Open access • Journal Article • DOI:10.1002/HUMU.21483

Comparative analysis of germline and somatic microlesion mutational spectra in 17 human tumor suppressor genes — Source link

Dobril Ivanov, Stephen E. Hamby, Peter D. Stenson, Andrew David Phillips ...+3 more authors

Institutions: Cardiff University, Nottingham Trent University, University of Ulm

Published on: 01 Jun 2011 - Human Mutation (Wiley Blackwell)

Topics: Germline mutation, Germline mosaicism, Germline, Missense mutation and Somatic cell

Comparative Analysis of Germline and Somatic Micro-lesion Mutational Spectra in 17 Human Tumour Suppressor Genes

Nadia Chuzhanova, Dobril Ivanov, Stephen Edward Hamby, Peter Stentson, Andrew D Phillips, Hildegard Kehrer-Sawatzki, David N. Cooper

Human Mutation

Comparative Analysis of Germline and Somatic Micro-lesion Mutational Spectra in 17 Human Tumour Suppressor Genes

Journal:	*Human Mutation*
Manuscript ID:	humu-2010-0507.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	10-Jan-2011
Complete List of Authors:	Chuzhanova, Nadia; Nottingham Trent University, School of Science and Technology Ivanov, Dobril; Cardiff University, Department of Psychological Medicine and Neurology Hamby, Stephen; Nottingham Trent University, School of Science and Technology Stentson, Peter Phillips, Andrew; Cardiff University, Institute of Medical Genetics, College of Medicine Kehrer-Sawatzki, Hildegard; University of Ulm Cooper, David; Cardiff University, Institute of Medical Genetics, College of Medicine
Key Words:	germline and somatic mutational spectra, tumour suppressor genes, recurrent mutation, mutation hotspot, non-B DNA, driver mutations

Comparative Analysis of Germline and Somatic Micro-lesion Mutational Spectra in 17 Human Tumour Suppressor Genes

Dobril Ivanov[1,2], Stephen E. Hamby[3], Peter D. Stenson[1], Andrew D. Phillips[1],
Hildegard Kehrer-Sawatzki[4], David N. Cooper[1] and Nadia Chuzhanova[3]

[1]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

[2]MRC Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine and Neurology, Biostatistics and Bioinformatics Unit, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

[3]School of Science and Technology, Nottingham Trent University, Nottingham, NG11 8NS, UK

[4]Institute of Human Genetics, University of Ulm, Albert-Einstein-Allee 11, 89081 Ulm, Germany

*All correspondence to: Prof. Nadia Chuzhanova, School of Science and Technology Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK

Tel: +44 (0) 0115 848 8304 E-mail: nadia.chuzhanova@ntu.ac.uk

Abstract

Mutations associated with tumorigenesis may either arise somatically or can be inherited through the germline. In this study, we performed a comparison of somatic, germline and shared (found in both soma and germline) mutational spectra for 17 human tumour suppressor genes which included missense single base-pair substitutions and micro-deletions/micro-insertions. Somatic and germline mutational spectra were similar in relation to C.G>T.A transitions but differed with respect to the frequency of A.T>G.C, A.T>T.A and C.G>A.T substitutions. Shared missense mutations were characterised by higher mutability rates, greater physicochemical differences between wild-type and mutant amino acid residues, and a tendency to occur in evolutionarily conserved amino acid residues and within CpG/CpHpG oligonucleotides. Mononucleotide runs (≥ 4 bp) were identified as hotspots for shared micro-deletions/micro-insertions. Both germline and somatic micro-deletions/micro-insertions were found to be significantly overrepresented within the 'indel hotspot' motif, GTAAGT. Using a naïve Bayes' classifier trained to discriminate between somatic, recurrent somatic, germline, shared and recurrent shared missense mutations, 63.1% of mutations in our dataset were correctly recognized. Using this classifier to analyse an independent dataset of probable driver mutations, we concluded that ~50% of these somatic missense mutations possess features consistent with their being either shared or recurrent, suggesting that a disproportionate number of such lesions are likely to be drivers of tumorigenesis.

Key Words: germline and somatic mutational spectra; tumour suppressor genes; recurrent mutation; mutation hotspot; non-B DNA; driver mutations

Introduction

A major distinction to be made between somatic and germline mutations is that the former occur during mitotic cell cycles whereas the latter are generally meiotic in origin. In addition, whilst somatic cancer-causing gene lesions come to clinical attention by conferring a growth advantage upon the affected cells or tissue, germ-line gene mutations causing inherited disease normally come to attention by conferring a disadvantage upon the individual, usually through haploinsufficiency. Finally, whereas inherited disease usually implies only one or two pathological mutations at a specific locus, cancer is often characterized by multiple somatic mutations distributed genome-wide. Those somatic mutations which confer a growth advantage on the cells in which they occur, which are positively selected for in the emerging tumour mass and which have therefore been causally implicated in tumorigenesis, are termed 'driver' mutations [Stratton *et al.*, 2009]. By contrast, those mutations which do not confer any growth advantage and have not been subject to selection during tumorigenesis, are termed 'passenger' mutations [Stratton *et al.*, 2009]. Such passenger mutations may arise at high frequency as a consequence either of increased genomic instability or simply due to the considerable number of cell divisions required to convert a single transformed cell into a clinically detectable tumour [Lengauer et al., 1998; Boland and Ricciardiello, 1999; Simpson 2008; Parmigiani et al. 2009; Stratton et al., 2009].

Despite these basic differences, the mutational spectra (and hence the underlying mutational mechanisms) associated with single base-pair substitutions [Krawczak et al., 1995; Schmutte and Jones, 1998; Cole et al., 2008; Lobo et al., 2009], micro-deletions and micro-insertions [Jego et al., 1993; Greenblatt et al. 1996] and gross gene rearrangements [Oldenburg et al., 2000; Kolomietz et al., 2002] in specific genes often appear to exhibit marked similarities between the germline and the soma. Further, certain triplet repeats associated with a number of inherited human conditions are known to be unstable in both the germline and somatic tissues, a finding which serves to explain not only the phenomenon of genetic anticipation characteristic of these disorders but also their inherent inter-individual clinical variability [Giovannone et al., 1997; Leeflang et al., 1999; Martorell et al., 2000; Sharma et al., 2002; Pollard et al., 2004]. However, by contrast, highly variable human minisatellites can display markedly different degrees of instability between the soma and the germline [Buard et al., 2000; Stead and Jeffreys, 2000; Shanks et al., 2008]. These studies notwithstanding, few attempts have so far been made to compare the nature, location and relative frequency of germline and somatic mutations.

Human cancer genes usually harbour either somatic or germline mutations [Goode et al., 2002; Futreal et al., 2004; Vogelstein and Kinzler, 2004]. There is, however, one category of cancer gene, broadly termed tumour suppressors, that by virtue of their being mutated in both the germline and the soma, provides us with an ideal model system to compare somatic vs. germline mutational spectra [Futreal et al., 2004]. Tumour suppressor genes, defined as “genes that sustain loss-of-function mutations in the development of cancer” [Haber and Harlow, 1997], are involved in the regulation of a diverse array of different cellular functions including cell cycle checkpoint control, detection and repair of DNA damage, protein ubiquitination and degradation, mitogenic signalling, cell specification, differentiation and migration, and tumour angiogenesis [Sherr, 2004]. They encode proteins with a regulatory role in cell cycle progression (e.g. Rb), DNA-binding transcription factors (e.g. p53) and inhibitors of cyclin-dependent kinases required for cell cycle progression (e.g. p16). In inherited cancer syndromes, the mutational inactivation of both tumour suppressor alleles is required to change the phenotype of the cell. This ‘two hit hypothesis’ provides the basis for our mechanistic understanding of tumour suppressor gene mutagenesis: a first (inherited) mutation in one tumour suppressor allele is followed by the somatic loss of the remaining wild-type allele via a number of different mutational mechanisms [Knudson, 2001]. Whereas the inherited lesion is usually fairly subtle, the second (somatic) hit may also involve the deletional loss of the entire gene or even a substantial portion of the chromosome involved. Alternatively, both ‘hits’ may constitute somatic mutations: whatever the actual mechanism, the end result is the same – the loss or inactivation of both gene copies. Some interplay may however occur between the soma and the germline in that the location of the germline mutation can in some instances influence the nature, frequency and location of the subsequent somatic mutation [Lamlum et al., 1999; Groves et al., 2002; Latchford et al., 2007; Dallosso et al., 2009].

Tumour suppressor genes are often somatically inactivated by mutational mechanisms that are almost exclusively confined to the soma and which are found only infrequently in the germline (e.g. gross mutations characterized by loss of heterozygosity, epi-mutations such as methylation-mediated promoter inactivation, and micro-lesions within highly repetitive sequence elements that are consequent to microsatellite instability). However, a typical spectrum of somatic mutations associated with tumorigenesis may also include gross rearrangements, copy number variation, and various types of micro-lesion (e.g. micro-deletions, micro-insertions and indels) including single base-pair substitutions [Loeb and Harris, 2008; Stratton et al., 2009]. Although the somatic micro-lesions are often quite similar to their germline counterparts, few studies of tumour suppressor genes have so far attempted to compare and contrast germline and somatic mutational spectra with respect to these relatively subtle types of mutation. However, such studies have indicated that germline and somatic micro-lesions can display remarkable similarities in terms of mutation type, location and relative frequency of occurrence, and hence by inference the putative underlying mechanisms of mutagenesis [Marshall et al., 1997; Ali et al., 1999; Gallou et al., 1999; Richter et al., 2003; Upadhyaya et al., 2004; Glazko et al., 2004; Tartaglia et al., 2006; Baser et al., 2006; Upadhyaya et al., 2008].

We attempt here a first formal comparison between germline and somatic micro-lesion mutational spectra for a total of 17 different human tumour suppressor genes [*APC* (MIM# 611731), *ATM* (MIM# 607585), *BRCA1* (MIM# 113705), *BRCA2* (MIM# 600185), *CDH1*

(MIM# 192090), *CDKN2A* (MIM# 600160), *NF1* (MIM# 162200), *NF2* (MIM# 607379), *PTCH1* (MIM# 601309), *PTEN* (MIM# 601728), *RB1* (MIM# 180200), *STK11* (MIM# 602216), *TP53* (MIM# 191170), *TSC1* (MIM# 605284), *TSC2* (MIM# 191092), *VHL* (MIM# 608537) and *WT1* (MIM# 607102)].

Materials and Methods

Sources of germline and somatic mutation data

Data on germline and somatic micro-lesions (viz. missense mutations, micro-deletions and micro-insertions involving ≤20 bp) were collated for 17 different human tumour suppressor genes. Germline mutation data were obtained from the Human Gene Mutation Database [HGMD; http://www.hgmd.org; Stenson et al., 2009]. Somatic mutation data were compiled from a number of different sources including online somatic mutational databases viz. *Catalogue of Somatic Mutations in Cancer* (http://www.sanger.ac.uk/genetics/CGP/cosmic; *RB1* and *PTEN*), the *Breast Cancer Information Core* (http://research.nhgri.nih.gov/bic; *BRCA1*), the *RB1 Gene Mutation Database* (http://www.verandi.de/joomla; *RB1*), the *International NF2 Mutation Database* (http://www.hgmd.cf.ac.uk/nf2; *NF2*), the *CDKN2A Database* (https://biodesktop.uvm.edu/perl/p16; *CDKN2A*) and the *IARC TP53 Mutation Database* (http://www-p53.iarc.fr; *TP53*), the *VHL Mutations Database* (http://www.umd.be/VHL/), and data privately communicated by Eamonn Maher (*VHL*) and Gareth Evans (*NF2*). Additional somatic mutation data [for *APC*, *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *PTCH1*, *STK11*, *TSC1*, *TSC2* and *WT1*] were obtained by searching PubMed.

To be regarded as *bona fide* somatic mutations, and therefore suitable for inclusion in this analysis, reported lesions had to have been shown not only to be present in a tumour tissue but also to be absent from a non-tumour tissue (usually blood) from the same patient. Hence, mutational data derived from 'sporadic' patients were not included unless a non-tumour tissue had also been examined in order to exclude the possibility that the lesions detected were constitutional in origin. Depending upon the number of independent occurrences, *f*, of a given somatic or shared mutation described in the literature, these mutation types were further subdivided into two categories: *recurrent mutations* (f>1) and *non-recurrent mutations* (f=1). At the time this study was initiated (October 2006), the number of available germline and somatic missense mutations for each of the 17 studied tumour suppressor genes were as listed in Table 1.

The analysis reported here focussed exclusively on missense mutations and micro-deletions/ micro-insertions. Nonsense mutations in tumour suppressor genes have already been addressed in the context of a general meta-analysis of this type of lesion [Mort et al., 2008]. Indels (representing a combination of micro-deletion and micro-insertion) were excluded from the analysis owing to their paucity.

Control datasets of potential mutations

For every tumour suppressor gene examined, all possible single base-pair substitutions in the gene coding sequence that (i) could potentially have given rise to a missense mutation and (ii) were not already included in either of the corresponding observed somatic and/or germline mutational spectra, were generated. These ‘potential missense mutations’ were used as a control dataset.

For each tumour suppressor gene, a matching control dataset of ‘potential micro-deletions’ was also generated by randomly selecting a first breakpoint and then choosing the length of the simulated micro-deletion (and therefore, the position of the second breakpoint) by reference to the probability distribution calculated for micro-deletions (from 1 bp to 20 bp) observed in the corresponding dataset of mutations. A matching dataset of micro-insertions was generated in similar fashion, with the sites of insertion being randomly selected. Since some of the micro-deletion/micro-insertion breakpoints occurred within an intron, extended cDNA sequences comprising exons and additional flanking intron sequences were used to generate corresponding control datasets.

Grantham scores

The ‘Grantham score’ or ‘Grantham difference’ [Grantham, 1974] measures the chemical difference between wild-type and mutated amino acid residues in terms of their side chain composition (i.e. the weight ratio of non-carbon components in end-groups or rings to carbons in side chains), polarity (i.e. basic, acidic or nonpolar depending upon side chain charge) and molecular volume.

On average, the physicochemical differences manifested by orthologous amino acid substitutions that have accumulated over evolutionary time will tend to be relatively small. By contrast, disease-causing substitutions are expected to exhibit higher Grantham scores, indicative of more dramatic physicochemical differences between the wild-type and mutated amino acid residues [Krawczak et al., 1998]. The values tabulated by Grantham [1974] were used in this study to calculate a median Grantham score for each set of missense mutations for each tumour suppressor gene.

Degree of evolutionary conservation

Amino acid residues that are highly conserved in orthologous proteins frequently represent sites of structural or functional importance. Hence, such highly conserved amino acid residues/protein regions often constitute hotspots for observed pathological mutations as a consequence of phenotype selection (rather than intrinsic mutability). To assess the degree of evolutionary conservation of those codons affected by somatic/germline mutations, orthologous tumour suppressor cDNA and protein sequences from different vertebrate species were retrieved from NCBI’s Entrez Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene). The species used as a source of these cDNA and protein sequences are listed in Supp. Table 1 for each tumour suppressor gene/protein. ClustalX (http://www.clustal.org/) was used to align the protein sequences. A program was written to replace all amino acids in the protein alignments by cDNA-derived codons, thereby avoiding the introduction of gaps within codons.

The evolutionary constraints acting upon the 17 human tumour suppressor genes at the codon level were inferred by calculating the $\frac{Ka}{Ka + Ks}$ ratio for each codon where *Ks* and *Ka* are respectively the relative numbers of synonymous and nonsynonymous substitutions between codons in two aligned sequences [Walker et al., 1999]. If two aligned codons required more than one substitution to be transformed into each other, then the minimum number of substitutions was assumed, and the most parsimonious path was determined using a PAM100 matrix and the Nei & Gojobori [1986] pathway method. Gaps inserted into the non-human vertebrate orthologous cDNA sequences during alignment were treated as being equivalent to a non-synonymous substitution. Codons that were not present in the human cDNA sequence were not considered. A value representing the median level of evolutionary conservation across all codons was then derived for each mutational spectrum.

Relative mutability rates

To assess the likelihood of observing a certain nucleotide change in a given position and in a specific context, two tabulated measures of the nearest neighbour-dependent mutation rate were employed. The first was derived from 20,200 single base-pair substitutions inferred from alignments of paired human gene/pseudogene sequences [Hess et al., 1994]. This was termed the *non-disease-associated mutability rate* and, since it approximates to the neutral mutation frequency, it should reflect the intrinsic mutability of the underlying DNA sequence. One would expect the non-disease-associated mutation rates associated with pathological mutations to be low implying that these specific substitutions are much less likely to occur as neutral substitutions.

The nearest neighbour-dependent mutation rates derived from germline single base-pair substitutions [using data from the Human Gene Mutation Database (HGMD); Stenson et al., 2009] by Krawczak et al. [1998] were used as an approximation of the *disease-associated mutability rate*. This mutation rate is a function of selection for loss of biological function as well as the underlying intrinsic mutability of the DNA sequence.

Repetitive sequence elements

A variety of repetitive sequence elements have been reported in association with human gene mutations causing both inherited disease and cancer. Direct and inverted repeats and symmetric elements [see Chuzhanova et al. 2003 for definitions] of length ≥8 bp, and less than 21 bp apart, capable of forming non-B DNA structures, were therefore sought within the extended cDNA sequences (comprising exons and up to ±85 bp of flanking sequence) using purposely designed software. In addition, DNA sequences were screened for the presence of mononucleotide runs of ≥4 bp.

Mutation descriptors

Each missense mutation was ascribed various descriptors indicating (a) the type of mutation [i.e. shared mutation (i.e. found to occur both somatically and in the germline); exclusively somatic; exclusively germline; shared recurrent mutation (i.e. found to occur not only in the germline but also somatically on more than one occasion; somatic recurrent mutations (recorded in the soma more than once, but not in the germline); potential mutation (as defined above)] and (b) its location [i.e. C→T and G→A within a CpG dinucleotide or within a CpHpG trinucleotide (where H=A, C or T) or in a repeat sequence (as described above)]. Mutations that have been reported as being exclusively somatic or exclusively germline will henceforth be referred to simply as ‘somatic’ and ‘germline’, respectively. The shared mutations, comprising the overlap between the somatic and germline mutations, may be visualized in the form of a Venn diagram (Figure 1). All somatic missense (including shared) mutations were further described as being either recurrent or non-recurrent (in the soma, see above; Figure 1). No such division was made for the relatively small number of recurrent micro-deletions and micro-insertions available; both recurrent and non-recurrent somatic mutations were therefore included in either the somatic or the shared datasets and labelled accordingly (Figure 1).

All micro-lesions (*viz.* missense mutations, micro-deletions and micro-insertions) in each gene were also labelled with respect to their occurrence within a region spanning a repetitive element or mononucleotide run including ±5 bp of flanking sequence. If a missense mutation (or at least one micro-deletion/micro-insertion breakpoint) was found to occur within this extended region, the micro-lesion was labelled as being found in association with the corresponding type of repeat.

Assessing the statistical significance of the results generated

To assess the similarity (or dissimilarity) of the germline and somatic mutational spectra with respect to (i) the frequency with which the missense mutations were located within CpG/non-CpG dinucleotides or CpHpG/non-CpHpG trinucleotides and (ii) the frequency with which the micro-deletions/micro-insertions were found within/outwith repeats, the various non-overlapping mutation datasets (bearing specific descriptors) were compared by means of the χ^2 test. Since the normality assumption did not hold for the datasets studied, the Wilcoxon rank-sum test was used to compare and contrast missense mutational spectra with respect to the Grantham score, degree of evolutionary conservation, and both the non-disease- and disease-associated mutability rates.

The permutation-based method [Olshen and Jain, 2002] was used to estimate the significance of our findings and to allow for multiple testing wherever appropriate. For each comparison, the null hypothesis [viz. no overall difference between two groups of mutations (e.g. somatic and potential) with respect to the specific property in question (e.g. occurrence in CpG or non-CpG nucleotides)], was tested for, either in the context of each gene or all genes combined. χ^2 or rank-sum statistics were calculated for the observed germline and somatic mutations as well as for 10,000 control sets of mutations created from the original sets by random permutation of the assigned mutational descriptors (e.g. randomly chosen mutations labelled as 'somatic' were re-labelled as 'germline'; randomly chosen mutations labelled as 'shared' were re-labelled as 'somatic', etc.). The test statistic (χ^2 or rank-sum) for the original datasets that exceeded the 95th percentile of χ^2 maxima for 10,000 control sets was deemed to be statistically significant; the corresponding p-value was termed the 'gene-wise' p-value. To allow for multiple testing in those cases where specific mutations in all genes were combined, a Bonferroni correction was applied; the corresponding p-value was termed the 'experiment-wise' p-value.

Naïve Bayes classifier

A decision tree classifier known as a Naïve Bayes tree [NBTree; Kohavi, 1996], implemented in the Weka machine learning package [Witten and Frank, 2005], was trained to discriminate between somatic, germline, shared, recurrent somatic and recurrent shared missense mutations. Each mutation was described by a total of six features including the degree of evolutionary conservation, the non-disease-associated and disease-associated relative mutability rates, Grantham score, and occurrence in CpG/CpHpG, non-CpG/non-CpHpG doublets/triplets or in repeats/mononucleotide runs. Ten-fold cross-validation was used to assess the accuracy of classification. The mutation datasets were balanced using random oversampling [Kotsiantis et al., 2006] by replicating random instances from the minority classes until all classes were represented by the same number of instances as the majority class.

Results and Discussion

The availability of both germline and somatic mutational spectra from tumour suppressor genes provides us with an ideal opportunity to study the nature of mutation of the same gene sequences in both the germline and the soma. The analysis reported here explores for the first time the similarities and differences exhibited by the germline, somatic (and shared) micro-lesion mutational spectra in 17 human tumour suppressor genes. The study presented here focussed upon missense mutations and micro-deletions as well as micro-insertions. Nonsense mutations in tumour suppressor genes have already been addressed elsewhere in the context of a general meta-analysis of this type of lesion [Mort et al., 2008].

Characteristics of germline and somatic missense mutations with respect to mutation type

Taken together, the combined mutational spectra for all 17 tumour suppressor genes contained twice as many somatic (61%) as germline (31%) mutations. For five genes (*APC*, *CDKN2A*, *NF2*, *PTEN* and *TP53*), a predominance of somatic over germline mutations was noted, with the *TP53* gene having the highest proportion of somatic mutations (92%). For the majority of genes, however (namely *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *PTCH1*, *RB1*, *STK11*, *TSC1*, *TSC2*, *VHL* and *WT1*), the analysed dataset included more germline than somatic mutations, with >97% of all mutations in the *BRCA1*, *NF1*, *TSC2* and *WT1* genes being germline in origin.

Shared mutations are of particular interest because identical mutational mechanisms operating in the germline and the soma may be inferred for such lesions. The expected number of shared mutations for each gene was calculated as $p_{\text{somatic}} \times p_{\text{germline}} \times (\text{total number of mutations})$, where *p* denotes the relative frequencies of somatic and germline mutations. Although the proportion of shared mutations varies markedly between genes (from 0% to 25% of the total), only two genes (*TP53* and *VHL*) were found to have a higher than expected number of shared mutations as calculated above.

Patterns of germline and somatic missense mutations by mutation type

Missense mutations were characterised by a predominance of transitions over transversions (Figure 2). The transition:transversion ratio was at its highest for shared recurrent mutations (3.5) and shared non-recurrent mutations (2.7). By contrast, the transition:transversion ratio for the control group (i.e. potential mutations) was 0.85. Significant differences in the transition:transversion ratio were observed between all mutation types (p<0.05) with the exception of germline vs. shared mutations (Figure 2).

Not surprisingly, a strong positive correlation was noted between somatic and shared mutational spectra (Pearson's correlation r=0.986, p= 2.91×10^{-4}) with respect to the frequencies of six mutational changes viz. A.T>C.G, A.T>G.C, A.T>T.A, C.G>A.T, C.G>G.C and C.G>T.A. Weaker negative correlations were found between somatic mutations and the control dataset of mutations (r= –0.887, p=0.019) and between shared and the control (r= –0.837, p=0.038) mutational spectra, indicative of the non-randomness of somatic mutation.

C.G>T.A transitions constituted the most frequent type of mutation in shared (46%), germline (29%) and somatic (25%) mutational spectra, significantly higher proportions than noted in the spectrum of mutations within our control dataset (13%, p<0.001) (Figure 2). Intriguingly, the number of A.T>G.C mutations was significantly higher (28%) in the germline as compared to the somatic (16%), shared (17%) and control (16%) mutational spectra (Figure 2). A.T>C.G mutations were significantly under-represented in the shared mutational spectrum (7%, p<0.001) as compared to the other spectra whereas A.T>T.A mutations were under-represented (7%, p<0.001) in both the germline and shared mutational spectra compared to both somatic and potential mutations (Figure 2). Finally, C.G>A.T mutations were significantly underrepresented in the germline mutational spectrum (10%) as compared to the somatic (16%, $p=1.2\times10^{-5}$) and potential (15%, $p=2.6\times10^{-5}$) spectra. Thus, the main similarity between the somatic and germline missense mutational spectra was in relation to C.G>T.A transitions whereas the main differences between these spectra involved the A.T>G.C, A.T>T.A and C.G>A.T mutations. It should be noted that the patterns of somatic nucleotide substitution exhibited by the 17 tumour suppressor genes studied here were markedly different from the genome-wide patterns of somatic nucleotide substitution previously observed in various cancer genome sequencing studies [Sjöblom et al., 2006; Greenman et al., 2007; Kan et al., 2010].

CpG- and CpHpG-located missense mutations

The CpG dinucleotide is a well known mutational hotspot in the human genome as a consequence of the spontaneous (and endogenous) deamination of 5-methylcytosine. In addition, Lister et al. [2009] reported abundant DNA methylation in CpHpG trinucleotides in the human genome, where H is either A, C or T, raising the possibility that CpHpG might also be a generalized mutation hotspot [Cooper et al., 2010].

The proportion of missense mutations that were either C>T or G>A within CpG or CpHpG oligonucleotides in the 17 tumour suppressor genes was found to vary between 0% and 100% (Table 2). This wide range in values may be attributed to the small size of some of the gene mutation datasets under study. Importantly, the CpG and CpHpG oligonucleotides were found to be disproportionately likely to harbour shared mutations; thus, 34% of shared recurrent mutations and 21% of shared non-recurrent mutations were C>T and G>A mutations in CpG dinucleotides with an additional 10% and 9% of mutations, respectively, occurring within CpHpG trinucleotides. Since driver mutations tend to occur disproportionately frequently within

CpG dinucleotides [Talavera et al., 2010], we postulate that missense mutations identified as being shared are highly likely to be driver mutations.

Significant differences were noted between the relative frequencies of CpG- and CpHpG-located mutations for somatic, germline, shared, somatic recurrent and shared recurrent missense mutations (Supp. Table 2).

We have previously shown that 18.2% and 9.9% of all missense/nonsense mutations recorded in the HGMD are C>T and G>A transitions in CpG and CpHpG oligonucleotides respectively [Cooper et al., 2010]. In the present study, we observed that the mutational spectra of shared and shared recurrent missense mutations in tumour suppressor genes were both found to be significantly enriched in CpG-located mutations (χ^2-test; p-values, 0.028 and 1.1×10^{-9} respectively). This implies that the CpG dinucleotide is a generalized mutation hotspot in both the soma and the germline as a consequence of the endogenous mutational mechanism of methylation-mediated deamination of 5-methylcytosine. By contrast, the number of CpG-located mutations was significantly underrepresented (χ^2-test; p-values$<5\times10^{-14}$) in the other mutational spectra (i.e. non-recurrent somatic, somatic recurrent and germline mutations) by comparison with HGMD data. To perform these comparisons, missense mutations (Table 2) and nonsense mutations [previously reported in Mort et al., 2008; see Table 6 therein] in all 17 tumour suppressor genes were combined. The proportion of shared recurrent missense mutations in tumour suppressor genes that were CpHpG-located was found to be significantly higher (p=0.023) than for mutations recorded in the HGMD whereas CpHpG-located somatic and recurrent somatic mutations were significantly under-represented ($p<4\times10^{-10}$). Significant enrichment in CpHpG-located mutations was observed for germline mutations as compared to somatic mutations ($p<3\times10^{-10}$) consistent with the reported decrease in CpHpG methylation in differentiated cells [Lister et al., 2009]. In summary, germline and shared missense mutations were found to be significantly enriched at CpG and CpHpG oligonucleotides.

The numbers of somatic and shared C>T and G>A transitions recorded within CpG dinucleotides for each gene (Table 2) did not correlate with the numbers of CpG dinucleotides found in these genes (r <–0.5, p>0.127) and hence do not simply reflect intragenic CpG frequency. A weak positive correlation between CpG-located mutations and the number of genic CpG dinucleotides was however noted for germline mutations (r= 0.489, p=0.046) indicating that CpG methylation is not entirely unrelated to the number of CpG dinucleotides, at least with respect to the germline; the relationship is however clearly more complex in the soma, possibly due to inter-tissue differences in gene methylation patterns [Tornaletti and Pfeifer, 1995] or transcription-coupled repair [Rubin and Green, 2009].

No correlation was found between the numbers of somatic, germline and shared mutations recorded within CpHpG trinucleotides and the corresponding numbers of CpHpG trinucleotides for these genes (r= -0.316, 0.373, -0.414; p-values 0.281, 0.216 and 0.098, respectively) indicating that mutation within CpHpG trinucleotides is likely to be very much a gene-specific phenomenon (presumably dependent on both the extent and the degree of spatial localization of CpHpG methylation in the germline and/or soma).

Finally, the number of CpG dinucleotides in the various tumour suppressor genes studied (Table 2) was not found to correlate with gene length (r= 0.3, p-value=0.241). By contrast, we found a significant correlation (r= 0.885, p-value=2.35×10^{-6}) between tumour suppressor gene length and the number of CpHpG trinucleotides (excluding those with mutations), indicating that the tumour suppressor genes under study possess a similar density of CpHpG trinucleotides per unit length. We surmise that the factors that govern the establishment of the methylation pattern of CpHpG trinucleotides are likely to be quite complex.

Evolutionary conservation of tumour suppressor genes *in relation to the sites of somatic and germline missense mutations*

For all 17 tumour suppressor genes, the degree of evolutionary conservation, as measured by Ka/Ks, was less than unity, indicating that these genes (and proteins) have been highly conserved evolutionarily as a consequence of the action of purifying selection. Indeed, the degree of evolutionary conservation displayed by most of the studied genes was markedly lower than the average (~0.18) noted in a comparison of 1880 human, rat and mouse gene orthologues [Makalowski and Boguski, 1998]. However, three genes (*CDKN2A, BRCA1* and *BRCA2*) were found to exhibit a higher rate of evolutionary conservation than the average between human and rodents.

The evolutionary conservation of each mutated codon was inferred by calculating the $\frac{Ka}{Ka+Ks}$ ratio; for each gene/spectrum, the mean value was then calculated across all mutations in the corresponding gene/spectrum. Shared recurrent missense mutations were found to occur disproportionately in highly conserved amino acid residues (mean degree of evolutionary conservation, 0.072) followed by shared non-recurrent mutations (0.138), somatic recurrent (0.169), germline (0.175), non-recurrent somatic (0.265), and control dataset mutations (0.255). The observed differences in the degree of evolutionary conservation for the different mutational spectra are shown in Supp. Table 2. These quite specific findings are consistent with the previously reported general tendency for cancer-associated mutations to occur frequently at evolutionarily conserved sites [Greenblatt et al., 2003; Tavtigian et al., 2009; Talavera et al., 2010].

Somatic non-recurrent mutations were found to occur in codons characterized by the highest mean value of $\frac{Ka}{Ka+Ks}$ ratios as compared not only to the shared recurrent and shared non-recurrent mutations (see above) but also to the mutations within the control dataset. This is consistent with the interpretation that a high proportion of non-recurrent somatic mutations, and most notably those which are located in less evolutionarily conserved regions, are likely to be 'passenger' mutations.

Missense mutations in relation to the disease- and non-disease-associated substitution rates

Employing alignments of paired human gene/pseudogene sequences, Hess et al. [1994] derived relative (non-disease-associated) nearest-neighbour-dependent mutability rates using the lowest frequency substitution type, C(T>G)A/T(A>C)G, as a baseline. These mutability rates were found to vary over a 52-fold range, with unity being assigned to the lowest frequency substitution type. This *non-disease-associated* mutability rate approximates to the neutral mutation frequency and hence reflects the intrinsic mutability of the underlying DNA sequence. Depending upon the observed nearest-neighbour context, we retrieved the corresponding non-disease-associated mutability rate (from the data of Hess et al. 1994) for each mutation (either observed or from the control dataset) and calculated the median value for each mutational spectrum. These median values are indicative of the relative mutability of each tumour suppressor gene. The median values were found to vary between 4 (*NF2*) and 8.9 (*STK11*) for somatic mutations, 4.1 (*TP53*) and 10.1 (*WT1*) for germline mutations, and 7.2 (*RB1*) and 11 (*PTEN*) for shared mutations (values given only for genes with more than three mutations in the corresponding category; see Supp. Table 3, indicating that many of the median values are quite low and hence the corresponding mutations are unlikely to be neutral.

When data from all 17 genes were combined, shared recurrent mutations were found to be characterised by intrinsically low non-disease-associated mutability (median=11), followed by even lower median mutability values for shared non-recurrent mutations (7.9), germline mutations (7.2), somatic recurrent and non-recurrent (4.7) and control dataset mutations (4.1). Such low median mutability values across all groups indicates that at least half of the mutations within observed triplets are unlikely to be neutral in the sense defined by Hess et al. [1994] and hence are not simply explicable in terms of intrinsic DNA mutability. The low median mutability values for the control dataset of mutations within tumour suppressor genes reflect the high level of evolutionary conservation manifested by tumour suppressor gene coding sequences across different species, implying that any mutation within a triplet characterized by a low non-disease-associated mutation rate is very likely to have pathological consequences and would thus be subject to purifying selection.

In contrast to the non-disease-associated mutability rate (which is purely a reflection of the intrinsic DNA mutability), the disease-associated mutability rate reflects (in addition to the intrinsic DNA mutability) the increased likelihood of coming to clinical attention conferred by the loss of biological function. The C(G>T)T mutation is one of the most frequent types of mutation associated with the loss of biological function [disease-associated mutability rate 10.255; Krawczak et al., 1998] but occurs much less frequently among neutral mutations [non-disease-associated mutability rate 4.4; Hess et al., 1994].

For each tumour suppressor gene and each mutational spectrum, the disease-associated median mutability values were calculated using mutability rates derived from Krawczak et al. [1998]. The disease-associated median value was found to be 0.85 for the germline mutations. The highest and lowest disease-associated median values for the mutation rates were noted for somatic mutations in the *STK11* gene (1.7; Supp. Table 3) and for germline mutations in the *TP53* (0.42) gene (values given only for genes with more than three mutations in the corresponding category). We found that shared recurrent and shared non-recurrent mutational spectra were characterized by higher median values of the disease-associated mutability rates (1.42 and 1.01 respectively) whereas somatic non-recurrent, somatic recurrent and control dataset mutations exhibited lower median mutability rates (0.5, 0.5 and 0.4 respectively) as compared to germline mutations (0.85). The finding that the shared mutations (which, by definition, occur in both the germline and the soma) are characterized by higher disease-associated mutability rates is not surprising since mutations that occur with the highest probability are among those most likely to be shared.

We postulated that those mutations which occur both in the germline and the soma, and which are characterised by higher disease-associated mutability rates are disproportionately likely to be drivers of tumour development. Consistent with this postulate, somatic recurrent and non-recurrent mutational spectra are characterized by lower median disease-associated mutability rates as compared to the germline spectrum. However, given that higher disease-associated mutability rates are a characteristic feature of driver mutations, a certain proportion of the somatic mutations, namely those characterised by higher disease-associated mutability rates, may correspond to functionally significant driver mutations.

In assessing the significance of our results, it was appropriate to consider the possibility that somatic mutations might display quite different nearest-neighbour-dependent disease-associated mutability rates from germline mutations. However, since a good correlation was observed between the mutability rates derived from inherited disease data [Krawczak et al., 1998] and the neighbour-dependent mutability rates calculated for the somatic mutations of the 17 tumour-suppressor genes studied here (Pearson's correlation r=0.703, p=6.6×10^{-30}), this *caveat* appears not to be an issue.

Distribution of Grantham scores with respect to tumour suppressor gene mutations

Shared recurrent mutations were found to exhibit the largest median chemical difference value (Grantham scores) between the wild-type and mutated amino acid residues (100) followed by shared non-recurrent mutations and germline mutations (both 93), somatic recurrent (85), somatic non-recurrent (80) and potential mutations (78). Since there was an obvious trend for shared recurrent and non-recurrent mutations to cause the most dramatic chemical changes of the affected codon, we may infer that these types of lesion are also more likely to be driver mutations. However, bearing in mind that the range of theoretically possible values varies between 5 (Leu ↔ Ile) and 215 (Cys ↔ Trp), less elevated median values may simply indicate that a proportion of the mutations in each mutational spectrum are likely to be chemically less dramatic (Grantham scores <100).

Missense mutations occurring within repeats and runs of identical nucleotides

A number of studies have noted that single base-pair substitutions associated with inherited disease occur disproportionately either within, or in close proximity to, repetitive sequences [Jego et al., 1993; Greenblatt et al., 1996; Tappino et al., 2009; Thomas et al., 2010; Leclercq et al., 2010]. Hence, we wished to assess whether either germline or somatic mutations occurred disproportionately either within, or in the vicinity (see *Mutation descriptors*) of, direct, inverted and symmetric repeats or mononucleotide runs in the 17 tumour suppressor genes under study (Table 3, Supplementary Tables 4-6).

On average, direct repeats of length ≥8 bp were found to cover 5.6% of the cDNA lengths of the 17 tumour suppressor genes, the coverage varying between 2.5% (*BRCA2*) and 17% (*PTEN*) of the respective gene sequences. The corresponding proportion of the cDNA lengths for inverted repeats ≥8 bp was 8.5%, with proportions varying between *PTCH1* (4.5%) and *RB1* (15.7%) while symmetric elements ≥8 bp were found to encompass 25% of the cDNA lengths (varying between 15.5% for *APC* and 44% for *PTEN*).

On average, mononucleotide runs ≥4 bp spanned 19.9% of the cDNA lengths, varying between 9.5% (*VHL*) and 29% (*TP53*). Approximately 24% of non-recurrent somatic and 20% of germline missense mutations were found in mononucleotide runs; these proportions were significantly higher than noted for shared non-recurrent missense mutations (4.9%, $p \leq 1.6 \times 10^{-4}$). A greater proportion of non-recurrent somatic missense mutations was found in direct repeats (7%) as compared to recurrent somatic missense mutations (2%, $p=8.8 \times 10^{-7}$), germline missense

(4%, p=0.028) and potential missense mutations (3.7%, $p=8.1\times10^{-7}$). This result may reflect the disproportionate number of CpG/CpHpG mutations among shared and recurrent somatic missense mutations. Further, for all mutational spectra examined (with the exception of the shared mutations), missense mutations were preferentially found in association with inverted and symmetric repeats as compared to the control dataset of mutations (p<0.05). However, no statistically significant differences were found between mutational spectra.

No correlation was observed between the number of mutations located within repeats and the fractional length of the cDNA covered by repeats, indicating that not every repeat sequence is mutation-prone. However, a strong correlation between the fractional length of the cDNA covered by repeats and cDNA length of genes (r >0.87 and $p<10^{-6}$) served to demonstrate that repeat density per unit length was approximately the same for all tumour suppressor genes studied.

Towards a classification of somatic and germline missense mutations

All observed mutations within each mutational spectrum were re-categorized (Supp. Table 7) with respect to the location of mutations within CpG/CpHpG oligonucleotides, within different types of repeat/mononucleotide runs, within both CpG/CpHpG oligonucleotides and repeats. 4×2 contingency tables were then used to measure the strength of the pairwise associations between the various mutational distributions presented in Supp. Table 7, the significance of the associations being assessed by means of a Chi-square test. Significant (p<0.002) pairwise differences were noted between somatic and germline, somatic and shared, and between germline and shared mutational spectra (p<0.002) with respect to the features listed above and each of four types of repeat, indicating that these features have great discriminant potential.

All somatic, germline, shared non-recurrent, recurrent somatic and shared recurrent missense mutations (each described by a combination of different features (i.e. degree of evolutionary conservation, non-disease- and disease-associated mutability rates, Grantham score, CpG/CpHpG location, occurrence within repeat/mononucleotide run) were then used to train a Naïve Bayes Tree classifier. 63.1% of somatic, germline, shared, recurrent somatic and shared recurrent mutations were correctly classified [the area under the Receiver Operating Characteristic (ROC) curve being 0.869, indicating a reasonably good classification] implying that the mutation groupings differ with respect to the different features in a consistent fashion. The complete Naïve Bayes Tree classifier is depicted in Supp. Figure 1.

An additional non-overlapping dataset of 568 missense somatic mutations, identified in the 17 tumour suppressor genes under study, were extracted from a collection of 2,488 mutations identified as being probable driver mutations [Carter et al., 2009]. Features such as the degree of evolutionary conservation, Grantham score, mutability rates, CpG/CpHpG location, occurrence within repeats/mononucleotide runs were again determined for each of these mutations. Employing our classifier, 7% and 10% respectively of these 568 mutations were found to possess features consistent with their being shared recurrent and shared non-recurrent mutations. In addition, 32% of these probable driver mutations were found to bear features characteristic of recurrent somatic mutations (i.e. mutations documented in different tumours). A further 25% of the probable (somatic) driver mutations were classified as possessing features characteristic of germline mutations and hence could conceivably be treated as shared mutations missing from the original training dataset. The remaining 25% of mutations were classified as non-recurrent somatic mutations. Using this classifier, which is based on a very modest number (6) of predictive features, to analyse an independent dataset of probable driver mutations, we were able to predict that ~50% of these somatic missense mutations exhibited features specific to either shared or recurrent mutations, indicating that a disproportionate number of such lesions are likely to be drivers of tumorigenesis. This percentage is certainly lower (79%) than that obtained by Carter et al., [2009] through the application of a Random Forest Classifier based on 500 trees and

>50 predictive features (using an out-of-the-bag error estimate similar to the cross-validation procedure) to the set of putative 2,488 driver mutations. However, based on the results of this study, we may conclude that, in general, the mutational spectrum of driver mutations is likely to contain a disproportionate number of somatic mutations that have germline counterparts (~17%) whilst an additional 32% of the driver mutations are likely to occur recurrently in the soma.

Truncating vs non-truncating mutations in the germline and soma

Somatic mutational spectra from the *BRCA2, CDKN2A, STK11*, *TP53* and *TSC1* genes were characterized by the predominance of non-truncating (i.e. missense) lesions over truncating lesions (i.e. nonsense mutations, frameshift micro-deletions, micro-insertions and indels) when nonsense mutations [reported in Mort et al. (2008)] and micro-indels (excluded from previous analyses) were also considered (Supp. Table 8). A similar predominance of non-truncating over truncating lesions was observed for the germline mutational spectra of the *CDKN2A*, *TP53*, *VHL* and *WT1* genes. In general, the ratio of non-truncating to truncating lesions was found to be significantly higher in the soma (0.85) than in the germline (0.30; p-value<2.20E-16). All other mutational spectra were characterized by the predominance of truncating mutations.

Occurrence of micro-deletions and micro-insertions within repeats and runs of identical nucleotides

The mutational spectrum of micro-deletions, combined for all 17 tumour suppressor genes, comprised 55% germline, 43% somatic and 2% shared mutations. The mutational spectrum of micro-insertions was similar to that of micro-deletions and comprised 60% germline, 38% somatic and 2% shared mutations. Approximately 77% somatic, 87% germline and 91% shared micro-deletions and micro-insertions were ≤4 bp in length. Strong (r = ~1) correlations were noted between the distributions of micro-deletions and micro-insertions with respect to the length of the deleted/inserted fragments, both gene-wise and for all genes combined ($r>0.9$, $p<10^{-8}$) for all mutational spectra.

Recent studies have revealed that simple repetitive DNA sequences are not only capable of adopting non-B DNA conformations and are highly mutagenic [Bacolla et al., 2004; Bacolla and Wells, 2004; Chuzhanova et al., 2009]. Indeed, both direct repeats and mononucleotide runs have long been known to be mutation hotspots in the *TP53* gene [Jego et al., 1993; Greenblatt et al., 1996]. The number of micro-lesions occurring in the vicinity (see *Mutation descriptors*) of direct, symmetric and inverted repeats (capable respectively of slipped, triplex and cruciform non-B structure formation), or within mononucleotide runs (which often mediate micro-deletions/micro-insertions) were therefore determined. The number of mutations found in the vicinity of all three types of repeat, and within mononucleotide runs, are given in Tables 3 and Supp. Tables 4-6.

The highest proportion of mutations in mononucleotide runs was found for the shared (39%), germline (30%) and somatic (25%) mutational spectra. Significant differences were observed between shared and germline (p=0.0002), somatic and shared (p=0.045), and between all mutational spectra and potential mutations (p<0.0001) with respect to their occurrence within mononucleotide runs, confirming that these simple repeats constitute an important hotspot for micro-deletions and micro-insertions in both the soma and the germline. The preponderance of such mutations in mononucleotide runs is unsurprising in the context of the shared mutations since all mutations that occur with high frequency within mutation hotspots are more likely to be shared between the germline and the soma (as previously noted for CpG and CpHpG mutations). No other types of repeat were disproportionately associated (after correction for multiple testing) with micro-deletions and micro-insertions.

Hotspots in somatic and germline mutational spectra

For the purposes of this analysis, a mutation hotspot was defined as a stretch of DNA of length ≤20 bp where four or more independent mutational events have been reported and a significant degree ($p \leq 0.05$) of clustering of these mutations was evident for a given stretch of DNA. In this definition of a hotspot, each recurrent mutation was considered only once. The order statistics, r-scans, as described by Karlin and Macken [1991] and applied in Bacolla et al. [2006], were used to detect significant clustering of mutations by comparison with a Poisson distribution of mutations along the gene sequence. Overlapping hotspot regions were considered as a single hotspot.

The only mutational hotspot for somatic missense mutations was observed in the *PTEN* gene and comprised 18 mutations in the region between nucleotide positions 269 and 286. Several germline mutational hotspots were however detected for missense mutational spectra in the *ATM*, *BRCA1*, *BRCA2*, *NF1*, *PTEN*, *RB1*, *STK11*, *TP53* and *WT1* genes (Table 4). Several somatic mutational hotspots were found for micro-deletions/micro-insertions in the *APC* gene, the largest of which contained 33 mutations (positions 4303-4398) and forms part of a previously reported mutation cluster region [Miyoshi et al., 1992]. Hotspots identified in different mutational spectra were however unique to that spectrum. The only overlap noted between mutational hotspots identified in germline and somatic micro-deletion/micro-insertion mutational spectra was observed for the *APC* gene (the overlapping region comprising nucleotide positions 3919-3933). This micro-deletion/micro-insertion hotspot also includes codon 1309 (cDNA positions 3925-3927) found to be frequently mutated in Greek and French patients with familial adenomatous polyposis [Fostira et al. 2010; Lagarde et al. 2010].

Inspection of hotspot regions revealed that they are rich in repetitive elements, runs of identical nucleotides and CpG/CpHpG oligonucleotides, offering immediate explanations for the elevated mutability.

Germline and somatic mutations located within specific hotspot motifs

The cDNA sequences of 17 tumour suppressor genes were screened for the presence of nine specific motifs (and their complements) previously reported as being hotspots for mutation. These motifs included the putative somatic (cancer) mutation hotspot, WKVNRRRNVWK [the 'THEMIS motif'; Makridakis et al., 2009], the RGYW motif that correlates with the DNA polymerase eta error spectrum [Rogozin et al., 2001] and several so-called 'super hotspot' motifs originally found in germline micro-insertions and micro-deletions [Ball et al., 2005] and indels [Chuzhanova et al., 2003]. For the purposes of this analysis, the shared mutations were added to both the germline and somatic mutational spectra. Both germline and somatic micro-deletions and micro-insertions were found to be significantly overrepresented ($p \leq 0.002$) in the 'indel super hotspot' motif GTAAGT and its complement. Somatic micro-deletions and micro-insertions were also significantly overrepresented ($p=0.009$) with respect to the micro-deletion/micro-insertion super hotspot AAATCT and its complement. The number of germline (but not somatic) micro-deletions/micro-insertions in the THEMIS motif were significantly overrepresented ($p=0.003$) as compared to the controls. No significant difference was however observed in the number of missense mutations occurring in any motifs analysed.

Conclusions

A number of important conclusions may be drawn from the results reported here. Firstly, it would appear that missense mutations that are found both in the soma and the germline (shared mutations) are disproportionately more likely to exert profound effects on tumour development and/or progression (i.e. more likely to be driver mutations) than exclusively somatic non-recurrent missense mutations (at least for the *TP53* and *CDKN2A* genes whose mutations contributed the bulk of the documented shared mutations in our tumour suppressor gene mutation dataset). Shared mutations also occur preferentially in CpG/CpHpG oligonucleotides and are characterised by higher mutability rates (both non-disease- and disease-associated). Further, we found that shared mutations tend to occur in those codons that have been more highly conserved evolutionarily, and are associated with more dramatic chemical differences between the substituted (wild-type) and substituting amino acids. Taken together, it would thus appear that shared mutations are influenced to a greater extent by the local nucleotide sequence context than either germline or somatic non-recurrent missense mutations. Since this implies that shared mutations (the mutation category most likely to harbour driver mutations) have a tendency to arise through the action of similar endogenous mutational mechanisms, we may infer that endogenous mechanisms of mutagenesis exert a disproportionate effect on tumorigenesis.

In an analysis of an unrelated dataset, we demonstrated that 17% of somatic missense mutations previously identified as being probable drivers [Carter et al., 2009] were found to possess the same features as shared (both recurrent and non-recurrent) mutations. A further 32% of these probable driver mutations shared the features expected of recurrent somatic mutations. Thus, we may conclude that ~50% of these somatic missense mutations possess features consistent with their being either shared or recurrent, suggesting that a disproportionate number of such lesions are likely to be drivers of tumorigenesis.

A sizeable proportion of shared (39%) and germline (30%) micro-lesions were found to be located in runs of identical nucleotides ≥4 bp, making mononucleotide runs a hotspot for micro-deletion and micro-insertions. The most likely underlying causative mechanism for these mutations is slipped mispairing at DNA replication mediating duplications and 'de-duplications' [Kondrashov & Rogozin, 2004]. With regard to missense mutations, CpG and CpHpG oligonucleotides were found to be hotspots for shared recurrent and shared non-recurrent missense mutations; 34% (10%) and 21% (9%) of respective mutations were found in CpG (CpHpG) oligonucleotides. Further, 12% of the 568 probable driver mutations [derived from Carter et al., 2009] were found to occur in CpG/CpHpG oligonucleotides. 41% of probable driver mutations were found in repeats that were capable of non-B DNA structure formation (cf. 23% for potential mutations). Several hotspot regions were found in the mutational spectra of various genes; one of these, in the *APC* gene, was a hotspot for both somatic and germline micro-deletions/micro-insertions and corresponded to a previously recognized mutation hotspot [Miyoshi et al., 1992].

Taken together, the results and analysis presented herein strongly suggest that algorithms that attempt to predict the relative impact of tumour-associated micro-lesions on (tumour suppressor) gene and protein function [Tavtigian et al., 2008; Couch et al., 2008; Thusberg and Vihinen, 2009], should take into consideration the origin (i.e. somatic, germline or shared) of the mutations, their sequence context and repetitivity, as well as their frequency of occurrence.

Acknowledgements

We are most grateful to Eamonn Maher (Birmingham, UK), Gareth Evans (Manchester, UK) and the late Michael Baser for their provision of somatic mutation data. We are also very grateful to Larry Brody (NIH, Bethesda, MD, USA), Shiu-Ru Lin (Kaohsiung Medical University, Taiwan) and Tone Lovig (University of Oslo, Norway) for providing us with further information on specific somatic mutations and Rachel Karchin (Johns Hopkins University, Baltimore, USA) for making available her dataset of probable driver mutations.

References

Ali IU, Schriml LM, Dean M. 1999. Mutational spectra of *PTEN/MMAC1* gene: a tumor suppressor with lipid phosphatase activity. J Natl Cancer Inst 91:1922-1932.

Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova NA, Abeysinghe SS, O'Connell CD, Cooper DN, Wells RD. 2004. Breakpoints of gross deletions coincide with non-B DNA conformations. Proc Natl Acad Sci USA 101:14162-14167.

Bacolla A, Wells RD. 2004. Non-B DNA conformations, genomic rearrangements, and human disease. J Biol Chem 279:47411-47414.

Bacolla A, Collins JR, Gold B, Chuzhanova N, Yi M, Stephens RM, Stefanov S, Olsh A, Jakupciak JP, Dean M, Lempicki RA, Cooper DN, Wells RD. 2006. Long homopurine•homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. Nucleic Acids Res 34: 2663-2675.

Ball EV, Stenson PD, Krawczak M, Cooper DN, Chuzhanova NA. 2005. Micro-deletions and micro-insertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. Human Mutat 26:205-213.

Baser ME, Contributors to the International NF2 Mutation Database. Hum. Mutat. 27:297-306.

Boland CR, Ricciardiello L. 1999. How many mutations does it take to make a tumor? Proc Natl Acad Sci USA 96:14675-14677.

Buard J, Collick A, Brown J, Jeffreys AJ. 2000. Somatic versus germline mutation processes at minisatellite CEB1 (D2S90) in humans and transgenic mice. Genomics 65**:**95-103.

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. 2009. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 69:6660-6667.

Chuzhanova NA, Anassis EJ, Ball E, Krawczak M, Cooper DN. 2003. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat 21:28-44.

Chuzhanova N, Chen JM, Bacolla A, Patrinos GP, Férec C, Wells RD, Cooper DN. 2009. Gene conversion causing human inherited disease: the evidence for involvement of recombination-associated motifs and non-B DNA-forming sequences in DNA breakage. Hum Mutat 30:1189-1198.

Cole DN, Carlson JA, Wilson VL. 2008. Human germline and somatic cells have similar *TP5*3 and Kirsten-RAS gene single base mutation frequencies. Environ. Mol. Mutagen. 2008 49:417-425.

Cooper DN, Mort M, Stenson PD, Ball EV, Chuzhanova NA. 2010. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides as well as in CpG dinucleotides. Hum Genomics 4:406-410..

Couch FJ, Rasmussen LJ, Hofstra R, Monteiro AN, Greenblatt MS, de Wind N; IARC Unclassified Genetic Variants Working Group. 2008. Assessment of functional effects of unclassified genetic variants. Hum Mutat 29:1314-1326.

Dallosso AR, Jones S, Azzopardi D, Moskvina V, Al-Tassan N, Williams GT, Idziaszczyk S, Davies DR, Milewski P, Williams S, Beynon J, Sampson JR, Cheadle JP. 2009. The *APC* variant p.Glu1317Gln predisposes to colorectal adenomas by a novel mechanism of relaxing the target for tumorigenic somatic *APC* mutations. Hum Mutat 30:1412-1418.

Fostira F, Thodi G, Sandaltzopoulos R, Fountzilas G, Yannoukakos D. 2010. Mutational spectrum of APC and genotype-phenotype correlations in Greek FAP patients. BMC Cancer 10:389.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. Nat Rev Cancer 4:177-183.

Gallou C, Joly D, Mejean A, Staroz F, Martin N, Tarlet G, Orfanelli MT, Bouvier R, Droz D, Chretien Y, Maréchal JM, Richard S, Junien C, Béroud C. 1999. Mutations of the *VHL* gene in sporadic renal cell carcinoma: definition of a risk factor for VHL patients to develop an RCC. Hum Mutat 13:464-475.

Giovannone B, Sabbadini G, Di Maio L, Calabrese O, Castaldo I, Frontali M, Novelleto A, Squitieri F. 1997. Analysis of (CAG)n size heterogeneity in somatic and sperm cell DNA from intermediate and expanded Huntington disease gene carriers. Hum Mutat 10:458-464.

Glazko GV, Koonin EV, Rogozin IB. 2004. Mutation hotspots in the p53 gene in tumors of different origin: correlation with evolutionary conservation and signs of positive selection. Biochim Biophys Acta 1679**:**95-106.

Goode EL, Ulrich CM, Potter JD. 2002. Polymorphisms in DNA repair genes and associations with cancer risk. Cancer Epidemiol Biomarkers Prev 11:1513-1530.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Greenblatt MS, Grollman AP, Harris CC. 1996. Deletions and insertions in the p53 tumor suppressor gene in human cancers: confirmation of the DNA polymerase slippage/misalignment model. Cancer Res 56:2130-2136.

Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP. 2003. Detailed computational study of p53 and p16: using evolutionary sequence analysis and disease-associated mutations to predict the functional consequences of allelic variants. Oncogene 22:1150-1163.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR. 2007. Patterns of somatic mutation in human cancer genomes. Nature 446:153-158.

Groves C, Lamlum H, Crabtree M, Williamson J, Taylor C, Bass S, Cuthbert-Heavens D, Hodgson S, Phillips R, Tomlinson I. 2002. Mutation cluster region, association between germline and somatic mutations and genotype-phenotype correlation in upper gastrointestinal familial adenomatous polyposis. Am J Pathol 160:2055-2061.

Haber D, Harlow E. 1997. Tumour-suppressor genes: evolving definitions in the genomic age. Nat Genet 16**:**320-322.

Hess ST, Blake JD, Blake RD. 1994. Wide variations in neighbor-dependent substitution rates. J Mol Biol 236:1022-1033.

Jego N, Thomas G, Hamelin R. 1993. Short direct repeats flanking deletions, and duplicating insertions in p53 gene in human cancers. Oncogene 8:209-213.

Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S, Davis DP, Carlton VE, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, de Sauvage FJ, Faham M, Seshagiri S. 2010. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466:869-873.

Karlin S, Macken C. 1991. Some statistical problems in the assessment of inhomogenesis of DNA sequence data. J Am Statist Assoc 86:27–35.

Knudson AG. 2001. Two genetic hits (more or less) to cancer. Nat Rev Cancer 1**:**157-162.

Kohavi R. 1996. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hHybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, pp 202-207**.**

Kolomietz E, Meyn MS, Pandita A, Squire JA. 2002. The role of *Alu* repeat clusters as mediators of recurrent chromosomal aberrations in tumors. Genes Chrom Cancer 35:97-112.

Kondrashov AS, Rogozin IB. 2004. Context of deletions and insertions in human coding sequences. Hum Mutat 23:177–185.

Kotsiantis S, Kanellopoulos D, Pintelas P. 2006. Handling imbalanced datasets: a review. GESTS International Transactions on Computer Science and Engineering 30:25-36.

Krawczak M, Smith-Sorensen B, Schmidtke J, Kakkar VV, Cooper DN, Hovig E. 1995. 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 5:48-57.

Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63:474-488.

Lagarde A, Rouleau E, Ferrari A, Noguchi T, Qiu J, Briaux A, Bourdon V, Rémy V, Gaildrat P, Adélaïde J, Birnbaum D, Lidereau R, Sobol H, Olschwang S. 2010. Germline *APC* mutation spectrum derived from 863 genomic variations identified through a 15-year medical genetics service to French patients with FAP. J Med Genet [Epub ahead of print] PubMed PMID: 20685668.

Lamlum H, Ilyas M, Rowan A, Clark S, Johnson V, Bell J, Frayling I, Efstathiou J, Pack K, Payne S, Roylance R, Gorman P, Sheer D, Neale K, Phillips R, Talbot I, Bodmer W, Tomlinson I. 1999. The type of somatic mutation at *APC* in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's 'two-hit' hypothesis. Nat Med 5:1071-1075.

Latchford A Volikos E, Johnson V, Rogers P, Suraweera N, Tomlinson I, Phillips R, Silver A. 2007. *APC* mutations in FAP-associated desmoid tumours are non-random but not 'just right'. Hum Mol Genet 16:78-82.

Leclercq S, Rivals E, Jarne P. 2010. DNA slippage occurs at microsatellite loci without minimal threshold length in human: a comparative genomic approach. Genome Biol Evol 2:325-335.

Leeflang EP, Tavare S, Marjoram P, Neal CO, Srinidhi J, MacFarlane H, MacDonald ME, Gusella JF, de Young M, Wexler NS, Arnheim N. 1999. Analysis of germline mutation spectra at the Huntington's disease locus supports a mitotic mutation mechanism. Hum Mol Genet 8:173-183.

Lengauer C, Kinzler KW, Vogelstein B. 1998. Genetic instabilities in human cancers. Nature 396:643-649.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 19:315-322.

Lobo GP, Waite KA, Planchon SM, Romigh T, Nassif NT, Eng C. 2009. Germline and somatic cancer-associated mutations in the ATP-binding motifs of PTEN influence its subcellular localization and tumor suppressive function. Hum. Mol. Genet. 18:2851-2862.

Loeb LA, Harris CC. 2008. Advances in chemical carcinogenesis: a historical review and prospective. Cancer Res 68: 6863-6872.

Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci USA 95:9407-9412.

Makridakis NM, Caldas Ferraz LF, Reichardt JK. 2009. Genomic analysis of cancer tissue reveals that somatic mutations commonly occur in a specific motif. Hum Mutat 30:39-48.

Marshall B, Isidro G, Carvalhas R, Boavida M. 1997. Germline versus somatic mutations of the *APC* gene: evidence for mechanistic differences. Hum Mutat 9**:**286-288.

Martorell L, Monckton DG, Gamez J, Baiget M. 2000. Complex patterns of male germline instability and somatic mosaicism in myotonic dystrophy type 1. Eur J Hum Genet 8: 423-430.

Miyoshi Y, Nagase H, Ando H, Horii A, Ichii S, Nakatsuru S, Aoki T, Miki Y, Mori T, Nakamura Y. 1992. Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. Hum Mol Genet 1:229-33.

Mort M, Ivanov D, Cooper DN, Chuzhanova NA. 2008. A meta-analysis of nonsense mutations causing human genetic disease. Hum Mutat 29:1037-1047.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3**:**418-426.

Oldenburg J, Rost S, El-Maarri O, Leuer M, Olek K, Muller CR, Schwaab R. 2000. *De novo* factor VIII gene intron 22 inversion in a female carrier presents as a somatic mosaicism. Blood 96: 2905-2906.

Olshen AB, Jain AN. 2002. Deriving quantitative conclusions from microarray expression data. Bioinformatics 18:961-970.

Parmigiani G, Boca S, Lin J, Kinzler KW, Velculescu V, Vogelstein B. 2009. Design and analysis issues in genome-wide somatic mutation studies of cancer. Genomics 93:17-21.

Pollard LM, Sharma R, Gomez M, Shah S, Delatycki MB, Pianese L, Monticelli A, Keats BJ, Bidichandani SI. 2004. Replication-mediated instability of the GAA triplet repeat mutation in Friedreich ataxia. Nucleic Acids Res 32:5962-5971.

Richter S, Vandezande K, Chen N, Zhang K, Sutherland J, Anderson J, Han L, Panton R, Branco P, Gallie B. 2003. Sensitive and efficient detection of *RB1* gene mutations enhances care for families with retinoblastoma. Am J Hum Genet 72:253-269.

Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. 2001. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nat Immunol 2:530-536.

Rubin AF, Green P. 2009. Mutation patterns in cancer genomes. Proc Natl Acad Sci USA 106: 21766-21770.

Sharma R, Bhatti S, Gomez M, Clark RM, Murray C, Ashizawa T, Bidichandani SI. 2002. The GAA triplet-repeat sequence in Friedreich ataxia shows a high level of somatic instability *in vivo*, with a significant predilection for large contractions. Hum Mol Genet 11:2175-2187.

Sherr CJ. 2004. Principles of tumor suppression. Cell 116:235-246.

Schmutte C, Jones PA. 1998. Involvement of DNA methylation in human carcinogenesis. Biol Chem 379:377-388.

Shanks ME, May CA, Dubrova YE, Balaresque P, Rosser ZH, Adams SM, Jobling MA. 2008. Complex germline and somatic mutation processes at a haploid human minisatellite shown by single-molecule analysis. Mutat. Res. 648:46-53.

Simpson AJ. 2009. Sequence-based advances in the definition of cancer-associated gene mutations. Curr Opin Oncol 21:47-52.

Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson

D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314:268-274.

Stead JD, Jeffreys AJ. 2000. Allele diversity and germline mutation at the insulin minisatellite. Hum Mol Genet 9:713-723.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The Human Gene Mutation Database: 2008 update. *Genome Med.* 1:13.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. Nature 458:719-724.

Talavera D, Taylor MS, Thornton JM. 2010. The (non)malignancy of cancer amino acidic substitutions. Proteins 78:518-529.

Tappino B, Chuzhanova NA, Regis S, Dardis A, Corsolini F, Stroppiano M, Tonoli E, Beccari T, Rosano C, Mucha J, Blanco M, Szlago M, Di Rocco M, Cooper DN, Filocamo M. 2009. Molecular characterization of 22 novel UDP-N-acetylglucosamine-1-phosphate transferase alpha- and beta-subunit (GNPTAB) gene mutations causing mucolipidosis types IIalpha/beta and IIIalpha/beta in 46 patients. Hum Mutat 30:E956-973.

Tartaglia M, Martinelli S, Stella L, Bocchinfuso G, Flex E, Cordeddu V, Zampino G, van der Burgt I, Palleschi A, Petrucci TC, Sorcini M, Schoch C, Foà R, Emanuel PD, Gelb BD. 2006. Diversity and functional consequences of germline and somatic *PTPN11* mutations in human disease. Am J Hum Genet 78:279-290.

Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB; IARC Unclassified Genetic Variants Working Group. 2008. *In silico* analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 29:1327-1336.

Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B, McKay-Chopin S, Thomas A, Vallée MP, Voegele C, Webb PM, Whiteman DC;

Australian Cancer Study; Breast Cancer Family Registries (BCFR); Kathleen Cuningham Foundation Consortium for Research into Familial Aspects of Breast Cancer (kConFab), Sangrajrang S, Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G. 2009. Am J Hum Genet 85:427-446.

Thomas L, Kluwe L, Chuzhanova N, Mautner V, Upadhyaya M. 2010. Analysis of *NF1* somatic mutations in cutaneous neurofibromas from patients with high tumor burden. Neurogenetics 11:391-400.

Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30:703-714.

Tornaletti S, Pfeifer GP. 1995. Complete and tissue-independent methylation of CpG sites in the *p53* gene: implications for mutations in human cancers. Oncogene 10:1493-1499.

Upadhyaya M, Han S, Consoli C, Majounie E, Horan M, Thomas NS, Potts C, Griffiths S, Ruggieri M, von Deimling A, Cooper DN. 2004. Characterization of the somatic mutational spectrum of the neurofibromatosis type 1 (*NF1*) gene in neurofibromatosis patients with benign and malignant tumors. Hum Mutat 23:134-136.

Upadhyaya M, Kluwe L, Spurlock G, Monem B, Majounie E, Mantripragada K, Ruggieri M, Chuzhanova N, Evans DG, Ferner R, Thomas N, Guha A, Mautner V. 2008. Germline and somatic *NF1* gene mutation spectrum in *NF1*-associated malignant peripheral nerve sheath tumors (MPNSTs). Hum Mutat 29:74-82.

Vogelstein B, Kinzler KW. 2004. Cancer genes and the pathways they control. Nat Med 10:789-799.

Walker DR, Bond JP, Tarone RE, Harris CC, Makalowski W, Boguski MS. Greenblatt MS. 1999. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. Oncogene 18:211-218.

Witten IH, Frank E. 2005. *Data mining: practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, San Francisco, pp. 365-483.

Table 1. Summary of mutational spectra in the 17 tumour suppressor genes studied

Gene symbol	Gene length (in bp)	Number of observed missense mutations						Number of observed micro-deletions and micro-insertions			
		somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	Total	somatic non-recurrent	germline	shared	Total
APC	8532	34	25	1	4	0	**64**	181	399	15	**595**
ATM	9171	10	81	0	1	0	**92**	5	157	0	**162**
BRCA1	5592	5	172	0	0	1	**178**	9	338	5	**352**
BRCA2	10257	19	91	2	2	0	**114**	9	332	3	**344**
CDH1	2649	14	19	1	0	0	**34**	15	20	0	**35**
CDKN2A	471	173	35	30	6	1	**245**	100	16	2	**118**
NF1	8457	2	85	0	0	0	**87**	16	323	3	**342**
NF2	1788	20	22	0	3	0	**45**	204	66	5	**275**
PTCH1	4344	13	25	1	0	0	**39**	20	74	0	**94**
PTEN	1212	154	23	11	49	12	**249**	192	41	10	**243**
RB1	2787	22	35	3	1	1	**62**	42	165	4	**211**
STK11	1302	16	28	4	3	0	**51**	4	69	2	**75**
TP53	1182	358	6	9	793	87	**1253**	738	11	12	**761**
TSC1	3495	2	7	0	0	0	**9**	1	78	0	**79**
TSC2	5424	0	93	1	0	1	**95**	5	156	0	**161**
VHL	642	41	98	39	5	9	**192**	209	86	14	**309**
WT1	1350	1	41	0	0	0	**42**	7	12	0	**19**
TOTAL	68655	**884**	**886**	**102**	**867**	**112**	**2851**	**1757**	**2343**	**75**	**4175**

Table 2. Missense mutations found in CpG and CpHpG oligonucleotides for the 17 tumour suppressor genes under study.

Gene symbol	Number of possible missense mutations		Number of observed CpG-located mutations						Number of observed CpHpG-located mutations					
	in CpG	in CpHpG	somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	total	somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	to
APC	177	300	1	6	0	0	0	**7**	0	1	0	0	0	
ATM	157	232	0	12	0	1	0	**13**	0	3	0	0	0	
BRCA1	70	192	0	12	0	0	0	**12**	0	6	0	0	0	
BRCA2	116	310	3	15	2	0	0	**20**	0	2	0	0	0	
CDH1	135	116	1	5	0	0	0	**6**	0	0	0	0	0	
CDKN2A	50	16	35	3	10	0	0	**48**	9	0	0	0	0	
NF1	226	275	0	6	0	0	0	**6**	0	1	3	0	0	
NF2	89	59	1	4	0	1	0	**6**	0	0	0	0	0	
PTCH1	345	213	2	4	0	0	0	**6**	0	0	0	0	0	
PTEN	14	33	1	1	0	5	4	**11**	2	0	0	0	0	
RB1	80	81	4	3	2	0	0	**9**	1	1	1	1	0	
STK11	137	60	4	3	2	2	0	**11**	0	0	0	0	0	
TP53	15	22	8	0	0	35	28	**71**	10	0	0	23	8	**4**
TSC1	147	139	0	1	0	0	0	**1**	0	0	0	0	0	
TSC2	454	238	0	19	1	0	1	**21**	0	7	1	0	1	
VHL	78	24	7	2	4	0	5	**18**	0	2	4	0	2	
WT1	143	70	0	9	0	0	0	**9**	0	4	0	0	0	
TOTAL	**2433**	2380	**67**	**105**	**21**	**44**	**38**	**275**	**22**	**27**	**9**	**24**	**11**	**9**

Table 3. Summary of mutations occurring in runs of identical nucleotides ≥4 bp in the 17 tumour suppressor genes.

Gene symbol	Proportion of gene length covered by runs (%)	Number of missense mutations found in runs						Number of micro-deletions and micro-insertions found in runs			
		somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	Total	somatic non-recurrent	germline	shared	Total
APC	13	5	3	0	2	0	**10**	74	108	6	**188**
ATM	26	2	20	0	0	0	**22**	3	55	0	**58**
BRCA1	16	3	37	0	0	0	**40**	2	120	3	**125**
BRCA2	19	4	27	0	0	0	**31**	5	151	2	**158**
CDH1	18	5	7	0	0	0	**12**	3	11	0	**14**
CDKN2A	17	42	7	2	0	1	**52**	30	5	0	**35**
NF1	24	1	15	0	0	0	**16**	5	74	2	**81**
NF2	19	3	2	0	0	0	**5**	40	8	0	**48**
PTCH1	15	4	7	0	0	0	**11**	6	24	0	**30**
PTEN	32	41	8	1	15	1	**66**	56	12	2	**70**
RB1	37	5	9	1	0	0	**15**	14	54	3	**71**
STK11	24	1	7	0	2	0	**10**	2	23	2	**27**
TP53	29	89	2	1	166	13	**271**	177	3	7	**187**
TSC1	15	1	2	0	0	0	**3**	0	15	0	**15**
TSC2	17	0	10	0	0	0	**10**	0	36	0	**36**
VHL	10	2	3	0	0	0	**5**	15	6	2	**23**
WT1	20	0	12	0	0	0	**12**	2	4	0	**6**
TOTAL	20	**208**	**178**	**5**	**185**	**15**	**591**	**434**	**709**	**29**	**1172**

Table 4. Mutational hotspots found in 17 tumour suppressor genes. The number of mutations within the hotspots is shown in parentheses. Shared overlapping hotspot regions for somatic and germline micro-deletions/insertions is shown in bold**.** Positions are given with respect to the corresponding cDNA sequences.

Gene symbol	Missense mutations		Micro-deletions/insertions	
	somatic	germline	somatic	germline
APC			3856-3882 (9) **3897-3933** (15) 3977-3989 (5) 4117-4140 (7) 4178-4200 (9) 4231-4271 (17) 4303-4398 (33) 4450-4495 (27) 4662-4669 (5)	1484-1492 (4) 1857-1882 (11) 2306-2313 (4) 2789-2821 (13) **3919-3935** (7)
ATM		8479-8494 (6)		
BRCA1		181-191 (6) 5085-5098 (8) 5201-5222 (9) 5236-5258 (8)		
BRCA2		8165-8182 (4)		6196-6203 (4) 6443-6450 (8)
NF1		2329-2352 (6) 2530-2543 (5) 4255-4274 (6)		6788-6798 (5)
PTEN	269-287 (18)	367-371 (4)		
RB1		1960-1970 (5)		202-220 (7)
STK11		526-545 (5)		150-197 (11) 737-757 (6)
TP53		832-848 (11)		
TSC1				2101-2112 (5)
TSC2				2059-2074 (5) 4247-4268 (5)
WT1		1174-1201 (13)		

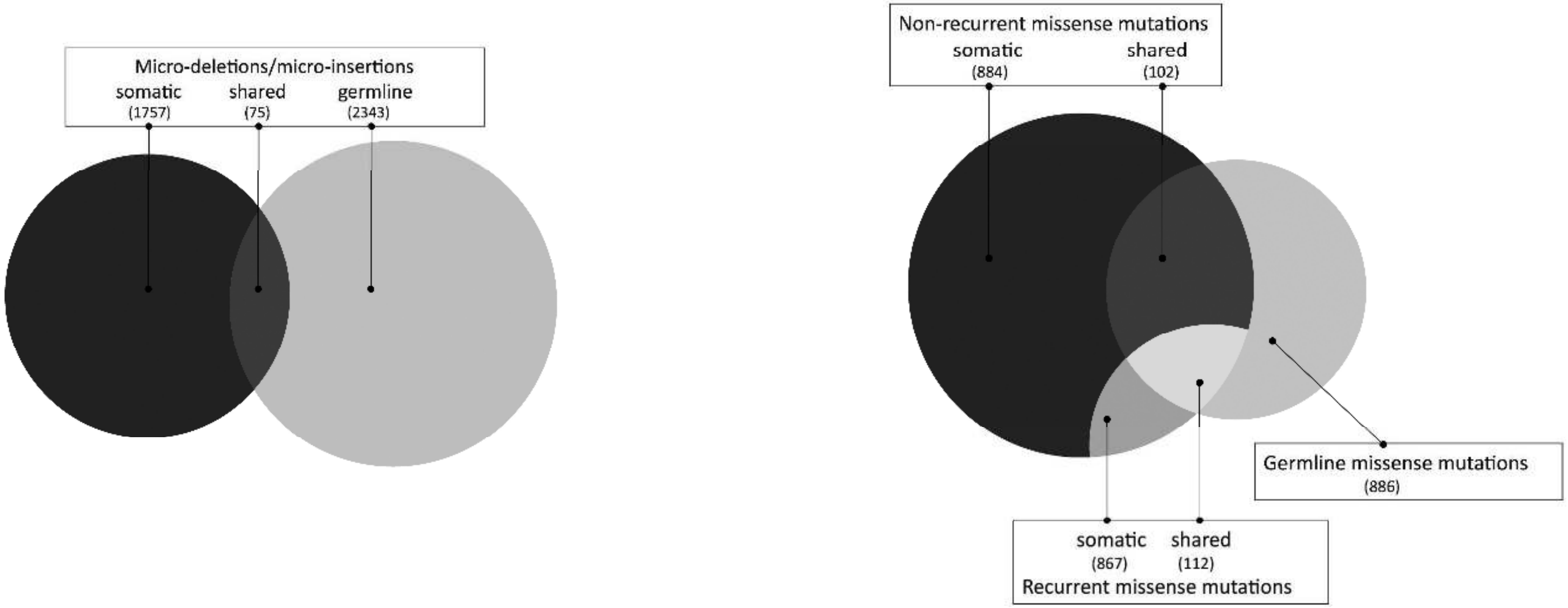
Micro-deletions/micro-insertions
somatic
(1757)
shared
(75)
germline
(2343)
Non-recurrent missense mutations
somatic
(884)
shared
(102)
Germline missense mutations
(886)
somatic
(867)
shared
(112)
Recurrent missense mutations

Figure 1. Diagrammatic representation of the number of various types of mutations analysed in the present study.

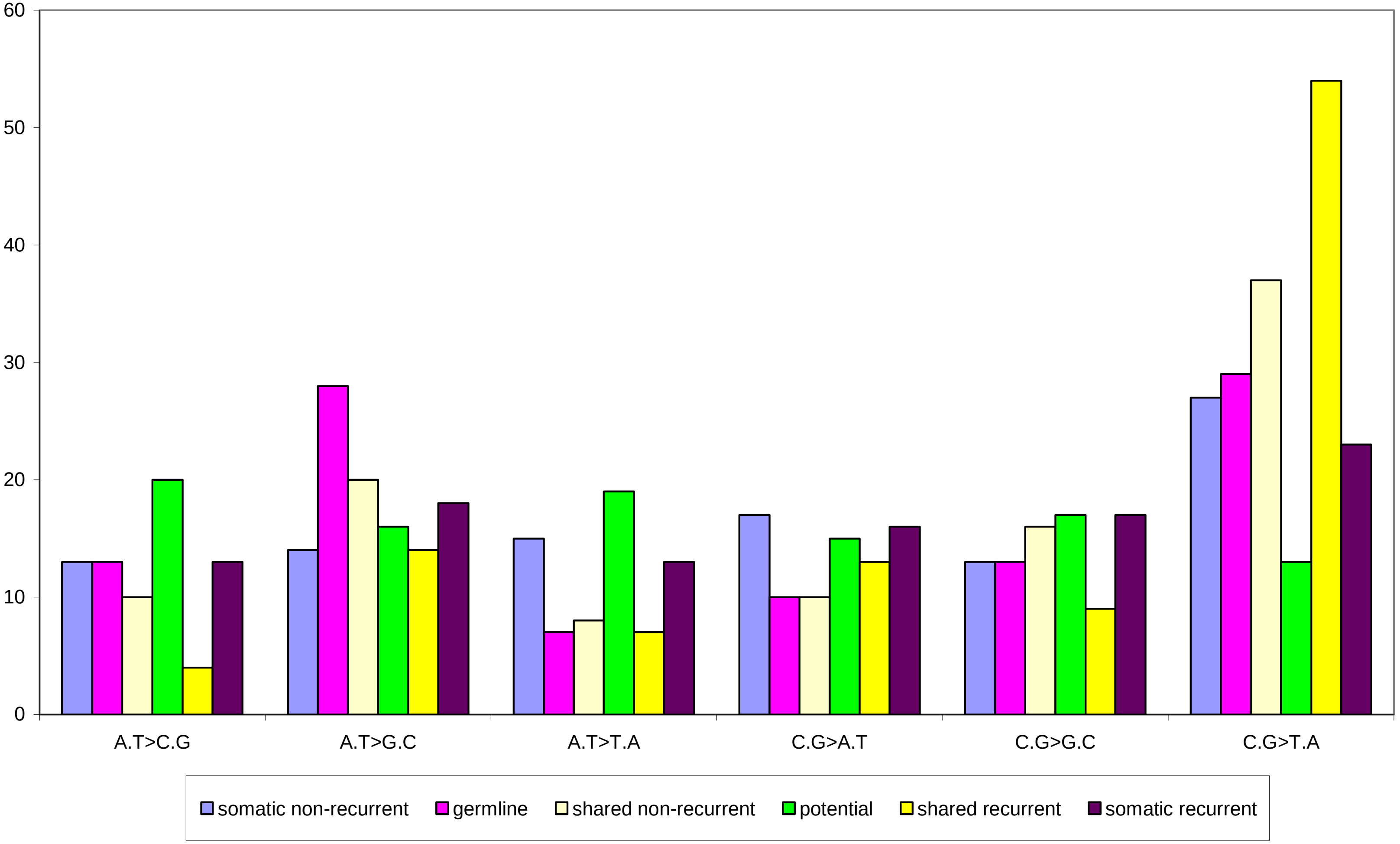

Figure 2. Nucleotide substitution patterns of missense mutations in 17 tumour suppressor genes.

Supplementary Table 1. Tumour suppressor gene orthologues used to estimate the degree of evolutionary conservation of the various gene coding sequences

Gene	*Species*	cDNA sequence identifier	Protein sequence identifier
APC	*Xenopus laevis*	U64442.1	AAB41671.1
	Bos taurus	XM_865627.1	XP_870720.1
	Rattus norvegicus	NM_012499.1	NP_036631.1
	Mus musculus	NM_007462.1	NP_031488.1
ATM	*Gallus gallus*	XM_417160.1	XP_417160.1
	Xenopus laevis	AY668954.1	AAT72929.1
	Rattus norvegicus	XM_236275.3	XP_236275.3
	Sus scrofa	AY587061	AAT01608.1
	Canis familiaris	XM_845871.1	XP_850964.1
	Mus musculus	NM_007499	NP_031525.1
BRCA1	*Gallus gallus*	*NM_204169.1*	*NP_989500.1*
	Xenopus laevis	*AF416868.1*	*AAL13037.1*
	Bos taurus	*NM_178573.1*	*NP_848668.1*
	Rattus norvegicus	*NM_012514.1*	*NP_036646.1*
	Canis familiaris	*NM_001013416.1*	*NP_001013434.1*
	Mus musculus	*NM_009764.2*	*NP_033894.2*
BRCA2	*Gallus gallus*	*NM_204276.1*	*NP_989607.1*
	Danio rerio	*XM_690042.1*	*XP_695134.1*
	Bos taurus	*XM_583622.2*	*XP_583622.2*
	Rattus norvegicus	*NM_031542.1*	*NP_113730.1*
	Canis familiaris	*NM_001006653.4*	*NP_001006654.2*
	Mus musculus	*NM_009765.1*	*NP_033895.1*
CDH1	*Xenopus laevis*	*BC068940.1*	*AAH68940.1*
	Danio rerio	*NM_131820.1*	*NP_571895.1*
	Bos taurus	*NM_001002763.1*	*NP_001002763.1*
	Rattus norvegicus	*NM_031334.1*	*NP_112624.1*
	Canis familiaris	*XM_536807.2*	*XP_536807.2*
	Mus musculus	*NM_009864.1*	*NP_033994.1*
CDKN2A	*Gallus gallus*	*NM_204433.1*	*NP_989764.1*
	Takifugu rubripes	*AJ250231.1*	*CAC12808.1*
	Bos taurus	*XM_868375.1*	*XP_873468.1*
	Rattus norvegicus	*NM_031550.1*	*NP_113738.1*
	Canis familiaris	*XM_538685.2*	*XP_538685.2*
	Mus musculus	*AF044336.1*	*AAC08963.1*
NF1	*Gallus gallus*	XM_415914.1	XP_415914.1
	Takifugu rubripes	AF064564.2	AAD15839.1
	Rattus norvegicus	NM_012609.1	NP_036741.1
	Canis familiaris	XM_537738.2	XP_537738.2
	Mus musculus	NM_010897.1	NP_035027.1
NF2	*Gallus gallus*	NM_204497.2	NP_989828.2
	Danio rerio	NM_212951.1	NP_998116.1
	Bos taurus	XM_611643.2	XP_611643.2
	Rattus norvegicus	XM_341248.2	XP_341249.2
	Canis familiaris	XM_534729.2	XP_534729.2
	Mus musculus	NM_010898.2	NP_035028.2
PTCH1	*Xenopus laevis*	AF302765.1	AAK15463.1
	Gallus gallus	NM_204960.1	NP_990291.1
	Danio rerio	NM_130988.1	NP_571063.1
	Meriones unguiculatus	AB188226.1	BAE78534.1
	Rattus norvegicus	NM_053566.1	NP_446018.1
	Mus musculus	NM_008957.1	NP_032983.1
PTEN	*Xenopus laevis*	AF144732.1	AAD46165.1

	Gallus gallus	XM_421555.1	XP_421555.1
	Bos taurus	XM_613125.2	XP_613125.2
	Canis familiaris	NM_001003192.1	NP_001003192.1
	Rattus norvegicus	NM_031606.1	NP_113794.1
	Mus musculus	NM_008960.2	NP_032986.1
RB1	*Gallus gallus*	NM_204419.1	NP_989750.1
	Rattus norvegicus	XM_344434.2	XP_344435.2
	Canis familiaris	XM_534118.2	XP_534118.2
	Mus musculus	NM_009029.1	NP_033055.1
	Oncorhynchus mykiss	AF102861.1	AAD13390.1
	Notophthalmus viridescens	Y09226.1	CAA70428.1
STK11	*Xenopus laevis*	U24435.1	AAC59904.1
	Danio rerio	NM_001017839.1	NP_001017839.1
	Rattus norvegicus	XM_234900.2	XP_234900.2
	Raja erinacea	AF486831.1	AAL92113.1
	Canis familiaris	XM_542206.2	XP_542206.2
	Mus musculus	NM_011492.1	NP_035622.1
TP53	*Gallus gallus*	NM_205264.1	NP_990595.1
	Danio rerio	NM_131327.1	NP_571402.1
	Bos taurus	NM_174201.2	NP_776626.1
	Rattus norvegicus	NM_030989.1	NP_112251.1
	Canis familiaris	NM_001003210.1	NP_001003210.1
	Mus musculus	NM_011640.1	NP_035770.1
TSC1	*Gallus gallus*	XM_415449.1	XP_415449.1
	Danio rerio	XM_691747.1	XP_696839.1
	Bos taurus	XM_612846.2	XP_612846.2
	Rattus norvegicus	NM_021854.1	NP_068626.1
	Canis familiaris	XM_537808.2	XP_537808.2
	Mus musculus	NM_022887.2	NP_075025.2
TSC2	*Gallus gallus*	XM_414853.1	XP_414853.1
	Takifugu rubripes	AF013614	AAB86682.1
	Bos taurus	XM_581197.2	XP_581197.2
	Rattus norvegicus	NM_012680.2	NP_036812.2
	Canis familiaris	XM_537008.2	XP_537008.2
	Mus musculus	NM_011647.2	NP_035777.2
VHL	*Gallus gallus*	XM_414447.1	XP_414447.1
	Danio rerio	XM_681176.1	XP_686268.1
	Bos taurus	XM_613870.2	XP_613870.2
	Rattus norvegicus	NM_052801.1	NP_434688.1
	Canis familiaris	NM_001008552.1	NP_001008552.1
	Mus musculus	NM_009507.2	NP_033533.1
WT1	*Xenopus laevis*	U42011.1	AAB53152.1
	Gallus gallus	NM_205216.1	NP_990547.1
	Rattus norvegicus	NM_031534.1	NP_113722.1
	Canis familiaris	XM_846479.1	XP_851572.1
	Sus scrofa	NM_001001264.1	NP_001001264.1
	Mus musculus	NM_144783.1	NP_659032.1

SupplementaryTable 2. Differences in distribution of parameters for somatic, germline, shared, somatic recurrent and shared recurrent missense mutations. Observed median and/or mean values are shown in brackets. DAVID: I prefer ‘with respect’ In my view according means that Hess and KR did the study

Parameter	**Observed trend** **(p<0.05)**
Median non-disease associated mutability rate according to Hess et al. [1994]	**shared recurrent >>shared non-recurrent >germline>>somatic~somatic recurrent*** [10.7] [7.9] [7.3] [4.7] [4.7]
Median disease-associated mutability rate according to Krawczak et al. [1998]	**shared recurrent>shared non-recurrent >germline>>somatic~somatic recurrent** [1.42] [1.01] [0.85] [0.53] [0.53]
Mean/median degree of evolutionary conservation	**shared recurrent < shared << somatic** [0.072/0] [0.138/0] [0.265/0.24] **somatic >> germline** [0.265/0.24] [0.18/0]
Mean Grantham score	**germline >somatic recurrent ~somatic non-recurrent** [93] [85] [80] **shared recurrent~shared non-recurrent >> somatic recurrent** [100] [93] [85]
Proportion of CpG-located mutations	**shared recurrent~shared >>germline>>somatic ~somatic recurrent** [0.34] [0.21] [0.12] [0.08] [0.05]
Proportion of CpHpG-located mutations	**shared recurrent~shared >> somatic recurrent** [0.098] [0.082] [0.028]
Proportion of mutations located within or in the vicinity of direct repeats	**somatic>>germline>>recurrent somatic** [0.07] [0.04] [0.02]

Proportion of mutations located within (or in the vicinity of) runs of identical nucleotides	**somatic>>shared** [0.24] [0.05] **germline>>shared** [0.20] [0.05] **somatic>>shared recurrent** [0.24] [0.16] **somatic recurrent>>shared** [0.21] [0.05]

*Inequality **shared>germline>somatic** implies that a significant difference (p<0.05) in the corresponding parameter was observed between each pair of mutational spectra, i.e. shared vs germline, shared vs somatic and germline vs somatic. Symbol '~' denotes the absence of any significant difference between any two mutational spectra with respect to a given parameter. Symbols '>>' or '<<' indicate experiment-wise statistical significance of the observed inequality whereas symbols '<' or '>' indicate gene-wise statistical significance.

Supplementary Table 3. Various parameters of gene-wise somatic and germline missense mutational spectra vs. potential mutational spectra exhibiting either gene-wise (p<0.05) or experiment-wise differences (p<0.05; shaded in light grey) with respect to the parameters measured.

	Non-disease associated mutation rate		**Disease-associated mutation rate**		**Evolutionary conservation rate**		**Grantham score**		**CpG-located missense mutations**		**CpHpG-located missense mutations**	
	Gene symbol	**Median**	**Gene symbol**	**Median**	**Gene symbol**	**Median**	**Gene symbol**	**Median**	**Gene symbol**	**%**	**Gene symbol**	**%**
Somatic mutations			*STK11*	1.66					*STK11*	25		
			PTCH1	1.06								
	APC	8.4	*CDKN2A*	1.01	*CDKN2A*	0.38			*CDKN2A*	20	*CDKN2A*	5.2
	CDKN2A	7.9	*APC*	0.83								
	PTEN	5.6	*PTEN*	0.53								
	TP53	4.6	*TP53*	0.5	*TP53*	0.17			*RB1*	18	*TP53*	2.8
					VHL	0.14			*BRCA2*	16		
									PTCH1	15		
for all 17 genes combined	somatic	4.7	somatic	0.53	somatic	0	somatic	78	somatic	8	somatic	2.5
	control	4.1	control	0.4	control	0.2	control	74	control	2	control	2
	germline	7.2	germline	0.85	germline	0	germline	94	germline	12	germline	3
Germline mutations												
	TSC2	7.2			*TSC2*	0			*BRCA1*	7	*BRCA1*	3.6
	NF1	7.3					*NF1*	98				
	RB1	7.6							*NF1*	7		
	ATM	7.9	*ATM*	0.79	*ATM*	0	*ATM*	98	*ATM*	15	*ATM*	3.8
	BRCA1	7.9	*BRCA1*	0.81	*VHL*	0	*VHL*	99	*BRCA1*	16		
	BRCA2	8.7	*BRCA2*	0.81					*NF1*	18		
			PTEN	0.92							*TSC2*	8.1
			RB1	0.99							*WT1*	10.8
			NF1	1.03								
			TSC2	1.03								

	WT1	10.1	*WT1*	1.22	*WT1*	0			*TSC2*	21		
			CDH1	1.27	*BRCA1*	0.14			*APC*	24		
					CDKN2A	0.29			*CDH1*	26		

Supplementary Table 4. Summary of mutations occurring in direct repeats of length ≥8 bp in the 17 tumour suppressor genes.

Gene symbol	Proportion of gene length covered by repeats (%)	Number of missense mutations found in repeats						Number of micro-deletions and micro-insertions found in repeats			
		somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	Total	somatic non-recurrent	germline	shared	Total
APC	4	3	0	0	0	0	**3**	17	21	1	**17**
ATM	7	2	0	0	0	0	**2**	0	11	0	**0**
BRCA1	5	0	9	0	0	0	**9**	1	8	0	**1**
BRCA2	2	0	0	0	0	0	**0**	1	12	0	**1**
CDH1	3	0	0	0	0	0	**0**	0	1	0	**0**
CDKN2A	17	25	8	3	0	0	**36**	28	2	0	**28**
NF1	7	0	2	0	0	0	**2**	0	15	0	**0**
NF2	3	0	0	0	0	0	**0**	1	1	0	**1**
PTCH1	3	0	0	0	0	0	**0**	0	0	0	**0**
PTEN	17	7	0	0	4	2	**13**	20	5	1	**20**
RB1	12	0	1	0	0	0	**1**	2	12	0	**2**
STK11	10	0	3	1	0	0	**4**	0	6	0	**0**
TP53	14	24	1	0	13	2	**40**	21	0	0	**21**
TSC1	5	0	1	0	0	0	**1**	0	4	0	**0**
TSC2	5	0	10	1	0	0	**11**	0	6	0	**0**
VHL	6	0	1	0	0	0	**1**	0	1	0	**0**
WT1	7	1	0	0	0	0	**1**	0	0	0	**0**
TOTAL	**6**	**62**	**36**	**5**	**17**	**4**	**124**	**91**	**105**	**2**	**91**

Supplementary Table 5. Summary of mutations occurring in inverted repeats of length ≥8 bp in the 17 tumour suppressor genes.

Gene symbol	Proportion of gene length covered by repeats (%)	Number of missense mutations found in repeats						Number of micro-deletions and micro-insertions found in repeats			
		somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	Total	somatic non-recurrent	germline	shared	Total
APC	6	5	4	1	1	0	**5**	21	27	2	**50**
ATM	13	1	14	0	0	0	**1**	1	16	0	**17**
BRCA1	6	0	15	0	0	0	**0**	0	22	1	**23**
BRCA2	7	3	1	0	0	0	**3**	1	27	0	**28**
CDH1	5	0	1	0	0	0	**0**	1	0	0	**1**
CDKN2A	8	30	5	6	2	1	**30**	13	2	1	**16**
NF1	11	0	3	0	0	0	**0**	1	24	0	**25**
NF2	10	1	3	0	0	0	**1**	11	6	0	**17**
PTCH1	5	1	0	0	0	0	**1**	0	2	0	**2**
PTEN	6	10	1	1	4	1	**10**	9	2	0	**11**
RB1	16	4	5	1	0	0	**4**	7	28	0	**35**
STK11	13	1	5	0	1	0	**1**	1	9	0	**10**
TP53	5	13	0	0	51	9	**13**	53	2	0	**55**
TSC1	5	0	1	0	0	0	**0**	0	7	0	**7**
TSC2	9	0	6	0	0	0	**0**	1	13	0	**14**
VHL	12	9	8	1	1	0	**9**	36	15	2	**53**
WT1	7	0	2	0	0	0	**0**	0	0	0	**0**
TOTAL	**9**	**78**	**74**	**10**	**60**	**11**	**78**	**156**	**202**	**6**	**364**

Supplementary Table 6. Summary of mutations occurring within symmetric repeats of length ≥8 bp in the 17 tumour suppressor genes.

Gene symbol	Proportion of gene length covered by repeats (%)	Number of missense mutations found in repeats						Number of micro-deletions and micro-insertions found in repeats			
		somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	Total	somatic non-recurrent	germline	shared	Total
APC	16	5	2	0	2	0	**9**	58	87	6	**151**
ATM	32	2	11	0	0	0	**13**	2	43	0	**45**
BRCA1	20	1	30	0	0	0	**31**	0	82	2	**84**
BRCA2	18	6	18	0	0	0	**24**	2	79	3	**84**
CDH1	24	4	0	0	0	0	**4**	5	8	0	**13**
CDKN2A	24	49	13	5	2	0	**69**	35	7	1	**43**
NF1	31	1	20	0	0	0	**21**	2	85	2	**89**
NF2	24	6	3	0	1	0	**10**	49	12	3	**64**
PTCH1	23	5	8	1	0	0	**14**	5	23	0	**28**
PTEN	44	27	3	1	9	0	**40**	42	13	1	**56**
RB1	48	3	10	1	0	0	**14**	4	41	1	**46**
STK11	33	3	6	0	2	0	**11**	1	20	1	**22**
TP53	30	60	2	1	132	23	**218**	147	1	0	**148**
TSC1	23	0	3	0	0	0	**3**	0	27	0	**27**
TSC2	23	0	13	0	0	0	**13**	1	29	0	**30**
VHL	17	3	9	2	0	2	**16**	25	7	2	**34**
WT1	26	0	6	0	0	0	**6**	3	4	0	**7**
TOTAL	**25**	**175**	**157**	**11**	**148**	**25**	**516**	**381**	**568**	**22**	**971**

Suplementary Table 7. Occurrence of missense mutations in repeats/runs of identical nucleotides and/or CpG/CpHpG oligonucleotides

Type of Repeats	Type of mutational spectrum	Number of mutations			Remaining number of mutations
		exclusively in repeats/runs	exclusively in CpG/CpHpG	in both repeats/runs and CpG/CpHpG	
Runs	somatic non-recurrent	184	58	24	618
	germline	151	100	27	608
	somatic recurrent	167	46	18	636
	shared non-recurrent	5	28	0	69
	shared recurrent	10	38	5	59
	potential	32861	3902	765	111495
Direct	somatic non-recurrent	52	72	10	750
	germline	31	122	5	728
	somatic recurrent	14	61	3	789
	shared non-recurrent	3	26	2	71
	shared recurrent	2	41	2	67
	potential	5252	4431	236	139104

Inverted	somatic non-recurrent	65	69	13	737
	germline	64	117	10	695
	somatic recurrent	55	59	5	748
	shared non-recurrent	8	26	2	66
	shared recurrent	7	39	4	62
	potential	10790	4314	353	133566
Symmetric	somatic non-recurrent	155	62	20	647
	germline	140	110	17	619
	somatic recurrent	137	53	11	666
	shared non-recurrent	7	24	4	67
	shared recurrent	16	34	9	53
	potential	28646	3752	915	115710

Supplementary Table 8. Truncating vs. non-truncating lesions

Gene		Missense	Nonsense	Micro-deletions	Micro-insertions	Micro-indels	Non-truncating lesions	Truncating lesions	Ratio of non-truncating to truncating lesions	Ratio of truncating somatic to truncating germline lesions
APC	Somatic	39	79	152	44	3	39	278	0.14	0.46
	Germline	23	180	299	115	12	23	606	0.04	
ATM	Somatic	11	7	4	1	0	11	12	0.92	0.05
	Germline	76	75	122	35	14	76	246	0.31	
BRCA1	Somatic	6	9	9	5	0	6	23	0.26	0.05
	Germline	170	121	259	85	12	170	477	0.36	
BRCA2	Somatic	21	1	8	4	0	21	13	1.62	0.03
	Germline	86	76	247	90	11	86	424	0.20	
CDH1	Somatic	15	7	13	2	0	15	22	0.68	0.69
	Germline	19	11	12	8	1	19	32	0.59	
CDKN2A	Somatic	198	18	77	25	8	198	128	1.55	4.74
	Germline	62	7	11	7	2	62	27	2.30	
NF1	Somatic	2	11	16	3	0	2	30	0.07	0.07
	Germline	83	115	221	105	8	83	449	0.18	
NF2	Somatic	23	42	182	28	6	23	258	0.09	2.22
	Germline	20	43	55	16	2	20	116	0.17	
PTCH1	Somatic	14	9	14	6	1	14	30	0.47	0.28
	Germline	24	27	42	32	8	24	109	0.22	
PTEN	Somatic	226	56	152	51	4	226	263	0.86	3.21
	Germline	45	28	29	22	3	45	82	0.55	
RB1	Somatic	25	27	34	12	3	25	76	0.33	0.30

	Germline	37	76	117	53	11	37	257	0.14	
STK11	Somatic	20	10	5	1	1	20	17	1.18	0.17
	Germline	30	27	47	24	3	30	101	0.30	
TP53	Somatic	1229	96	512	238	0	1229	846	1.45	24.89
	Germline	94	10	16	5	3	94	34	2.76	
TSC1	Somatic	2	1	1	0	0	2	2	1.00	0.02
	Germline	7	37	53	25	4	7	119	0.06	
TSC2	Somatic	2	1	3	2	1	2	7	0.29	0.03
	Germline	89	74	110	46	3	89	233	0.38	
VHL	Somatic	88	15	180	44	1	88	240	0.37	1.82
	Germline	143	27	63	37	5	143	132	1.08	
WT1	Somatic	1	3	4	3	0	1	10	0.10	0.37
	Germline	40	14	8	4	1	40	27	1.48	
Total	Somatic	1922	392	1366	469	28	1922	2255	0.85	0.65
	Germline	1048	948	1711	709	103	1048	3471	0.30	

Supplementary Figure 1. Naive Bayes Tree Classifier. Number in parenthesis shows the probability of a mutations being somatic non-recurrent, germline, shared non-recurrent, somatic recurrent and shared recurrent respectively.

```
Attributes:
              Mut_Type
              Hess_value
              Krawczak_value
              Evol
              Grantham_score
              CpG/CHG
              Repeats
Test mode:    10-fold cross-validation

NBTree
------------------

Evol <= 0.205
|   Repeats = 0
|   |   CpG/CHG = 0
|   |   |   Krawczak_value <= 1.0465
|   |   |   |   Evol <= 0.155
|   |   |   |   |   Evol <= 0.12
|   |   |   |   |   |   Krawczak_value <= 0.811
|   |   |   |   |   |   |   Krawczak_value <= 0.099
|   |   |   |   |   |   |   |   Hess_value <= 3.1:          (0.42) (0.08) (0.08) (0.33) (0.08)
|   |   |   |   |   |   |   |   Hess_value > 3.1:           (0.23) (0.13) (0.03) (0.10) (0.52)
|   |   |   |   |   |   |   Krawczak_value > 0.099
|   |   |   |   |   |   |   |   Hess_value <= 2.5
|   |   |   |   |   |   |   |   |   Grantham_score <= 146.5
|   |   |   |   |   |   |   |   |   |   Hess_value <= 2.15: (0.27) (0.47) (0.02) (0.22) (0.02)
|   |   |   |   |   |   |   |   |   |   Hess_value > 2.15:  (0.14) (0.24) (0.05) (0.52) (0.05)
|   |   |   |   |   |   |   |   |   Grantham_score > 146.5: (0.47) (0.07) (0.07) (0.33) (0.07)
|   |   |   |   |   |   |   |   Hess_value > 2.5
|   |   |   |   |   |   |   |   |   Hess_value <= 5.45
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 30.5
|   |   |   |   |   |   |   |   |   |   |   Hess_value <= 5.2
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.55
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 2.75: (0.27) (0.09) (0.09) (0.45) (0.09)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 2.75:  (0.25) (0.43) (0.03) (0.28) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.55:      (0.29) (0.08) (0.04) (0.54) (0.04)
|   |   |   |   |   |   |   |   |   |   |   Hess_value > 5.2:           (0.12) (0.12) (0.06) (0.12) (0.59)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 30.5
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.411
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.35
|   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.3775
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.1975
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 146:  (0.23) (0.13) (0.03) (0.57) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 146:   (0.28) (0.16)  (0.4) (0.12) (0.04)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.1975
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.22: (0.11) (0.04) (0.26) (0.11) (0.48)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.22
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 2.85
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 147.5: (0.21) (0.14) (0.28) (0.34) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 147.5
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 2.75:  (0.21) (0.04) (0.29) (0.08) (0.38)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 2.75:   (0.05) (0.05) (0.79) (0.05) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 2.85
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 155.5
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 3.95:  (0.18) (0.15) (0.03) (0.61) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 3.95:   (0.10) (0.14) (0.14) (0.43) (0.19)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 155.5:  (0.23) (0.06) (0.49)  (0.2) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.3775:      (0.12) (0.32) (0.04) (0.48) (0.04)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.35
|   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 100.5
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.2455: (0.09) (0.45) (0.09) (0.27) (0.09)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.2455:  (0.42) (0.29) (0.03) (0.23) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 100.5:       (0.23) (0.14) (0.05) (0.32) (0.27)
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.411
|   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 105.5
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.85
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 100
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 63: (0.04) (0.04) (0.77) (0.13) (0.02)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 63:  (0.21) (0.26) (0.05) (0.42) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 100:     (0.04) (0.04) (0.78) (0.09) (0.04)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 70.5:   (0.26) (0.16) (0.05) (0.47) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 70.5:    (0.13) (0.10) (0.63) (0.10) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.85:  (0.31) (0.38) (0.08) (0.15) (0.08)
|   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 105.5
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 3.05: (0.28)  (0.1) (0.45) (0.14) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 3.05:  (0.18) (0.32) (0.04) (0.32) (0.14)
|   |   |   |   |   |   |   |   |   Hess_value > 5.45
|   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.336: (0.06) (0.06) (0.63) (0.22) (0.03)
|   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.336:  (0.13) (0.46) (0.19) (0.20) (0.01)
|   |   |   |   |   |   Krawczak_value > 0.811
|   |   |   |   |   |   |   Grantham_score <= 78.5
|   |   |   |   |   |   |   |   Grantham_score <= 37.5:   (0.27) (0.27) (0.05) (0.05) (0.36)
|   |   |   |   |   |   |   |   Grantham_score > 37.5:    (0.51) (0.17) (0.02) (0.27) (0.02)
|   |   |   |   |   |   |   Grantham_score > 78.5
|   |   |   |   |   |   |   |   Hess_value <= 10.95
|   |   |   |   |   |   |   |   |   Grantham_score <= 129: (0.03) (0.28) (0.03) (0.15) (0.51)
|   |   |   |   |   |   |   |   |   Grantham_score > 129:  (0.35) (0.13) (0.04) (0.04) (0.43)
|   |   |   |   |   |   |   |   Hess_value > 10.95:      (0.22) (0.39) (0.06) (0.28) (0.06)
|   |   |   |   |   Evol > 0.12
|   |   |   |   |   |   Evol <= 0.135: (0.08) (0.15) (0.62) (0.08) (0.08)
|   |   |   |   |   |   Evol > 0.135
|   |   |   |   |   |   |   Krawczak_value <= 0.5255
|   |   |   |   |   |   |   |   Hess_value <= 4.3:   (0.03) (0.40) (0.27) (0.27) (0.03)
|   |   |   |   |   |   |   |   Hess_value > 4.3:    (0.06) (0.06) (0.75) (0.06) (0.06)
|   |   |   |   |   |   |   Krawczak_value > 0.5255: (0.22) (0.04) (0.04) (0.13) (0.57)
|   |   |   |   Evol > 0.155
|   |   |   |   |   Evol <= 0.175:  (0.38) (0.24) (0.05) (0.29) (0.05)
|   |   |   |   |   Evol > 0.175:   (0.17)  (0.1) (0.03) (0.41) (0.28)
|   |   |   Krawczak_value > 1.0465
|   |   |   |   Hess_value <= 12.35
|   |   |   |   |   Krawczak_value <= 1.1575: (0.03) (0.06) (0.68) (0.21) (0.03)
|   |   |   |   |   Krawczak_value > 1.1575
|   |   |   |   |   |   Hess_value <= 7.05:   (0.07) (0.24) (0.03) (0.38) (0.28)
|   |   |   |   |   |   Hess_value > 7.05
|   |   |   |   |   |   |   Krawczak_value <= 1.838
|   |   |   |   |   |   |   |   Krawczak_value <= 1.725
|   |   |   |   |   |   |   |   |   Krawczak_value <= 1.27
|   |   |   |   |   |   |   |   |   |   Hess_value <= 7.6: (0.04) (0.15) (0.42) (0.04) (0.35)
|   |   |   |   |   |   |   |   |   |   Hess_value > 7.6:  (0.16) (0.21) (0.05) (0.05) (0.53)
|   |   |   |   |   |   |   |   |   Krawczak_value > 1.27
|   |   |   |   |   |   |   |   |   |   Krawczak_value <= 1.5585
|   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 60:  (0.19) (0.14) (0.05) (0.29) (0.33)
|   |   |   |   |   |   |   |   |   |   |   Grantham_score > 60:   (0.15)  (0.3) (0.05) (0.45) (0.05)
|   |   |   |   |   |   |   |   |   |   Krawczak_value > 1.5585
|   |   |   |   |   |   |   |   |   |   |   Hess_value <= 8.65
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 7.5: (0.20) (0.07)  (0.6) (0.07) (0.07)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 7.5:  (0.04) (0.15) (0.31) (0.08) (0.42)
|   |   |   |   |   |   |   |   |   |   |   Hess_value > 8.65:     (0.38) (0.38) (0.06) (0.13) (0.06)
|   |   |   |   |   |   |   |   Krawczak_value > 1.725:    (0.09) (0.05) (0.27) (0.55) (0.05)
|   |   |   |   |   |   |   Krawczak_value > 1.838
|   |   |   |   |   |   |   |   Hess_value <= 11.5:  (0.04) (0.34) (0.35) (0.09) (0.18)
|   |   |   |   |   |   |   |   Hess_value > 11.5:   (0.03) (0.18) (0.46)  (0.1) (0.23)
|   |   |   |   Hess_value > 12.35
|   |   |   |   |   Grantham_score <= 86
|   |   |   |   |   |   Hess_value <= 13.8:  (0.15) (0.15) (0.03) (0.38) (0.29)
|   |   |   |   |   |   Hess_value > 13.8:   (0.13) (0.09) (0.52) (0.04) (0.22)
|   |   |   |   |   Grantham_score > 86
|   |   |   |   |   |   Hess_value <= 13.15: (0.03) (0.41) (0.03) (0.03)  (0.5)
|   |   |   |   |   |   Hess_value > 13.15:  (0.13)  (0.2) (0.03)  (0.2) (0.43)
|   |   CpG/CHG = 1
|   |   |   Hess_value <= 59.5
|   |   |   |   Grantham_score <= 44.5: (0.03) (0.04) (0.18) (0.07) (0.68)
|   |   |   |   Grantham_score > 44.5:  (0.03) (0.12) (0.41) (0.01) (0.44)
|   |   |   Hess_value > 59.5:          (0.20) (0.60) (0.03) (0.14) (0.03)
|   Repeats = 1
|   |   CpG/CHG = 0
|   |   |   Hess_value <= 4.35
|   |   |   |   Evol <= 0.18
|   |   |   |   |   Evol <= 0.065
|   |   |   |   |   |   Krawczak_value <= 0.232
|   |   |   |   |   |   |   Grantham_score <= 134.5
|   |   |   |   |   |   |   |   Grantham_score <= 112.5
|   |   |   |   |   |   |   |   |   Grantham_score <= 54: (0.33) (0.11) (0.06) (0.11) (0.39)
|   |   |   |   |   |   |   |   |   Grantham_score > 54:  (0.23) (0.23) (0.03) (0.48) (0.03)
|   |   |   |   |   |   |   |   Grantham_score > 112.5:   (0.44) (0.06) (0.06) (0.06) (0.38)
|   |   |   |   |   |   |   Grantham_score > 134.5:       (0.13) (0.07) (0.07) (0.67) (0.07)
|   |   |   |   |   |   Krawczak_value > 0.232
|   |   |   |   |   |   |   Hess_value <= 3.3
|   |   |   |   |   |   |   |   Krawczak_value <= 0.341
|   |   |   |   |   |   |   |   |   Grantham_score <= 84:   (0.24) (0.04) (0.56) (0.12) (0.04)
|   |   |   |   |   |   |   |   |   Grantham_score > 84
|   |   |   |   |   |   |   |   |   |   Hess_value <= 2.65: (0.09) (0.52) (0.04)  (0.3) (0.04)
|   |   |   |   |   |   |   |   |   |   Hess_value > 2.65:  (0.27) (0.14) (0.05)  (0.5) (0.05)
|   |   |   |   |   |   |   |   Krawczak_value > 0.341
|   |   |   |   |   |   |   |   |   Krawczak_value <= 0.463:(0.38) (0.46) (0.04) (0.08) (0.04)
|   |   |   |   |   |   |   |   |   Krawczak_value > 0.463: (0.21) (0.31) (0.03) (0.41) (0.03)
|   |   |   |   |   |   |   Hess_value > 3.3:               (0.20) (0.27) (0.01) (0.51) (0.01)
|   |   |   |   |   Evol > 0.065:   (0.36)  (0.5) (0.05) (0.05) (0.05)
|   |   |   |   Evol > 0.18: (      (0.10) (0.05) (0.05) (0.76) (0.05)
|   |   |   Hess_value > 4.35
|   |   |   |   Evol <= 0.045
|   |   |   |   |   Grantham_score <= 30.5
|   |   |   |   |   |   Hess_value <= 5.55:         (0.43) (0.18) (0.04) (0.32) (0.04)
|   |   |   |   |   |   Hess_value > 5.55
|   |   |   |   |   |   |   Grantham_score <= 26.5: (0.18) (0.44) (0.03) (0.32) (0.03)
|   |   |   |   |   |   |   Grantham_score > 26.5:  (0.11) (0.11) (0.05) (0.68) (0.05)
|   |   |   |   |   Grantham_score > 30.5
|   |   |   |   |   |   Grantham_score <= 118.5
|   |   |   |   |   |   |   Grantham_score <= 95.5
|   |   |   |   |   |   |   |   Hess_value <= 10.6
|   |   |   |   |   |   |   |   |   Grantham_score <= 75.5
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 69.5
|   |   |   |   |   |   |   |   |   |   |   Hess_value <= 7.05
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.65
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.55: (0.07) (0.23) (0.03) (0.13) (0.53)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.55:  (0.30) (0.30) (0.30) (0.05) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.65:      (0.07) (0.21) (0.03) (0.31) (0.38)
|   |   |   |   |   |   |   |   |   |   |   Hess_value > 7.05:          (0.23) (0.02) (0.02) (0.32) (0.41)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 69.5:          (0.10) (0.10) (0.33) (0.02) (0.45)
|   |   |   |   |   |   |   |   |   Grantham_score > 75.5
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 92.5:    (0.13) (0.29) (0.04)  (0.5) (0.04)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 92.5:     (0.18) (0.32) (0.41) (0.05) (0.05)
|   |   |   |   |   |   |   |   Hess_value > 10.6:      (0.26) (0.23) (0.03) (0.46) (0.03)
|   |   |   |   |   |   |   Grantham_score > 95.5
|   |   |   |   |   |   |   |   Hess_value <= 5.55
|   |   |   |   |   |   |   |   |   Hess_value <= 4.65: (0.27) (0.45) (0.09) (0.09) (0.09)
|   |   |   |   |   |   |   |   |   Hess_value > 4.65:  (0.03) (0.06) (0.03)  (0.2) (0.69)
|   |   |   |   |   |   |   |   Hess_value > 5.55
|   |   |   |   |   |   |   |   |   Grantham_score <= 102.5: (0.18) (0.56) (0.02) (0.13) (0.11)
|   |   |   |   |   |   |   |   |   Grantham_score > 102.5:  (0.06)  (0.2) (0.03) (0.37) (0.34)
|   |   |   |   |   |   Grantham_score > 118.5
|   |   |   |   |   |   |   Grantham_score <= 149.5:         (0.08) (0.13) (0.18) (0.04) (0.57)
|   |   |   |   |   |   |   Grantham_score > 149.5
|   |   |   |   |   |   |   |   Hess_value <= 10.45
|   |   |   |   |   |   |   |   |   Krawczak_value <= 0.428: (0.36) (0.09) (0.09) (0.36) (0.09)
|   |   |   |   |   |   |   |   |   Krawczak_value > 0.428:  (0.07) (0.26) (0.56) (0.09) (0.02)
|   |   |   |   |   |   |   |   Hess_value > 10.45:          (0.04) (0.16) (0.24) (0.06) (0.50)
|   |   |   |   Evol > 0.045:    (0.33) (0.37) (0.04) (0.22) (0.04)
|   |   CpG/CHG = 1
|   |   |   Grantham_score <= 99.5
|   |   |   |   Hess_value <= 10.05
|   |   |   |   |   Grantham_score <= 86: (0.07) (0.14) (0.07) (0.21)  (0.5)
|   |   |   |   |   Grantham_score > 86:  (0.03) (0.03) (0.88) (0.03) (0.03)
|   |   |   |   Hess_value > 10.05
|   |   |   |   |   Evol <= 0.07
|   |   |   |   |   |   Krawczak_value <= 12.275
|   |   |   |   |   |   |   Krawczak_value <= 9.211
|   |   |   |   |   |   |   |   Krawczak_value <= 8.5135
|   |   |   |   |   |   |   |   |   Krawczak_value <= 7.551: (0.45) (0.27) (0.09) (0.09) (0.09)
|   |   |   |   |   |   |   |   |   Krawczak_value > 7.551:  (0.03) (0.14) (0.03) (0.03) (0.76)
|   |   |   |   |   |   |   |   Krawczak_value > 8.5135:     (0.47) (0.35) (0.06) (0.06) (0.06)
|   |   |   |   |   |   |   Krawczak_value > 9.211
|   |   |   |   |   |   |   |   Hess_value <= 46.4: (0.26) (0.11) (0.05) (0.05) (0.53)
|   |   |   |   |   |   |   |   Hess_value > 46.4:  (0.02) (0.02) (0.22) (0.06) (0.68)
|   |   |   |   |   |   Krawczak_value > 12.275:    (0.08) (0.03) (0.72) (0.03) (0.14)
|   |   |   |   |   Evol > 0.07:   (0.07) (0.03) (0.03) (0.03) (0.83)
|   |   |   Grantham_score > 99.5
|   |   |   |   Krawczak_value <= 7.519:   (0.03) (0.03) (0.03)  (0.1) (0.82)
|   |   |   |   Krawczak_value > 7.519
|   |   |   |   |   Grantham_score <= 113: (0.02) (0.19) (0.02) (0.06) (0.70)
|   |   |   |   |   Grantham_score > 113:  (0.13) (0.57) (0.04) (0.22) (0.04)
Evol > 0.205
|   Hess_value <= 9.65
|   |   Repeats = 0
|   |   |   Hess_value <= 8.8
|   |   |   |   Grantham_score <= 40.5
|   |   |   |   |   Hess_value <= 2.65: (0.60) (0.07) (0.07) (0.20) (0.07)
|   |   |   |   |   Hess_value > 2.65
|   |   |   |   |   |   Krawczak_value <= 1.083
|   |   |   |   |   |   |   Krawczak_value <= 0.269:     (0.11) (0.39) (0.06) (0.39) (0.06)
|   |   |   |   |   |   |   Krawczak_value > 0.269
|   |   |   |   |   |   |   |   Krawczak_value <= 0.6155
|   |   |   |   |   |   |   |   |   Hess_value <= 4:     (0.68) (0.05) (0.05) (0.16) (0.05)
|   |   |   |   |   |   |   |   |   Hess_value > 4:      (0.48) (0.28) (0.04) (0.16) (0.04)
|   |   |   |   |   |   |   |   Krawczak_value > 0.6155: (0.22)  (0.5) (0.06) (0.17) (0.06)
|   |   |   |   |   |   Krawczak_value > 1.083:          (0.17) (0.08) (0.08) (0.58) (0.08)
|   |   |   |   Grantham_score > 40.5
|   |   |   |   |   Hess_value <= 5.05
|   |   |   |   |   |   Grantham_score <= 194.5
|   |   |   |   |   |   |   Krawczak_value <= 0.365
|   |   |   |   |   |   |   |   Hess_value <= 3.95
|   |   |   |   |   |   |   |   |   Grantham_score <= 66.5
|   |   |   |   |   |   |   |   |   |   Hess_value <= 2.65:      (0.21) (0.07) (0.38) (0.07) (0.28)
|   |   |   |   |   |   |   |   |   |   Hess_value > 2.65
|   |   |   |   |   |   |   |   |   |   |   Evol <= 0.275:       (0.05) (0.05) (0.79) (0.05) (0.05)
|   |   |   |   |   |   |   |   |   |   |   Evol > 0.275:        (0.32) (0.08) (0.36)  (0.2) (0.04)
|   |   |   |   |   |   |   |   |   Grantham_score > 66.5
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 159.5: (0.36) (0.37) (0.12) (0.15) (0.01)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 159.5:  (0.19) (0.04)  (0.3) (0.07) (0.41)
|   |   |   |   |   |   |   |   Hess_value > 3.95
|   |   |   |   |   |   |   |   |   Krawczak_value <= 0.229:     (0.26) (0.19) (0.04) (0.48) (0.04)
|   |   |   |   |   |   |   |   |   Krawczak_value > 0.229:      (0.39) (0.07) (0.04) (0.04) (0.46)
|   |   |   |   |   |   |   Krawczak_value > 0.365
|   |   |   |   |   |   |   |   Hess_value <= 4.55
|   |   |   |   |   |   |   |   |   Hess_value <= 4.3
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 105.5: (0.51) (0.14) (0.03) (0.29) (0.03)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 105.5
|   |   |   |   |   |   |   |   |   |   |   Hess_value <= 3.3:   (0.50) (0.33) (0.06) (0.06) (0.06)
|   |   |   |   |   |   |   |   |   |   |   Hess_value > 3.3:    (0.36) (0.16) (0.04) (0.28) (0.16)
|   |   |   |   |   |   |   |   |   Hess_value > 4.3:            (0.06) (0.24) (0.06) (0.29) (0.35)
|   |   |   |   |   |   |   |   Hess_value > 4.55:               (0.39) (0.04) (0.04) (0.48) (0.04)
|   |   |   |   |   |   Grantham_score > 194.5:   (0.09) (0.09) (0.73) (0.05) (0.05)
|   |   |   |   |   Hess_value > 5.05
|   |   |   |   |   |   Grantham_score <= 45.5:   (0.04) (0.11) (0.54) (0.29) (0.04)
|   |   |   |   |   |   Grantham_score > 45.5
|   |   |   |   |   |   |   Evol <= 0.51
|   |   |   |   |   |   |   |   Hess_value <= 7.25
|   |   |   |   |   |   |   |   |   Evol <= 0.28: (0.07) (0.43) (0.07) (0.36) (0.07)
|   |   |   |   |   |   |   |   |   Evol > 0.28:  (0.27) (0.27) (0.24) (0.18) (0.03)
|   |   |   |   |   |   |   |   Hess_value > 7.25
|   |   |   |   |   |   |   |   |   Hess_value <= 7.6:        (0.09) (0.18) (0.09) (0.55) (0.09)
|   |   |   |   |   |   |   |   |   Hess_value > 7.6
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 69: (0.57) (0.09) (0.04) (0.26) (0.04)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 69:  (0.04) (0.29) (0.04) (0.58) (0.04)
|   |   |   |   |   |   |   Evol > 0.51
|   |   |   |   |   |   |   |   Grantham_score <= 88.5
|   |   |   |   |   |   |   |   |   Krawczak_value <= 1.005:  (0.25) (0.33) (0.08) (0.25) (0.08)
|   |   |   |   |   |   |   |   |   Krawczak_value > 1.005
|   |   |   |   |   |   |   |   |   |   Evol <= 0.61:         (0.06) (0.03) (0.85) (0.03) (0.03)
|   |   |   |   |   |   |   |   |   |   Evol > 0.61:          (0.50) (0.25) (0.08) (0.08) (0.08)
|   |   |   |   |   |   |   |   Grantham_score > 88.5:        (0.27) (0.09) (0.09) (0.45) (0.09)
|   |   |   Hess_value > 8.8
|   |   |   |   Krawczak_value <= 1.1745
|   |   |   |   |   Krawczak_value <= 0.862: (0.69) (0.08) (0.08) (0.08) (0.08)
|   |   |   |   |   Krawczak_value > 0.862:  (0.13) (0.16) (0.03) (0.09) (0.59)
|   |   |   |   Krawczak_value > 1.1745:     (0.58) (0.05) (0.05) (0.26) (0.05)
|   |   Repeats = 1
|   |   |   Grantham_score <= 123
|   |   |   |   Evol <= 0.285
|   |   |   |   |   Evol <= 0.255: (0.47) (0.06) (0.03) (0.25) (0.19)
|   |   |   |   |   Evol > 0.255:  (0.09) (0.06) (0.42) (0.03) (0.39)
|   |   |   |   Evol > 0.285
|   |   |   |   |   Krawczak_value <= 1.27
|   |   |   |   |   |   Hess_value <= 8.55
|   |   |   |   |   |   |   CpG/CHG = 0
|   |   |   |   |   |   |   |   Hess_value <= 6.75
|   |   |   |   |   |   |   |   |   Evol <= 0.415
|   |   |   |   |   |   |   |   |   |   Evol <= 0.355
|   |   |   |   |   |   |   |   |   |   |   Evol <= 0.295
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 2.75: (0.32) (0.05) (0.42) (0.16) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 2.75:  (0.65) (0.23) (0.04) (0.04) (0.04)
|   |   |   |   |   |   |   |   |   |   |   Evol > 0.295:           (0.25) (0.19) (0.06) (0.44) (0.06)
|   |   |   |   |   |   |   |   |   |   Evol > 0.355
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.5455
|   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.284
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 3.55: (0.18) (0.04) (0.71) (0.04) (0.04)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 3.55:  (0.07) (0.21)  (0.5) (0.14) (0.07)
|   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.284: (0.27) (0.32) (0.05) (0.32) (0.05)
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.5455:    (0.05) (0.05) (0.67) (0.05) (0.19)
|   |   |   |   |   |   |   |   |   Evol > 0.415
|   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.4675
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.417
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.8
|   |   |   |   |   |   |   |   |   |   |   |   |   Evol <= 0.585:    (0.38) (0.18) (0.03) (0.38) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Evol > 0.585:     (0.70) (0.14) (0.03) (0.11) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.8:     (0.71) (0.07) (0.07) (0.07) (0.07)
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.417:   (0.08) (0.31) (0.08) (0.46) (0.08)
|   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.4675
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.5205: (0.24) (0.06) (0.53) (0.12) (0.06)
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.5205:  (0.78) (0.07) (0.04) (0.07) (0.04)
|   |   |   |   |   |   |   |   Hess_value > 6.75
|   |   |   |   |   |   |   |   |   Grantham_score <= 57: (0.11) (0.53) (0.05) (0.26) (0.05)
|   |   |   |   |   |   |   |   |   Grantham_score > 57:  (0.47) (0.22) (0.03) (0.25) (0.03)
|   |   |   |   |   |   |   |   CpG/CHG = 1:          (0.40) (0.10) (0.10) (0.30) (0.10)
|   |   |   |   |   |   Hess_value > 8.55
|   |   |   |   |   |   |   Evol <= 0.54:         (0.27) (0.20) (0.07) (0.40) (0.07)
|   |   |   |   |   |   |   Evol > 0.54:          (0.03) (0.03) (0.84) (0.06) (0.03)
|   |   |   |   |   Krawczak_value > 1.27
|   |   |   |   |   |   Grantham_score <= 86:     (0.52) (0.04) (0.04) (0.37) (0.04)
|   |   |   |   |   |   Grantham_score > 86:      (0.40) (0.40) (0.03) (0.13) (0.03)
|   |   |   Grantham_score > 123
|   |   |   |   Evol <= 0.445
|   |   |   |   |   Hess_value <= 3.45
|   |   |   |   |   |   Krawczak_value <= 0.4665: (0.03) (0.19) (0.03) (0.16) (0.59)
|   |   |   |   |   |   Krawczak_value > 0.4665:  (0.25) (0.08) (0.08) (0.50) (0.08)
|   |   |   |   |   Hess_value > 3.45:            (0.43) (0.05) (0.05) (0.43) (0.05)
|   |   |   |   Evol > 0.445:                     (0.44) (0.09) (0.03) (0.41) (0.03)
|   Hess_value > 9.65
|   |   Hess_value <= 42.75
|   |   |   Hess_value <= 12.1
|   |   |   |   Repeats = 0
|   |   |   |   |   Evol <= 0.325:         (0.32) (0.39) (0.21) (0.04) (0.04)
|   |   |   |   |   Evol > 0.325
|   |   |   |   |   |   Hess_value <= 11.4
|   |   |   |   |   |   |   Evol <= 0.705: (0.26) (0.33) (0.04) (0.33) (0.04)
|   |   |   |   |   |   |   Evol > 0.705:  (0.06) (0.75) (0.06) (0.06) (0.06)
|   |   |   |   |   |   Hess_value > 11.4: (0.18) (0.23) (0.05) (0.14) (0.41)
|   |   |   |   Repeats = 1
|   |   |   |   |   Grantham_score <= 91.5
|   |   |   |   |   |   Grantham_score <= 85
|   |   |   |   |   |   |   Hess_value <= 11.4: (0.18) (0.24) (0.47) (0.08) (0.03)
|   |   |   |   |   |   |   Hess_value > 11.4:  (0.05) (0.32) (0.05) (0.14) (0.45)
|   |   |   |   |   |   Grantham_score > 85:    (0.20) (0.45) (0.05) (0.25) (0.05)
|   |   |   |   |   Grantham_score > 91.5:      (0.33) (0.17) (0.03) (0.03) (0.43)
|   |   |   Hess_value > 12.1
|   |   |   |   Evol <= 0.51
|   |   |   |   |   Repeats = 0
|   |   |   |   |   |   Grantham_score <= 44.5:   (0.20) (0.07) (0.07) (0.60) (0.07)
|   |   |   |   |   |   Grantham_score > 44.5
|   |   |   |   |   |   |   Grantham_score <= 51: (0.09) (0.27) (0.09) (0.09) (0.45)
|   |   |   |   |   |   |   Grantham_score > 51:  (0.24) (0.53) (0.03) (0.18) (0.03)
|   |   |   |   |   Repeats = 1:                  (0.32) (0.45) (0.05) (0.14) (0.05)
|   |   |   |   Evol > 0.51
|   |   |   |   |   Hess_value <= 13.35: (0.27) (0.15) (0.04)  (0.5) (0.04)
|   |   |   |   |   Hess_value > 13.35:  (0.28) (0.44) (0.04)  (0.2) (0.04)
|   |   Hess_value > 42.75
|   |   |   Repeats = 0
|   |   |   |   Evol <= 0.59
|   |   |   |   |   Evol <= 0.255:       (0.08) (0.12) (0.73) (0.04) (0.04)
|   |   |   |   |   Evol > 0.255
|   |   |   |   |   |   Evol <= 0.375:   (0.18) (0.03) (0.03) (0.28) (0.49)
|   |   |   |   |   |   Evol > 0.375:    (0.40) (0.13) (0.07) (0.33) (0.07)
|   |   |   |   Evol > 0.59
|   |   |   |   |   Grantham_score <= 139:  (0.02) (0.20) (0.75) (0.02) (0.02)
|   |   |   |   |   Grantham_score > 139:   (0.36) (0.43) (0.07) (0.07) (0.07)
|   |   |   Repeats = 1
|   |   |   |   Hess_value <= 59.5
|   |   |   |   |   Hess_value <= 50.35:    (0.40) (0.15) (0.05) (0.35) (0.05)
|   |   |   |   |   Hess_value > 50.35:     (0.67) (0.13) (0.04) (0.13) (0.04)
|   |   |   |   Hess_value > 59.5:          (0.19) (0.63) (0.06) (0.06) (0.06)
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2797               63.1377 %
Incorrectly Classified Instances      1633               36.8623 %
Kappa statistic                          0.5392
Mean absolute error                      0.1878
Root mean squared error                  0.3177
Relative absolute error                 58.6858 %
```

```
Root relative squared error             79.4156 %
Total Number of Instances             4430

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.505     0.106      0.544     0.505     0.523      0.826    1
                 0.426     0.082      0.566     0.426     0.486      0.778    2
                 0.894     0.091      0.712     0.894     0.792      0.967    3
                 0.475     0.109      0.52      0.475     0.497      0.809    4
                 0.858     0.073      0.745     0.858     0.797      0.964    5
Weighted Avg.    0.631     0.092      0.617     0.631     0.619      0.869

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 447 125  63 207  44 |   a = 1
 170 377  89 153  97 |   b = 2
  12   9 792   9  64 |   c = 3
 181 144  85 421  55 |   d = 4
  12  11  84  19 760 |   e = 5
```

Comparative Analysis of Germline and Somatic Micro-lesion Mutational Spectra in 17 Human Tumour Suppressor Genes

Dobril Ivanov[1,2], Stephen E. Hamby[3], Peter D. Stenson[1], Andrew D. Phillips[1],
Hildegard Kehrer-Sawatzki[4], David N. Cooper[1] and Nadia Chuzhanova[3]

[1]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

[2]MRC Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine and Neurology, Biostatistics and Bioinformatics Unit, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

[3]School of Science and Technology, Nottingham Trent University, Nottingham, NG11 8NS, UK

[4]Institute of Human Genetics, University of Ulm, Albert-Einstein-Allee 11, 89081 Ulm, Germany

*All correspondence to: Prof. Nadia Chuzhanova, School of Science and Technology Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK

Tel: +44 (0) 0115 848 8304 E-mail: nadia.chuzhanova@ntu.ac.uk

Abstract

Mutations associated with tumorigenesis may either arise somatically or can be inherited through the germline. We performed a comparison of somatic, germline, shared (found in both soma and germline) and somatic recurrent mutational spectra for 17 human tumour suppressor genes which included missense single base-pair substitutions and micro-deletions/micro-insertions. Somatic and germline mutational spectra were similar in relation to C.G>T.A transitions but differed with respect to the frequency of A.T>G.C, A.T>T.A and C.G>A.T substitutions. Shared missense mutations were characterised by higher mutability rates, greater physicochemical differences between wild-type and mutant residues, and a tendency to occur in evolutionarily conserved residues and within CpG/CpHpG oligonucleotides. Mononucleotide runs (≥ 4 bp) were identified as hotspots for shared micro-deletions/micro-insertions. Both germline and somatic micro-deletions/micro-insertions were found to be significantly overrepresented within the 'indel-hotspot' motif, GTAAGT. Using a naïve Bayes' classifier trained to discriminate between five missense mutation groups, 63% of mutations in our dataset were on average correctly recognized. Applying this classifier to an independent dataset of probable driver mutations, we concluded that ~50% of these somatic missense mutations possess features consistent with their being either shared or recurrent, suggesting that a disproportionate number of such lesions are likely to be drivers of tumorigenesis.

Deleted: In this study

Key Words: germline and somatic mutational spectra; tumour suppressor genes; recurrent mutation; mutation hotspot; non-B DNA; driver mutations

Introduction

A major distinction to be made between somatic and germline mutations is that the former occur during mitotic cell cycles whereas the latter are generally meiotic in origin. In addition, whilst somatic cancer-causing gene lesions come to clinical attention by conferring a growth advantage upon the affected cells or tissue, germ-line gene mutations causing inherited disease normally come to attention by conferring a disadvantage upon the individual, usually through haploinsufficiency. Finally, whereas inherited disease usually implies only one or two pathological mutations at a specific locus, cancer is often characterized by multiple somatic mutations distributed genome-wide. Those somatic mutations which confer a growth advantage on the cells in which they occur, which are positively selected for in the emerging tumour mass and which have therefore been causally implicated in tumorigenesis, are termed 'driver' mutations [Stratton *et al.*, 2009]. By contrast, those mutations which do not confer any growth advantage and have not been subject to selection during tumorigenesis, are termed 'passenger' mutations [Stratton *et al.*, 2009]. Such passenger mutations may arise at high frequency as a consequence either of increased genomic instability or simply due to the considerable number of cell divisions required to convert a single transformed cell into a clinically detectable tumour [Lengauer et al., 1998; Boland and Ricciardiello, 1999; Simpson 2008; Parmigiani et al. 2009; Stratton et al., 2009].

Despite these basic differences, the mutational spectra (and hence the underlying mutational mechanisms) associated with single base-pair substitutions [Krawczak et al., 1995; Schmutte and Jones, 1998; Cole et al., 2008; Lobo et al., 2009], micro-deletions and micro-insertions [Jego et al., 1993; Greenblatt et al. 1996] and gross gene rearrangements [Oldenburg et al., 2000; Kolomietz et al., 2002] in specific genes often appear to exhibit marked similarities between the germline and the soma. Further, certain triplet repeats associated with a number of inherited human conditions are known to be unstable in both the germline and somatic tissues, a finding which serves to explain not only the phenomenon of genetic anticipation characteristic of these disorders but also their inherent inter-individual clinical variability [Giovannone et al., 1997; Leeflang et al., 1999; Martorell et al., 2000; Sharma et al., 2002; Pollard et al., 2004]. However, by contrast, highly variable human minisatellites can display markedly different degrees of instability between the soma and the germline [Buard et al., 2000; Stead and Jeffreys, 2000; Shanks et al., 2008]. These studies notwithstanding, few attempts have so far been made to compare the nature, location and relative frequency of germline and somatic mutations.

Human cancer genes usually harbour either somatic or germline mutations [Goode et al., 2002; Futreal et al., 2004; Vogelstein and Kinzler, 2004]. There is, however, one category of cancer gene, broadly termed tumour suppressors, that by virtue of their being mutated in both the germline and the soma, provides us with an ideal model system to compare somatic vs. germline mutational spectra [Futreal et al., 2004]. Tumour suppressor genes, defined as “genes that sustain loss-of-function mutations in the development of cancer” [Haber and Harlow, 1997], are involved in the regulation of a diverse array of different cellular functions including cell cycle checkpoint control, detection and repair of DNA damage, protein ubiquitination and degradation, mitogenic signalling, cell specification, differentiation and migration, and tumour angiogenesis [Sherr, 2004]. They encode proteins with a regulatory role in cell cycle progression (e.g. Rb), DNA-binding transcription factors (e.g. p53) and inhibitors of cyclin-dependent kinases required for cell cycle progression (e.g. p16). In inherited cancer syndromes, the mutational inactivation of both tumour suppressor alleles is required to change the phenotype of the cell. This ‘two hit hypothesis’ provides the basis for our mechanistic understanding of tumour suppressor gene mutagenesis: a first (inherited) mutation in one tumour suppressor allele is followed by the somatic loss of the remaining wild-type allele via a number of different mutational mechanisms [Knudson, 2001]. Whereas the inherited lesion is usually fairly subtle, the second (somatic) hit may also involve the deletional loss of the entire gene or even a substantial portion of the chromosome involved. Alternatively, both ‘hits’ may constitute somatic mutations: whatever the actual mechanism, the end result is the same – the loss or inactivation of both gene copies. Some interplay may however occur between the soma and the germline in that the location of the germline mutation can in some instances influence the nature, frequency and location of the subsequent somatic mutation [Lamlum et al., 1999; Groves et al., 2002; Latchford et al., 2007; Dallosso et al., 2009; Dworkin et al., 2010].

Tumour suppressor genes are often somatically inactivated by mutational mechanisms that are almost exclusively confined to the soma and which are found only infrequently in the germline (e.g. gross mutations characterized by loss of heterozygosity, epi-mutations such as methylation-mediated promoter inactivation, and micro-lesions within highly repetitive sequence elements that are consequent to microsatellite instability). However, a typical spectrum of somatic mutations associated with tumorigenesis may also include gross rearrangements, copy number variation, and various types of micro-lesion (e.g. micro-deletions, micro-insertions and indels) including single base-pair substitutions [Loeb and Harris, 2008; Stratton et al., 2009]. Although the somatic micro-lesions are often quite similar to their germline counterparts, few studies of tumour suppressor genes have so far attempted to compare and contrast germline and somatic mutational spectra with respect to these relatively subtle types of mutation. However, several such studies have indicated that germline and somatic micro-lesions can display remarkable similarities in terms of mutation type, location and relative frequency of occurrence, and hence by inference the putative underlying mechanisms of mutagenesis [Marshall et al., 1997; Ali et al., 1999; Gallou et al., 1999; Richter et al., 2003; Upadhyaya et al., 2004; Glazko et al., 2004; Tartaglia et al., 2006; Baser et al., 2006; Upadhyaya et al., 2008].

We attempt here a first formal comparison between germline and somatic micro-lesion mutational spectra for a total of 17 different human tumour suppressor genes [*APC* (MIM# 611731), *ATM* (MIM# 607585), *BRCA1* (MIM# 113705), *BRCA2* (MIM# 600185), *CDH1*

(MIM# 192090), *CDKN2A* (MIM# 600160), *NF1* (MIM# 162200), *NF2* (MIM# 607379), *PTCH1* (MIM# 601309), *PTEN* (MIM# 601728), *RB1* (MIM# 180200), *STK11* (MIM# 602216), *TP53* (MIM# 191170), *TSC1* (MIM# 605284), *TSC2* (MIM# 191092), *VHL* (MIM# 608537) and *WT1* (MIM# 607102)].

Materials and Methods

Sources of germline and somatic mutation data

Data on germline and somatic micro-lesions (viz. missense mutations, micro-deletions and micro-insertions involving ≤20 bp) were collated for 17 different human tumour suppressor genes. Germline mutation data were obtained from the Human Gene Mutation Database [HGMD; http://www.hgmd.org; Stenson et al., 2009]. HGMD lists mutations for which there is direct evidence for a pathological effect but includes only one example of every lesion. Apart from this, no specific filters were applied to the available data. Somatic mutation data were compiled from a number of different sources including online somatic mutational databases viz. *Catalogue of Somatic Mutations in Cancer* (http://www.sanger.ac.uk/genetics/CGP/cosmic; *RB1* and *PTEN*), the *Breast Cancer Information Core* (http://research.nhgri.nih.gov/bic; *BRCA1*), the *RB1 Gene Mutation Database* (http://www.verandi.de/joomla; *RB1*), the *International NF2 Mutation Database* (http://www.hgmd.cf.ac.uk/nf2; *NF2*), the *CDKN2A Database* (https://biodesktop.uvm.edu/perl/p16; *CDKN2A*) and the *IARC TP53 Mutation Database* (http://www-p53.iarc.fr; *TP53*), the *VHL Mutations Database* (http://www.umd.be/VHL/), and data privately communicated by Eamonn Maher (*VHL*) and Gareth Evans (*NF2*). Additional somatic mutation data [for *APC*, *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *PTCH1*, *STK11*, *TSC1*, *TSC2* and *WT1*] were obtained by searching PubMed.

To be regarded as *bona fide* somatic mutations, and therefore suitable for inclusion in this analysis, reported lesions had to have been shown not only to be present in a tumour tissue but also to be absent from a non-tumour tissue (usually blood) from the same patient. Hence, mutational data derived from 'sporadic' patients were not included unless a non-tumour tissue had also been examined in order to exclude the possibility that the lesions detected were constitutional in origin. Depending upon the number of independent occurrences, *f*, of a given somatic or shared mutation described in the literature, these mutation types were further subdivided into two categories: *recurrent mutations* (*f*>1) and *non-recurrent mutations* (*f*=1). At the time this study was initiated (October 2006), the number of available germline and somatic missense mutations for each of the 17 studied tumour suppressor genes were as listed in Table 1.

The analysis reported here focussed exclusively on missense mutations and micro-deletions/ micro-insertions. Nonsense mutations in tumour suppressor genes have already been addressed in the context of a general meta-analysis of this type of lesion [Mort et al., 2008]. Indels (complex lesions representing combined micro-deletion micro-insertions) were excluded fro the analysis owing to their paucity.

Deleted: s

Deleted: ed

Deleted: /

Deleted: s

Control datasets of potential mutations

For every tumour suppressor gene examined, all possible single base-pair substitutions in the gene coding sequence that (i) could potentially have given rise to a missense mutation and (ii) were not already included in either of the corresponding observed somatic and/or germline mutational spectra, were generated. These 'potential missense mutations' were used as a control dataset.

For each tumour suppressor gene, a matching control dataset of 'potential micro-deletions' was also generated by randomly selecting a first breakpoint and then choosing the length of the simulated micro-deletion (and hence the position of the second breakpoint) by reference to the probability distribution calculated for micro-deletions (from 1 bp to 20 bp) observed in the corresponding dataset of mutations. A matching dataset of micro-insertions was generated in similar fashion, with the sites of insertion being randomly selected. Since some of the micro-deletion/micro-insertion breakpoints occurred within an intron, extended cDNA sequences comprising exons and additional flanking intronic sequence were used to generate corresponding control datasets.

Grantham scores

The 'Grantham score' or 'Grantham difference' [Grantham, 1974] measures the chemical difference between wild-type and mutated amino acid residues in terms of their side chain composition (i.e. the weight ratio of non-carbon components in end-groups or rings to carbons in side chains), polarity (i.e. basic, acidic or nonpolar depending upon side chain charge) and molecular volume.

On average, the physicochemical differences manifested by orthologous amino acid substitutions that have accumulated over evolutionary time will tend to be relatively small. By contrast, disease-causing substitutions are expected to exhibit higher Grantham scores, indicative of more dramatic physicochemical differences between the wild-type and mutated amino acid residues [Krawczak et al., 1998]. The values tabulated by Grantham [1974] were used in this study to calculate a median Grantham score for each set of missense mutations for each tumour suppressor gene.

Degree of evolutionary conservation

Amino acid residues that are highly conserved in orthologous proteins frequently represent sites of structural or functional importance. Hence, such highly conserved amino acid residues/protein regions often constitute hotspots for observed pathological mutations as a consequence of phenotype selection (rather than intrinsic mutability). To assess the degree of evolutionary conservation of those codons affected by somatic/germline mutations, orthologous tumour suppressor cDNA and protein sequences from different vertebrate species were retrieved from NCBI’s Entrez Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene). The species used as a source of these cDNA and protein sequences are listed in Supp. Table 1 for each tumour suppressor gene/protein. ClustalX (http://www.clustal.org/) was used to align the protein sequences. A program was written to replace all amino acids in the protein alignments by cDNA-derived codons, thereby avoiding the introduction of gaps within codons.

The evolutionary constraints acting upon the 17 human tumour suppressor genes at the codon level were inferred by calculating the $\frac{Ka}{Ka+Ks}$ ratio for each codon where Ks and Ka are respectively the relative numbers of synonymous and nonsynonymous substitutions between codons in two aligned sequences [Walker et al., 1999]. If two aligned codons required more than one substitution to be transformed into each other, then the minimum number of substitutions was assumed, and the most parsimonious path was determined using a PAM100 matrix and the Nei & Gojobori [1986] pathway method. Gaps inserted into the non-human vertebrate orthologous cDNA sequences during alignment were treated as being equivalent to a non-synonymous substitution. Codons that were not present in the human cDNA sequence were not considered. A value representing the median level of evolutionary conservation across all codons was then derived for each mutational spectrum; the higher values correspond to less conserved genes whereas the lower values refer to more highly conserved ones.

Relative mutability rates

To assess the likelihood of observing a certain nucleotide change in a given position and in a specific context, two tabulated measures of the nearest neighbour-dependent mutation rate were employed. The first was derived from 20,200 single base-pair substitutions inferred from alignments of paired human gene/pseudogene sequences [Hess et al., 1994]. This was termed the *non-disease-associated mutability rate* and, since it approximates to the neutral mutation frequency, it should reflect the intrinsic mutability of the underlying DNA sequence. One would expect the non-disease-associated mutation rates associated with pathological mutations to be low implying that these specific substitutions are much less likely to occur as neutral substitutions.

The nearest neighbour-dependent mutation rates derived from germline single base-pair substitutions [using data from the Human Gene Mutation Database (HGMD); Stenson et al., 2009] by Krawczak et al. [1998] were used as an approximation of the *disease-associated mutability rate*. This mutation rate is a function of selection for loss of biological function as well as the underlying intrinsic mutability of the DNA sequence. This mutability rate varies between 0.032 for the C(A>T)G mutation and 13.023 for the C(G>A)G mutation [Krawczak et al., 1998].

Repetitive sequence elements

A variety of repetitive sequence elements have been reported in association with human gene mutations causing both inherited disease and cancer. Direct and inverted repeats and symmetric elements [see Chuzhanova et al. 2003 for definitions] of length ≥8 bp, and less than 21 bp apart, capable of forming non-B DNA structures, were therefore sought within the extended cDNA sequences (comprising exons and up to ±85 bp of flanking sequence) using purposely designed software. In addition, DNA sequences were screened for the presence of mononucleotide runs of ≥4 bp.

Mutation descriptors

Each missense mutation was ascribed various descriptors indicating (a) the type of mutation [i.e. shared mutation (i.e. found to occur both somatically and in the germline); exclusively somatic; exclusively germline; shared recurrent mutation (i.e. found to occur not only in the germline but also somatically on more than one occasion; somatic recurrent mutation (recorded in the soma more than once, but not in the germline); potential mutation (as defined above)] and (b) its location [i.e. C→T and G→A within a CpG dinucleotide or within a CpHpG trinucleotide (where H=A, C or T) or in a repeat sequence (as described above)]. Mutations that have been reported as being exclusively somatic or exclusively germline will henceforth be referred to simply as 'somatic' and 'germline', respectively. The shared mutations, comprising the overlap between the somatic and germline mutations, may be visualized in the form of a Venn diagram (Figure 1). All somatic missense (including shared) mutations were further described as being either recurrent or non-recurrent (in the soma, see above; Figure 1). No such division was made for the relatively small number of recurrent micro-deletions and micro-insertions available; both recurrent and non-recurrent somatic mutations were therefore included in either the somatic or the shared datasets and labelled accordingly (Figure 1).

All micro-lesions (*viz.* missense mutations, micro-deletions and micro-insertions) in each gene were also labelled with respect to their occurrence within a region spanning a repetitive element or mononucleotide run including ±5 bp of flanking sequence. If a missense mutation (or at least one micro-deletion/micro-insertion breakpoint) was found to occur within this extended region, the micro-lesion was labelled as being found in association with the corresponding type of repeat.

Assessing the statistical significance of the results generated

To assess the similarity (or dissimilarity) of the germline and somatic mutational spectra with respect to (i) the frequency with which the missense mutations were located within CpG/non-CpG dinucleotides or CpHpG/non-CpHpG trinucleotides and (ii) the frequency with which the micro-deletions/micro-insertions were found within/outwith repeats, the various non-overlapping mutation datasets (bearing specific descriptors) were compared by means of the χ^2 test. Since the normality assumption did not hold for the datasets studied, the Wilcoxon rank-sum test was used to compare and contrast missense mutational spectra with respect to the Grantham score, degree of evolutionary conservation, and both the non-disease- and disease-associated mutability rates.

The permutation-based method [Olshen and Jain, 2002] was used to estimate the significance of our findings and to allow for multiple testing wherever appropriate. For each comparison, the null hypothesis [viz. no overall difference between two groups of mutations (e.g. somatic and potential) with respect to the specific property in question (e.g. occurrence in CpG or non-CpG nucleotides)], was tested for, either in the context of each gene or all genes combined. χ^2 or rank-sum statistics were calculated for the observed germline and somatic mutations as well as for 10,000 control sets of mutations created from the original sets by random permutation of the assigned mutational descriptors (e.g. randomly chosen mutations labelled as 'somatic' were re-labelled as 'germline'; randomly chosen mutations labelled as 'shared' were re-labelled as 'somatic', etc.). The test statistic (χ^2 or rank-sum) for the original datasets that exceeded the 95th percentile of χ^2 maxima for 10,000 control sets was deemed to be statistically significant; the corresponding p-value was termed the 'gene-wise' p-value. To allow for multiple testing in those cases where specific mutations in all genes were combined, a Bonferroni correction was applied; the corresponding p-value was termed the 'experiment-wise' p-value.

Power calculations for the χ^2 tests were performed using the Pwr.Chisq.test package, part of the R Statistical Language (http://cran.r-project.org/). A data based simulation method [Walters 2004] was used to perform power calculations for the Wilcoxon rank-sum tests. Only results showing ≥80% power to detect experiment- or gene-wise significance were reported.

Naïve Bayes classifier

A decision tree classifier known as a Naïve Bayes tree [NBTree; Kohavi, 1996], implemented in the Weka machine learning package [Witten and Frank, 2005], was trained to discriminate between somatic, germline, shared, recurrent somatic and recurrent shared missense mutations. Each mutation was described by a total of six features including the degree of evolutionary conservation, the non-disease-associated and disease-associated relative mutability rates, Grantham score, and occurrence in CpG/CpHpG, non-CpG/non-CpHpG doublets/triplets or in repeats/mononucleotide runs. Ten-fold cross-validation was used to assess the accuracy of classification. The mutation datasets were balanced using random oversampling [Kotsiantis et al., 2006] by replicating random instances from the minority classes until all classes were represented by the same number of instances as the majority class.

Results and Discussion

The availability of both germline and somatic mutational spectra from tumour suppressor genes provides us with an ideal opportunity to study the nature of mutation of the same gene sequences in both the germline and the soma. The analysis reported here explores for the first time the similarities and differences exhibited by the germline, somatic (and shared) micro-lesion mutational spectra in 17 human tumour suppressor genes. The study presented here focussed upon missense mutations and micro-deletions as well as micro-insertions. Nonsense mutations in tumour suppressor genes have already been addressed elsewhere in the context of a general meta-analysis of this type of lesion [Mort et al., 2008].

Characteristics of germline and somatic missense mutations with respect to mutation type

Taken together, the combined mutational spectra for all 17 tumour suppressor genes containe twice as many somatic (61%) as germline (31%) mutations. Further details are provided in th Supplementary Text online.

Deleted: For five genes (*APC*, *CDKN2A*, *NF2*, *PTEN* and *TP53*), a predominance of somatic over germline mutations was noted, with the *TP53* gene having the highest proportion of somatic mutations (92%). For the majority of genes, however (namely *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *PTCH1*, *RB1*, *STK11*, *TSC1*, *TSC2*, *VHL* and *WT1*), the analysed dataset included more germline than somatic mutations, with >97% of all mutations in the *BRCA1*, *NF1*, *TSC2* and *WT1* genes being germline in origin.

Shared mutations are of particular interest because identical mutational mechanisms operating in the germline and the soma may be inferred for such lesions. The expected number of shared mutations for each gene was calculated as $p_{\text{somatic}} \times p_{\text{germline}} \times (\text{total number of mutations})$, where *p* denotes the relative frequencies of somatic and germline mutations. Although the proportion of shared mutations varies markedly between genes (from 0% to 25% of the total), only two genes (*TP53* and *VHL*) were found to have a higher than expected number of shared mutations as calculated above.

Patterns of germline and somatic missense mutations by mutation type

Missense mutations were characterised by a predominance of transitions over transversions (Figure 2). The transition:transversion ratio was at its highest for shared recurrent mutations (3.5) and shared non-recurrent mutations (2.7). By contrast, the transition:transversion ratio for the control group (i.e. potential mutations) was 0.85. Significant differences in the transition:transversion ratio were observed between all mutation types ($p<0.05$) with the exception of germline vs. shared mutations (Figure 2).

Not surprisingly, a strong positive correlation was noted between somatic and shared mutational spectra (Pearson's correlation $r=0.986$, $p= 2.91\times10^{-4}$) with respect to the frequencies of six mutational changes viz. A.T>C.G, A.T>G.C, A.T>T.A, C.G>A.T, C.G>G.C and C.G>T.A. Weaker negative correlations were found between somatic mutations and the control dataset of mutations ($r= -0.887$, $p=0.019$) and between the shared and control ($r= -0.837$, $p=0.038$) mutational spectra, indicative of the non-randomness of somatic mutation.

C.G>T.A transitions constituted the most frequent type of mutation in shared (46%), germline (29%) and somatic (25%) mutational spectra, significantly higher proportions than noted in the spectrum of mutations within our control dataset (13%, $p<0.001$) (Figure 2). Intriguingly, the number of A.T>G.C mutations was significantly higher (28%) in the germline as compared to the somatic (16%), shared (17%) and control (16%) mutational spectra (Figure 2). A.T>C.G mutations were significantly under-represented in the shared mutational spectrum (7%, $p<0.001$) as compared to the other spectra whereas A.T>T.A mutations were under-represented (7%, $p<0.001$) in both the germline and shared mutational spectra compared to both somatic and potential mutations (Figure 2). Finally, C.G>A.T mutations were significantly underrepresented in the germline mutational spectrum (10%) as compared to the somatic (16%, $p=1.2\times10^{-5}$) and potential (15%, $p=2.6\times10^{-5}$) spectra. Thus, the main similarity between the somatic and germline missense mutational spectra was in relation to C.G>T.A transitions whereas the main differences between these spectra involved the A.T>G.C, A.T>T.A and C.G>A.T mutations. In passing, it should be noted that the patterns of somatic nucleotide substitution exhibited by the 17 tumour suppressor genes studied here were markedly different from the genome-wide patterns of somatic nucleotide substitution observed in various cancer genome sequencing studies [Sjöblom et al., 2006; Greenman et al., 2007; Kan et al., 2010]; these mutation datasets are likely to differ quite dramatically with respect to their relative proportions of 'passenger' mutations.

Deleted: previously

CpG- and CpHpG-located missense mutations

The CpG dinucleotide is a well known mutational hotspot in the human genome as a consequence of the spontaneous (and endogenous) deamination of 5-methylcytosine. In addition, Lister et al. [2009] reported abundant DNA methylation in CpHpG trinucleotides in the human genome, where H is either A, C or T, raising the possibility that CpHpG might also be a generalized mutation hotspot [Cooper et al., 2010].

The proportion of missense mutations that were either C>T or G>A within CpG or CpHpG oligonucleotides in the 17 tumour suppressor genes was found to vary between 0% and 100% (Table 2). This wide range in values may be attributed to the small size of some of the gene mutation datasets under study. Importantly, the CpG and CpHpG oligonucleotides were found to be disproportionately likely to harbour shared mutations; thus, 34% of shared recurrent mutations and 21% of shared non-recurrent mutations were C>T and G>A mutations in CpG dinucleotides with an additional 10% and 9% of mutations, respectively, occurring within CpHpG trinucleotides. Since driver mutations tend to occur disproportionately frequently within CpG dinucleotides [Talavera et al., 2010], we postulate that missense mutations identified as being shared are highly likely to be driver mutations.

Significant differences were noted between the relative frequencies of CpG- and CpHpG-located mutations for somatic, germline, shared, somatic recurrent and shared recurrent missense mutations (Supp. Table 2).

We have previously shown that 18.2% and 9.9% of all missense/nonsense mutations recorded in the HGMD are C>T and G>A transitions in CpG and CpHpG oligonucleotides respectively [Cooper et al., 2010]. In the present study, we observed that the mutational spectra of shared and shared recurrent missense mutations in tumour suppressor genes were both found to be significantly enriched in CpG-located mutations (χ^2-test; p-values, 0.028 and 1.1×10^{-9} respectively). This implies that the CpG dinucleotide is a generalized mutation hotspot in both the soma and the germline as a consequence of the endogenous mutational mechanism of methylation-mediated deamination of 5-methylcytosine. By contrast, the number of CpG-located mutations was significantly underrepresented (χ^2-test; p-values$<5\times10^{-14}$) in the other mutational spectra (i.e. non-recurrent somatic, somatic recurrent and germline mutations) by comparison with HGMD data. To perform these comparisons, missense mutations (Table 2) and nonsense mutations [previously reported in Mort et al., 2008; see Table 6 therein] in all 17 tumour suppressor genes were combined. The proportion of shared recurrent missense mutations in tumour suppressor genes that were CpHpG-located was found to be significantly higher (p=0.023) than for mutations recorded in the HGMD whereas CpHpG-located somatic and recurrent somatic mutations were significantly under-represented ($p<4\times10^{-10}$). Significant enrichment in CpHpG-located mutations was observed for germline mutations as compared to somatic mutations ($p<3\times10^{-10}$) consistent with the reported decrease in CpHpG methylation in differentiated cells [Lister et al., 2009]. In summary, germline and shared missense mutations were found to be significantly enriched at CpG and CpHpG oligonucleotides.

The numbers of somatic and shared C>T and G>A transitions recorded within CpG dinucleotides for each gene (Table 2) did not correlate with the numbers of CpG dinucleotides found in these genes (r <–0.5, p>0.127) and hence do not simply reflect intragenic CpG frequency. A weak positive correlation between CpG-located mutations and the number of genic CpG dinucleotides was however noted for germline mutations (r= 0.489, p=0.046) indicating that CpG methylation is not entirely unrelated to the number of CpG dinucleotides, at least with respect to the germline; the relationship is however clearly more complex in the soma, possibly due to inter-tissue differences in gene methylation patterns [Tornaletti and Pfeifer, 1995] or transcription-coupled repair [Rubin and Green, 2009].

No correlation was found between the numbers of somatic, germline and shared mutations recorded within CpHpG trinucleotides and the corresponding numbers of CpHpG trinucleotides for these genes (r= -0.316, 0.373, -0.414; p-values 0.281, 0.216 and 0.098, respectively) indicating that mutation within CpHpG trinucleotides is likely to be very much a gene-specific phenomenon (presumably dependent on both the extent and the degree of spatial localization of CpHpG methylation in the germline and/or soma).

Finally, the number of CpG dinucleotides in the various tumour suppressor genes studied (Table 2) was not found to correlate with gene length (r= 0.3, p-value=0.241). By contrast, we found a significant correlation (r= 0.885, p-value=2.35×10^{-6}) between tumour suppressor gene length and the number of CpHpG trinucleotides (excluding those with mutations), indicating that the tumour suppressor genes under study possess a similar density of CpHpG trinucleotides per unit length. We surmise that the factors that govern the establishment of the methylation pattern of CpHpG trinucleotides are likely to be quite complex.

Evolutionary conservation of tumour suppressor genes *in relation to the sites of somatic and germline missense mutations*

For all 17 tumour suppressor genes, the degree of evolutionary conservation, as measured by Ka/Ks, was less than unity, indicating that these genes (and proteins) have been highly conserved evolutionarily as a consequence of the action of purifying selection. Indeed, the degree of evolutionary conservation displayed by most of the studied genes was markedly lower than the average (~0.18) noted in a comparison of 1880 human, rat and mouse gene orthologues [Makalowski and Boguski, 1998]. However, three genes (*CDKN2A, BRCA1* and *BRCA2*) were found to exhibit a higher rate of evolutionary conservation than the average between human and rodents.

The evolutionary conservation of each mutated codon was inferred by calculating the $\frac{Ka}{Ka+Ks}$ ratio; for each gene/spectrum, the mean value was then calculated across all mutations in the corresponding gene/spectrum. Shared recurrent missense mutations were found to occur disproportionately in highly conserved amino acid residues (mean degree of evolutionary conservation, 0.072) followed by shared non-recurrent mutations (0.138), somatic recurrent (0.169), germline (0.175), non-recurrent somatic (0.265), and control dataset mutations (0.255). The observed differences in the degree of evolutionary conservation for the different mutational spectra are shown in Supp. Table 2. These quite specific findings are consistent with the previously reported general tendency for cancer-associated mutations to occur frequently at evolutionarily conserved sites [Greenblatt et al., 2003; Tavtigian et al., 2009; Talavera et al., 2010].

Somatic non-recurrent mutations were found to occur in codons characterized by the highest mean value of $\frac{Ka}{Ka+Ks}$ ratios as compared not only to the shared recurrent and shared non-recurrent mutations (see above) but also to the mutations within the control dataset. This is consistent with the interpretation that a high proportion of non-recurrent somatic mutations, and most notably those which are located in less evolutionarily conserved regions (characterised by higher values of the degree of evolutionary conservation), are likely to be ‘passenger’ mutations.

Missense mutations in relation to the disease- and non-disease-associated substitution rates

Employing alignments of paired human gene/pseudogene sequences, Hess et al. [1994] derived relative (non-disease-associated) nearest-neighbour-dependent mutability rates using the lowest frequency substitution type, C(T>G)A/T(A>C)G, as a baseline. These mutability rates were found to vary over a 52-fold range, with unity being assigned to the lowest frequency substitution type. This *non-disease-associated* mutability rate approximates to the neutral mutation frequency and hence reflects the intrinsic mutability of the underlying DNA sequence. Depending upon the observed nearest-neighbour context, we retrieved the corresponding non-disease-associated mutability rate (from the data of Hess et al. 1994) for each mutation (either observed or from the control dataset) and calculated the median value for each mutational spectrum. These median values are indicative of the relative mutability of each tumour suppressor gene. Further details are provided in the Supplementary Text online.

Deleted: The median values were found to vary between 4 (*NF2*) and 8.9 (*STK11*) for somatic mutations, 4.1 (*TP53*) and 10.1 (*WT1*) for germline mutations, and 7.2 (*RB1*) and 11 (*PTEN*) for shared mutations (values given only for genes with more than three mutations in the corresponding category; see Supp. Table 3, indicating that many of the median values are quite low and hence the corresponding mutations are unlikely to be neutral.¶

When data from all 17 genes were combined, shared recurrent mutations were found to be characterised by intrinsically low non-disease-associated mutability (median=11), followed b even lower median mutability values for shared non-recurrent mutations (7.9), germline mutations (7.2), somatic recurrent and non-recurrent (4.7) and control dataset mutations (4.1). Such low median mutability values across all groups indicates that at least half of the mutations within observed triplets are unlikely to be neutral in the sense defined by Hess et al. [1994] and hence are not simply explicable in terms of intrinsic DNA mutability. The low median mutability values for the control dataset of mutations within tumour suppressor genes reflect the high level of evolutionary conservation manifested by tumour suppressor gene coding sequences across different species, implying that any mutation within a triplet characterized by a low non-disease-associated mutation rate is very likely to have pathological consequences and would thus be subject to purifying selection.

In contrast to the non-disease-associated mutability rate (which is purely a reflection of the intrinsic DNA mutability), the disease-associated mutability rate reflects (in addition to the intrinsic DNA mutability) the increased likelihood of coming to clinical attention conferred by the loss of biological function. The C(G>T)T mutation is one of the most frequent types of mutation associated with the loss of biological function [disease-associated mutability rate 10.255; Krawczak et al., 1998] but occurs much less frequently among neutral mutations [non-disease-associated mutability rate 4.4; Hess et al., 1994].

For each tumour suppressor gene and each mutational spectrum, the disease-associated median mutability values were calculated using mutability rates derived from Krawczak et al. [1998]. The disease-associated median value was found to be 0.85 for the germline mutations. Further details are provided in the Supplementary Text online. We found that shared recurrent and shared non-recurrent mutational spectra were characterized by higher median values of the disease-associated mutability rates (1.42 and 1.01 respectively) whereas somatic non-recurrent, somatic recurrent and control dataset mutations exhibited lower median mutability rates (0.5, 0.5 and 0.4 respectively) as compared to germline mutations (0.85). The finding that the shared mutations (which, by definition, occur in both the germline and the soma) are characterized by higher disease-associated mutability rates is not surprising since mutations that occur with the highest probability are among those most likely to be shared.

Deleted: The highest and lowest disease-associated median values for the mutation rates were noted for somatic mutations in the *STK11* gene (1.7; Supp. Table 3) and for germline mutations in the *TP53* (0.42) gene (values given only for genes with more than three mutations in the corresponding category).

We postulated that those mutations which occur both in the germline and the soma, and which are characterised by higher disease-associated mutability rates are disproportionately likely to be drivers of tumour development. Consistent with this postulate, somatic recurrent and non-recurrent mutational spectra are characterized by lower median disease-associated mutability rates as compared to the germline spectrum. However, given that higher disease-associated mutability rates are a characteristic feature of driver mutations, a certain proportion of the somatic mutations, namely those characterised by higher disease-associated mutability rates, may correspond to functionally significant driver mutations.

In assessing the significance of our results, it was appropriate to consider the possibility that somatic mutations might display quite different nearest-neighbour-dependent disease-associated mutability rates from germline mutations. However, since a good correlation was observed between the mutability rates derived from inherited disease data [Krawczak et al., 1998] and the neighbour-dependent mutability rates calculated for the somatic mutations of the 17 tumour-suppressor genes studied here (Pearson's correlation r=0.703, $p=6.6\times10^{-30}$), this *caveat* appears not to be an issue.

Distribution of Grantham scores with respect to tumour suppressor gene mutations

Shared recurrent mutations were found to exhibit the largest median chemical difference value (Grantham scores) between the wild-type and mutated amino acid residues (100) followed by shared non-recurrent mutations and germline mutations (both 93), somatic recurrent (85), somatic non-recurrent (80) and potential mutations (78). Since there was an obvious trend for shared recurrent and non-recurrent mutations to cause the most dramatic chemical changes of the affected codon, we may infer that these types of lesion are also more likely to be driver mutations. However, bearing in mind that the range of theoretically possible values varies between 5 (Leu ↔ Ile) and 215 (Cys ↔ Trp), less elevated median values may simply indicate that a proportion of the mutations in each mutational spectrum are likely to be chemically less dramatic (Grantham scores <100).

Missense mutations occurring within repeats and runs of identical nucleotides

A number of studies have noted that single base-pair substitutions associated with inherited disease occur disproportionately either within, or in close proximity to, repetitive sequences [Jego et al., 1993; Greenblatt et al., 1996; Tappino et al., 2009; Thomas et al., 2010; Leclercq et al., 2010]. Hence, we wished to assess whether either germline or somatic mutations occurred disproportionately either within, or in the vicinity (see *Mutation descriptors*) of, direct, inverted and symmetric repeats or mononucleotide runs in the 17 tumour suppressor genes under study (Table 3, Supplementary Tables 4-6).

On average, direct repeats of length ≥8 bp were found to cover 5.6% of the cDNA lengths of the 17 tumour suppressor genes. Further details are provided in the Supplementary Text onlin

Deleted: , the coverage varying between 2.5% (*BRCA2*) and 17% (*PTEN*) of the respective gene sequences. The corresponding proportion of the cDNA lengths for inverted repeats ≥8 bp was 8.5%, with proportions varying between *PTCH1* (4.5%) and *RB1* (15.7%) while symmetric elements ≥8 bp were found to encompass 25% of the cDNA lengths (varying between 15.5% for *APC* and 44% for *PTEN*). ¶

On average, mononucleotide runs ≥4 bp spanned 19.9% of the cDNA lengths. Approximate 24% of non-recurrent somatic and 20% of germline missense mutations were found in mononucleotide runs; these proportions were significantly higher than noted for shared non-recurrent missense mutations (4.9%, $p \leq 1.6\times10^{-4}$). A greater proportion of non-recurrent somatic missense mutations was found in direct repeats (7%) as compared to recurrent somatic missense mutations (2%, $p=8.8\times10^{-7}$), germline missense (4%, p=0.028) and potential missense mutations (3.7%, $p=8.1\times10^{-7}$). This result may reflect the disproportionate number of CpG/CpHpG mutations among shared and recurrent somatic missense mutations. Further, for all mutational spectra examined (with the exception of the shared mutations), missense mutations were preferentially found in association with inverted and symmetric repeats as compared to the control dataset of mutations (p<0.05). However, no statistically significant differences were found between mutational spectra. Further details are provided in the Supplementary Text online.

Deleted: , varying between 9.5% (*VHL*) and 29% (*TP53*)

Deleted: No correlation was observed between the number of mutations located within repeats and the fractional length of the cDNA covered by repeats, indicating that not every repeat sequence is mutation-prone. However, a strong correlation between the fractional length of the cDNA covered by repeats and cDNA length of genes ($r > 0.87$ and $p<10^{-6}$) served to demonstrate that repeat density per unit length was approximately the same for all tumour suppressor genes studied. ¶

Towards a classification of somatic and germline missense mutations

All observed mutations within each mutational spectrum were re-categorized (Supp. Table 7) with respect to the location of mutations within CpG/CpHpG oligonucleotides, within different types of repeat/mononucleotide runs, within both CpG/CpHpG oligonucleotides and repeats. 4×2 contingency tables were then used to measure the strength of the pairwise associations between the various mutational distributions presented in Supp. Table 7, the significance of the associations being assessed by means of a Chi-square test. Significant ($p<0.002$) pairwise differences were noted between somatic and germline, somatic and shared, and between germline and shared mutational spectra ($p<0.002$) with respect to the features listed above and each of four types of repeat, indicating that these features have great discriminant potential.

All somatic, germline, shared non-recurrent, recurrent somatic and shared recurrent missense mutations (each described by a combination of different features (i.e. degree of evolutionary conservation, non-disease- and disease-associated mutability rates, Grantham score, CpG/CpHpG location, occurrence within repeat/mononucleotide run) were then used to train a Naïve Bayes Tree classifier. On average, 63.1% of somatic, germline, shared, recurrent somatic and shared recurrent mutations were correctly classified [the area under the Receiver Operating Characteristic (ROC) curve being 0.869, indicating a reasonably good classification] with 71% and 75% respectively of shared and shared recurrent mutations being correctly recognized implying that the mutation groupings differ with respect to the different features in a consistent fashion. One would expect 20% of mutations to be assigned to each of the five groups by chance alone. Indeed, the average percentage did not exceed 20% when randomly selected datasets matching the number of somatic, germline, shared, recurrent somatic and shared mutations were drawn from the set of potential mutations; the average was taken over 10 matching datasets. The complete Naïve Bayes Tree classifier is depicted in Supp. Figure 1.

An additional non-overlapping dataset of 568 missense somatic mutations, identified in the 17 tumour suppressor genes under study, were extracted from a collection of 2,488 mutations identified as being probable driver mutations [Carter et al., 2009]. Features such as the degree of evolutionary conservation, Grantham score, mutability rates, CpG/CpHpG location, occurrence within repeats/mononucleotide runs were again determined for each of these mutations. Employing our classifier, 7% and 10% respectively of these 568 mutations were found to possess features consistent with their being shared recurrent and shared non-recurrent mutations. In addition, 32% of these probable driver mutations were found to bear features characteristic of recurrent somatic mutations (i.e. mutations documented in different tumours). A further 25% of the probable (somatic) driver mutations were classified as possessing features characteristic of germline mutations and hence could conceivably be treated as shared mutations missing from the original training dataset. The remaining 25% of mutations were classified as non-recurrent somatic mutations. Using this classifier, which is based on a very modest number (6) of predictive features, to analyse an independent dataset of probable driver mutations, we were able to predict that ~50% of these somatic missense mutations exhibited features specific to either shared or recurrent mutations, indicating that a disproportionate number of such lesions are likely to be drivers of tumorigenesis. This percentage is certainly lower (79%) than that obtained by Carter et al., [2009] through the application of a Random Forest Classifier based on 500 trees and >50 predictive features (using an ‘out-of-the-bag’ error estimate similar to the cross-validation procedure) to the set of putative 2,488 driver mutations. However, based on the results of this study, we may conclude that, in general, the mutational spectrum of driver mutations is likely to contain a disproportionate number of somatic mutations that have germline counterparts (~17%) whilst an additional 32% of the driver mutations are likely to occur recurrently in the soma.

Occurrence of micro-deletions and micro-insertions within repeats and runs of identical nucleotides

The mutational spectrum of micro-deletions, combined for all 17 tumour suppressor genes, comprised 55% germline, 43% somatic and 2% shared mutations. The mutational spectrum o micro-insertions was similar to that of micro-deletions and comprised 60% germline, 38% somatic and 2% shared mutations. Approximately 77% somatic, 87% germline and 91% shar micro-deletions and micro-insertions were ≤4 bp in length. Strong (r = ~1) correlations were noted between the distributions of micro-deletions and micro-insertions with respect to the le of the deleted/inserted fragments, both gene-wise and for all genes combined (r>0.9, $p<10^{-8}$) for all mutational spectra.

Recent studies have revealed that simple repetitive DNA sequences are not only capable of adopting non-B DNA conformations and are highly mutagenic [Bacolla et al., 2004; Bacolla and Wells, 2004; Chuzhanova et al., 2009]. Indeed, both direct repeats and mononucleotide runs have long been known to be mutation hotspots in the *TP53* gene [Jego et al., 1993; Greenblatt et al., 1996]. The number of micro-lesions occurring in the vicinity (see *Mutation descriptors*) of direct, symmetric and inverted repeats (capable respectively of slipped, triplex and cruciform non-B structure formation), or within mononucleotide runs (which often mediate micro-deletions/micro-insertions) were therefore determined. The number of mutations found in the vicinity of all three types of repeat, and within mononucleotide runs, are given in Tables 3 and Supp. Tables 4-6.

The highest proportion of mutations in mononucleotide runs was found for the shared (39%), germline (30%) and somatic (25%) mutational spectra. Significant differences were observed between shared and germline (p=0.0002), somatic and shared (p=0.045), and between all mutational spectra and potential mutations (p<0.0001) with respect to their occurrence within mononucleotide runs, confirming that these simple repeats constitute an important hotspot for

Deleted: *Truncating vs non-truncating mutations in the germline and soma¶*
Somatic mutational spectra from the *BRCA2, CDKN2A, STK11, TP53* and *TSC1* genes were characterized by the predominance of non-truncating (i.e. missense) lesions over truncating lesions (i.e. nonsense mutations, frameshift micro-deletions, micro-insertions and indels) when nonsense mutations [reported in Mort et al. (2008)] and micro-indels (excluded from previous analyses) were also considered (Supp. Table 8). A similar predominance of non-truncating over truncating lesions was observed for the germline mutational spectra of the *CDKN2A, TP53, VHL* and *WT1* genes. In general, the ratio of non-truncating to truncating lesions was found to be significantly higher in the soma (0.85) than in the germline (0.30; p-value<2.20E-16). All other mutational spectra were characterized by the predominance of truncating mutations. ¶
¶

micro-deletions and micro-insertions in both the soma and the germline. The preponderance of such mutations in mononucleotide runs is unsurprising in the context of the shared mutations since all mutations that occur with high frequency within mutation hotspots are more likely to be shared between the germline and the soma (as previously noted for CpG and CpHpG mutations). No other types of repeat were disproportionately associated (after correction for multiple testing) with micro-deletions and micro-insertions.

Regional hotspots in somatic and germline mutational spectra

For the purposes of the following analysis, a regional mutation hotspot was defined as a stretch of DNA of length ≤20 bp where four or more independent mutational events have been reported and a significant degree ($p \leq 0.05$) of clustering of these mutations was evident for a given stretch of DNA. In this definition of a regional hotspot, each recurrent mutation was considered only once. The order statistics, r-scans, as described by Karlin and Macken [1991] and applied in Bacolla et al. [2006], were used to detect significant clustering of mutations by comparison with a Poisson distribution of mutations along the gene sequence. Overlapping hotspot regions were considered as a single regional hotspot.

The only regional mutational hotspot for somatic missense mutations was observed in the *PTEN* gene and comprised 18 mutations in the region between nucleotide positions 269 and 286. Several germline regional mutational hotspots were however detected for missense mutational spectra in the *ATM*, *BRCA1*, *BRCA2*, *NF1*, *PTEN*, *RB1*, *STK11*, *TP53* and *WT1* genes (Table 4). Several somatic regional mutational hotspots were found for micro-deletions/micro-insertions in the *APC* gene, the largest of which contained 33 mutations (positions 4303-4398) and forms part of a previously reported mutation cluster region [Miyoshi et al., 1992]. Regional hotspots identified in different mutational spectra were however unique to that spectrum. The only overlap noted between regional mutational hotspots identified in germline and somatic micro-deletion/micro-insertion mutational spectra was observed for the *APC* gene (the overlapping region comprising nucleotide positions 3919-3933). This micro-deletion/micro-insertion hotspot also includes codon 1309 (cDNA positions 3925-3927) found to be frequently mutated in Greek and French patients with familial adenomatous polyposis [Fostira et al. 2010; Lagarde et al. 2010].

Inspection of regional hotspot sequences revealed that they are rich in repetitive elements, runs of identical nucleotides and CpG/CpHpG oligonucleotides, offering immediate explanations for the elevated mutability.

Germline and somatic mutations located within specific hotspot motifs

The cDNA sequences of 17 tumour suppressor genes were screened for the presence of nine specific motifs (and their complements) previously reported as being hotspots for mutation. These motifs included the putative somatic (cancer) mutation hotspot, WKVNRRRNVWK [the 'THEMIS motif'; Makridakis et al., 2009], the RGYW motif that correlates with the DNA polymerase eta error spectrum [Rogozin et al., 2001] and several so-called 'super hotspot' motifs originally found in germline micro-insertions and micro-deletions [Ball et al., 2005] and indels [Chuzhanova et al., 2003]. For the purposes of this analysis, the shared mutations were added to both the germline and somatic mutational spectra. Both germline and somatic micro-deletions and micro-insertions were found to be significantly overrepresented ($p \leq 0.002$) in the 'indel super hotspot' motif GTAAGT and its complement. Somatic micro-deletions and micro-insertions were also significantly overrepresented ($p=0.009$) with respect to the micro-deletion/micro-insertion super hotspot AAATCT and its complement. The number of germline (but not somatic) micro-deletions/micro-insertions in the THEMIS motif were significantly overrepresented ($p=0.003$) as compared to the controls. No significant difference was however observed in the number of missense mutations occurring in any motifs analysed.

Conclusions

Several conclusions may be drawn from the results reported here. Firstly, it would appear that those missense mutations that are found both in the soma and the germline ('shared mutations') are disproportionately more likely to exert an effect on tumour development and/or progressi (i.e. more likely to be driver mutations) than exclusively somatic non-recurrent missense mutations (at least for the *TP53* and *CDKN2A* genes whose mutations contributed the bulk of the documented shared mutations in our tumour suppressor gene mutation dataset). Shared mutations also occur preferentially in CpG/CpHpG oligonucleotides and are characterised by higher mutability rates (both non-disease- and disease-associated). Further, we found that shared mutations tend to occur in those codons that have been more highly conserved evolutionarily, and are associated with more dramatic chemical differences between the substituted (wild-type) and substituting amino acids. Taken together, it would thus appear that shared mutations are influenced to a greater extent by the local nucleotide sequence context than either germline or somatic non-recurrent missense mutations. Since this implies that shared mutations (the mutation category most likely to harbour driver mutations) have a tendency to arise through the action of similar endogenous mutational mechanisms, we may infer that endogenous mechanisms of mutagenesis exert a disproportionate effect on tumorigenesis.

Deleted: A number of important

Deleted: profound

Deleted: s

In an analysis of an unrelated dataset, we demonstrated that 17% of somatic missense mutations previously identified as being probable drivers [Carter et al., 2009] were found to possess the same features as shared (both recurrent and non-recurrent) mutations. A further 32% of these probable driver mutations shared the features expected of recurrent somatic mutations. Thus, we may conclude that ~50% of these somatic missense mutations possess features consistent with their being either shared or recurrent, suggesting that a disproportionate number of such lesions are likely to be drivers of tumorigenesis.

A sizeable proportion of shared (39%) and germline (30%) micro-lesions were found to be located in runs of identical nucleotides ≥4 bp, making mononucleotide runs a hotspot for micro-deletion and micro-insertions. The most likely underlying causative mechanism for these mutations is slipped mispairing at DNA replication mediating duplications and 'de-duplications' [Kondrashov & Rogozin, 2004]. With regard to missense mutations, CpG and CpHpG oligonucleotides were found to be hotspots for shared recurrent and shared non-recurrent missense mutations; 34% (10%) and 21% (9%) of respective mutations were found in CpG (CpHpG) oligonucleotides. Further, 12% of the 568 probable driver mutations [derived from Carter et al., 2009] were found to occur in CpG/CpHpG oligonucleotides. 41% of probable driver mutations were found in repeats that were capable of non-B DNA structure formation (cf. 23% for potential mutations). Several regional mutation hotspots were found in the mutational spectra of various genes; one of these, in the *APC* gene, was a regional hotspot for both somatic and germline micro-deletions/micro-insertions and corresponded to a previously recognized mutation hotspot [Miyoshi et al., 1992].

Taken together, the results and analysis presented herein strongly suggest that algorithms that attempt to predict the relative impact of tumour-associated micro-lesions on (tumour suppressor) gene and protein function [Tavtigian et al., 2008; Couch et al., 2008; Thusberg and Vihinen, 2009], should take into consideration the origin (i.e. somatic, germline or shared) of the mutations, their sequence context and repetitivity, as well as their frequency of occurrence.

Acknowledgements

We are most grateful to Eamonn Maher (Birmingham, UK), Gareth Evans (Manchester, UK) and the late Michael Baser for their provision of somatic mutation data. We are also very grateful to Larry Brody (NIH, Bethesda, MD, USA), Shiu-Ru Lin (Kaohsiung Medical University, Taiwan) and Tone Lovig (University of Oslo, Norway) for providing us with further information on specific somatic mutations and Rachel Karchin (Johns Hopkins University, Baltimore, USA) for making available her dataset of probable driver mutations.

References

Ali IU, Schriml LM, Dean M. 1999. Mutational spectra of *PTEN/MMAC1* gene: a tumor suppressor with lipid phosphatase activity. J Natl Cancer Inst 91:1922-1932.

Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova NA, Abeysinghe SS, O'Connell CD, Cooper DN, Wells RD. 2004. Breakpoints of gross deletions coincide with non-B DNA conformations. Proc Natl Acad Sci USA 101:14162-14167.

Bacolla A, Wells RD. 2004. Non-B DNA conformations, genomic rearrangements, and human disease. J Biol Chem 279:47411-47414.

Bacolla A, Collins JR, Gold B, Chuzhanova N, Yi M, Stephens RM, Stefanov S, Olsh A, Jakupciak JP, Dean M, Lempicki RA, Cooper DN, Wells RD. 2006. Long homopurine•homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. Nucleic Acids Res 34: 2663-2675.

Ball EV, Stenson PD, Krawczak M, Cooper DN, Chuzhanova NA. 2005. Micro-deletions and micro-insertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. Human Mutat 26:205-213.

Baser ME, Contributors to the International NF2 Mutation Database. Hum. Mutat. 27:297-306.

Boland CR, Ricciardiello L. 1999. How many mutations does it take to make a tumor? Proc Natl Acad Sci USA 96:14675-14677.

Buard J, Collick A, Brown J, Jeffreys AJ. 2000. Somatic versus germline mutation processes at minisatellite CEB1 (D2S90) in humans and transgenic mice. Genomics 65**:**95-103.

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. 2009. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 69:6660-6667.

Chuzhanova NA, Anassis EJ, Ball E, Krawczak M, Cooper DN. 2003. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat 21:28-44.

Chuzhanova N, Chen JM, Bacolla A, Patrinos GP, Férec C, Wells RD, Cooper DN. 2009. Gene conversion causing human inherited disease: the evidence for involvement of recombination-associated motifs and non-B DNA-forming sequences in DNA breakage. Hum Mutat 30:1189-1198.

Cole DN, Carlson JA, Wilson VL. 2008. Human germline and somatic cells have similar *TP5*3 and Kirsten-RAS gene single base mutation frequencies. Environ. Mol. Mutagen. 2008 49:417-425.

Cooper DN, Mort M, Stenson PD, Ball EV, Chuzhanova NA. 2010. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides as well as in CpG dinucleotides. Hum Genomics 4:406-410.

Couch FJ, Rasmussen LJ, Hofstra R, Monteiro AN, Greenblatt MS, de Wind N; IARC Unclassified Genetic Variants Working Group. 2008. Assessment of functional effects of unclassified genetic variants. Hum Mutat 29:1314-1326.

Dallosso AR, Jones S, Azzopardi D, Moskvina V, Al-Tassan N, Williams GT, Idziaszczyk S, Davies DR, Milewski P, Williams S, Beynon J, Sampson JR, Cheadle JP. 2009. The *APC* variant p.Glu1317Gln predisposes to colorectal adenomas by a novel mechanism of relaxing the target for tumorigenic somatic *APC* mutations. Hum Mutat 30:1412-1418.

Dworkin AM, Ridd K, Bautista D, Allain DC, Iwenofu OH, Roy R, Bastian BC, Toland AE. 2010. Germline variation controls the architecture of somatic alterations in tumors. PLoS Genet 6: e1001136.

Fostira F, Thodi G, Sandaltzopoulos R, Fountzilas G, Yannoukakos D. 2010. Mutational spectrum of *APC* and genotype-phenotype correlations in Greek FAP patients. BMC Cancer 10:389.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. Nat Rev Cancer 4:177-183.

Gallou C, Joly D, Mejean A, Staroz F, Martin N, Tarlet G, Orfanelli MT, Bouvier R, Droz D, Chretien Y, Maréchal JM, Richard S, Junien C, Béroud C. 1999. Mutations of the *VHL* gene in sporadic renal cell carcinoma: definition of a risk factor for VHL patients to develop an RCC. Hum Mutat 13:464-475.

Giovannone B, Sabbadini G, Di Maio L, Calabrese O, Castaldo I, Frontali M, Novelleto A, Squitieri F. 1997. Analysis of (CAG)n size heterogeneity in somatic and sperm cell DNA from intermediate and expanded Huntington disease gene carriers. Hum Mutat 10:458-464.

Glazko GV, Koonin EV, Rogozin IB. 2004. Mutation hotspots in the p53 gene in tumors of different origin: correlation with evolutionary conservation and signs of positive selection. Biochim Biophys Acta 1679**:**95-106.

Goode EL, Ulrich CM, Potter JD. 2002. Polymorphisms in DNA repair genes and associations with cancer risk. Cancer Epidemiol Biomarkers Prev 11:1513-1530.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Greenblatt MS, Grollman AP, Harris CC. 1996. Deletions and insertions in the p53 tumor suppressor gene in human cancers: confirmation of the DNA polymerase slippage/misalignment model. Cancer Res 56:2130-2136.

Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP. 2003. Detailed computational study of p53 and p16: using evolutionary sequence analysis and disease-associated mutations to predict the functional consequences of allelic variants. Oncogene 22:1150-1163.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR. 2007. Patterns of somatic mutation in human cancer genomes. Nature 446:153-158.

Groves C, Lamlum H, Crabtree M, Williamson J, Taylor C, Bass S, Cuthbert-Heavens D, Hodgson S, Phillips R, Tomlinson I. 2002. Mutation cluster region, association between germline and somatic mutations and genotype-phenotype correlation in upper gastrointestinal familial adenomatous polyposis. Am J Pathol 160:2055-2061.

Haber D, Harlow E. 1997. Tumour-suppressor genes: evolving definitions in the genomic age. Nat Genet 16**:**320-322.

Hess ST, Blake JD, Blake RD. 1994. Wide variations in neighbor-dependent substitution rates. J Mol Biol 236:1022-1033.

Jego N, Thomas G, Hamelin R. 1993. Short direct repeats flanking deletions, and duplicating insertions in p53 gene in human cancers. Oncogene 8:209-213.

Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S, Davis DP, Carlton VE, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard

DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, de Sauvage FJ, Faham M, Seshagiri S. 2010. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466:869-873.

Karlin S, Macken C. 1991. Some statistical problems in the assessment of inhomogenesis of DNA sequence data. J Am Statist Assoc 86:27–35.

Knudson AG. 2001. Two genetic hits (more or less) to cancer. Nat Rev Cancer 1**:**157-162.

Kohavi R. 1996. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hHybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, pp 202-207**.**

Kolomietz E, Meyn MS, Pandita A, Squire JA. 2002. The role of *Alu* repeat clusters as mediators of recurrent chromosomal aberrations in tumors. Genes Chrom Cancer 35:97-112.

Kondrashov AS, Rogozin IB. 2004. Context of deletions and insertions in human coding sequences. Hum Mutat 23:177–185.

Kotsiantis S, Kanellopoulos D, Pintelas P. 2006. Handling imbalanced datasets: a review. GESTS International Transactions on Computer Science and Engineering 30:25-36.

Krawczak M, Smith-Sorensen B, Schmidtke J, Kakkar VV, Cooper DN, Hovig E. 1995. 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 5:48-57.

Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63:474-488.

Lagarde A, Rouleau E, Ferrari A, Noguchi T, Qiu J, Briaux A, Bourdon V, Rémy V, Gaildrat P, Adélaïde J, Birnbaum D, Lidereau R, Sobol H, Olschwang S. 2010. Germline *APC* mutation spectrum derived from 863 genomic variations identified through a

15-year medical genetics service to French patients with FAP. J Med Genet 47:721-722.

Lamlum H, Ilyas M, Rowan A, Clark S, Johnson V, Bell J, Frayling I, Efstathiou J, Pack K, Payne S, Roylance R, Gorman P, Sheer D, Neale K, Phillips R, Talbot I, Bodmer W, Tomlinson I. 1999. The type of somatic mutation at *APC* in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's 'two-hit' hypothesis. Nat Med 5:1071-1075.

Latchford A Volikos E, Johnson V, Rogers P, Suraweera N, Tomlinson I, Phillips R, Silver A. 2007. *APC* mutations in FAP-associated desmoid tumours are non-random but not 'just right'. Hum Mol Genet 16:78-82.

Leclercq S, Rivals E, Jarne P. 2010. DNA slippage occurs at microsatellite loci without minimal threshold length in human: a comparative genomic approach. Genome Biol Evol 2:325-335.

Leeflang EP, Tavare S, Marjoram P, Neal CO, Srinidhi J, MacFarlane H, MacDonald ME, Gusella JF, de Young M, Wexler NS, Arnheim N. 1999. Analysis of germline mutation spectra at the Huntington's disease locus supports a mitotic mutation mechanism. Hum Mol Genet 8:173-183.

Lengauer C, Kinzler KW, Vogelstein B. 1998. Genetic instabilities in human cancers. Nature 396:643-649.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 19:315-322.

Lobo GP, Waite KA, Planchon SM, Romigh T, Nassif NT, Eng C. 2009. Germline and somatic cancer-associated mutations in the ATP-binding motifs of PTEN influence its subcellular localization and tumor suppressive function. Hum. Mol. Genet. 18:2851-2862.

Loeb LA, Harris CC. 2008. Advances in chemical carcinogenesis: a historical review and prospective. Cancer Res 68: 6863-6872.

Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci USA 95:9407-9412.

Makridakis NM, Caldas Ferraz LF, Reichardt JK. 2009. Genomic analysis of cancer tissue reveals that somatic mutations commonly occur in a specific motif. Hum Mutat 30:39-48.

Marshall B, Isidro G, Carvalhas R, Boavida M. 1997. Germline versus somatic mutations of the *APC* gene: evidence for mechanistic differences. Hum Mutat 9**:**286-288.

Martorell L, Monckton DG, Gamez J, Baiget M. 2000. Complex patterns of male germline instability and somatic mosaicism in myotonic dystrophy type 1. Eur J Hum Genet 8: 423-430.

Miyoshi Y, Nagase H, Ando H, Horii A, Ichii S, Nakatsuru S, Aoki T, Miki Y, Mori T, Nakamura Y. 1992. Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. Hum Mol Genet 1:229-33.

Mort M, Ivanov D, Cooper DN, Chuzhanova NA. 2008. A meta-analysis of nonsense mutations causing human genetic disease. Hum Mutat 29:1037-1047.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3**:**418-426.

Oldenburg J, Rost S, El-Maarri O, Leuer M, Olek K, Muller CR, Schwaab R. 2000. *De novo* factor VIII gene intron 22 inversion in a female carrier presents as a somatic mosaicism. Blood 96: 2905-2906.

Olshen AB, Jain AN. 2002. Deriving quantitative conclusions from microarray expression data. Bioinformatics 18:961-970.

Parmigiani G, Boca S, Lin J, Kinzler KW, Velculescu V, Vogelstein B. 2009. Design and analysis issues in genome-wide somatic mutation studies of cancer. Genomics 93:17-21.

Pollard LM, Sharma R, Gomez M, Shah S, Delatycki MB, Pianese L, Monticelli A, Keats BJ, Bidichandani SI. 2004. Replication-mediated instability of the GAA triplet repeat mutation in Friedreich ataxia. Nucleic Acids Res 32:5962-5971.

Richter S, Vandezande K, Chen N, Zhang K, Sutherland J, Anderson J, Han L, Panton R, Branco P, Gallie B. 2003. Sensitive and efficient detection of *RB1* gene mutations enhances care for families with retinoblastoma. Am J Hum Genet 72:253-269.

Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. 2001. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nat Immunol 2:530-536.

Rubin AF, Green P. 2009. Mutation patterns in cancer genomes. Proc Natl Acad Sci USA 106: 21766-21770.

Sharma R, Bhatti S, Gomez M, Clark RM, Murray C, Ashizawa T, Bidichandani SI. 2002. The GAA triplet-repeat sequence in Friedreich ataxia shows a high level of somatic instability *in vivo*, with a significant predilection for large contractions. Hum Mol Genet 11:2175-2187.

Sherr CJ. 2004. Principles of tumor suppression. Cell 116:235-246.

Schmutte C, Jones PA. 1998. Involvement of DNA methylation in human carcinogenesis. Biol Chem 379:377-388.

Shanks ME, May CA, Dubrova YE, Balaresque P, Rosser ZH, Adams SM, Jobling MA. 2008. Complex germline and somatic mutation processes at a haploid human minisatellite shown by single-molecule analysis. Mutat Res 648:46-53.

Simpson AJ. 2009. Sequence-based advances in the definition of cancer-associated gene mutations. Curr Opin Oncol 21:47-52.

Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314:268-274.

Stead JD, Jeffreys AJ. 2000. Allele diversity and germline mutation at the insulin minisatellite. Hum Mol Genet 9:713-723.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The Human Gene Mutation Database: 2008 update. *Genome Med.* 1:13.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. Nature 458:719-724.

Talavera D, Taylor MS, Thornton JM. 2010. The (non)malignancy of cancer amino acidic substitutions. Proteins 78:518-529.

Tappino B, Chuzhanova NA, Regis S, Dardis A, Corsolini F, Stroppiano M, Tonoli E, Beccari T, Rosano C, Mucha J, Blanco M, Szlago M, Di Rocco M, Cooper DN, Filocamo M. 2009. Molecular characterization of 22 novel UDP-N-acetylglucosamine-1-phosphate transferase alpha- and beta-subunit (*GNPTAB*) gene mutations causing mucolipidosis types IIalpha/beta and IIIalpha/beta in 46 patients. Hum Mutat 30:E956-973.

Tartaglia M, Martinelli S, Stella L, Bocchinfuso G, Flex E, Cordeddu V, Zampino G, van der Burgt I, Palleschi A, Petrucci TC, Sorcini M, Schoch C, Foà R, Emanuel PD, Gelb BD. 2006. Diversity and functional consequences of germline and somatic *PTPN11* mutations in human disease. Am J Hum Genet 78:279-290.

Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB; IARC Unclassified Genetic Variants Working Group. 2008. *In silico* analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 29:1327-1336.

Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B, McKay-Chopin S, Thomas A, Vallée MP, Voegele C, Webb PM, Whiteman DC; Australian Cancer Study; Breast Cancer Family Registries (BCFR); Kathleen Cuningham Foundation Consortium for Research into Familial Aspects of Breast Cancer (kConFab), Sangrajrang S, Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G. 2009. Am J Hum Genet 85:427-446.

Thomas L, Kluwe L, Chuzhanova N, Mautner V, Upadhyaya M. 2010. Analysis of *NF1* somatic mutations in cutaneous neurofibromas from patients with high tumor burden. Neurogenetics 11:391-400.

Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30:703-714.

Tornaletti S, Pfeifer GP. 1995. Complete and tissue-independent methylation of CpG sites in the *p53* gene: implications for mutations in human cancers. Oncogene 10:1493-1499.

Upadhyaya M, Han S, Consoli C, Majounie E, Horan M, Thomas NS, Potts C, Griffiths S, Ruggieri M, von Deimling A, Cooper DN. 2004. Characterization of the somatic mutational spectrum of the neurofibromatosis type 1 (*NF1*) gene in neurofibromatosis patients with benign and malignant tumors. Hum Mutat 23:134-136.

Upadhyaya M, Kluwe L, Spurlock G, Monem B, Majounie E, Mantripragada K, Ruggieri M, Chuzhanova N, Evans DG, Ferner R, Thomas N, Guha A, Mautner V. 2008. Germline and somatic *NF1* gene mutation spectrum in *NF1*-associated malignant peripheral nerve sheath tumors (MPNSTs). Hum Mutat 29:74-82.

Vogelstein B, Kinzler KW. 2004. Cancer genes and the pathways they control. Nat Med 10:789-799.

Walker DR, Bond JP, Tarone RE, Harris CC, Makalowski W, Boguski MS. Greenblatt MS. 1999. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. Oncogene 18:211-218.

Walters, S. J. 2004. Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36. *Health Qual Life Outcomes* 2, p. 26.

Witten IH, Frank E. 2005. *Data mining: practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, San Francisco, pp. 365-483.

Figure Legends

Figure 1. Diagrammatic representation of the number of various types of mutations analysed in the present study.

Figure 2. Nucleotide substitution patterns of missense mutations in 17 tumour suppressor genes.

Comparative Analysis of Germline and Somatic Micro-lesion Mutational Spectra in 17 Human Tumour Suppressor Genes

(Supplementary Text)

Dobril Ivanov[1,2], Stephen E. Hamby[3], Peter D. Stenson[1], Andrew D. Phillips[1],
Hildegard Kehrer-Sawatzki[4], David N. Cooper[1] and Nadia Chuzhanova[3]

[1]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

[2]MRC Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine and Neurology, Biostatistics and Bioinformatics Unit, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK

[3]School of Science and Technology, Nottingham Trent University, Nottingham, NG11 8NS, UK

[4]Institute of Human Genetics, University of Ulm, Albert-Einstein-Allee 11, 89081 Ulm, Germany

*All correspondence to: Prof. Nadia Chuzhanova, School of Science and Technology Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK

Tel: +44 (0) 0115 848 8304 E-mail: nadia.chuzhanova@ntu.ac.uk

Gene-wise characteristics of germline and somatic missense mutations with respect to mutation type

Taken together, the combined mutational spectra for all 17 tumour suppressor genes contained twice as many somatic (61%) as germline (31%) mutations. For five genes (*APC*, *CDKN2A*, *NF2*, *PTEN* and *TP53*), a predominance of somatic over germline mutations was noted, with the *TP53* gene having the highest proportion of somatic mutations (92%). For the majority of genes, however (namely *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *PTCH1*, *RB1*, *STK11*, *TSC1*, *TSC2*, *VHL* and *WT1*), the analysed dataset included more germline than somatic mutations, with >97% of all mutations in the *BRCA1*, *NF1*, *TSC2* and *WT1* genes being germline in origin.

Gene-wise characteristics of missense mutations in relation to the disease- and non-disease-associated substitution rates

The median values were found to vary between 4 (*NF2*) and 8.9 (*STK11*) for somatic mutations, 4.1 (*TP53*) and 10.1 (*WT1*) for germline mutations, and 7.2 (*RB1*) and 11 (*PTEN*) for shared mutations (values given only for genes with more than three mutations in the corresponding category; see Supp. Table 3, indicating that many of the median values are quite low and hence the corresponding mutations are unlikely to be neutral.

The highest and lowest disease-associated median values for the mutation rates were noted for somatic mutations in the *STK11* gene (1.7; Supp. Table 3) and for germline mutations in the *TP53* (0.42) gene (values given only for genes with more than three mutations in the corresponding category).

Gene-wise occurrence of missense mutations within repeats and runs of identical nucleotides

On average, the coverage of the respective gene sequences by direct repeats of length ≥8 bp was found to vary between 2.5% (*BRCA2*) and 17% (*PTEN*). The corresponding proportion of the cDNA lengths for inverted repeats ≥8 bp was found to vary between 4.5% (*PTCH1)* and *RB1* 15.7% (*RB1*) while symmetric elements ≥8 bp were found to vary between 15.5% for *APC* and 44% for *PTEN* genes.

On average, mononucleotide runs ≥4 bp spanned 19.9% of the cDNA lengths, varying between 9.5% (*VHL*) and 29% (*TP53*).

No correlation was observed between the number of mutations located within repeats and the fractional length of the cDNA covered by repeats, indicating that not every repeat sequence is mutation-prone. However, a strong correlation between the fractional length of the cDNA covered by repeats and cDNA length of genes ($r > 0.87$ and $p < 10^{-6}$) served to demonstrate that repeat density per unit length was approximately the same for all tumour suppressor genes studied.

Truncating vs non-truncating mutations in the germline and soma

Somatic mutational spectra from the *BRCA2, CDKN2A, STK11*, *TP53* and *TSC1* genes were characterized by the predominance of non-truncating (i.e. missense) lesions over truncating lesions (i.e. nonsense mutations, frameshift micro-deletions, micro-insertions and indels) when nonsense mutations [reported in Mort et al. (2008)] and micro-indels (excluded from previous analyses) were also considered (Supp. Table 8). A similar predominance of non-truncating over truncating lesions was observed for the germline mutational spectra of the *CDKN2A*, *TP53*, *VHL* and *WT1* genes. In general, the ratio of non-truncating to truncating lesions was found to be significantly higher in the soma (0.85) than in the germline (0.30; p-value<2.20E-16). All other mutational spectra were characterized by the predominance of truncating mutations.

References

Mort M, Ivanov D, Cooper DN, Chuzhanova NA. 2008. A meta-analysis of nonsense mutations causing human genetic disease. Hum Mutat 29:1037-1047.

Supplementary Figure 1. Naive Bayes Tree Classifier. Number in parenthesis shows the probability of a mutations being somatic non-recurrent, germline, shared non-recurrent, somatic recurrent and shared recurrent respectively.

```
Attributes:
              Mut_Type
              Hess_value
              Krawczak_value
              Evol
              Grantham_score
              CpG/CHG
              Repeats
Test mode:    10-fold cross-validation

NBTree
------------------

Evol <= 0.205
|   Repeats = 0
|   |   CpG/CHG = 0
|   |   |   Krawczak_value <= 1.0465
|   |   |   |   Evol <= 0.155
|   |   |   |   |   Evol <= 0.12
|   |   |   |   |   |   Krawczak_value <= 0.811
|   |   |   |   |   |   |   Krawczak_value <= 0.099
|   |   |   |   |   |   |   |   Hess_value <= 3.1:          (0.42) (0.08) (0.08) (0.33) (0.08)
|   |   |   |   |   |   |   |   Hess_value > 3.1:           (0.23) (0.13) (0.03) (0.10) (0.52)
|   |   |   |   |   |   |   Krawczak_value > 0.099
|   |   |   |   |   |   |   |   Hess_value <= 2.5
|   |   |   |   |   |   |   |   |   Grantham_score <= 146.5
|   |   |   |   |   |   |   |   |   |   Hess_value <= 2.15: (0.27) (0.47) (0.02) (0.22) (0.02)
|   |   |   |   |   |   |   |   |   |   Hess_value > 2.15:  (0.14) (0.24) (0.05) (0.52) (0.05)
|   |   |   |   |   |   |   |   |   Grantham_score > 146.5: (0.47) (0.07) (0.07) (0.33) (0.07)
|   |   |   |   |   |   |   |   Hess_value > 2.5
|   |   |   |   |   |   |   |   |   Hess_value <= 5.45
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 30.5
|   |   |   |   |   |   |   |   |   |   |   Hess_value <= 5.2
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.55
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 2.75: (0.27) (0.09) (0.09) (0.45) (0.09)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 2.75:  (0.25) (0.43) (0.03) (0.28) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.55:      (0.29) (0.08) (0.04) (0.54) (0.04)
|   |   |   |   |   |   |   |   |   |   |   Hess_value > 5.2:           (0.12) (0.12) (0.06) (0.12) (0.59)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 30.5
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.411
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.35
|   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.3775
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.1975
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 146:  (0.23) (0.13) (0.03) (0.57) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 146:   (0.28) (0.16)  (0.4) (0.12) (0.04)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.1975
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.22: (0.11) (0.04) (0.26) (0.11) (0.48)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.22
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 2.85
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 147.5: (0.21) (0.14) (0.28) (0.34) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 147.5
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 2.75:  (0.21) (0.04) (0.29) (0.08) (0.38)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 2.75:   (0.05) (0.05) (0.79) (0.05) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 2.85
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 155.5
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 3.95:  (0.18) (0.15) (0.03) (0.61) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 3.95:   (0.10) (0.14) (0.14) (0.43) (0.19)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 155.5:  (0.23) (0.06) (0.49)  (0.2) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.3775:      (0.12) (0.32) (0.04) (0.48) (0.04)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.35
|   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 100.5
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.2455: (0.09) (0.45) (0.09) (0.27) (0.09)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.2455:  (0.42) (0.29) (0.03) (0.23) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 100.5:       (0.23) (0.14) (0.05) (0.32) (0.27)
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.411
|   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 105.5
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.85
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 100
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 63: (0.04) (0.04) (0.77) (0.13) (0.02)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 63:  (0.21) (0.26) (0.05) (0.42) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 100:     (0.04) (0.04) (0.78) (0.09) (0.04)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 70.5:   (0.26) (0.16) (0.05) (0.47) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 70.5:    (0.13) (0.10) (0.63) (0.10) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.85:  (0.31) (0.38) (0.08) (0.15) (0.08)
|   |   |   |   |   |   |   |   |   |   |   |   Grantham_score > 105.5
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 3.05: (0.28)  (0.1) (0.45) (0.14) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 3.05:  (0.18) (0.32) (0.04) (0.32) (0.14)
|   |   |   |   |   |   |   |   |   Hess_value > 5.45
|   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.336: (0.06) (0.06) (0.63) (0.22) (0.03)
|   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.336:  (0.13) (0.46) (0.19) (0.20) (0.01)
|   |   |   |   |   |   Krawczak_value > 0.811
|   |   |   |   |   |   |   Grantham_score <= 78.5
|   |   |   |   |   |   |   |   Grantham_score <= 37.5:   (0.27) (0.27) (0.05) (0.05) (0.36)
|   |   |   |   |   |   |   |   Grantham_score > 37.5:    (0.51) (0.17) (0.02) (0.27) (0.02)
|   |   |   |   |   |   |   Grantham_score > 78.5
|   |   |   |   |   |   |   |   Hess_value <= 10.95
|   |   |   |   |   |   |   |   |   Grantham_score <= 129: (0.03) (0.28) (0.03) (0.15) (0.51)
|   |   |   |   |   |   |   |   |   Grantham_score > 129:  (0.35) (0.13) (0.04) (0.04) (0.43)
|   |   |   |   |   |   |   |   Hess_value > 10.95:      (0.22) (0.39) (0.06) (0.28) (0.06)
|   |   |   |   |   Evol > 0.12
|   |   |   |   |   |   Evol <= 0.135: (0.08) (0.15) (0.62) (0.08) (0.08)
|   |   |   |   |   |   Evol > 0.135
|   |   |   |   |   |   |   Krawczak_value <= 0.5255
|   |   |   |   |   |   |   |   Hess_value <= 4.3:   (0.03) (0.40) (0.27) (0.27) (0.03)
|   |   |   |   |   |   |   |   Hess_value > 4.3:    (0.06) (0.06) (0.75) (0.06) (0.06)
|   |   |   |   |   |   |   Krawczak_value > 0.5255: (0.22) (0.04) (0.04) (0.13) (0.57)
|   |   |   |   Evol > 0.155
|   |   |   |   |   Evol <= 0.175:  (0.38) (0.24) (0.05) (0.29) (0.05)
|   |   |   |   |   Evol > 0.175:   (0.17)  (0.1) (0.03) (0.41) (0.28)
|   |   |   Krawczak_value > 1.0465
|   |   |   |   Hess_value <= 12.35
|   |   |   |   |   Krawczak_value <= 1.1575: (0.03) (0.06) (0.68) (0.21) (0.03)
|   |   |   |   |   Krawczak_value > 1.1575
|   |   |   |   |   |   Hess_value <= 7.05:   (0.07) (0.24) (0.03) (0.38) (0.28)
|   |   |   |   |   |   Hess_value > 7.05
|   |   |   |   |   |   |   Krawczak_value <= 1.838
|   |   |   |   |   |   |   |   Krawczak_value <= 1.725
|   |   |   |   |   |   |   |   |   Krawczak_value <= 1.27
|   |   |   |   |   |   |   |   |   |   Hess_value <= 7.6: (0.04) (0.15) (0.42) (0.04) (0.35)
|   |   |   |   |   |   |   |   |   |   Hess_value > 7.6:  (0.16) (0.21) (0.05) (0.05) (0.53)
|   |   |   |   |   |   |   |   |   Krawczak_value > 1.27
|   |   |   |   |   |   |   |   |   |   Krawczak_value <= 1.5585
|   |   |   |   |   |   |   |   |   |   |   Grantham_score <= 60:  (0.19) (0.14) (0.05) (0.29) (0.33)
|   |   |   |   |   |   |   |   |   |   |   Grantham_score > 60:   (0.15)  (0.3) (0.05) (0.45) (0.05)
|   |   |   |   |   |   |   |   |   |   Krawczak_value > 1.5585
|   |   |   |   |   |   |   |   |   |   |   Hess_value <= 8.65
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 7.5: (0.20) (0.07)  (0.6) (0.07) (0.07)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 7.5:  (0.04) (0.15) (0.31) (0.08) (0.42)
|   |   |   |   |   |   |   |   |   |   |   Hess_value > 8.65:     (0.38) (0.38) (0.06) (0.13) (0.06)
|   |   |   |   |   |   |   |   Krawczak_value > 1.725:    (0.09) (0.05) (0.27) (0.55) (0.05)
|   |   |   |   |   |   |   Krawczak_value > 1.838
|   |   |   |   |   |   |   |   Hess_value <= 11.5:  (0.04) (0.34) (0.35) (0.09) (0.18)
|   |   |   |   |   |   |   |   Hess_value > 11.5:   (0.03) (0.18) (0.46)  (0.1) (0.23)
|   |   |   |   Hess_value > 12.35
|   |   |   |   |   Grantham_score <= 86
|   |   |   |   |   |   Hess_value <= 13.8:  (0.15) (0.15) (0.03) (0.38) (0.29)
|   |   |   |   |   |   Hess_value > 13.8:   (0.13) (0.09) (0.52) (0.04) (0.22)
|   |   |   |   |   Grantham_score > 86
|   |   |   |   |   |   Hess_value <= 13.15: (0.03) (0.41) (0.03) (0.03)  (0.5)
|   |   |   |   |   |   Hess_value > 13.15:  (0.13)  (0.2) (0.03)  (0.2) (0.43)
|   |   CpG/CHG = 1
|   |   |   Hess_value <= 59.5
|   |   |   |   Grantham_score <= 44.5: (0.03) (0.04) (0.18) (0.07) (0.68)
|   |   |   |   Grantham_score > 44.5:  (0.03) (0.12) (0.41) (0.01) (0.44)
|   |   |   Hess_value > 59.5:          (0.20) (0.60) (0.03) (0.14) (0.03)
|   Repeats = 1
|   |   CpG/CHG = 0
|   |   |   Hess_value <= 4.35
|   |   |   |   Evol <= 0.18
|   |   |   |   |   Evol <= 0.065
|   |   |   |   |   |   Krawczak_value <= 0.232
|   |   |   |   |   |   |   Grantham_score <= 134.5
|   |   |   |   |   |   |   |   Grantham_score <= 112.5
|   |   |   |   |   |   |   |   |   Grantham_score <= 54: (0.33) (0.11) (0.06) (0.11) (0.39)
|   |   |   |   |   |   |   |   |   Grantham_score > 54:  (0.23) (0.23) (0.03) (0.48) (0.03)
|   |   |   |   |   |   |   |   Grantham_score > 112.5:   (0.44) (0.06) (0.06) (0.06) (0.38)
|   |   |   |   |   |   |   Grantham_score > 134.5:       (0.13) (0.07) (0.07) (0.67) (0.07)
|   |   |   |   |   |   Krawczak_value > 0.232
|   |   |   |   |   |   |   Hess_value <= 3.3
|   |   |   |   |   |   |   |   Krawczak_value <= 0.341
|   |   |   |   |   |   |   |   |   Grantham_score <= 84:   (0.24) (0.04) (0.56) (0.12) (0.04)
|   |   |   |   |   |   |   |   |   Grantham_score > 84
|   |   |   |   |   |   |   |   |   |   Hess_value <= 2.65: (0.09) (0.52) (0.04)  (0.3) (0.04)
|   |   |   |   |   |   |   |   |   |   Hess_value > 2.65:  (0.27) (0.14) (0.05)  (0.5) (0.05)
|   |   |   |   |   |   |   |   Krawczak_value > 0.341
|   |   |   |   |   |   |   |   |   Krawczak_value <= 0.463:(0.38) (0.46) (0.04) (0.08) (0.04)
|   |   |   |   |   |   |   |   |   Krawczak_value > 0.463: (0.21) (0.31) (0.03) (0.41) (0.03)
|   |   |   |   |   |   |   Hess_value > 3.3:              (0.20) (0.27) (0.01) (0.51) (0.01)
|   |   |   |   |   Evol > 0.065:   (0.36)  (0.5) (0.05) (0.05) (0.05)
|   |   |   |   Evol > 0.18: (      (0.10) (0.05) (0.05) (0.76) (0.05)
|   |   |   Hess_value > 4.35
|   |   |   |   Evol <= 0.045
|   |   |   |   |   Grantham_score <= 30.5
|   |   |   |   |   |   Hess_value <= 5.55:         (0.43) (0.18) (0.04) (0.32) (0.04)
|   |   |   |   |   |   Hess_value > 5.55
|   |   |   |   |   |   |   Grantham_score <= 26.5: (0.18) (0.44) (0.03) (0.32) (0.03)
|   |   |   |   |   |   |   Grantham_score > 26.5:  (0.11) (0.11) (0.05) (0.68) (0.05)
|   |   |   |   |   Grantham_score > 30.5
|   |   |   |   |   |   Grantham_score <= 118.5
|   |   |   |   |   |   |   Grantham_score <= 95.5
|   |   |   |   |   |   |   |   Hess_value <= 10.6
|   |   |   |   |   |   |   |   |   Grantham_score <= 75.5
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 69.5
|   |   |   |   |   |   |   |   |   |   |   Hess_value <= 7.05
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.65
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.55: (0.07) (0.23) (0.03) (0.13) (0.53)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.55:  (0.30) (0.30) (0.30) (0.05) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.65:      (0.07) (0.21) (0.03) (0.31) (0.38)
|   |   |   |   |   |   |   |   |   |   |   Hess_value > 7.05:          (0.23) (0.02) (0.02) (0.32) (0.41)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 69.5:          (0.10) (0.10) (0.33) (0.02) (0.45)
|   |   |   |   |   |   |   |   |   Grantham_score > 75.5
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 92.5:    (0.13) (0.29) (0.04)  (0.5) (0.04)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 92.5:     (0.18) (0.32) (0.41) (0.05) (0.05)
|   |   |   |   |   |   |   |   Hess_value > 10.6:      (0.26) (0.23) (0.03) (0.46) (0.03)
|   |   |   |   |   |   |   Grantham_score > 95.5
|   |   |   |   |   |   |   |   Hess_value <= 5.55
|   |   |   |   |   |   |   |   |   Hess_value <= 4.65: (0.27) (0.45) (0.09) (0.09) (0.09)
|   |   |   |   |   |   |   |   |   Hess_value > 4.65:  (0.03) (0.06) (0.03)  (0.2) (0.69)
|   |   |   |   |   |   |   |   Hess_value > 5.55
|   |   |   |   |   |   |   |   |   Grantham_score <= 102.5: (0.18) (0.56) (0.02) (0.13) (0.11)
|   |   |   |   |   |   |   |   |   Grantham_score > 102.5:  (0.06)  (0.2) (0.03) (0.37) (0.34)
|   |   |   |   |   |   Grantham_score > 118.5
|   |   |   |   |   |   |   Grantham_score <= 149.5:         (0.08) (0.13) (0.18) (0.04) (0.57)
|   |   |   |   |   |   |   Grantham_score > 149.5
|   |   |   |   |   |   |   |   Hess_value <= 10.45
|   |   |   |   |   |   |   |   |   Krawczak_value <= 0.428: (0.36) (0.09) (0.09) (0.36) (0.09)
|   |   |   |   |   |   |   |   |   Krawczak_value > 0.428:  (0.07) (0.26) (0.56) (0.09) (0.02)
|   |   |   |   |   |   |   |   Hess_value > 10.45:          (0.04) (0.16) (0.24) (0.06) (0.50)
|   |   |   |   Evol > 0.045:    (0.33) (0.37) (0.04) (0.22) (0.04)
|   |   CpG/CHG = 1
|   |   |   Grantham_score <= 99.5
|   |   |   |   Hess_value <= 10.05
|   |   |   |   |   Grantham_score <= 86: (0.07) (0.14) (0.07) (0.21)  (0.5)
|   |   |   |   |   Grantham_score > 86:  (0.03) (0.03) (0.88) (0.03) (0.03)
|   |   |   |   Hess_value > 10.05
|   |   |   |   |   Evol <= 0.07
|   |   |   |   |   |   Krawczak_value <= 12.275
|   |   |   |   |   |   |   Krawczak_value <= 9.211
|   |   |   |   |   |   |   |   Krawczak_value <= 8.5135
|   |   |   |   |   |   |   |   |   Krawczak_value <= 7.551: (0.45) (0.27) (0.09) (0.09) (0.09)
|   |   |   |   |   |   |   |   |   Krawczak_value > 7.551:  (0.03) (0.14) (0.03) (0.03) (0.76)
|   |   |   |   |   |   |   |   Krawczak_value > 8.5135:     (0.47) (0.35) (0.06) (0.06) (0.06)
|   |   |   |   |   |   |   Krawczak_value > 9.211
|   |   |   |   |   |   |   |   Hess_value <= 46.4: (0.26) (0.11) (0.05) (0.05) (0.53)
|   |   |   |   |   |   |   |   Hess_value > 46.4:  (0.02) (0.02) (0.22) (0.06) (0.68)
|   |   |   |   |   |   Krawczak_value > 12.275:    (0.08) (0.03) (0.72) (0.03) (0.14)
|   |   |   |   |   Evol > 0.07:   (0.07) (0.03) (0.03) (0.03) (0.83)
|   |   |   Grantham_score > 99.5
|   |   |   |   Krawczak_value <= 7.519:   (0.03) (0.03) (0.03)  (0.1) (0.82)
|   |   |   |   Krawczak_value > 7.519
|   |   |   |   |   Grantham_score <= 113: (0.02) (0.19) (0.02) (0.06) (0.70)
|   |   |   |   |   Grantham_score > 113:  (0.13) (0.57) (0.04) (0.22) (0.04)
Evol > 0.205
|   Hess_value <= 9.65
|   |   Repeats = 0
|   |   |   Hess_value <= 8.8
|   |   |   |   Grantham_score <= 40.5
|   |   |   |   |   Hess_value <= 2.65: (0.60) (0.07) (0.07) (0.20) (0.07)
|   |   |   |   |   Hess_value > 2.65
|   |   |   |   |   |   Krawczak_value <= 1.083
|   |   |   |   |   |   |   Krawczak_value <= 0.269:     (0.11) (0.39) (0.06) (0.39) (0.06)
|   |   |   |   |   |   |   Krawczak_value > 0.269
|   |   |   |   |   |   |   |   Krawczak_value <= 0.6155
|   |   |   |   |   |   |   |   |   Hess_value <= 4:     (0.68) (0.05) (0.05) (0.16) (0.05)
|   |   |   |   |   |   |   |   |   Hess_value > 4:      (0.48) (0.28) (0.04) (0.16) (0.04)
|   |   |   |   |   |   |   |   Krawczak_value > 0.6155: (0.22)  (0.5) (0.06) (0.17) (0.06)
|   |   |   |   |   |   Krawczak_value > 1.083:          (0.17) (0.08) (0.08) (0.58) (0.08)
|   |   |   |   Grantham_score > 40.5
|   |   |   |   |   Hess_value <= 5.05
|   |   |   |   |   |   Grantham_score <= 194.5
|   |   |   |   |   |   |   Krawczak_value <= 0.365
|   |   |   |   |   |   |   |   Hess_value <= 3.95
|   |   |   |   |   |   |   |   |   Grantham_score <= 66.5
|   |   |   |   |   |   |   |   |   |   Hess_value <= 2.65:      (0.21) (0.07) (0.38) (0.07) (0.28)
|   |   |   |   |   |   |   |   |   |   Hess_value > 2.65
|   |   |   |   |   |   |   |   |   |   |   Evol <= 0.275:       (0.05) (0.05) (0.79) (0.05) (0.05)
|   |   |   |   |   |   |   |   |   |   |   Evol > 0.275:        (0.32) (0.08) (0.36)  (0.2) (0.04)
|   |   |   |   |   |   |   |   |   Grantham_score > 66.5
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 159.5: (0.36) (0.37) (0.12) (0.15) (0.01)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 159.5:  (0.19) (0.04)  (0.3) (0.07) (0.41)
|   |   |   |   |   |   |   |   Hess_value > 3.95
|   |   |   |   |   |   |   |   |   Krawczak_value <= 0.229:     (0.26) (0.19) (0.04) (0.48) (0.04)
|   |   |   |   |   |   |   |   |   Krawczak_value > 0.229:      (0.39) (0.07) (0.04) (0.04) (0.46)
|   |   |   |   |   |   |   Krawczak_value > 0.365
|   |   |   |   |   |   |   |   Hess_value <= 4.55
|   |   |   |   |   |   |   |   |   Hess_value <= 4.3
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 105.5: (0.51) (0.14) (0.03) (0.29) (0.03)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 105.5
|   |   |   |   |   |   |   |   |   |   |   Hess_value <= 3.3:   (0.50) (0.33) (0.06) (0.06) (0.06)
|   |   |   |   |   |   |   |   |   |   |   Hess_value > 3.3:    (0.36) (0.16) (0.04) (0.28) (0.16)
|   |   |   |   |   |   |   |   |   Hess_value > 4.3:            (0.06) (0.24) (0.06) (0.29) (0.35)
|   |   |   |   |   |   |   |   Hess_value > 4.55:               (0.39) (0.04) (0.04) (0.48) (0.04)
|   |   |   |   |   |   Grantham_score > 194.5:   (0.09) (0.09) (0.73) (0.05) (0.05)
|   |   |   |   |   Hess_value > 5.05
|   |   |   |   |   |   Grantham_score <= 45.5:   (0.04) (0.11) (0.54) (0.29) (0.04)
|   |   |   |   |   |   Grantham_score > 45.5
|   |   |   |   |   |   |   Evol <= 0.51
|   |   |   |   |   |   |   |   Hess_value <= 7.25
|   |   |   |   |   |   |   |   |   Evol <= 0.28: (0.07) (0.43) (0.07) (0.36) (0.07)
|   |   |   |   |   |   |   |   |   Evol > 0.28:  (0.27) (0.27) (0.24) (0.18) (0.03)
|   |   |   |   |   |   |   |   Hess_value > 7.25
|   |   |   |   |   |   |   |   |   Hess_value <= 7.6:        (0.09) (0.18) (0.09) (0.55) (0.09)
|   |   |   |   |   |   |   |   |   Hess_value > 7.6
|   |   |   |   |   |   |   |   |   |   Grantham_score <= 69: (0.57) (0.09) (0.04) (0.26) (0.04)
|   |   |   |   |   |   |   |   |   |   Grantham_score > 69:  (0.04) (0.29) (0.04) (0.58) (0.04)
|   |   |   |   |   |   |   Evol > 0.51
|   |   |   |   |   |   |   |   Grantham_score <= 88.5
|   |   |   |   |   |   |   |   |   Krawczak_value <= 1.005:  (0.25) (0.33) (0.08) (0.25) (0.08)
|   |   |   |   |   |   |   |   |   Krawczak_value > 1.005
|   |   |   |   |   |   |   |   |   |   Evol <= 0.61:         (0.06) (0.03) (0.85) (0.03) (0.03)
|   |   |   |   |   |   |   |   |   |   Evol > 0.61:          (0.50) (0.25) (0.08) (0.08) (0.08)
|   |   |   |   |   |   |   |   Grantham_score > 88.5:        (0.27) (0.09) (0.09) (0.45) (0.09)
|   |   |   Hess_value > 8.8
|   |   |   |   Krawczak_value <= 1.1745
|   |   |   |   |   Krawczak_value <= 0.862: (0.69) (0.08) (0.08) (0.08) (0.08)
|   |   |   |   |   Krawczak_value > 0.862:  (0.13) (0.16) (0.03) (0.09) (0.59)
|   |   |   |   Krawczak_value > 1.1745:     (0.58) (0.05) (0.05) (0.26) (0.05)
|   |   Repeats = 1
|   |   |   Grantham_score <= 123
|   |   |   |   Evol <= 0.285
|   |   |   |   |   Evol <= 0.255: (0.47) (0.06) (0.03) (0.25) (0.19)
|   |   |   |   |   Evol > 0.255:  (0.09) (0.06) (0.42) (0.03) (0.39)
|   |   |   |   Evol > 0.285
|   |   |   |   |   Krawczak_value <= 1.27
|   |   |   |   |   |   Hess_value <= 8.55
|   |   |   |   |   |   |   CpG/CHG = 0
|   |   |   |   |   |   |   |   Hess_value <= 6.75
|   |   |   |   |   |   |   |   |   Evol <= 0.415
|   |   |   |   |   |   |   |   |   |   Evol <= 0.355
|   |   |   |   |   |   |   |   |   |   |   Evol <= 0.295
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 2.75: (0.32) (0.05) (0.42) (0.16) (0.05)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 2.75:  (0.65) (0.23) (0.04) (0.04) (0.04)
|   |   |   |   |   |   |   |   |   |   |   Evol > 0.295:           (0.25) (0.19) (0.06) (0.44) (0.06)
|   |   |   |   |   |   |   |   |   |   Evol > 0.355
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.5455
|   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.284
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 3.55: (0.18) (0.04) (0.71) (0.04) (0.04)
|   |   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 3.55:  (0.07) (0.21)  (0.5) (0.14) (0.07)
|   |   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.284: (0.27) (0.32) (0.05) (0.32) (0.05)
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.5455:    (0.05) (0.05) (0.67) (0.05) (0.19)
|   |   |   |   |   |   |   |   |   Evol > 0.415
|   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.4675
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.417
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value <= 4.8
|   |   |   |   |   |   |   |   |   |   |   |   |   Evol <= 0.585:    (0.38) (0.18) (0.03) (0.38) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   |   Evol > 0.585:     (0.70) (0.14) (0.03) (0.11) (0.03)
|   |   |   |   |   |   |   |   |   |   |   |   Hess_value > 4.8:     (0.71) (0.07) (0.07) (0.07) (0.07)
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.417:   (0.08) (0.31) (0.08) (0.46) (0.08)
|   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.4675
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value <= 0.5205: (0.24) (0.06) (0.53) (0.12) (0.06)
|   |   |   |   |   |   |   |   |   |   |   Krawczak_value > 0.5205:  (0.78) (0.07) (0.04) (0.07) (0.04)
|   |   |   |   |   |   |   |   Hess_value > 6.75
|   |   |   |   |   |   |   |   |   Grantham_score <= 57: (0.11) (0.53) (0.05) (0.26) (0.05)
|   |   |   |   |   |   |   |   |   Grantham_score > 57:  (0.47) (0.22) (0.03) (0.25) (0.03)
|   |   |   |   |   |   |   |   CpG/CHG = 1:          (0.40) (0.10) (0.10) (0.30) (0.10)
|   |   |   |   |   |   Hess_value > 8.55
|   |   |   |   |   |   |   Evol <= 0.54:         (0.27) (0.20) (0.07) (0.40) (0.07)
|   |   |   |   |   |   |   Evol > 0.54:          (0.03) (0.03) (0.84) (0.06) (0.03)
|   |   |   |   |   Krawczak_value > 1.27
|   |   |   |   |   |   Grantham_score <= 86:     (0.52) (0.04) (0.04) (0.37) (0.04)
|   |   |   |   |   |   Grantham_score > 86:      (0.40) (0.40) (0.03) (0.13) (0.03)
|   |   |   Grantham_score > 123
|   |   |   |   Evol <= 0.445
|   |   |   |   |   Hess_value <= 3.45
|   |   |   |   |   |   Krawczak_value <= 0.4665: (0.03) (0.19) (0.03) (0.16) (0.59)
|   |   |   |   |   |   Krawczak_value > 0.4665:  (0.25) (0.08) (0.08) (0.50) (0.08)
|   |   |   |   |   Hess_value > 3.45:            (0.43) (0.05) (0.05) (0.43) (0.05)
|   |   |   |   Evol > 0.445:                     (0.44) (0.09) (0.03) (0.41) (0.03)
|   Hess_value > 9.65
|   |   Hess_value <= 42.75
|   |   |   Hess_value <= 12.1
|   |   |   |   Repeats = 0
|   |   |   |   |   Evol <= 0.325:         (0.32) (0.39) (0.21) (0.04) (0.04)
|   |   |   |   |   Evol > 0.325
|   |   |   |   |   |   Hess_value <= 11.4
|   |   |   |   |   |   |   Evol <= 0.705: (0.26) (0.33) (0.04) (0.33) (0.04)
|   |   |   |   |   |   |   Evol > 0.705:  (0.06) (0.75) (0.06) (0.06) (0.06)
|   |   |   |   |   |   Hess_value > 11.4: (0.18) (0.23) (0.05) (0.14) (0.41)
|   |   |   |   Repeats = 1
|   |   |   |   |   Grantham_score <= 91.5
|   |   |   |   |   |   Grantham_score <= 85
|   |   |   |   |   |   |   Hess_value <= 11.4: (0.18) (0.24) (0.47) (0.08) (0.03)
|   |   |   |   |   |   |   Hess_value > 11.4:  (0.05) (0.32) (0.05) (0.14) (0.45)
|   |   |   |   |   |   Grantham_score > 85:    (0.20) (0.45) (0.05) (0.25) (0.05)
|   |   |   |   |   Grantham_score > 91.5:      (0.33) (0.17) (0.03) (0.03) (0.43)
|   |   |   Hess_value > 12.1
|   |   |   |   Evol <= 0.51
|   |   |   |   |   Repeats = 0
|   |   |   |   |   |   Grantham_score <= 44.5:   (0.20) (0.07) (0.07) (0.60) (0.07)
|   |   |   |   |   |   Grantham_score > 44.5
|   |   |   |   |   |   |   Grantham_score <= 51: (0.09) (0.27) (0.09) (0.09) (0.45)
|   |   |   |   |   |   |   Grantham_score > 51:  (0.24) (0.53) (0.03) (0.18) (0.03)
|   |   |   |   |   Repeats = 1:                  (0.32) (0.45) (0.05) (0.14) (0.05)
|   |   |   |   Evol > 0.51
|   |   |   |   |   Hess_value <= 13.35: (0.27) (0.15) (0.04)  (0.5) (0.04)
|   |   |   |   |   Hess_value > 13.35:  (0.28) (0.44) (0.04)  (0.2) (0.04)
|   |   Hess_value > 42.75
|   |   |   Repeats = 0
|   |   |   |   Evol <= 0.59
|   |   |   |   |   Evol <= 0.255:       (0.08) (0.12) (0.73) (0.04) (0.04)
|   |   |   |   |   Evol > 0.255
|   |   |   |   |   |   Evol <= 0.375:   (0.18) (0.03) (0.03) (0.28) (0.49)
|   |   |   |   |   |   Evol > 0.375:    (0.40) (0.13) (0.07) (0.33) (0.07)
|   |   |   |   Evol > 0.59
|   |   |   |   |   Grantham_score <= 139:  (0.02) (0.20) (0.75) (0.02) (0.02)
|   |   |   |   |   Grantham_score > 139:   (0.36) (0.43) (0.07) (0.07) (0.07)
|   |   |   Repeats = 1
|   |   |   |   Hess_value <= 59.5
|   |   |   |   |   Hess_value <= 50.35:    (0.40) (0.15) (0.05) (0.35) (0.05)
|   |   |   |   |   Hess_value > 50.35:     (0.67) (0.13) (0.04) (0.13) (0.04)
|   |   |   |   Hess_value > 59.5:          (0.19) (0.63) (0.06) (0.06) (0.06)
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2797               63.1377 %
Incorrectly Classified Instances      1633               36.8623 %
Kappa statistic                          0.5392
Mean absolute error                      0.1878
Root mean squared error                  0.3177
Relative absolute error                 58.6858 %
```

```
Root relative squared error             79.4156 %
Total Number of Instances             4430

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.505     0.106      0.544     0.505     0.523      0.826    1
                 0.426     0.082      0.566     0.426     0.486      0.778    2
                 0.894     0.091      0.712     0.894     0.792      0.967    3
                 0.475     0.109      0.52      0.475     0.497      0.809    4
                 0.858     0.073      0.745     0.858     0.797      0.964    5
Weighted Avg.    0.631     0.092      0.617     0.631     0.619      0.869

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 447 125  63 207  44 |   a = 1
 170 377  89 153  97 |   b = 2
  12   9 792   9  64 |   c = 3
 181 144  85 421  55 |   d = 4
 12  11  84  19 760 |  e = 5
```

Supplementary Table 1. Tumour suppressor gene orthologues used to estimate the degree of evolutionary conservation of the various gene coding sequences

Gene	*Species*	cDNA sequence identifier	Protein sequence identifier
APC	*Xenopus laevis*	U64442.1	AAB41671.1
	Bos taurus	XM_865627.1	XP_870720.1
	Rattus norvegicus	NM_012499.1	NP_036631.1
	Mus musculus	NM_007462.1	NP_031488.1
ATM	*Gallus gallus*	XM_417160.1	XP_417160.1
	Xenopus laevis	AY668954.1	AAT72929.1
	Rattus norvegicus	XM_236275.3	XP_236275.3
	Sus scrofa	AY587061	AAT01608.1
	Canis familiaris	XM_845871.1	XP_850964.1
	Mus musculus	NM_007499	NP_031525.1
BRCA1	*Gallus gallus*	*NM_204169.1*	*NP_989500.1*
	Xenopus laevis	*AF416868.1*	*AAL13037.1*
	Bos taurus	*NM_178573.1*	*NP_848668.1*
	Rattus norvegicus	*NM_012514.1*	*NP_036646.1*
	Canis familiaris	*NM_001013416.1*	*NP_001013434.1*
	Mus musculus	*NM_009764.2*	*NP_033894.2*
BRCA2	*Gallus gallus*	*NM_204276.1*	*NP_989607.1*
	Danio rerio	*XM_690042.1*	*XP_695134.1*
	Bos taurus	*XM_583622.2*	*XP_583622.2*
	Rattus norvegicus	*NM_031542.1*	*NP_113730.1*
	Canis familiaris	*NM_001006653.4*	*NP_001006654.2*
	Mus musculus	*NM_009765.1*	*NP_033895.1*
CDH1	*Xenopus laevis*	*BC068940.1*	*AAH68940.1*
	Danio rerio	*NM_131820.1*	*NP_571895.1*
	Bos taurus	*NM_001002763.1*	*NP_001002763.1*
	Rattus norvegicus	*NM_031334.1*	*NP_112624.1*
	Canis familiaris	*XM_536807.2*	*XP_536807.2*
	Mus musculus	*NM_009864.1*	*NP_033994.1*
CDKN2A	*Gallus gallus*	*NM_204433.1*	*NP_989764.1*
	Takifugu rubripes	*AJ250231.1*	*CAC12808.1*
	Bos taurus	*XM_868375.1*	*XP_873468.1*
	Rattus norvegicus	*NM_031550.1*	*NP_113738.1*
	Canis familiaris	*XM_538685.2*	*XP_538685.2*
	Mus musculus	*AF044336.1*	*AAC08963.1*
NF1	*Gallus gallus*	XM_415914.1	XP_415914.1
	Takifugu rubripes	AF064564.2	AAD15839.1
	Rattus norvegicus	NM_012609.1	NP_036741.1
	Canis familiaris	XM_537738.2	XP_537738.2
	Mus musculus	NM_010897.1	NP_035027.1
NF2	*Gallus gallus*	NM_204497.2	NP_989828.2
	Danio rerio	NM_212951.1	NP_998116.1
	Bos taurus	XM_611643.2	XP_611643.2
	Rattus norvegicus	XM_341248.2	XP_341249.2
	Canis familiaris	XM_534729.2	XP_534729.2
	Mus musculus	NM_010898.2	NP_035028.2
PTCH1	*Xenopus laevis*	AF302765.1	AAK15463.1
	Gallus gallus	NM_204960.1	NP_990291.1
	Danio rerio	NM_130988.1	NP_571063.1
	Meriones unguiculatus	AB188226.1	BAE78534.1
	Rattus norvegicus	NM_053566.1	NP_446018.1
	Mus musculus	NM_008957.1	NP_032983.1
PTEN	*Xenopus laevis*	AF144732.1	AAD46165.1

	Gallus gallus	XM_421555.1	XP_421555.1
	Bos taurus	XM_613125.2	XP_613125.2
	Canis familiaris	NM_001003192.1	NP_001003192.1
	Rattus norvegicus	NM_031606.1	NP_113794.1
	Mus musculus	NM_008960.2	NP_032986.1
RB1	*Gallus gallus*	NM_204419.1	NP_989750.1
	Rattus norvegicus	XM_344434.2	XP_344435.2
	Canis familiaris	XM_534118.2	XP_534118.2
	Mus musculus	NM_009029.1	NP_033055.1
	Oncorhynchus mykiss	AF102861.1	AAD13390.1
	Notophthalmus viridescens	Y09226.1	CAA70428.1
STK11	*Xenopus laevis*	U24435.1	AAC59904.1
	Danio rerio	NM_001017839.1	NP_001017839.1
	Rattus norvegicus	XM_234900.2	XP_234900.2
	Raja erinacea	AF486831.1	AAL92113.1
	Canis familiaris	XM_542206.2	XP_542206.2
	Mus musculus	NM_011492.1	NP_035622.1
TP53	*Gallus gallus*	NM_205264.1	NP_990595.1
	Danio rerio	NM_131327.1	NP_571402.1
	Bos taurus	NM_174201.2	NP_776626.1
	Rattus norvegicus	NM_030989.1	NP_112251.1
	Canis familiaris	NM_001003210.1	NP_001003210.1
	Mus musculus	NM_011640.1	NP_035770.1
TSC1	*Gallus gallus*	XM_415449.1	XP_415449.1
	Danio rerio	XM_691747.1	XP_696839.1
	Bos taurus	XM_612846.2	XP_612846.2
	Rattus norvegicus	NM_021854.1	NP_068626.1
	Canis familiaris	XM_537808.2	XP_537808.2
	Mus musculus	NM_022887.2	NP_075025.2
TSC2	*Gallus gallus*	XM_414853.1	XP_414853.1
	Takifugu rubripes	AF013614	AAB86682.1
	Bos taurus	XM_581197.2	XP_581197.2
	Rattus norvegicus	NM_012680.2	NP_036812.2
	Canis familiaris	XM_537008.2	XP_537008.2
	Mus musculus	NM_011647.2	NP_035777.2
VHL	*Gallus gallus*	XM_414447.1	XP_414447.1
	Danio rerio	XM_681176.1	XP_686268.1
	Bos taurus	XM_613870.2	XP_613870.2
	Rattus norvegicus	NM_052801.1	NP_434688.1
	Canis familiaris	NM_001008552.1	NP_001008552.1
	Mus musculus	NM_009507.2	NP_033533.1
WT1	*Xenopus laevis*	U42011.1	AAB53152.1
	Gallus gallus	NM_205216.1	NP_990547.1
	Rattus norvegicus	NM_031534.1	NP_113722.1
	Canis familiaris	XM_846479.1	XP_851572.1
	Sus scrofa	NM_001001264.1	NP_001001264.1
	Mus musculus	NM_144783.1	NP_659032.1

SupplementaryTable 2. Differences in distribution of parameters for somatic, germline, shared, somatic recurrent and shared recurrent missense mutations. Observed median and/or mean values are shown in brackets. (Note that the higher values correspond to less conserved genes whereas the low values refer to highly conserved ones).

Parameter	**Observed trend** **(p<0.05)**
Median non-disease associated mutability rate according to Hess et al. [1994]	**shared recurrent >>shared non-recurrent >germline>>somatic~somatic recurrent*** [10.7] [7.9] [7.3] [4.7] [4.7]
Median disease-associated mutability rate according to Krawczak et al. [1998]	**shared recurrent>shared non-recurrent >germline>>somatic~somatic recurrent** [1.42] [1.01] [0.85] [0.53] [0.53]
Mean/median degree of evolutionary conservation	**shared recurrent < shared non-recurrent << somatic non-recurrent** [0.072/0] [0.138/0] [0.265/0.24] **somatic non-recurrent >> germline** [0.265/0.24] [0.18/0]
Mean Grantham score	**germline >somatic recurrent ~somatic non-recurrent** [93] [85] [80] **shared recurrent~shared non-recurrent >> somatic recurrent** [100] [93] [85]
Proportion of CpG-located mutations	**shared recurrent~shared >>germline>>somatic ~somatic recurrent** [0.34] [0.21] [0.12] [0.08] [0.05]
Proportion of CpHpG-located mutations	**shared recurrent~shared >> somatic recurrent** [0.098] [0.082] [0.028]
Proportion of mutations located within or in the vicinity of direct repeats	**somatic>>germline>>recurrent somatic** [0.07] [0.04] [0.02]

Proportion of mutations located within (or in the vicinity of) runs of identical nucleotides	**somatic>>shared** **somatic>>shared recurrent** [0.24] [0.05] [0.24] [0.16] **germline>>shared** **somatic recurrent>>shared** [0.20] [0.05] [0.21] [0.05]

*Inequality **shared>germline>somatic** implies that a significant difference (p<0.05) in the corresponding parameter was observed between each pair of mutational spectra, i.e. shared vs germline, shared vs somatic and germline vs somatic. Symbol '~' denotes the absence of any significant difference between any two mutational spectra with respect to a given parameter. Symbols '>>' or '<<' indicate experiment-wise statistical significance of the observed inequality whereas symbols '<' or '>' indicate gene-wise statistical significance.

Supplementary Table 3. Various parameters of gene-wise somatic and germline missense mutational spectra vs. potential mutational spectra exhibiting either gene-wise (p<0.05) or experiment-wise differences (p<0.05; shaded in light grey) with respect to the parameters measured.

	Non-disease associated mutation rate		**Disease-associated mutation rate**		**Evolutionary conservation rate**		**Grantham score**		**CpG-located missense mutations**		**CpHpG-located missense mutations**	
	Gene symbol	**Median**	**Gene symbol**	**Median**	**Gene symbol**	**Median**	**Gene symbol**	**Median**	**Gene symbol**	**%**	**Gene symbol**	**%**
Somatic mutations			*STK11*	1.66					*STK11*	25		
			PTCH1	1.06								
	APC	8.4	*CDKN2A*	1.01	*CDKN2A*	0.38			*CDKN2A*	20	*CDKN2A*	5.2
	CDKN2A	7.9	*APC*	0.83								
	PTEN	5.6	*PTEN*	0.53								
	TP53	4.6	*TP53*	0.5	*TP53*	0.17			*RB1*	18	*TP53*	2.8
					VHL	0.14			*BRCA2*	16		
									PTCH1	15		
for all 17 genes combined	somatic	4.7	somatic	0.53	somatic	0	somatic	78	somatic	8	somatic	2.5
	control	4.1	control	0.4	control	0.2	control	74	control	2	control	2
	germline	7.2	germline	0.85	germline	0	germline	94	germline	12	germline	3
Germline mutations												
	TSC2	7.2			*TSC2*	0			*BRCA1*	7	*BRCA1*	3.6
	NF1	7.3					*NF1*	98				
	RB1	7.6							*NF1*	7		
	ATM	7.9	*ATM*	0.79	*ATM*	0	*ATM*	98	*ATM*	15	*ATM*	3.8
	BRCA1	7.9	*BRCA1*	0.81	*VHL*	0	*VHL*	99	*BRCA1*	16		
	BRCA2	8.7	*BRCA2*	0.81					*NF1*	18		
			PTEN	0.92							*TSC2*	8.1
			RB1	0.99							*WT1*	10.8
			NF1	1.03								
			TSC2	1.03								

	WT1	10.1	*WT1*	1.22	*WT1*	0			*TSC2*	21		
			CDH1	1.27	*BRCA1*	0.14			*APC*	24		
					CDKN2A	0.29			*CDH1*	26		

Supplementary Table 4. Summary of mutations occurring in direct repeats of length ≥8 bp in the 17 tumour suppressor genes.

Gene symbol	Proportion of gene length covered by repeats (%)	Number of missense mutations found in repeats						Number of micro-deletions and micro-insertions found in repeats			
		somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	Total	somatic non-recurrent	germline	shared	Total
APC	4	3	0	0	0	0	**3**	17	21	1	**17**
ATM	7	2	0	0	0	0	**2**	0	11	0	**0**
BRCA1	5	0	9	0	0	0	**9**	1	8	0	**1**
BRCA2	2	0	0	0	0	0	**0**	1	12	0	**1**
CDH1	3	0	0	0	0	0	**0**	0	1	0	**0**
CDKN2A	17	25	8	3	0	0	**36**	28	2	0	**28**
NF1	7	0	2	0	0	0	**2**	0	15	0	**0**
NF2	3	0	0	0	0	0	**0**	1	1	0	**1**
PTCH1	3	0	0	0	0	0	**0**	0	0	0	**0**
PTEN	17	7	0	0	4	2	**13**	20	5	1	**20**
RB1	12	0	1	0	0	0	**1**	2	12	0	**2**
STK11	10	0	3	1	0	0	**4**	0	6	0	**0**
TP53	14	24	1	0	13	2	**40**	21	0	0	**21**
TSC1	5	0	1	0	0	0	**1**	0	4	0	**0**
TSC2	5	0	10	1	0	0	**11**	0	6	0	**0**
VHL	6	0	1	0	0	0	**1**	0	1	0	**0**
WT1	7	1	0	0	0	0	**1**	0	0	0	**0**
TOTAL	**6**	**62**	**36**	**5**	**17**	**4**	**124**	**91**	**105**	**2**	**91**

Supplementary Table 5. Summary of mutations occurring in inverted repeats of length ≥8 bp in the 17 tumour suppressor genes.

Gene symbol	Proportion of gene length covered by repeats (%)	Number of missense mutations found in repeats						Number of micro-deletions and micro-insertions found in repeats			
		somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	Total	somatic non-recurrent	germline	shared	Total
APC	6	5	4	1	1	0	**5**	21	27	2	**50**
ATM	13	1	14	0	0	0	**1**	1	16	0	**17**
BRCA1	6	0	15	0	0	0	**0**	0	22	1	**23**
BRCA2	7	3	1	0	0	0	**3**	1	27	0	**28**
CDH1	5	0	1	0	0	0	**0**	1	0	0	**1**
CDKN2A	8	30	5	6	2	1	**30**	13	2	1	**16**
NF1	11	0	3	0	0	0	**0**	1	24	0	**25**
NF2	10	1	3	0	0	0	**1**	11	6	0	**17**
PTCH1	5	1	0	0	0	0	**1**	0	2	0	**2**
PTEN	6	10	1	1	4	1	**10**	9	2	0	**11**
RB1	16	4	5	1	0	0	**4**	7	28	0	**35**
STK11	13	1	5	0	1	0	**1**	1	9	0	**10**
TP53	5	13	0	0	51	9	**13**	53	2	0	**55**
TSC1	5	0	1	0	0	0	**0**	0	7	0	**7**
TSC2	9	0	6	0	0	0	**0**	1	13	0	**14**
VHL	12	9	8	1	1	0	**9**	36	15	2	**53**
WT1	7	0	2	0	0	0	**0**	0	0	0	**0**
TOTAL	**9**	**78**	**74**	**10**	**60**	**11**	**78**	**156**	**202**	**6**	**364**

Supplementary Table 6. Summary of mutations occurring within symmetric repeats of length ≥8 bp in the 17 tumour suppressor genes.

Gene symbol	Proportion of gene length covered by repeats (%)	Number of missense mutations found in repeats						Number of micro-deletions and micro-insertions found in repeats			
		somatic non-recurrent	germline	shared non-recurrent	somatic recurrent	shared recurrent	Total	somatic non-recurrent	germline	shared	Total
APC	16	5	2	0	2	0	**9**	58	87	6	**151**
ATM	32	2	11	0	0	0	**13**	2	43	0	**45**
BRCA1	20	1	30	0	0	0	**31**	0	82	2	**84**
BRCA2	18	6	18	0	0	0	**24**	2	79	3	**84**
CDH1	24	4	0	0	0	0	**4**	5	8	0	**13**
CDKN2A	24	49	13	5	2	0	**69**	35	7	1	**43**
NF1	31	1	20	0	0	0	**21**	2	85	2	**89**
NF2	24	6	3	0	1	0	**10**	49	12	3	**64**
PTCH1	23	5	8	1	0	0	**14**	5	23	0	**28**
PTEN	44	27	3	1	9	0	**40**	42	13	1	**56**
RB1	48	3	10	1	0	0	**14**	4	41	1	**46**
STK11	33	3	6	0	2	0	**11**	1	20	1	**22**
TP53	30	60	2	1	132	23	**218**	147	1	0	**148**
TSC1	23	0	3	0	0	0	**3**	0	27	0	**27**
TSC2	23	0	13	0	0	0	**13**	1	29	0	**30**
VHL	17	3	9	2	0	2	**16**	25	7	2	**34**
WT1	26	0	6	0	0	0	**6**	3	4	0	**7**
TOTAL	**25**	**175**	**157**	**11**	**148**	**25**	**516**	**381**	**568**	**22**	**971**

Suplementary Table 7. Occurrence of missense mutations in repeats/runs of identical nucleotides and/or CpG/CpHpG oligonucleotides

Type of Repeats	Type of mutational spectrum	Number of mutations			Remaining number of mutations
		exclusively in repeats/runs	exclusively in CpG/CpHpG	in both repeats/runs and CpG/CpHpG	
Runs	somatic non-recurrent	184	58	24	618
	germline	151	100	27	608
	somatic recurrent	167	46	18	636
	shared non-recurrent	5	28	0	69
	shared recurrent	10	38	5	59
	potential	32861	3902	765	111495
Direct	somatic non-recurrent	52	72	10	750
	germline	31	122	5	728
	somatic recurrent	14	61	3	789
	shared non-recurrent	3	26	2	71
	shared recurrent	2	41	2	67
	potential	5252	4431	236	139104

Inverted	somatic non-recurrent	65	69	13	737
	germline	64	117	10	695
	somatic recurrent	55	59	5	748
	shared non-recurrent	8	26	2	66
	shared recurrent	7	39	4	62
	potential	10790	4314	353	133566
Symmetric	somatic non-recurrent	155	62	20	647
	germline	140	110	17	619
	somatic recurrent	137	53	11	666
	shared non-recurrent	7	24	4	67
	shared recurrent	16	34	9	53
	potential	28646	3752	915	115710

Supplementary Table 8. Truncating vs. non-truncating lesions

Gene		Missense	Nonsense	Micro-deletions	Micro-insertions	Micro-indels	Non-truncating lesions	Truncating lesions	Ratio of non-truncating to truncating lesions	Ratio of truncating somatic to truncating germline lesions
APC	Somatic	39	79	152	44	3	39	278	0.14	0.46
	Germline	23	180	299	115	12	23	606	0.04	
ATM	Somatic	11	7	4	1	0	11	12	0.92	0.05
	Germline	76	75	122	35	14	76	246	0.31	
BRCA1	Somatic	6	9	9	5	0	6	23	0.26	0.05
	Germline	170	121	259	85	12	170	477	0.36	
BRCA2	Somatic	21	1	8	4	0	21	13	1.62	0.03
	Germline	86	76	247	90	11	86	424	0.20	
CDH1	Somatic	15	7	13	2	0	15	22	0.68	0.69
	Germline	19	11	12	8	1	19	32	0.59	
CDKN2A	Somatic	198	18	77	25	8	198	128	1.55	4.74
	Germline	62	7	11	7	2	62	27	2.30	
NF1	Somatic	2	11	16	3	0	2	30	0.07	0.07
	Germline	83	115	221	105	8	83	449	0.18	
NF2	Somatic	23	42	182	28	6	23	258	0.09	2.22
	Germline	20	43	55	16	2	20	116	0.17	
PTCH1	Somatic	14	9	14	6	1	14	30	0.47	0.28
	Germline	24	27	42	32	8	24	109	0.22	
PTEN	Somatic	226	56	152	51	4	226	263	0.86	3.21
	Germline	45	28	29	22	3	45	82	0.55	
RB1	Somatic	25	27	34	12	3	25	76	0.33	0.30

	Germline	37	76	117	53	11	37	257	0.14	
STK11	Somatic	20	10	5	1	1	20	17	1.18	0.17
	Germline	30	27	47	24	3	30	101	0.30	
TP53	Somatic	1229	96	512	238	0	1229	846	1.45	24.89
	Germline	94	10	16	5	3	94	34	2.76	
TSC1	Somatic	2	1	1	0	0	2	2	1.00	0.02
	Germline	7	37	53	25	4	7	119	0.06	
TSC2	Somatic	2	1	3	2	1	2	7	0.29	0.03
	Germline	89	74	110	46	3	89	233	0.38	
VHL	Somatic	88	15	180	44	1	88	240	0.37	1.82
	Germline	143	27	63	37	5	143	132	1.08	
WT1	Somatic	1	3	4	3	0	1	10	0.10	0.37
	Germline	40	14	8	4	1	40	27	1.48	
Total	Somatic	1922	392	1366	469	28	1922	2255	0.85	0.65
	Germline	1048	948	1711	709	103	1048	3471	0.30	