | 8                         | 3                         | 42                | [2]       |
| Chronic lymphocytic leukemia     | GWS                           | 4048                | 0.52                        | 19                        | ns                        | 4                 | [3]       |
| Colorectal_1                     | EWS                           | 1153                | 0.81                        | >3                        | ns                        | 38                | [2]       |
| Colorectal_2                     | EWS                           | 686                 | 0.79                        | >3                        | ns                        | 36                | ICGC^2    |
| Gastric                          | EWS                           | 4703                | 0.66                        | 24                        | ns                        | 22                | [4]       |
| Glioblastoma                     | EWS                           | 2245                | 0.91                        | 9                         | ns                        | 27                | [5]       |
| Head_neck SCC (2 datasets)       | EWS                           | 9974                | 0.80                        | 15                        | 4                         | 91                | [6,7]     |
| Liver_ncc                        | GWS                           | 142,791             | 0.50                        | 23                        | ns                        | 11                | ICGC^3    |
| Liver_riken                      | GWS                           | 175,094             | 0.45                        | 20                        | 5                         | 16                | ICGC^4    |
| Lung_nsc                         | GWS                           | 50,666              | 0.78                        | 10                        | 9                         | 1                 | [8]       |
| Lung_sc                          | GWS                           | 22,905              | 0.65                        | 22                        | 2                         | 1                 | [9]       |
| Melanoma_ews                     | EWS                           | 2809                | 0.88                        | 2                         | 4                         | 14                | [10]      |
| Melanoma_gws                     | GWS                           | 33,339              | 0.83                        | ns                        | 6                         | 1                 | [11]      |
| Multiple myeloma                 | EWS                           | 1204                | 0.68                        | 16                        | ns                        | 38                | [12]      |
| Ovarian carcinoma                | EWS                           | 18,121              | 0.73                        | 22                        | ns                        | 316               | [13]      |
| Pancreatic_au                    | EWS                           | 1864                | 0.70                        | 14                        | ns                        | 87                | ICGC^5    |
| Pancreatic_ca                    | GWS                           | 28,041              | 0.53                        | 24                        | ns                        | 5                 | ICGC^6    |
| Prostate                         | GWS                           | 28,189              | 0.56                        | 23                        | ns                        | 7                 | [14]      |
| Mixed                            | EWS                           | 2472                | 0.78                        | 13                        | 3                         | 393               | [15]      |
| Inherited missense/nonsense mutations | na                           | 67,513              | 0.68                        | 26                        | ns                        | BIOBASE^7        |
| Inherited splicing mutations     | na                            | 9907                | 0.65                        | 2                          | 2                         | BIOBASE^7        |
| 1000 Genomes Project             | GWS                           | 16,744,224          | 0.56                        | 38                        | 3^9                       | 1000 GP^8        |

Mutation datasets were derived from sixteen cancer genome studies published between 2007 and 2011, five cancer genome studies reported by the ICGC, two datasets of germline pathogenic mutations from HGMD and the “rs” entries (i.e. entries in the dbSNP membership, build 129) from the 1000 Genomes Project. \(-\log P\) values were obtained from two-tailed t-tests (z-tests for the colorectal cancer datasets) using the NGNN counts from AgilentV2 (EWS) and Duke35 (GWS).

SBS, single base-pair substitutions;
^1http://dcc.icgc.org/ (JHU, US);
^2http://dcc.icgc.org/ (RIKEN, JP);
^3http://dcc.icgc.org/ (NCC, JP);
^4http://dcc.icgc.org/ (QCMG, AU);
^5http://dcc.icgc.org/ (OICR, CA);
^6http://www.biobase-international.com/product/hgmd (HGMD Professional Human Gene Mutation Database, release 2011.3);
^7http://www.1000genomes.org/ (2011 release).

doi:10.1371/journal.pgen.1003816.t001
Figure 1. SBSSs and VIPs. Panel A, whisker plot of the fractions of SBSSs at G+C bp for the EWS and GWS datasets computed using AgilentV2 and Duke35 mappability counts, respectively; red line, mean; black line, median; green lines, average GC-contents in the mappable AgilentV2 (EWS) and Duke35 (GWS) sets. Panel B, NGRA sequences are enriched in SBSSs in melanoma. y-axis, for each 4-member sequence combination with matching P1–P3 bases, the fraction of mutations at P4-A was divided by the average fraction of mutations at P4-(C/T/G); x-axis, P3 base composition; R, pyrimidine; Y, purine; P-value from two-tailed t-test. Panel C, the in normal fractions of mutated DGN (D = A/G/T) sequences, F, for the seven cancer datasets with \( \log P \approx 3 \) for DGRN>DGYN (Table 1) were combined and plotted as a function of the average absolute free energy of base stacking, \( \Delta G(v) \), for each of the 4G DGN sequences. Panel D, three-dimensional model of the \( (5'-GGG-3')-(5'-CCC-3') \) trinucleotide showing the LUMO (low unoccupied beta molecular orbital) of the ionized sequence. Panel E, plot of the normalized fractions (log fi) of mutated DGN sequences (Duke35 counts) for the Lung, rsc cancer dataset vs. VIPs; outer circle, 5’D base; inner circle, 3’N base; blue, adenine; green, guanine; red, thymine; yellow, cytosine. Panel F, agglomerative hierarchical clustering of 14 cancer genome datasets obtained from linear correlations with ln VIP values, as obtained from T. hg19 counts; colored boxes, elements found to be clustered at the 90% confidence interval.
doi:10.1371/journal.pgen.1003816.g001

displaying the largest differences) and germline mutation datasets (Table S4, Panels B-D), with \( -\log P \) values ranging from 8 (Breast; 9.3±3.3x10^{-3} for CGNN vs. 4.0±2.6x10^{-3} for DGNN, AgilentV2 counts) to 41 (1000GP; 6.9±1.3x10^{-3} for CGNN vs. 7.5±1.5x10^{-3} for DGNN, CRG50 counts) (Table 1). Exceptions were the two Melanoma datasets, for which such differences were modest due to considerable variability in the data (2.2±3.4x10^{-4} for CGNN vs. 0.9±1.2x10^{-4} for DGNN, P-value 0.02 for Melanoma_Ews, AgilentV2 counts; 4.9±7.5x10^{-3} for CGNN vs. 2.4±2.8x10^{-3} for DGNN, P-value 0.06 for Melanoma_gws, Duke35 counts). Thus, with the notable exception of the melanomas, the CpG dinucleotide, a substrate for cytosine methylation, represents a strong mutation hotspot, both in the soma and the germline.

For the DGN sequences, a P3-purine significantly increased the proportions of P2 SBSSs, \( \beta(DGRN) \), compared to a pyrimidine, \( \beta(DGYN) \) in 7 cancer types \( -\log P \) values 3-9, including lung, head and neck, and melanoma (Table 1), for which associations with exposure to either cigarette smoke or sunlight have been documented. An additional dataset, Lung_sc, from the established cell-line NCI-H209 displayed modest \( \beta(DGRN) > \beta(DGYN) \) (P-value ~0.04). Again, in no cases were P-values contradictory based upon the mapping method used (Table S4, Panels B-D). The data for the melanomas were particularly striking since P3-A increased the fractions of mutation at P2-G by ~10-fold relative to DGBN (B = C/G/T) (14.3 vs. 6.9x10^{-5} for Melanoma_gws, P-value 9.1x10^{-5}; and 68.4 vs. 6.4x10^{-5} for Melanoma_ews, P-value 2.4x10^{-3}; according to Duke35 and AgilentV2, respectively, Table S3). In addition, the CGAN motifs displayed ~3-fold higher mutation fractions than the DGAN motifs (14.3 vs. 4.9x10^{-5}; P-value 0.019; and 68.4 vs. 20.5x10^{-5}; P-value 2.8x10^{-3} in Melanoma_gws and Melanoma_ews, respectively, although a mutagenic role for CpG methylation was not apparent (i.e. the CGBN and DGBN fractions were indistinguishable, P-values ~0.6-0.7). In addition, for the two melanoma datasets, P4-A significantly increased (2.2±0.2 fold) mutation at P2-G (Fig. 1B) when P2 and P4 were separated by a purine (P-value 5.6x10^{-7}). Additional analyses in four melanoma datasets [17,21,34,40] confirmed this finding (ratio for the eight possible NGRA/NGRB groups in these four datasets was 2.2±0.5 and 2.2±0.3 for the combined six datasets, Table S6). The increase in mutation at P2-G, relative to P4-C and P4-T, was also observed for P4-G;
however, this effect was less consistent than P4-A and was observed more frequently when P3 was occupied by a guanine (18/23 cases) rather than an adenine (5/23 cases), Table S6.

In summary, SBSs at P2-G were dependent upon the sequence composition of the 3’-nearest neighbor in a number of different cancer types; in melanoma, this effect extended to the next 3’ base when bridged by a purine. Thus, GpR and GpRpA sequences constitute mutational hotspots that render the 5’ G sensitive to mutation. Further analyses performed in individual cancer samples (Text S1, Figure S1 and Table S4, Panels B–E) indicated that biological mechanisms, rather than differences in variant-calling algorithms or variability between individual samples, were the likely causes of such mutational patterns.

Electron Transfer and Sequence-Dependent SBSs

Guanine is the most readily oxidized base [41] and its ionization energy, i.e. the energy required to abstract an electron, depends upon the identity of the flanking nucleotides [42–45]. Substantial work performed with model DNA sequences in vitro has shown that, following one-electron oxidation reactions, the sites of electron loss (hole) migrate efficiently (rate constants ~10^7 s^-1) from the original locations to distant sites, where they become trapped in troughs of low ionization energy, most often at GG and GGG sequences [36,44,46–48]. Because oxidative DNA damage occurs spontaneously in the cell, we tested whether the sequence-dependent SBS patterns were consistent with a mutagenesis model that included: a) loss of an electron within the NGNN sequences; b) hole migration to the P2-guanine; and c) chemical modification of the P2-guanines leading to base substitutions [22].

Absolute Free Energy of Base Stacking

The binding energy of single-stranded stacked bases is presumed to be dependent upon the affinity of interactions, or the extent of electron sharing, between π orbitals across bases [49]. The free energy of base stacking, rather than hydrogen bonding, has been reported to be the major source of stability in duplex DNA [50]. Hence, we expected that strongly interacting bases would be more prone to one-electron oxidation and, hence, to higher SBS rates than weakly interacting bases. To this end, we used the absolute free-energy values of base stacking between non-bonded bases, ΔG(0), derived from a theoretical study [51] using a continuum solvation model and Amber force field to assess the relationships with ΔG(DGN) values, as we previously employed [52]. For 5/7 datasets with ΔG(DGRN)/ΔG(DYN) (Table 1), i.e. two melanomas, Lung_nsc, Liver_riken and Mixed, a significant positive correlation existed between the fraction of mutated DGN sequences and free energies of base stacking (Table S7, Panel A; r² 0.10–0.71; P-values <0.001–0.031). The normalized mutation fractions for the combined 7 datasets also displayed significant correlation (r² 0.47; P<0.001; P(0.05) = 1.000; Fig. 1C and Table S7, Panel A; ΔG(DGN) were according to Duke35 and AgilentV2 mappability).

Vertical Ionization Potentials (VIPs)

VIP, the minimum energy required to abstract an electron, is commonly used as a measure of one-electron oxidation reactivity [41]. We modeled the susceptibilities of G-centered DGN double-stranded trinmers to oxidation via quantum chemical computations of VIPs. These analyses are expected to compare favorably with data obtained from the computationally more demanding tetramers; the VIPs for tetramers would be expected to be lower than for trinmers while maintaining similar sequence-dependent rankings for P2 [53–56]. The VIP values for the 12 trinmers were ~30–40% lower than that of an isolated guanine (Table 2) whose VIP estimate was close to the experimentally determined lowest band maximum [57]. The trimer with the lowest VIP was GGG, in agreement with prior calculations [42–45] and all trimers containing a GG doublet had lower VIPs than those with a single G. In addition, a purine at the 3’ position was consistently associated with lower VIPs than a pyrimidine at the 3’ position. Thus, DNA sequence context affects VIPs, in accordance with guanine reactivities to oxidative reactions in vitro [42,44].

| Sequence | Vertical Ionization Potential (eV) |
|----------|----------------------------------|
| GGG      | 5.39                             |
| GGA      | 5.50                             |
| GGT      | 5.54                             |
| GCC      | 5.63                             |
| AGG      | 5.57                             |
| TGG      | 5.59                             |
| TGA      | 5.64                             |
| AGA      | 5.66                             |
| AGT      | 5.81                             |
| TGT      | 5.90                             |
| AGC      | 5.95                             |
| TGC      | 5.97                             |
| G        | 8.02±                             |
| G (exp)  | 8.26±                             |

VIPs for the centrally (italicized) guanine computed at the M06-2X/6-311G(d) level of theory;
*VIP of free unalkylated guanine;
from reference [57].

Inspection of the lowest unoccupied beta molecular orbital (LUMO) for each DNA trimer cation, [DGN]+, in which the ionization state was modeled by removing an electron, showed that the electron hole invariably had π character with high densities at the central guanine (Fig. 1D), or at the 5’G in the GGH (H = A, C, T) sequences, consistent with previous work [42,55,56,58,59], implying that P2G was a frequent site for one-electron oxidation reactions. Analyses between ΔG(DGN) and VIP values displayed significant correlations for 4/5 datasets that also revealed a correlation with ΔG(0), i.e. melanomas, Lung_nsc and Liver_riken (as per Duke35 and AgilentV2 mappability; Figure 1E, Figure S2 and Table S7, Panel B; r² 0.54–0.75; P-values<0.001–0.007). Notably, robust correlation was also evident when the ΔG(DGN) data from all 18 cancer datasets were normalized and then computed as average values (Table S7, Panel B and Figure S2, Panel D, r² 0.40; P-value 0.026; P(0.05) = 0.615). The regression coefficients obtained using the T_hg19 and T_exons mappability data for the datasets with >2,000 SBSs were also used to perform hierarchical clustering based on absolute Manhattan distances (Fig. 1F). At >90% confidence interval, this yielded three clusters, the largest of which contained the same cancer datasets, with the exception of Ovarian carcinoma, that also displayed ΔG(DGRN)>ΔG(DYN) ratios (Table 1). In summary, both base stacking and VIP data support the conclusion that electron transfer in DNA represents
a significant mechanism for sequence context-dependent mutagenesis in cancer.

Sequence-Dependent SBSs in Cancer-Associated Genes

Driver mutations include non-synonymous (NS) substitutions that play a key role in cancer initiation and progression. To assess whether bona fide driver mutations also occurred in a sequence context-dependent manner, we examined the NS substitutions that altered the same genomic coordinate in more than one patient sample, and the genes affected (Text S1, Figure S3 and Table S8). For the 224 recurrent NS substitutions at G>C bps, we calculated the relative enrichment E for each of the 64 NGNN motifs, a value which is expected to approximate to 1 if the base substitutions are completely independent of flanking sequence. E values were greater for the CGNN than for the DGNN (D=A/G/T) sequences (Figure 2, Panel A). Among the DGNN sequences, P3-purines were associated with significantly more mutations than P3-pyrimidines, a difference that was attributable to the presence of P3-A [DGAN>DGBN, B=C/G/T]. This trend remained unaffected after the 45 entries from the two melanoma datasets (which showed [DGAN]>[DGBP] in the respective EWS and GWS screens (Figure 1, Panel B)) were removed (EDGAN) = 0.92±0.52; EDGBN) = 0.39±0.43; P = 0.001). Thus, recurrent NS substitutions occurred preferentially at GpG and GpA dinucleotides in cancer genomes, mirroring the sequence context-dependent pattern of SBSs observed between genome-wide and exome-wide (Figure 1 and Table 1). Of the 18 codon changes that occurred >12 times, 7/10 affected NGNN sequences and 5/10 occurred at CGNN sequences, all in well-established cancer genes (Table S9, Panel A). Likewise, the most commonly mutated CGNN (Table S9, Panel B) and DGAN (Table S9, Panel C) motifs affected known driver mutations, alongside several novel candidate genes and driver mutations (Table S9), including p53R248G, which has been reported to alter protein function (http://www.p53.iarc.fr), and GRHL3, WNK3, EPHB1, ADCY2, GSK3B and LRRN3, which are not currently listed in the cancer gene census (http://www.sanger.ac.uk/genetics/CGP/Census).

Tissue Distribution and Networks Affected

To examine whether recurrent NS substitutions occurred equally in all tumor tissue types, we determined the relative distributions of the most frequently mutated genomic coordinates after normalizing for both tissue representation and the total number of SBSs per dataset; in the absence of any bias, each tissue would contribute 12.5%. The four genes with ≥12 recurrently mutated genic coordinates (TP53, KRAS, PIK3CA and Braf) (Table S9, Panel A) were predominantly of breast (34%), intestine (24%) and lung (18%) origin (Figure 2, Panel B). The three most commonly mutated CGNN sequences (CGTG, CGGA and CGTG; S=9.5, 6.0 and 5.7, respectively) were found in genes mutationally altered in the intestine (26%), ovary (19%) and breast (15%), whereas the most commonly mutated DGAN sequences (TGAT, GGAA and TGAA; S=2.8, 2.0 and 2.0, respectively) were found predominantly in genes altered in melanoma (37%), breast and lung (17% each) (Table S9). By contrast, these mutated motifs were underrepresented in the liver (≤2%). Of the 64 codons affected, 6 (in TP53, 2 in PIK3CA and 1 in GNAS) are known driver mutations, 4 introduced stop codons into TP53, and 26 occurred within genes whose involvement in cancer is strongly suspected (Table S9 and http://www.sanger.ac.uk/genetics/CGP/Census). Thus, although high-confidence driver mutations occurred preferentially at CGNN and DGAN motifs, their occurrence between tissues was highly asymmetrical, with DGAN mutations occurring predominantly in tumors of the skin.

Finally, we used pathway analyses to survey the 150 recurrently mutated genes (Figure 2, Panel C). In all tumor tissues, 18 pathways/networks related to cell-cycle checkpoints and the DNA damage response were found to be compromised in all types of tumor, the sole exception being melanomas in which only the MAP kinase signaling pathway was consistently altered. A similar pattern was revealed when all NS substitutions were analyzed, irrespective of whether the data from all patients were merged (Figure S4, Panel A) or plotted separately (Figure S4, Panels B and C). The highest-ranking pathways were dominated by TP53 mutations in most tumor types (Figure S4, Panels B and C), with the exception of pancreatic cancers in which KRAS mutations dominated. Both p53 and KRAS proteins are known to act on parallel signaling cascades that regulate TERT, the active reverse transcriptase component of telomerase that controls the stability of chromosome ends (http://www.biocarta.com/pathfiles/h_telPathway.asp). Hence, although critical pathways represent common targets for oncogenic transformation, the altered genes may vary between different patients or organ/tissue types. In summary, a distinction emerged between melanoma and the other types of cancer, both with regard to the sequence contexts targeted by driver mutations, DGAN vs. CGNN sequences, and to the pathways that hosted these mutations, MAP kinase vs. p53-associated signaling pathways.

Guanine Is Preferentially Targeted by Pathogenic Germline Mutations

In the HGMD missense/nonsense mutation dataset, approximately 68% of SBSs occurred at G>C bps, a proportion similar to the EWS cancer datasets, although correlations with ΔG(v) or VIPs were absent (Table S7, Panel B) and no enrichment for DGRN sequences was apparent (Table 1). However, x(DGNN), which measured the fraction of mutated DGRN motifs relative to the direction of transcription, revealed that the P2 position was more likely to contain a guanine on the non-transcribed strand, relative to the transcribed strand, when stacking interactions with neighboring bases were high (r2 = 0.32; P-value<0.001; P(99.75) = 0.991; Figure 3, Panel A). No such behavior was evident in the cancer datasets (not shown), whereas limited bias was observed in the 1000 Genomes Project dataset (Figure 3, Panel A). Thus, transcription led to a pattern of sequence context-dependent SBSs among pathogenic germline mutations, which mirrored that observed in several cancer genomes. The HGMD dataset of inherited splicing mutations contained 9,907 SBSs that may be assumed to adversely affect RNA processing; 8,308 (84%) of these mapped to within 5 bases of donor and acceptor splice junctions (Figure 3, Panel B). Strikingly, although the canonical GT and AG intronic splice junctions at the donor AG-GTAA GT and acceptor CAG-GT sequences were frequent (3,460 and 1,050, respectively), no enrichment for DGRN sequences was apparent (Table 1). However, x(DGNN), which measured the fraction of mutated DGRN motifs relative to the direction of transcription, revealed that the P2 position was more likely to contain a guanine on the non-transcribed strand, when stacking interactions with neighboring bases were high (r2 = 0.32; P-value<0.001; P(99.75) = 0.991; Figure 3, Panel A). No such behavior was evident in the cancer datasets (not shown), whereas limited bias was observed in the 1000 Genomes Project dataset (Figure 3, Panel A). Thus, transcription led to a pattern of sequence context-dependent SBSs among pathogenic germline mutations, which mirrored that observed in several cancer genomes.
Discussion

Large-scale next-generation sequencing projects of cancer genomes are providing an unprecedented opportunity to address the key issue of the nature of the underlying mechanisms of base substitution in tumorigenesis. This issue is generally approached by analyzing the types of base substitution specific to the cancer tissue, i.e., mutation spectra, based on the assumption that different types of base substitution originate via different mutational mechanisms, as assessed by animal model systems [4,6,8,9,16–18,21,60–63]. We have chosen the alternative approach of addressing the sequence-context dependency of single base substitution, with the expectation of shedding light on the earliest step(s) in the process of mutagenesis, i.e., the susceptibility of DNA to base modification. Once modified, a base would then undergo various types of substitution based upon the type of modification it incurred, its interactions with DNA repair and replication systems and, possibly, the tissue of origin [62].

As revealed by correlations with VIPs and absolute free energies of base stacking, we uncovered a direct correlation between electronic coupling along the DNA chain, leading to electron transfer, and sequence-dependent SBSs, both in human cancers and as a cause of inherited disease. Thus, charge transfer appears to be the earliest event in the mutational mechanism acting along...
Figure 3. Germline mutations are affected by transcription. Panel A, HGMD dataset; y-axis, as in Figure 1C: x-axis, ratio of mutated NGNN sequences in protein coding genes containing the P2-guanine base on the non-transcribed (NT) vs. transcribed (T) strand; solid circles, HGMD dataset \( r^2 = 0.32, P(x) < 0.001, P(y) = 0.991 \); open circles, 1000 Genomes Project dataset. Panel B, inherited splicing mutations dataset; top, cartoon of exon-intron boundaries showing the conserved GT and AG bases at the donor (ds) and acceptor (as) splice sites; bottom, graph of splicing mutations; y-axis, number of SBSs; x-axis, position of SBSs relative to \(+/-20\) nt of splice junctions; Panel C, model for sequence-dependent SBSs in cancer and human inherited disease. In the first step, an electron is lost from within a tetranucleotide sequence, leaving a hole. In the second step, the hole migrates reversibly to various competing sites, including flanking or more distant bases, as well as other molecules and contacting chromatin-associated amino acids (not shown), eventually being trapped by a guanine base. The resulting guanine radical cation either causes DNA-protein crosslinking or undergoes subsequent chemical modifications. If the modified base is not corrected by DNA repair, it may give rise to a mutation (X-Y base-pair) as a result of error-prone DNA polymerase synthesis during DNA replication (dashed arrow).

Figure 3 legend:

Panel A: HGMD dataset; y-axis, as in Figure 1C: x-axis, ratio of mutated NGNN sequences in protein coding genes containing the P2-guanine base on the non-transcribed (NT) vs. transcribed (T) strand; solid circles, HGMD dataset \( r^2 = 0.32, P(x) < 0.001, P(y) = 0.991 \); open circles, 1000 Genomes Project dataset. Panel B: inherited splicing mutations dataset; top, cartoon of exon-intron boundaries showing the conserved GT and AG bases at the donor (ds) and acceptor (as) splice sites; bottom, graph of splicing mutations; y-axis, number of SBSs; x-axis, position of SBSs relative to \(+/-20\) nt of splice junctions; Panel C: model for sequence-dependent SBSs in cancer and human inherited disease. In the first step, an electron is lost from within a tetranucleotide sequence, leaving a hole. In the second step, the hole migrates reversibly to various competing sites, including flanking or more distant bases, as well as other molecules and contacting chromatin-associated amino acids (not shown), eventually being trapped by a guanine base. The resulting guanine radical cation either causes DNA-protein crosslinking or undergoes subsequent chemical modifications. If the modified base is not corrected by DNA repair, it may give rise to a mutation (X-Y base-pair) as a result of error-prone DNA polymerase synthesis during DNA replication (dashed arrow).

do:i:10.1371/journal.pgen.1003816.g003
The CpG dinucleotide was found to be a consistent mutational hotspot, both in cancer and the germline, a result that generalizes the conclusions drawn from previous studies [4,9,16–18,63]. The high frequency of C→T transitions at CpG dinucleotides is generally attributed to high rates of deamination of 5-methylcytosine resulting from methylation of CpG sites [9,19,20,62]. However, other mechanisms have been proposed [60,62], such as enhanced susceptibility of methylated CpG sites to damage by physical and chemical genotoxic agents [62]. This latter interpretation would be consistent with our finding that electronic coupling is an important factor in establishing the hierarchy for base modification in DNA. During the course of normalizing mutation fractions by genome mappability, we noted an enrichment of CGNN sequences in Segmental Duplications. Nakken et al. [70] reported a higher density of CpG islands in Segmental Duplications than in unique chromosomal regions, whereas Xie et al. found methylation-associated SNP clusters to be more prevalent in Segmental Duplications than in unique regions [71]. Thus, the prevalence of CGNN-associated SBSs may well be greater than our study indicates.

A confounding factor in our analyses is the relatively small number of SBSs, particularly in EWS datasets, which caused large variations in \( f_i \) values. Indeed, three of the four datasets that displayed high-confidence (i.e., \( P < 0.05 \) and \( P(\chi^2_{0,05}) > 0.800 \)) correlations between SBSs and AG(\( v \)) VIPs or VIPs were obtained from genome-wide studies. Combining all \( f_i \) values into a single group (Figure S2, Panel D) only alleviates the problem, since the \( f_i \) values for each dataset are given the same weight. Nevertheless, the ensuing “cautiously significant correlation” is consistent with a role for electronic coupling in cancer-related mutagenesis. A second confounding factor is the multiple roles that the GpA dinucleotide plays in mutagenesis, as elucidated to earlier. In the case of melanomas, if the numbers of mutated NGAN (and NGA) sequences were dominant, this might cause chance correlation with AG(\( v \)) or VIPs, when in fact most mutations could arise from CPDs on the complementary strand. Correlations for both the Melanoma_gws and Melanoma_ews datasets remained highly significant (\( P < 0.002; P(\chi^2_{0.05}) 0.920–1.000 \)) when the \( f_i \) values for the NGAN (or NGA) sequences were excluded from the analyses, thereby confirming a role for charge transfer. This conclusion is further supported by the observation that electronic coupling and photo-induced energy transfer reactions at pyrimidine dimers occur simultaneously and impinge on one another [72–74].

In cancer, the subset of mutational changes resulting from NS substitutions that occurred in different patient samples displayed the same enrichment of mutations at CpG and GpA sequences as the exome-wide and genome-wide sequence alterations, supporting the notion of common underlying causes, i.e., cytosine methylation, electron transfer (this study), enzymatic cytosine deamination and CPD formation (in melanoma) [10,11,17]. These commonalities suggest that the mechanisms involved in generating “driver” tumor initiating mutations are likely to be similar to those involved in generating the bulk of subsequent “passenger” mutations. Hodis et al. [75] reached a similar conclusion using a quite different approach. Thus, electron transfer appears to be involved in both the early (driver mutations) and late (passenger mutations) phases of tumorigenesis, particularly in tissues of epithelial origin. Recurrent NS substitutions were observed predominantly in gene networks associated with p53 function in all tumor types, the exception being melanoma where a preponderance of mutations at GpA segregated with genes of the MAP kinase signaling pathway. The reason for this distinction remains unclear; however, the critical role played by the MAP kinase signaling pathway in melanocyte proliferation in response to UV damage [76] suggests that positive selection may have been a contributory factor.

The results of the HGMD data analysis support the occurrence of electronic transfer in germline mutagenesis associated with human inherited disease, although sequence context-dependent mutagenesis was evident only when mutations were mapped onto the non-transcribed strands of genes. Guanines modified by oxidative DNA damage are repaired predominantly by base excision repair (BER) [77,78]. Since oxidative DNA damage occurs more efficiently in single-stranded DNA than in double-stranded DNA [79,90], oxidative guanine lesions may have formed more frequently on the single-stranded, non-transcribed, strand than on the DNA-RNA duplex during transcription. Thus, a greater number of lesions would be expected to escape BER on the non-transcribed strand than on the transcribed strand. In cancer cells, the large number of mutations that generally accumulate during tumor growth could have masked this bias. An alternative or additional possibility is that transcription-coupled nucleotide excision repair, a mechanism that processes bulky DNA adducts and which selectively corrects errors on the transcribed DNA strand [81], might have contributed to the strand asymmetric mutations observed in the HGMD dataset [9,11,12,82,83]. In similar vein, we interpret the selectivity of mutations at purines on the non-transcribed strands of splice junctions as a consequence of oxidative damage, whose effect could have been prolonged by the pausing of transcription-coupled splicing at splice junctions [84]. With the number of sequenced genomes rapidly increasing, it will be of great interest to ascertain whether electron transfer constitutes a general mutational mechanism that is common to all forms of life.

**Materials and Methods**

**Datasets**

We collected the publicly available data from cancer genome studies reported in PubMed from 2007 through December 2011 [1–15] together with the 5 largest datasets available from the International Cancer Genome Consortium (ICGC). The cancer genome datasets varied widely in terms of sequencing strategies, mapping techniques and variant-calling algorithms, implying that the power to detect SBSs may differ depending upon the datasets and methodologies used [85]. However, all studies excluded base variants present in matched-control tissues, such that the reported SBSs were changes attributed to somatic mutations in the tumor tissue. Matched controls were used for all patient samples. On average, between 6 [15] and 1834 [1] tumor-specific SBSs were reported in the EWS studies (between 1012 [3] and >50,000 [8] in the GWS studies) (Table 1), which is ~1–3 orders of magnitude lower than the numbers of non-synonymous and splice-site variants noted on average in whole-exome studies [86]. In addition to normal-tumor matched samples, single nucleotide polymorphisms present in dbSNP databases or in the Venter and Watson genomes [1,10,11] were also used to exclude common base variants. Differences in variant-calling power were mitigated in our study since we examined relative proportions of mutated sequences, rather than absolute mutation fractions.

A second source of variation in detecting SBSs among the cancer genome studies was the sequencing instrument used. Illumina sequencers have been reported to yield systematic base-call errors, especially at the last base of context-specific GGC and GGT sequences, which affect either the forward or reverse strand, and at inverted repeats [87,88]. The sequencing technologies employed included Illumina genome analyzers, SOLiD next-generation DNA sequencing, ion semiconductor sequencing,
Mappable Mutations

Approximately half of the human genome sequence comprises highly homologous repetitive DNA elements (4A repeats, LINE elements etc.) and simple repeats, and an additional ~3.6% contains Segmental Duplications, i.e. segments of >1 kb in length that are present at multiple loci and which share ~90–98% sequence similarity (http://genome.ucsc.edu). Thus, because only the mappable genome may be scored for mutations, we used various methods to estimate the total number of mappable NGNN sequences to use as denominators in the \( f_i \) fractions (see below). Three methods were used for the GWS studies: 1) the entries with a mappability index of 1 (representing unique sequences) from file wgEncodeDukeMapabilityUniqueness55kb.bigWig (http://hgdownload.soe.ucsc.edu.goldenPath/hg19/database/) generated for the ENCODE project by the Duke University Institute for Genome Sciences and Policy (IGSP) and at the European Bioinformatics Institute (EBI), which we refer to as Duke35; 2) we selected sequences from the mappability file wgEncodeCrgMapabilityAlign30mer.bed (http://hgdownload.soe.ucsc.edu.goldenPath/hg19/database/) [89] (Donohue et al., unpublished data), referred to as CRG50; and 3) we retrieved all NGNN sequences in the GRCh37/hg19 release of the human genome assembly (chromFa.tar.gz file at http://hgdownload.soe.ucsc.edu.goldenPath/hg19/database/) [90] (Donohue et al., unpublished data), referred to as CRG50.

For the EWS studies, the SureSelect Human All Exon Kit (http://www.genomics.agilent.com) was the most common platform reported [2,4,7,13]. A custom RefSeq CCDS PCR primer library was used to generate the Glioblastoma dataset [5] and a set of 1,507 genes (oncogenes, tumor suppressors, “druggable targets”) were targeted in the Mixed cancer dataset [15]. Hence, three methods were used to estimate the number of mappable NGNN sequences in the EWS studies: i.e. the NGNN counts from 1) the file S0293689_Covered.bed (http://www.Agilent.com), listing the coordinates of exons targeted by the SureSelect Human All Exon Kit (AgilentV2); 2) the RefSeq exons sequences of CRG50 (CRG50.exon) and 3) the total RefSeq exons from file seq_gene.md at ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/mapview/seq_gene.md.gz (T_exons).

Definitions

We defined \( f_j = m_j / t_j \), where \( m_j \) was the number of mutations at a specific NGNN\(N\)NCN sequence (henceforth designated as NGNN) and \( t_j \) was the total number of that sequence in one of the six “mappable” sets described above. The total number of NGNN sequences was doubled for the self-complementary AGCT, CCGG, GGGC and TGCA sequences since, like all NGCN sequences, they contain two mutations sites, one on the forward and one on the reverse strand. In relation to the counts of mutated NGNN motifs, if the G... occurred at the same genomic coordinate more than once within a cancer dataset, or if it was a homozygous mutation, it was considered as one count. Custom shell and FORTRAN scripts were used to obtain the total numbers of mappable NGNN and \( f_j \) fractions (see Text S1 for sample scripts). The normalized fractions of mutated DGN sequences were defined as \( F_j = \sum F_j \), thus, \( \sum F_j \) scaled to 1. \( N \) indicates any base (A/C/G/T); \( G \) indicates A/G/T; \( C \) indicates C/G/T. As mentioned, sequence designation implies double-stranded DNA (i.e. AGTC = (5’-AGTC-3’) and 5’-GACTC-3’). The average base stacking free energies \( \Delta G(\psi) \) were obtained from Friedman and Houig [51] by using the \( \Delta G(\psi) \) (\( \psi = 2 \)) values for the three base steps (DpG + GpN + NpN)/3. The free energy of base stacking \( \Delta G(\psi) \) is an estimate of the absolute contribution of base stacking to nucleic acid stability in the absence of hydrogen bonding interactions, and contains a contribution from nonpolar plus electrostatic forces, as assessed from a theoretical approach using the Amber force field and a continuum solvation model of water. The largest contribution to \( \Delta G(\psi) \) was found to arise from nonpolar [51], as opposed to electrostatic, interactions. Nonpolar interactions were contributed for the most part by enhancement in the Lennard-Jones component as a result of close packing, and to a smaller extent from hydrophobic interactions. Thus, the \( \Delta G(\psi) \) values follow the same trend as the nonpolar contributions to free energies of base stacking \( \Delta G_{\text{nonp}}(\psi) \), i.e. purine-purine >> purine-pyrimidine > pyrimidine-purine > pyrimidine-pyrimidine, in qualitative agreement with experimental determinations [51]. The relative enrichment \( E \) of sequence \( i \), \( E_i \) was defined as the ratio \( D_i/T_i \), where \( D_i = d_i / \sum d_i \), the being the number of times sequence \( i \) was mutated at least twice at the same hg19 coordinate and \( T_i = t_i / \sum t_i \), the being the total number of occurrences of sequence \( i \) exome-wide (\( T_{\text{exons}} \)). Finally, \( S = \sum s_i \), \( t_i = t_i / \sum t_i \) being the number of times the combined (top) sequences were recurrently mutated and \( c_i \) being the total number of NS substitutions in a particular type of cancer.

Molecular Modeling

Three-dimensional structures of the 12 possible double-stranded DGN trinucleotides were constructed using w3DNA [90]. Hydrogen atoms, atomic charges and four neutralizing Na\(^+\) counterions were assigned to each sequence according to the amber99 force field [91], using UCSF CHIMERA [92]. Na\(^+\) counterions were positioned next to the four DNA backbone phosphates. Each trinucleotide was energy minimized in vacuo using the 10,000 steps steepest descent algorithm and the amber99 force field in GROMACS 4.5.1 [93]. Ten and 14 Å cutoffs were used for Coulomb and van der Waals interactions, respectively.

Vertical Ionization Potentials (Vips)

VIPS were computed using Kohn-Sham density functional theory (DFT) [94] employing the Minnesota M06-2X functional [95,96] with all-electron 6-31G(d) basis sets [97,98], as implemented in the GAMESS electronic structure package [99], and including backbone phosphate groups and sodium counter ions in addition to the DGN double-stranded bases. The M06-2X functional was used since this method provides accurate descriptions of hydrogen bonding and stacking interactions between basepairs. We reasoned that the DGN set would provide the same type of information as the computationally more demanding NDGN set. Molecular orbitals were depicted using the MacMolPlt graphics program [100].

Pathway Analysis

For individual patient samples, mutations were collated and sorted into lists of genes carrying mutations using customized R scripts (http://www.r-project.org/). The gene lists for each sample were entered into our pattern extraction pipeline analysis (PPEP).

Mechanisms of Mutation and Human Genetic Disease
Hierarchical Clustering

Agglomerative hierarchical clustering dendrograms [103] were built using either the regression coefficients, $r$, between the fractions of mutated DGN sequences, $f(DGN)$, and the VIP values, or the absolute orthogonal distances (Manhattan distances) between each $f(NGNN)$ data point for all datasets. All-to-all comparisons were performed, allowing the relative estimation of all components of the systems, including the reference VIP branch.

Supporting Information

Figure S1 Individual samples from the same cancer dataset and cancer types share mutational patterns. Panel A, analysis of variance for the normalized fractions of CGNN vs. DGN and DGRN vs. DGYN for the 14 Melanoma_ews samples computed from the AgilentV2 mappability counts; full circles, Holm-Sidak test on the difference of means (normality test by Shapiro-Wilk, P = 0.657); triangles, Kruskal-Wallis one-way ANOVA on ranks (H = 18.66). Panel B, hierarchical clustering of the 20 largest datasets (Table S1) plus 5 individual samples comprising the Pancreatic_ca dataset, computed from the sum of the absolute difference (Manhattan distance) in the scaled frequencies of Pancreatic_ca dataset, computed from the sum of the absolute differences (Manhattan distances) between each $f(NGNN)$ data point for all datasets. All-to-all comparisons were performed, allowing the relative estimation of all components of the systems, including the reference VIP branch. (TIF)

Figure S2 Plots of the correlation between $f(DGN)$ and VIPs, $x$-axis, normalized mutations fractions; $y$-axis, VIP values; Panel A, $f_i$ for Melanoma_ews computed using the AgilentV2 mappability counts; Panels B and C, $f_i$ for the Melanoma_gws and Liver_riken, respectively, computed using the Duke35 mappability counts; Panel D, averaged $f_i$ values for 18 cancer datasets (Table S6). See Legend to Fig. 1 for details. (TIF)

Figure S3 Genes with NS substitutions and mutation spectrum. Panel A, Plot of the number of NS substitutions per 100 amino acids (H) for the 29 genes in the combined cancer datasets with ≥24 NS substitutions. Panel B, as in Panel A, for genes with ≥4 recurrent NS substitutions; dotted line, median protein size in amino acid residues, as derived from http://www.genecards.org/. Panel C, percentage of the six possible types of SBS for the 35,480 NS substitutions in the combined cancer genomes. (TIF)

Figure S4 Pathway-level heatmaps. Panel A, pathway-level heatmap for the highest-ranking (from top down) pathways (www.biocarta.com) based on the “fractional representation” of genes hit by NS substitutions across all combined individual patient samples for all tumor types. Rows, pathways; columns: tumor types; Fractional representation, color-gradient of the fraction of patient samples in each tumor type that have at least one gene-hit in the corresponding pathway. Panel B, patient sample-specific pathway-level heatmap. Gene-level heatmap displaying the individual patient samples (rows) carrying NS substitutions in genes belonging to a common pathway (columns) ranked in order of their “popularity” for the top pathways shown in Panel A. Panel C, Telomeres-Aging term gene-level heatmap. The detailed gene-level heatmap for the top pathway “Telomeres-Aging” shown in Panel B. Columns: individual patient samples arranged by tumor types; rows: genes hit by NS substitutions ranked in order of number of mutations. (TIF)

Figure S5 Nucleotide conservation and variation at splice junctions. Top panel, fractional conservation within +/−25 nt in 295,093 unique donor and acceptor splice junctions from annotated genes in the GRCh37/hg19 human genome assembly. Middle panel, number of single nucleotide variants (SNV) in the “rs” set of the 1000 Genomes Project, mapped as shown in the top panel. Bottom panel, number of SBSs in the TCGARN cancer dataset mapping to the splice junctions shown on the top panel. (TIF)

Table S1 NGNN counts and relative abundance in the whole genome and the mappable genome. Panel A, whole genome; Panel B, exome; raw NGNN counts, number of NGNN sequences in each set; relative NGNN percentage, raw numbers divided by total ×100; grouping, relative percentages for CGNN, DGN, NGNN containing at least 3 guanines (3G), DGRN and DGYN; comparison between sets, values obtained from set A/set B×100−100. (XLSX)

Table S2 Statistical significance of NGNN representation in mappable genome and exome. Two-tailed $t$-tests for the “comparison between sets” data from Table S1. (DOCX)

Table S3 List of chromosomal positions (GRCh37/hg19) affected by SBSs in the cancer datasets. Only non-redundant positions within each dataset are recorded. (XLSX)

Table S4 $f_i$ fractions and associated statistics. Panel A, counts of SBSs occurring at each NGNN sequence in all datasets and ratios of SBSs affecting G-C base-pairs; Panel B, $f_i$ fractions for the GWS datasets according to three methods and associated statistics (two-tailed paired $t$-test); Panel C, $f_i$ fractions for the EWS datasets according to three methods and associated statistics (two-tailed paired $t$-test); Panel D, $f_i$ fractions for the 2 small EWS datasets (Colorectal_1 and Colorectal_2) and the 14 Melanoma_ews samples according to three methods and associated statistics ($z$-test; $P_{0.05}$ statistics); Panel E, chromosomal positions and counts of SBSs occurring at each NGNN sequence in the 14 patient samples from the Melanoma_ews dataset (*T series) and in the 5 patient samples from the Pancreatic_ca dataset (PCSGP series). (XLSX)

Table S5 $f_i$ fractions for NGNN sequences with P3-A or P3-B (C/G/T) in melanoma. Statistical analyses, two-tailed $t$-tests. (DOCX)

Table S6 Normalized mutation fractions at NGNN sequences in melanoma. EWS or GWS $f(NGNN)$ fractions within each dataset were normalized by dividing for $f(CGAA)$, which displayed the
highest value in all datasets. \( f_i \) fractions were computed using the T_hg19 and T_exons counts. Green background, the \( b \) combinations of the NGRA sequence; tan background, \( f_{NGB} \) values that were lower than the corresponding \( f_{NGRA} \) values; turquoise background, \( f_{NGB} \) values that were higher than the corresponding \( f_{NGRA} \) values. \( \frac{\text{Wt}}{\text{[10]}} \); Pleasance, [11]; Berger, [17]; Nikolaev, [21]; Stark, [40]; Knutshammer, [34].

**Table S7** Correlations with free energy of base stacking and VIPs. **Panel A**, correlations of \( \ln(DGNN) \) values with free energies of base stacking; **Panel B**, correlations of \( \ln(DGNN) \) values with VIPs. For the combined datasets, the fractions were normalized, \( f_i \) values were computed using the Duke35 (GWS) and AgilentV2 (EWS) mappability counts.

**Table S8** Fractional representation of genes with NS substitutions across cancer tissue types and their Aggregate Score. List of genes found to contain non-synonymous substitutions and their combined fractional representation, expressed as Aggregate Score, across cancer types.

**Table S9** Recurrent NS substitutions and driver mutations. Hg19 coordinates containing the largest number of recurrent NS substitutions for various types of motifs and their potential as driver mutations; 1 sequences are reported with the mutated base (A or G) at P2; 2 support from PubMed for a driver mutation following searches for “gene_name & mutated_codon” or “gene_name & cancer” and a manual review of the articles found; *** mutated codon that has been reported as a driver mutation; ** [it] mutated codon in p53 that has not been reported to harbor a driver mutation; ** strong support for gene mutation or change in gene expression being involved in tumorigenesis; * weak support for gene mutation or change in gene expression being involved in tumorigenesis; **nk, insufficient information to assess whether a gene mutation or change in gene expression was involved in tumorigenesis.

**Acknowledgments**

We thank Frederick D. Lewis for comments and suggestions and members of the Texas Advanced Computing Center, the European Bioinformatics Institute and the ICGC for their assistance.

**Author Contributions**

Conceived and designed the experiments: AB KMV DNC. Performed the experiments: AB NAT MY JI RZC DED USM GW AJ. Analyzed the data: BTI RMS JRC KMV AB. Contributed reagents/materials/analysis tools: NV EVB. Wrote the paper: AB KMV DNC.

**References**

1. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456: 66–72.
2. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, et al. (2007) The genomic landscapes of human breast and colorectal cancers. Science 316: 1108–1113.
3. Puente XS, Pinyol M, Queraltada V, Condé L, Ordóñez GR, et al. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukemia. Nature 475: 101–105.
4. Wang K, Kan J, Yuan ST, Shi ST, Chu KM, et al. (2011). Exome sequencing identifies frequent mutation of ABHD1 in molecular subtypes of gastric cancer. Nat Genet 43: 1219–1225.
5. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. Science 321: 1097–1101.
6. Stranksy N, Egold AM, Tward AD, Kostic AD, Cibulskis K, et al. (2011). The mutational landscape of head and neck squamous cell carcinoma. Science 333: 1157–1160.
7. Agrawal N, Frederick MJ, Pickering CR, Bettgeowths C, Chang K, et al. (2011). Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in MEF2C. Science 333: 1154–1157.
8. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, et al. (2010) The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 465: 473–477.
9. Pleasance ED, Stephens PJ, O’Meara S, McBride DJ, Meynert A, et al. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463: 180–184.
10. Wei X, Walia V, Lin JC, Teer JK, Prickett TD, et al. (2011) Exome sequencing identifies GRN22 as frequently mutated in melanoma. Nat Genet 43: 442–446.
11. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463: 191–196.
12. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, et al. (2011) Initial genome sequencing and analysis of multiple myeloma. Nature 471: 467–472.
13. TCGAR (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615.
14. Berger MF, Lawrence MS, Demichelis F, Diey J, Cibulskis K, et al. (2011). The genomic complexity of primary human prostate cancer. Nature 470: 214–220.
15. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, et al. (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466: 689–693.
16. Tostiy T, Tsugumo K, Yamanoto S, Arai Y, Hosoda F, et al. (2011). High-resolution characterization of a hepatocellular carcinoma genome. Nat Genet 43: 464–469.
17. Berger MF, Hodis E, Hellerman TP, Deribe YL, Lawrance MS, et al. (2012). Melanoma genome sequencing reveals frequent PREX2 mutations. Nature 45: 502–506.
18. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, et al. (2012). Mutation processes molding the genomes of 21 breast cancers. Cell 149: 979–993.
19. Iovene D, Hambley SE, Stenson PD, Phillips AD, Kehler-Sawatzki H, et al. (2011) Comparative analysis of germline and somatic microresection mutational spectra in 17 tumor suppressor genes. Hum Mutat 32: 620–632.
20. Baele G, Van de Peer Y, Vanstaelandt S (2008) A model-based approach to study nearest-neighbor inferences reveals complex substitution patterns in non-coding sequences. Syst Biol 57: 675–692.
21. Nikolaev SI, Rimoldi D, Iceli C, Valesia A, Bobry D, et al. (2012) Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet 44: 153–159.
22. Burrows CJ, Muller JJ (1990) Oxidative nucleotide modifications leading to strand scission. Chem Rev 90: 1109–1132.
23. Dizdaroglu M (2012) Oxidatively induced DNA damage: Mechanisms, repair and disease. Cancer Lett 327: 28–47.
24. Turajlic S, Furney SJ, Lambros MB, Mitsopoulos C, Kozarewa I, et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell 149: 979–993.
25. Pfeifer GP, Besaratinia A (2012) UV wavelength-dependent DNA damage and DNA repair. Adv Exp Med Biol 754: 609–615.
32. Roberts SA, Sterling J, Thompson C, Harris S, Mav D, et al. (2012) Clustered
31. Baele G, Peer YV, Vansteelandt S (2009) Efficient context-dependent model
30. Zhang W, Bouffard GG, Wallace SS, Bond JP (2007) Estimation of DNA
29. Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent
structure rates by maximum likelihood. Mol Biol Evol 21: 468–488.
30. Zhang W, Boulard GG, Wallace SS, Bond JP (2007) Estimation of DNA
sequence-context-dependent mutation rates using primate genomic sequences.
31. Baele G, Peer YV, Vansteelandt S (2009) Efficient context-dependent model
based on clustering posterior distributions for non-coding sequences.
32. Roberts SA, Sterling J, Thompson C, Harris S, Mav D, et al. (2012) Clustered
33. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, et al. (2013) APOBEC3B
is an enzymatic source of mutation in breast cancer. Nature 494: 366–370.
34. Krauthammer M, Kong Y, Ha BH, Evans P, Bacchiocchi A, et al. (2012) Exome
sequencing identifies recurrent somatic RAC1 mutations in melanoma.
35. Nature Genet 44: 1006–1014.
36. Bocca A, Wang X, Yang J, Ahsan H, Chaturvedi AK, et al. (2010) Non-B DNA-forming
sequences and WRN deficiency independently increase the frequency of base substitution in human cells. J Biol Chem 286: 10017–10026.
37. Muren NB, Olmorn ED, Barton JK (2012) Solution, surface, and single molecule platform for the study of DNA-mediated charge transport. J Phys Chem Phys 14: 13734–13771.
38. Carmeli M, Zaidan TA, Kelley RF, Mi Q, Lewis FD, et al. (2009) Excited state, charge transfer, and spin dynamics in DNA hairpin conjugates with diradical dimers. J Phys Chem A 113: 4691–4700.
39. Wu Y, Yuan H, Tan S, Chen JQ, Tian D, et al. (2011) Increased complexity of gene structure and base composition in vertebrates. J Genet Genomics 38: 297–303.
40. Elder ES, Linton LM, Birken B, Nausha C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
41. Stark MS, Woods SL, Gartside MG, Bonazzi VF, Dutton-Regester K, et al. (2012) Frequent somatic mutations in MAP3K5 in metastatic melanoma. Nature Commun 3: 827.
42. Senthilkumar K, Grozema FC, Guerra CF, Bickelhaupt FM, Siebbeles LDA (2012) Mapping of the hallmarks of lung adenocarcinoma with massively parallel sequencing and analysis of the human genome. Nature 409: 860–921.
43. Jostter J, Bixor M, Rosch C, Spemann M, Wessely S (2001) Direct observation of hole transfer through DNA by hopping between adenine bases and by tunnelling. Nature 412: 318–320.
44. Nakken S, Rodland EA, Rogers T, Hoving E (2009) Large-scale inference of the point-mutation spectrum in human segmental duplications. BMC Genomics 10: 43.
45. Xie H, Wang M, Bischof J, Bonaldo M de F, Soares MB (2009) SNP-based prediction of the human germl cell methylation landscape. Genomics 93: 434–440.
46. Pan Z, Haribaran M, Arkin JD, Jalalvand AS, McCullagh M, et al. (2011) Electron donor-acceptor interactions with flanking purines influence the efficiency of thymine photodimerization. J Am Chem Soc 133: 20793–20798.
47. Bocca A, Wang X, Yang J, Ahsan H, Chaturvedi AK, et al. (2010) Non-B DNA-forming sequences and WRN deficiency independently increase the frequency of base substitution in human cells. J Biol Chem 286: 10017–10026.
48. Bocca A, Wang X, Yang J, Ahsan H, Chaturvedi AK, et al. (2010) Non-B DNA-forming sequences and WRN deficiency independently increase the frequency of base substitution in human cells. J Biol Chem 286: 10017–10026.
49. Saito I, Nakamura T, Nakatani K, Yoshioka Y, Yamaguchi K, et al. (1998) Mapping of the hot spots for DNA damage by one-electron oxidation: Efficacy of GG doublets and GGG triplets as a trap in long-range hole migration. J Am Chem Soc 120: 12686–12687.
50. Yokojima S, Yoshiki N, Yanoi W, Okada A (2009) Solvent effects on ionization potentials of guanine runs and chemically modified guanine in duplex DNA: effect of electrostatic interaction and its reduction due to solvent. J Phys Chem B 113: 16304–16309.
51. Friedman RA, Honig B (1993) A free energy analysis of nucleic acid base-pairing and length of simple repeats in vertebrate genomes are determined by their structural properties. Genome Res 18: 1545–1553.
52. Savici I, Pages F, Garcia-Fernández Del Llado I, Bautista M, González-Díez A, et al. (2002) Influence of the sequence dependent ionization potentials of guanines on the luminescence quenching of Ru(II)labeled oligonucleotides: A theoretical and experimental study. J Phys Chem B 106: 2763–2768.
53. Conwell EM, Basko DM (2001) Hole traps in DNA. J Am Chem Soc 123: 11441–11445.
54. Sugiyama H, Saito I (1996) Theoretical studies of GC-specific photocleavage of DNA via electron transfer: Significant lowering of ionization potential and 5′-localization of HOMO of stacked GCs in B-form DNA. J Am Chem Soc 118: 7063–7068.
55. Zaysteva IL, Trofimov AB, Schirmer J, Pleano K, Feyer V, et al. (2009) Theoretical and experimental study of valence-shell ionization spectra of DNA. J Phys Chem A 113: 15142–15149.
56. Park JH, Choi HY, Conwell EA (2004) Hole traps in DNA calculated with exponential-electron-lattice coupling. J Phys Chem B 108: 19438–19446.
57. Yoshihiko Y, Kitagawa Y, Takano Y, Yamagushi K, Nakamura T, et al. (1999) Excited and theoretical studies on the selectivity of GGG triplets toward one-electron oxidation in B-form DNA. J Am Chem Soc 121: 8712–8719.
58. Rubin AF, Green P (2009) Mutation patterns in cancer genomes. Proc Natl Acad Sci USA 106: 21766–21770.
59. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, et al. (2007) Paints of somatic mutation in human cancer genomes. Nature 446: 153–158.
60. Rubin AF, Green P (2009) Mutation patterns in cancer genomes. Proc Natl Acad Sci USA 106: 21766–21770.
61. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214–218.
62. Pfeifer GP, Besaratinia A (2009) Mutational spectra of human cancer. Hum Genet 125: 493–506.
63. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214–218.
64. Pfeifer GP, Besaratinia A (2009) Mutational spectra of human cancer. Hum Genet 125: 493–506.
65. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214–218.
66. Pfeifer GP, Besaratinia A (2009) Mutational spectra of human cancer. Hum Genet 125: 493–506.
67. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214–218.
68. Pfeifer GP, Besaratinia A (2009) Mutational spectra of human cancer. Hum Genet 125: 493–506.
85. O’Rawe J, Jiang T, Sun G, Wu Y, Wang W, et al. (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5: 28.
86. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20: 490–497.
87. Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, et al. (2011) Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics 12: 651.
88. Nakamura K, Ohshima T, Morimoto T, Ieda S, Yoshikawa H, et al. (2011) Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39: e90.
89. Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, et al. (2012) Fast computation and applications of genome mappability. PLoS One 7: e30577.
90. Zheng G, Lu XJ, Olson WK (2009) Web 3DNA—a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. Nucleic Acids Res 37: W240–246.
91. Wang J, Cieplak P, Kollman PA (2008) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J Comput Chem 21: 1049–1074.
92. Yang Z, Lasker K, Schneidman-Duhovny D, Webb B, Huang CC, et al. (2012) UCSF Chimera, MODELLER, and IMP: An integrated modeling system. J Struct Biol 179: 269–278.
93. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4: 417–427.
94. Kohn W, Becke AD, Parr RG (1996) Density functional theory of electronic structure. J Phys Chem 100: 12974–12980.
95. Zhao Y, Truhlar DG (2008) The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. Theor Chem Acc 120: 215–241.
96. Zhao Y, Truhlar DG (2008) Density functionals with broad applicability in chemistry. Acc Chem Res 41: 157–167.
97. Hariraha PC, Pople JA (1973) Influence of polarization functions on molecular-orbital hydrogenation energies. Theor Chim Acta 20: 213–222.
98. Fraud MM, Pietro WJ, Hicher WJ, Bankley JS, Gordon MS, et al. (1982) Self-consistent molecular-orbital methods. 23. A polarization-type basis set for 2nd-row elements. J Chem Phys 77: 3654–3665.
99. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, et al. (1993) General atomic and molecular electronic-structure system. J Comput Chem 14: 1347–1363.
100. Bode BM, Gordon MS (1998) MacMolPlt: A graphical user interface for GAMESS. J Mol Graphics Mod 16: 133–136.
101. Yi M, Mudunuri U, Che A, Stephens RM (2009) Seeking unique and common biological themes in multiple gene lists or datasets: pathway pattern extraction pipeline for pathway-level comparative analysis. BMC Bioinformatics 10: 200.
102. Yi M, Horton JD, Cohen JC, Hobbs HH, Stephens RM (2006) Whole-PathwayScope: a comprehensive pathway-based analysis tool for high-throughput data. BMC Bioinformatics 7: 30.
103. Fernandez A, Gomez S (2008) Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. J Classif 25: 43–65.