

 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

Related papers:

- [On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease.](#)
- [Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes](#)
- [Human DNA methylomes at base resolution show widespread epigenomic differences](#)
- [Contrasting Determinants of Mutation Rates in Germline and Soma](#)
- [Codon-level co-occurrences of germline variants and somatic mutations in cancer are rare but often lead to incorrect variant annotation and underestimated impact prediction.](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/comparative-analysis-of-germline-and-somatic-microlesion-j5pk4qv64w>



HAL
open science

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

► **To cite this version:**

Nadia Chuzhanova, Dobril Ivanov, Stephen Edward Hamby, Peter Stentson, Andrew D Phillips, et al.. Comparative Analysis of Germline and Somatic Micro-lesion Mutational Spectra in 17 Human Tumour Suppressor Genes. *Human Mutation*, Wiley, 2011, 10.1002/humu.21483 . hal-00625570

HAL Id: hal-00625570

<https://hal.archives-ouvertes.fr/hal-00625570>

Submitted on 22 Sep 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



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

Journal:	<i>Human Mutation</i>
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

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

Dobril Ivanov^{1,2}, Stephen E. Hamby³, Peter D. Stenson¹, Andrew D. Phillips¹,
Hildegard Kehrer-Sawatzki⁴, David N. Cooper¹ and Nadia Chuzhanova³

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

²MRC 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

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

⁴Institute 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

1
2
3 which serves to explain not only the phenomenon of genetic anticipation characteristic of these
4 disorders but also their inherent inter-individual clinical variability [Giovannone et al., 1997;
5
6
7
8 Leeflang et al., 1999; Martorell et al., 2000; Sharma et al., 2002; Pollard et al., 2004]. However,
9
10 by contrast, highly variable human minisatellites can display markedly different degrees of
11
12 instability between the soma and the germline [Buard et al., 2000; Stead and Jeffreys, 2000;
13
14
15 Shanks et al., 2008]. These studies notwithstanding, few attempts have so far been made to
16
17 compare the nature, location and relative frequency of germline and somatic mutations.
18

19
20 Human cancer genes usually harbour either somatic or germline mutations [Goode et al., 2002;
21
22 Futreal et al., 2004; Vogelstein and Kinzler, 2004]. There is, however, one category of cancer
23
24 gene, broadly termed tumour suppressors, that by virtue of their being mutated in both the
25
26 germline and the soma, provides us with an ideal model system to compare somatic vs. germline
27
28 mutational spectra [Futreal et al., 2004]. Tumour suppressor genes, defined as “genes that sustain
29
30 loss-of-function mutations in the development of cancer” [Haber and Harlow, 1997], are
31
32 involved in the regulation of a diverse array of different cellular functions including cell cycle
33
34 checkpoint control, detection and repair of DNA damage, protein ubiquitination and degradation,
35
36 mitogenic signalling, cell specification, differentiation and migration, and tumour angiogenesis
37
38 [Sherr, 2004]. They encode proteins with a regulatory role in cell cycle progression (e.g. Rb),
39
40 DNA-binding transcription factors (e.g. p53) and inhibitors of cyclin-dependent kinases required
41
42 for cell cycle progression (e.g. p16). In inherited cancer syndromes, the mutational inactivation
43
44 of both tumour suppressor alleles is required to change the phenotype of the cell. This ‘two hit
45
46 hypothesis’ provides the basis for our mechanistic understanding of tumour suppressor gene
47
48 mutagenesis: a first (inherited) mutation in one tumour suppressor allele is followed by the
49
50 somatic loss of the remaining wild-type allele via a number of different mutational mechanisms
51
52 [Knudson, 2001]. Whereas the inherited lesion is usually fairly subtle, the second (somatic) hit
53
54 may also involve the deletional loss of the entire gene or even a substantial portion of the
55
56
57
58
59
60

1
2
3 chromosome involved. Alternatively, both 'hits' may constitute somatic mutations: whatever the
4
5 actual mechanism, the end result is the same – the loss or inactivation of both gene copies. Some
6
7 interplay may however occur between the soma and the germline in that the location of the
8
9 germline mutation can in some instances influence the nature, frequency and location of the
10
11 subsequent somatic mutation [Lamlum et al., 1999; Groves et al., 2002; Latchford et al., 2007;
12
13 Dallosso et al., 2009].

14
15
16
17 Tumour suppressor genes are often somatically inactivated by mutational mechanisms that are
18
19 almost exclusively confined to the soma and which are found only infrequently in the germline
20
21 (e.g. gross mutations characterized by loss of heterozygosity, epi-mutations such as methylation-
22
23 mediated promoter inactivation, and micro-lesions within highly repetitive sequence elements
24
25 that are consequent to microsatellite instability). However, a typical spectrum of somatic
26
27 mutations associated with tumorigenesis may also include gross rearrangements, copy number
28
29 variation, and various types of micro-lesion (e.g. micro-deletions, micro-insertions and indels)
30
31 including single base-pair substitutions [Loeb and Harris, 2008; Stratton et al., 2009]. Although
32
33 the somatic micro-lesions are often quite similar to their germline counterparts, few studies of
34
35 tumour suppressor genes have so far attempted to compare and contrast germline and somatic
36
37 mutational spectra with respect to these relatively subtle types of mutation. However, such
38
39 studies have indicated that germline and somatic micro-lesions can display remarkable
40
41 similarities in terms of mutation type, location and relative frequency of occurrence, and hence
42
43 by inference the putative underlying mechanisms of mutagenesis [Marshall et al., 1997; Ali et
44
45 al., 1999; Gallou et al., 1999; Richter et al., 2003; Upadhyaya et al., 2004; Glazko et al., 2004;
46
47 Tartaglia et al., 2006; Baser et al., 2006; Upadhyaya et al., 2008].

48
49
50
51 We attempt here a first formal comparison between germline and somatic micro-lesion
52
53 mutational spectra for a total of 17 different human tumour suppressor genes [*APC* (MIM#
54
55 611731), *ATM* (MIM# 607585), *BRCA1* (MIM# 113705), *BRCA2* (MIM# 600185), *CDHI*

1
2
3 (MIM# 192090), *CDKN2A* (MIM# 600160), *NF1* (MIM# 162200), *NF2* (MIM# 607379),
4
5 *PTCH1* (MIM# 601309), *PTEN* (MIM# 601728), *RBI* (MIM# 180200), *STK11* (MIM# 602216),
6
7 *TP53* (MIM# 191170), *TSC1* (MIM# 605284), *TSC2* (MIM# 191092), *VHL* (MIM# 608537) and
8
9 *WT1* (MIM# 607102)].

14 15 **Materials and Methods**

16 17 *Sources of germline and somatic mutation data*

18
19
20 Data on germline and somatic micro-lesions (viz. missense mutations, micro-deletions and
21
22 micro-insertions involving ≤ 20 bp) were collated for 17 different human tumour suppressor
23
24 genes. Germline mutation data were obtained from the Human Gene Mutation Database
25
26 [HGMD; <http://www.hgmd.org>; Stenson et al., 2009]. Somatic mutation data were compiled
27
28 from a number of different sources including online somatic mutational databases viz. *Catalogue*
29
30 *of Somatic Mutations in Cancer* (<http://www.sanger.ac.uk/genetics/CGP/cosmic>; *RBI* and
31
32 *PTEN*), the *Breast Cancer Information Core* (<http://research.nhgri.nih.gov/bic>; *BRCA1*), the *RBI*
33
34 *Gene Mutation Database* (<http://www.verandi.de/joomla>; *RBI*), the *International NF2 Mutation*
35
36 *Database* (<http://www.hgmd.cf.ac.uk/nf2>; *NF2*), the *CDKN2A Database*
37
38 (<https://biodesktop.uvm.edu/perl/p16>; *CDKN2A*) and the *IARC TP53 Mutation Database*
39
40 (<http://www-p53.iarc.fr>; *TP53*), the *VHL Mutations Database* (<http://www.umd.be/VHL/>), and
41
42 data privately communicated by Eamonn Maher (*VHL*) and Gareth Evans (*NF2*). Additional
43
44 somatic mutation data [for *APC*, *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *PTCH1*, *STK11*, *TSC1*,
45
46 *TSC2* and *WT1*] were obtained by searching PubMed.

47
48 To be regarded as *bona fide* somatic mutations, and therefore suitable for inclusion in this
49
50 analysis, reported lesions had to have been shown not only to be present in a tumour tissue but
51
52 also to be absent from a non-tumour tissue (usually blood) from the same patient. Hence,
53
54 mutational data derived from 'sporadic' patients were not included unless a non-tumour tissue
55
56
57
58
59
60

1
2
3 had also been examined in order to exclude the possibility that the lesions detected were
4 constitutional in origin. Depending upon the number of independent occurrences, f , of a given
5 somatic or shared mutation described in the literature, these mutation types were further
6 subdivided into two categories: *recurrent mutations* ($f > 1$) and *non-recurrent mutations* ($f = 1$). At
7 the time this study was initiated (October 2006), the number of available germline and somatic
8 missense mutations for each of the 17 studied tumour suppressor genes were as listed in Table 1.
9

10
11
12
13
14
15
16
17
18 The analysis reported here focussed exclusively on missense mutations and micro-deletions/
19 micro-insertions. Nonsense mutations in tumour suppressor genes have already been addressed
20 in the context of a general meta-analysis of this type of lesion [Mort et al., 2008]. Indels
21 (representing a combination of micro-deletion and micro-insertion) were excluded from the
22 analysis owing to their paucity.
23
24
25
26
27
28
29
30
31

32 *Control datasets of potential mutations*

33
34 For every tumour suppressor gene examined, all possible single base-pair substitutions in the
35 gene coding sequence that (i) could potentially have given rise to a missense mutation and (ii)
36 were not already included in either of the corresponding observed somatic and/or germline
37 mutational spectra, were generated. These 'potential missense mutations' were used as a control
38 dataset.
39
40
41
42
43
44
45

46 For each tumour suppressor gene, a matching control dataset of 'potential micro-deletions' was
47 also generated by randomly selecting a first breakpoint and then choosing the length of the
48 simulated micro-deletion (and therefore, the position of the second breakpoint) by reference to
49 the probability distribution calculated for micro-deletions (from 1 bp to 20 bp) observed in the
50 corresponding dataset of mutations. A matching dataset of micro-insertions was generated in
51 similar fashion, with the sites of insertion being randomly selected. Since some of the micro-
52 deletion/micro-insertion breakpoints occurred within an intron, extended cDNA sequences
53
54
55
56
57
58
59
60

1
2
3 comprising exons and additional flanking intron sequences were used to generate corresponding
4
5 control datasets.
6
7
8
9

10 *Grantham scores*

11
12 The ‘Grantham score’ or ‘Grantham difference’ [Grantham, 1974] measures the chemical
13
14 difference between wild-type and mutated amino acid residues in terms of their side chain
15
16 composition (i.e. the weight ratio of non-carbon components in end-groups or rings to carbons in
17
18 side chains), polarity (i.e. basic, acidic or nonpolar depending upon side chain charge) and
19
20 molecular volume.
21
22
23

24
25 On average, the physicochemical differences manifested by orthologous amino acid
26
27 substitutions that have accumulated over evolutionary time will tend to be relatively small. By
28
29 contrast, disease-causing substitutions are expected to exhibit higher Grantham scores, indicative
30
31 of more dramatic physicochemical differences between the wild-type and mutated amino acid
32
33 residues [Krawczak et al., 1998]. The values tabulated by Grantham [1974] were used in this
34
35 study to calculate a median Grantham score for each set of missense mutations for each tumour
36
37 suppressor gene.
38
39
40
41
42

43 *Degree of evolutionary conservation*

44
45 Amino acid residues that are highly conserved in orthologous proteins frequently represent sites
46
47 of structural or functional importance. Hence, such highly conserved amino acid residues/protein
48
49 regions often constitute hotspots for observed pathological mutations as a consequence of
50
51 phenotype selection (rather than intrinsic mutability). To assess the degree of evolutionary
52
53 conservation of those codons affected by somatic/germline mutations, orthologous tumour
54
55 suppressor cDNA and protein sequences from different vertebrate species were retrieved from
56
57 NCBI’s Entrez Gene database (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>). The species
58
59
60

1
2
3 used as a source of these cDNA and protein sequences are listed in Supp. Table 1 for each
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 $\frac{K_a}{K_a + K_s}$ ratio for each codon where K_s and K_a 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

1
2
3 low implying that these specific substitutions are much less likely to occur as neutral
4
5 substitutions.
6

7
8 The nearest neighbour-dependent mutation rates derived from germline single base-pair
9
10 substitutions [using data from the Human Gene Mutation Database (HGMD); Stenson et al.,
11
12 2009] by Krawczak et al. [1998] were used as an approximation of the *disease-associated*
13
14 *mutability rate*. This mutation rate is a function of selection for loss of biological function as
15
16 well as the underlying intrinsic mutability of the DNA sequence.
17
18

19 20 21 22 *Repetitive sequence elements*

23
24 A variety of repetitive sequence elements have been reported in association with human gene
25
26 mutations causing both inherited disease and cancer. Direct and inverted repeats and symmetric
27
28 elements [see Chuzhanova et al. 2003 for definitions] of length ≥ 8 bp, and less than 21 bp apart,
29
30 capable of forming non-B DNA structures, were therefore sought within the extended cDNA
31
32 sequences (comprising exons and up to ± 85 bp of flanking sequence) using purposely designed
33
34 software. In addition, DNA sequences were screened for the presence of mononucleotide runs of
35
36 ≥ 4 bp.
37
38
39
40
41
42

43 44 *Mutation descriptors*

45
46 Each missense mutation was ascribed various descriptors indicating (a) the type of mutation [i.e.
47
48 shared mutation (i.e. found to occur both somatically and in the germline); exclusively somatic;
49
50 exclusively germline; shared recurrent mutation (i.e. found to occur not only in the germline but
51
52 also somatically on more than one occasion; somatic recurrent mutations (recorded in the soma
53
54 more than once, but not in the germline); potential mutation (as defined above)] and (b) its
55
56 location [i.e. C \rightarrow T and G \rightarrow A within a CpG dinucleotide or within a CpHpG trinucleotide
57
58 (where H=A, C or T) or in a repeat sequence (as described above)]. Mutations that have been
59
60

1
2
3 reported as being exclusively somatic or exclusively germline will henceforth be referred to
4 simply as 'somatic' and 'germline', respectively. The shared mutations, comprising the overlap
5 between the somatic and germline mutations, may be visualized in the form of a Venn diagram
6 (Figure 1). All somatic missense (including shared) mutations were further described as being
7 either recurrent or non-recurrent (in the soma, see above; Figure 1). No such division was made
8 for the relatively small number of recurrent micro-deletions and micro-insertions available; both
9 recurrent and non-recurrent somatic mutations were therefore included in either the somatic or
10 the shared datasets and labelled accordingly (Figure 1).
11
12
13
14
15
16
17
18
19
20
21

22 All micro-lesions (*viz.* missense mutations, micro-deletions and micro-insertions) in each gene
23 were also labelled with respect to their occurrence within a region spanning a repetitive element
24 or mononucleotide run including ± 5 bp of flanking sequence. If a missense mutation (or at least
25 one micro-deletion/micro-insertion breakpoint) was found to occur within this extended region,
26 the micro-lesion was labelled as being found in association with the corresponding type of
27 repeat.
28
29
30
31
32
33
34
35
36
37
38

39 *Assessing the statistical significance of the results generated*

40
41 To assess the similarity (or dissimilarity) of the germline and somatic mutational spectra with
42 respect to (i) the frequency with which the missense mutations were located within CpG/non-
43 CpG dinucleotides or CpHpG/non-CpHpG trinucleotides and (ii) the frequency with which the
44 micro-deletions/micro-insertions were found within/outwith repeats, the various non-overlapping
45 mutation datasets (bearing specific descriptors) were compared by means of the χ^2 test. Since the
46 normality assumption did not hold for the datasets studied, the Wilcoxon rank-sum test was used
47 to compare and contrast missense mutational spectra with respect to the Grantham score, degree
48 of evolutionary conservation, and both the non-disease- and disease-associated mutability rates.
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 The permutation-based method [Olshen and Jain, 2002] was used to estimate the significance
4 of our findings and to allow for multiple testing wherever appropriate. For each comparison, the
5 null hypothesis [viz. no overall difference between two groups of mutations (e.g. somatic and
6 potential) with respect to the specific property in question (e.g. occurrence in CpG or non-CpG
7 nucleotides)], was tested for, either in the context of each gene or all genes combined. χ^2 or rank-
8 sum statistics were calculated for the observed germline and somatic mutations as well as for
9 10,000 control sets of mutations created from the original sets by random permutation of the
10 assigned mutational descriptors (e.g. randomly chosen mutations labelled as ‘somatic’ were re-
11 labelled as ‘germline’; randomly chosen mutations labelled as ‘shared’ were re-labelled as
12 ‘somatic’, etc.). The test statistic (χ^2 or rank-sum) for the original datasets that exceeded the 95th
13 percentile of χ^2 maxima for 10,000 control sets was deemed to be statistically significant; the
14 corresponding p-value was termed the ‘gene-wise’ p-value. To allow for multiple testing in those
15 cases where specific mutations in all genes were combined, a Bonferroni correction was applied;
16 the corresponding p-value was termed the ‘experiment-wise’ p-value.
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38

39 *Naïve Bayes classifier*

40
41 A decision tree classifier known as a Naïve Bayes tree [NBTree; Kohavi, 1996], implemented in
42 the Weka machine learning package [Witten and Frank, 2005], was trained to discriminate
43 between somatic, germline, shared, recurrent somatic and recurrent shared missense mutations.
44
45 Each mutation was described by a total of six features including the degree of evolutionary
46 conservation, the non-disease-associated and disease-associated relative mutability rates,
47 Grantham score, and occurrence in CpG/CpHpG, non-CpG/non-CpHpG doublets/triplets or in
48 repeats/mononucleotide runs. Ten-fold cross-validation was used to assess the accuracy of
49 classification. The mutation datasets were balanced using random oversampling [Kotsiantis et
50
51
52
53
54
55
56
57
58
59
60

1
2
3 al., 2006] by replicating random instances from the minority classes until all classes were
4
5 represented by the same number of instances as the majority class.
6
7
8
9

10 11 12 13 **Results and Discussion**

14
15 The availability of both germline and somatic mutational spectra from tumour suppressor genes
16 provides us with an ideal opportunity to study the nature of mutation of the same gene sequences
17 in both the germline and the soma. The analysis reported here explores for the first time the
18 similarities and differences exhibited by the germline, somatic (and shared) micro-lesion
19 mutational spectra in 17 human tumour suppressor genes. The study presented here focussed
20 upon missense mutations and micro-deletions as well as micro-insertions. Nonsense mutations in
21 tumour suppressor genes have already been addressed elsewhere in the context of a general
22 meta-analysis of this type of lesion [Mort et al., 2008].
23
24
25
26
27
28
29
30
31
32
33
34
35
36

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

38
39 Taken together, the combined mutational spectra for all 17 tumour suppressor genes contained
40 twice as many somatic (61%) as germline (31%) mutations. For five genes (*APC*, *CDKN2A*,
41 *NF2*, *PTEN* and *TP53*), a predominance of somatic over germline mutations was noted, with the
42 *TP53* gene having the highest proportion of somatic mutations (92%). For the majority of genes,
43 however (namely *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *PTCH1*, *RBI*, *STK11*, *TSC1*, *TSC2*, *VHL*
44 and *WT1*), the analysed dataset included more germline than somatic mutations, with >97% of all
45 mutations in the *BRCA1*, *NF1*, *TSC2* and *WT1* genes being germline in origin.
46
47
48
49
50
51
52
53
54

55
56 Shared mutations are of particular interest because identical mutational mechanisms operating
57 in the germline and the soma may be inferred for such lesions. The expected number of shared
58 mutations for each gene was calculated as $p_{\text{somatic}} \times p_{\text{germline}} \times (\text{total number of mutations})$,
59
60

1
2
3 where p denotes the relative frequencies of somatic and germline mutations. Although the
4
5 proportion of shared mutations varies markedly between genes (from 0% to 25% of the total),
6
7 only two genes (*TP53* and *VHL*) were found to have a higher than expected number of shared
8
9 mutations as calculated above.
10
11

12 13 14 15 *Patterns of germline and somatic missense mutations by mutation type*

16
17 Missense mutations were characterised by a predominance of transitions over transversions
18
19 (Figure 2). The transition:transversion ratio was at its highest for shared recurrent mutations (3.5)
20
21 and shared non-recurrent mutations (2.7). By contrast, the transition:transversion ratio for the
22
23 control group (i.e. potential mutations) was 0.85. Significant differences in the
24
25 transition:transversion ratio were observed between all mutation types ($p < 0.05$) with the
26
27 exception of germline vs. shared mutations (Figure 2).
28
29
30

31
32 Not surprisingly, a strong positive correlation was noted between somatic and shared
33
34 mutational spectra (Pearson's correlation $r = 0.986$, $p = 2.91 \times 10^{-4}$) with respect to the frequencies
35
36 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
37
38 C.G>T.A. Weaker negative correlations were found between somatic mutations and the control
39
40 dataset of mutations ($r = -0.887$, $p = 0.019$) and between shared and the control ($r = -0.837$,
41
42 $p = 0.038$) mutational spectra, indicative of the non-randomness of somatic mutation.
43
44
45

46
47 C.G>T.A transitions constituted the most frequent type of mutation in shared (46%), germline
48
49 (29%) and somatic (25%) mutational spectra, significantly higher proportions than noted in the
50
51 spectrum of mutations within our control dataset (13%, $p < 0.001$) (Figure 2). Intriguingly, the
52
53 number of A.T>G.C mutations was significantly higher (28%) in the germline as compared to
54
55 the somatic (16%), shared (17%) and control (16%) mutational spectra (Figure 2). A.T>C.G
56
57 mutations were significantly under-represented in the shared mutational spectrum (7%, $p < 0.001$)
58
59 as compared to the other spectra whereas A.T>T.A mutations were under-represented (7%,
60

1
2
3
4 p<0.001) in both the germline and shared mutational spectra compared to both somatic and
5
6 potential mutations (Figure 2). Finally, C.G>A.T mutations were significantly underrepresented
7
8 in the germline mutational spectrum (10%) as compared to the somatic (16%, $p=1.2\times 10^{-5}$) and
9
10 potential (15%, $p=2.6\times 10^{-5}$) spectra. Thus, the main similarity between the somatic and germline
11
12 missense mutational spectra was in relation to C.G>T.A transitions whereas the main differences
13
14 between these spectra involved the A.T>G.C, A.T>T.A and C.G>A.T mutations. It should be
15
16 noted that the patterns of somatic nucleotide substitution exhibited by the 17 tumour suppressor
17
18 genes studied here were markedly different from the genome-wide patterns of somatic nucleotide
19
20 substitution previously observed in various cancer genome sequencing studies [Sjöblom et al.,
21
22
23
24
25 2006; Greenman et al., 2007; Kan et al., 2010].
26
27
28

29 *CpG- and CpHpG-located missense mutations*

30
31
32 The CpG dinucleotide is a well known mutational hotspot in the human genome as a
33
34 consequence of the spontaneous (and endogenous) deamination of 5-methylcytosine. In addition,
35
36 Lister et al. [2009] reported abundant DNA methylation in CpHpG trinucleotides in the human
37
38 genome, where H is either A, C or T, raising the possibility that CpHpG might also be a
39
40 generalized mutation hotspot [Cooper et al., 2010].
41
42

43
44 The proportion of missense mutations that were either C>T or G>A within CpG or CpHpG
45
46 oligonucleotides in the 17 tumour suppressor genes was found to vary between 0% and 100%
47
48 (Table 2). This wide range in values may be attributed to the small size of some of the gene
49
50 mutation datasets under study. Importantly, the CpG and CpHpG oligonucleotides were found to
51
52 be disproportionately likely to harbour shared mutations; thus, 34% of shared recurrent
53
54 mutations and 21% of shared non-recurrent mutations were C>T and G>A mutations in CpG
55
56 dinucleotides with an additional 10% and 9% of mutations, respectively, occurring within
57
58 CpHpG trinucleotides. Since driver mutations tend to occur disproportionately frequently within
59
60

1
2
3 CpG dinucleotides [Talavera et al., 2010], we postulate that missense mutations identified as
4
5
6 being shared are highly likely to be driver mutations.
7

8 Significant differences were noted between the relative frequencies of CpG- and CpHpG-
9
10 located mutations for somatic, germline, shared, somatic recurrent and shared recurrent missense
11
12 mutations (Supp. Table 2).
13

14
15 We have previously shown that 18.2% and 9.9% of all missense/nonsense mutations recorded
16
17 in the HGMD are C>T and G>A transitions in CpG and CpHpG oligonucleotides respectively
18
19 [Cooper et al., 2010]. In the present study, we observed that the mutational spectra of shared and
20
21 shared recurrent missense mutations in tumour suppressor genes were both found to be
22
23 significantly enriched in CpG-located mutations (χ^2 -test; p-values, 0.028 and 1.1×10^{-9}
24
25 respectively). This implies that the CpG dinucleotide is a generalized mutation hotspot in both
26
27 the soma and the germline as a consequence of the endogenous mutational mechanism of
28
29 methylation-mediated deamination of 5-methylcytosine. By contrast, the number of CpG-located
30
31 mutations was significantly underrepresented (χ^2 -test; p-values $< 5 \times 10^{-14}$) in the other mutational
32
33 spectra (i.e. non-recurrent somatic, somatic recurrent and germline mutations) by comparison
34
35 with HGMD data. To perform these comparisons, missense mutations (Table 2) and nonsense
36
37 mutations [previously reported in Mort et al., 2008; see Table 6 therein] in all 17 tumour
38
39 suppressor genes were combined. The proportion of shared recurrent missense mutations in
40
41 tumour suppressor genes that were CpHpG-located was found to be significantly higher
42
43 ($p=0.023$) than for mutations recorded in the HGMD whereas CpHpG-located somatic and
44
45 recurrent somatic mutations were significantly under-represented ($p < 4 \times 10^{-10}$). Significant
46
47 enrichment in CpHpG-located mutations was observed for germline mutations as compared to
48
49 somatic mutations ($p < 3 \times 10^{-10}$) consistent with the reported decrease in CpHpG methylation in
50
51 differentiated cells [Lister et al., 2009]. In summary, germline and shared missense mutations
52
53 were found to be significantly enriched at CpG and CpHpG oligonucleotides.
54
55
56
57
58
59
60

1
2
3 The numbers of somatic and shared C>T and G>A transitions recorded within CpG
4 dinucleotides for each gene (Table 2) did not correlate with the numbers of CpG dinucleotides
5 found in these genes ($r < -0.5$, $p > 0.127$) and hence do not simply reflect intragenic CpG
6 frequency. A weak positive correlation between CpG-located mutations and the number of genic
7 CpG dinucleotides was however noted for germline mutations ($r = 0.489$, $p = 0.046$) indicating that
8 CpG methylation is not entirely unrelated to the number of CpG dinucleotides, at least with
9 respect to the germline; the relationship is however clearly more complex in the soma, possibly
10 due to inter-tissue differences in gene methylation patterns [Tornaletti and Pfeifer, 1995] or
11 transcription-coupled repair [Rubin and Green, 2009].
12
13
14
15
16
17
18
19
20
21
22
23

24 No correlation was found between the numbers of somatic, germline and shared mutations
25 recorded within CpHpG trinucleotides and the corresponding numbers of CpHpG trinucleotides
26 for these genes ($r = -0.316, 0.373, -0.414$; p -values 0.281, 0.216 and 0.098, respectively)
27 indicating that mutation within CpHpG trinucleotides is likely to be very much a gene-specific
28 phenomenon (presumably dependent on both the extent and the degree of spatial localization of
29 CpHpG methylation in the germline and/or soma).
30
31
32
33
34
35
36
37

38 Finally, the number of CpG dinucleotides in the various tumour suppressor genes studied
39 (Table 2) was not found to correlate with gene length ($r = 0.3$, p -value=0.241). By contrast, we
40 found a significant correlation ($r = 0.885$, p -value= 2.35×10^{-6}) between tumour suppressor gene
41 length and the number of CpHpG trinucleotides (excluding those with mutations), indicating that
42 the tumour suppressor genes under study possess a similar density of CpHpG trinucleotides per
43 unit length. We surmise that the factors that govern the establishment of the methylation pattern
44 of CpHpG trinucleotides are likely to be quite complex.
45
46
47
48
49
50
51
52
53
54
55
56
57

58 *Evolutionary conservation of tumour suppressor genes in relation to the sites of somatic and*
59 *germline missense mutations*
60

1
2
3 For all 17 tumour suppressor genes, the degree of evolutionary conservation, as measured by
4
5 Ka/Ks , was less than unity, indicating that these genes (and proteins) have been highly
6
7
8 conserved evolutionarily as a consequence of the action of purifying selection. Indeed, the
9
10 degree of evolutionary conservation displayed by most of the studied genes was markedly lower
11
12 than the average (~0.18) noted in a comparison of 1880 human, rat and mouse gene orthologues
13
14 [Makalowski and Boguski, 1998]. However, three genes (*CDKN2A*, *BRCA1* and *BRCA2*) were
15
16 found to exhibit a higher rate of evolutionary conservation than the average between human and
17
18 rodents.
19
20
21
22

23
24 The evolutionary conservation of each mutated codon was inferred by calculating the $\frac{Ka}{Ka + Ks}$
25
26 ratio; for each gene/spectrum, the mean value was then calculated across all mutations in the
27
28 corresponding gene/spectrum. Shared recurrent missense mutations were found to occur
29
30 disproportionately in highly conserved amino acid residues (mean degree of evolutionary
31
32 conservation, 0.072) followed by shared non-recurrent mutations (0.138), somatic recurrent
33
34 (0.169), germline (0.175), non-recurrent somatic (0.265), and control dataset mutations (0.255).
35
36 The observed differences in the degree of evolutionary conservation for the different mutational
37
38 spectra are shown in Supp. Table 2. These quite specific findings are consistent with the
39
40 previously reported general tendency for cancer-associated mutations to occur frequently at
41
42 evolutionarily conserved sites [Greenblatt et al., 2003; Tavgigian et al., 2009; Talavera et al.,
43
44 2010].
45
46
47
48
49

50
51 Somatic non-recurrent mutations were found to occur in codons characterized by the highest
52
53 mean value of $\frac{Ka}{Ka + Ks}$ ratios as compared not only to the shared recurrent and shared non-
54
55 recurrent mutations (see above) but also to the mutations within the control dataset. This is
56
57 consistent with the interpretation that a high proportion of non-recurrent somatic mutations, and
58
59
60

1
2
3 most notably those which are located in less evolutionarily conserved regions, are likely to be
4
5 'passenger' mutations.
6
7

8
9
10 *Missense mutations in relation to the disease- and non-disease-associated substitution rates*

11
12 Employing alignments of paired human gene/pseudogene sequences, Hess et al. [1994] derived
13
14 relative (non-disease-associated) nearest-neighbour-dependent mutability rates using the lowest
15
16 frequency substitution type, C(T>G)A/T(A>C)G, as a baseline. These mutability rates were
17
18 found to vary over a 52-fold range, with unity being assigned to the lowest frequency
19
20 substitution type. This *non-disease-associated* mutability rate approximates to the neutral
21
22 mutation frequency and hence reflects the intrinsic mutability of the underlying DNA sequence.
23
24 Depending upon the observed nearest-neighbour context, we retrieved the corresponding non-
25
26 disease-associated mutability rate (from the data of Hess et al. 1994) for each mutation (either
27
28 observed or from the control dataset) and calculated the median value for each mutational
29
30 spectrum. These median values are indicative of the relative mutability of each tumour
31
32 suppressor gene. The median values were found to vary between 4 (*NF2*) and 8.9 (*STK11*) for
33
34 somatic mutations, 4.1 (*TP53*) and 10.1 (*WT1*) for germline mutations, and 7.2 (*RBI*) and 11
35
36 (*PTEN*) for shared mutations (values given only for genes with more than three mutations in the
37
38 corresponding category; see Supp. Table 3, indicating that many of the median values are quite
39
40 low and hence the corresponding mutations are unlikely to be neutral.
41
42
43
44
45
46
47

48
49 When data from all 17 genes were combined, shared recurrent mutations were found to be
50
51 characterised by intrinsically low non-disease-associated mutability (median=11), followed by
52
53 even lower median mutability values for shared non-recurrent mutations (7.9), germline
54
55 mutations (7.2), somatic recurrent and non-recurrent (4.7) and control dataset mutations (4.1).
56
57 Such low median mutability values across all groups indicates that at least half of the mutations
58
59 within observed triplets are unlikely to be neutral in the sense defined by Hess et al. [1994] and
60

1
2
3 hence are not simply explicable in terms of intrinsic DNA mutability. The low median mutability
4 values for the control dataset of mutations within tumour suppressor genes reflect the high level
5 of evolutionary conservation manifested by tumour suppressor gene coding sequences across
6 different species, implying that any mutation within a triplet characterized by a low non-disease-
7 associated mutation rate is very likely to have pathological consequences and would thus be
8 subject to purifying selection.
9

10
11 In contrast to the non-disease-associated mutability rate (which is purely a reflection of the
12 intrinsic DNA mutability), the disease-associated mutability rate reflects (in addition to the
13 intrinsic DNA mutability) the increased likelihood of coming to clinical attention conferred by
14 the loss of biological function. The C(G>T)T mutation is one of the most frequent types of
15 mutation associated with the loss of biological function [disease-associated mutability rate
16 10.255; Krawczak et al., 1998] but occurs much less frequently among neutral mutations [non-
17 disease-associated mutability rate 4.4; Hess et al., 1994].
18
19

20
21 For each tumour suppressor gene and each mutational spectrum, the disease-associated median
22 mutability values were calculated using mutability rates derived from Krawczak et al. [1998].
23 The disease-associated median value was found to be 0.85 for the germline mutations. The
24 highest and lowest disease-associated median values for the mutation rates were noted for
25 somatic mutations in the *STK11* gene (1.7; Supp. Table 3) and for germline mutations in the
26 *TP53* (0.42) gene (values given only for genes with more than three mutations in the
27 corresponding category). We found that shared recurrent and shared non-recurrent mutational
28 spectra were characterized by higher median values of the disease-associated mutability rates
29 (1.42 and 1.01 respectively) whereas somatic non-recurrent, somatic recurrent and control
30 dataset mutations exhibited lower median mutability rates (0.5, 0.5 and 0.4 respectively) as
31 compared to germline mutations (0.85). The finding that the shared mutations (which, by
32 definition, occur in both the germline and the soma) are characterized by higher disease-
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 associated mutability rates is not surprising since mutations that occur with the highest
4
5 probability are among those most likely to be shared.
6
7

8 We postulated that those mutations which occur both in the germline and the soma, and which
9
10 are characterised by higher disease-associated mutability rates are disproportionately likely to be
11
12 drivers of tumour development. Consistent with this postulate, somatic recurrent and non-
13
14 recurrent mutational spectra are characterized by lower median disease-associated mutability
15
16 rates as compared to the germline spectrum. However, given that higher disease-associated
17
18 mutability rates are a characteristic feature of driver mutations, a certain proportion of the
19
20 somatic mutations, namely those characterised by higher disease-associated mutability rates, may
21
22 correspond to functionally significant driver mutations.
23
24
25

26
27 In assessing the significance of our results, it was appropriate to consider the possibility that
28
29 somatic mutations might display quite different nearest-neighbour-dependent disease-associated
30
31 mutability rates from germline mutations. However, since a good correlation was observed
32
33 between the mutability rates derived from inherited disease data [Krawczak et al., 1998] and the
34
35 neighbour-dependent mutability rates calculated for the somatic mutations of the 17 tumour-
36
37 suppressor genes studied here (Pearson's correlation $r=0.703$, $p=6.6\times 10^{-30}$), this *caveat* appears
38
39 not to be an issue.
40
41
42
43
44
45

46 *Distribution of Grantham scores with respect to tumour suppressor gene mutations*

47

48 Shared recurrent mutations were found to exhibit the largest median chemical difference value
49
50 (Grantham scores) between the wild-type and mutated amino acid residues (100) followed by
51
52 shared non-recurrent mutations and germline mutations (both 93), somatic recurrent (85),
53
54 somatic non-recurrent (80) and potential mutations (78). Since there was an obvious trend for
55
56 shared recurrent and non-recurrent mutations to cause the most dramatic chemical changes of the
57
58 affected codon, we may infer that these types of lesion are also more likely to be driver
59
60

1
2
3 mutations. However, bearing in mind that the range of theoretically possible values varies
4
5 between 5 (Leu ↔ Ile) and 215 (Cys ↔ Trp), less elevated median values may simply indicate
6
7 that a proportion of the mutations in each mutational spectrum are likely to be chemically less
8
9 dramatic (Grantham scores <100).
10
11

12 13 14 15 *Missense mutations occurring within repeats and runs of identical nucleotides*

16
17 A number of studies have noted that single base-pair substitutions associated with inherited
18
19 disease occur disproportionately either within, or in close proximity to, repetitive sequences
20
21 [Jego et al., 1993; Greenblatt et al., 1996; Tappino et al., 2009; Thomas et al., 2010; Leclercq et
22
23 al., 2010]. Hence, we wished to assess whether either germline or somatic mutations occurred
24
25 disproportionately either within, or in the vicinity (see *Mutation descriptors*) of, direct, inverted
26
27 and symmetric repeats or mononucleotide runs in the 17 tumour suppressor genes under study
28
29 (Table 3, Supplementary Tables 4-6).
30
31
32

33
34 On average, direct repeats of length ≥ 8 bp were found to cover 5.6% of the cDNA lengths of
35
36 the 17 tumour suppressor genes, the coverage varying between 2.5% (*BRCA2*) and 17% (*PTEN*)
37
38 of the respective gene sequences. The corresponding proportion of the cDNA lengths for
39
40 inverted repeats ≥ 8 bp was 8.5%, with proportions varying between *PTCH1* (4.5%) and *RBI*
41
42 (15.7%) while symmetric elements ≥ 8 bp were found to encompass 25% of the cDNA lengths
43
44 (varying between 15.5% for *APC* and 44% for *PTEN*).
45
46
47

48
49 On average, mononucleotide runs ≥ 4 bp spanned 19.9% of the cDNA lengths, varying between
50
51 9.5% (*VHL*) and 29% (*TP53*). Approximately 24% of non-recurrent somatic and 20% of
52
53 germline missense mutations were found in mononucleotide runs; these proportions were
54
55 significantly higher than noted for shared non-recurrent missense mutations (4.9%, $p \leq 1.6 \times 10^{-4}$).
56
57 A greater proportion of non-recurrent somatic missense mutations was found in direct repeats
58
59 (7%) as compared to recurrent somatic missense mutations (2%, $p = 8.8 \times 10^{-7}$), germline missense
60

1
2
3 (4%, $p=0.028$) and potential missense mutations (3.7%, $p=8.1\times 10^{-7}$). This result may reflect the
4
5 disproportionate number of CpG/CpHpG mutations among shared and recurrent somatic
6
7 missense mutations. Further, for all mutational spectra examined (with the exception of the
8
9 shared mutations), missense mutations were preferentially found in association with inverted and
10
11 symmetric repeats as compared to the control dataset of mutations ($p<0.05$). However, no
12
13 statistically significant differences were found between mutational spectra.
14
15

16
17 No correlation was observed between the number of mutations located within repeats and the
18
19 fractional length of the cDNA covered by repeats, indicating that not every repeat sequence is
20
21 mutation-prone. However, a strong correlation between the fractional length of the cDNA
22
23 covered by repeats and cDNA length of genes ($r >0.87$ and $p<10^{-6}$) served to demonstrate that
24
25 repeat density per unit length was approximately the same for all tumour suppressor genes
26
27 studied.
28
29
30
31
32
33

34 *Towards a classification of somatic and germline missense mutations*

35
36 All observed mutations within each mutational spectrum were re-categorized (Supp. Table 7)
37
38 with respect to the location of mutations within CpG/CpHpG oligonucleotides, within different
39
40 types of repeat/mononucleotide runs, within both CpG/CpHpG oligonucleotides and repeats. 4×2
41
42 contingency tables were then used to measure the strength of the pairwise associations between
43
44 the various mutational distributions presented in Supp. Table 7, the significance of the
45
46 associations being assessed by means of a Chi-square test. Significant ($p<0.002$) pairwise
47
48 differences were noted between somatic and germline, somatic and shared, and between
49
50 germline and shared mutational spectra ($p<0.002$) with respect to the features listed above and
51
52 each of four types of repeat, indicating that these features have great discriminant potential.
53
54
55
56

57 All somatic, germline, shared non-recurrent, recurrent somatic and shared recurrent missense
58
59 mutations (each described by a combination of different features (i.e. degree of evolutionary
60

1
2
3 conservation, non-disease- and disease-associated mutability rates, Grantham score,
4 CpG/CpHpG location, occurrence within repeat/mononucleotide run) were then used to train a
5 Naïve Bayes Tree classifier. 63.1% of somatic, germline, shared, recurrent somatic and shared
6 recurrent mutations were correctly classified [the area under the Receiver Operating
7 Characteristic (ROC) curve being 0.869, indicating a reasonably good classification] implying
8 that the mutation groupings differ with respect to the different features in a consistent fashion.
9 The complete Naïve Bayes Tree classifier is depicted in Supp. Figure 1.

10
11
12
13
14
15
16
17
18
19
20 An additional non-overlapping dataset of 568 missense somatic mutations, identified in the 17
21 tumour suppressor genes under study, were extracted from a collection of 2,488 mutations
22 identified as being probable driver mutations [Carter et al., 2009]. Features such as the degree of
23 evolutionary conservation, Grantham score, mutability rates, CpG/CpHpG location, occurrence
24 within repeats/mononucleotide runs were again determined for each of these mutations.
25 Employing our classifier, 7% and 10% respectively of these 568 mutations were found to possess
26 features consistent with their being shared recurrent and shared non-recurrent mutations. In
27 addition, 32% of these probable driver mutations were found to bear features characteristic of
28 recurrent somatic mutations (i.e. mutations documented in different tumours). A further 25% of
29 the probable (somatic) driver mutations were classified as possessing features characteristic of
30 germline mutations and hence could conceivably be treated as shared mutations missing from the
31 original training dataset. The remaining 25% of mutations were classified as non-recurrent
32 somatic mutations. Using this classifier, which is based on a very modest number (6) of
33 predictive features, to analyse an independent dataset of probable driver mutations, we were able
34 to predict that ~50% of these somatic missense mutations exhibited features specific to either
35 shared or recurrent mutations, indicating that a disproportionate number of such lesions are likely
36 to be drivers of tumorigenesis. This percentage is certainly lower (79%) than that obtained by
37 Carter et al., [2009] through the application of a Random Forest Classifier based on 500 trees and
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 >50 predictive features (using an out-of-the-bag error estimate similar to the cross-validation
4 procedure) to the set of putative 2,488 driver mutations. However, based on the results of this
5
6 study, we may conclude that, in general, the mutational spectrum of driver mutations is likely to
7
8 contain a disproportionate number of somatic mutations that have germline counterparts (~17%)
9
10 whilst an additional 32% of the driver mutations are likely to occur recurrently in the soma.
11
12
13
14
15
16

17 *Truncating vs non-truncating mutations in the germline and soma*

18
19 Somatic mutational spectra from the *BRCA2*, *CDKN2A*, *STK11*, *TP53* and *TSC1* genes were
20 characterized by the predominance of non-truncating (i.e. missense) lesions over truncating
21 lesions (i.e. nonsense mutations, frameshift micro-deletions, micro-insertions and indels) when
22 nonsense mutations [reported in Mort et al. (2008)] and micro-indels (excluded from previous
23 analyses) were also considered (Supp. Table 8). A similar predominance of non-truncating over
24 truncating lesions was observed for the germline mutational spectra of the *CDKN2A*, *TP53*, *VHL*
25 and *WT1* genes. In general, the ratio of non-truncating to truncating lesions was found to be
26 significantly higher in the soma (0.85) than in the germline (0.30; p-value<2.20E-16). All other
27 mutational spectra were characterized by the predominance of truncating mutations.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

43 *Occurrence of micro-deletions and micro-insertions within repeats and runs of identical* 44 *nucleotides*

45
46
47 The mutational spectrum of micro-deletions, combined for all 17 tumour suppressor genes,
48 comprised 55% germline, 43% somatic and 2% shared mutations. The mutational spectrum of
49 micro-insertions was similar to that of micro-deletions and comprised 60% germline, 38%
50 somatic and 2% shared mutations. Approximately 77% somatic, 87% germline and 91% shared
51 micro-deletions and micro-insertions were ≤ 4 bp in length. Strong ($r = \sim 1$) correlations were
52
53
54
55
56
57
58
59
60 noted between the distributions of micro-deletions and micro-insertions with respect to the length

1
2
3 of the deleted/inserted fragments, both gene-wise and for all genes combined ($r>0.9$, $p<10^{-8}$) for
4 all mutational spectra.
5
6

7
8 Recent studies have revealed that simple repetitive DNA sequences are not only capable of
9 adopting non-B DNA conformations and are highly mutagenic [Bacolla et al., 2004; Bacolla and
10 Wells, 2004; Chuzhanova et al., 2009]. Indeed, both direct repeats and mononucleotide runs have
11 long been known to be mutation hotspots in the *TP53* gene [Jego et al., 1993; Greenblatt et al.,
12 1996]. The number of micro-lesions occurring in the vicinity (see *Mutation descriptors*) of
13 direct, symmetric and inverted repeats (capable respectively of slipped, triplex and cruciform
14 non-B structure formation), or within mononucleotide runs (which often mediate micro-
15 deletions/micro-insertions) were therefore determined. The number of mutations found in the
16 vicinity of all three types of repeat, and within mononucleotide runs, are given in Tables 3 and
17 Supp. Tables 4-6.
18
19
20
21
22
23
24
25
26
27
28
29
30

31 The highest proportion of mutations in mononucleotide runs was found for the shared (39%),
32 germline (30%) and somatic (25%) mutational spectra. Significant differences were observed
33 between shared and germline ($p=0.0002$), somatic and shared ($p=0.045$), and between all
34 mutational spectra and potential mutations ($p<0.0001$) with respect to their occurrence within
35 mononucleotide runs, confirming that these simple repeats constitute an important hotspot for
36 micro-deletions and micro-insertions in both the soma and the germline. The preponderance of
37 such mutations in mononucleotide runs is unsurprising in the context of the shared mutations
38 since all mutations that occur with high frequency within mutation hotspots are more likely to be
39 shared between the germline and the soma (as previously noted for CpG and CpHpG mutations).
40 No other types of repeat were disproportionately associated (after correction for multiple testing)
41 with micro-deletions and micro-insertions.
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Hotspots in somatic and germline mutational spectra

1
2
3 For the purposes of this analysis, a mutation hotspot was defined as a stretch of DNA of length
4
5 ≤ 20 bp where four or more independent mutational events have been reported and a significant
6
7 degree ($p \leq 0.05$) of clustering of these mutations was evident for a given stretch of DNA. In this
8
9 definition of a hotspot, each recurrent mutation was considered only once. The order statistics, r-
10
11 scans, as described by Karlin and Macken [1991] and applied in Bacolla et al. [2006], were used
12
13 to detect significant clustering of mutations by comparison with a Poisson distribution of
14
15 mutations along the gene sequence. Overlapping hotspot regions were considered as a single
16
17 hotspot.
18
19

20
21
22 The only mutational hotspot for somatic missense mutations was observed in the *PTEN* gene
23
24 and comprised 18 mutations in the region between nucleotide positions 269 and 286. Several
25
26 germline mutational hotspots were however detected for missense mutational spectra in the
27
28 *ATM*, *BRCA1*, *BRCA2*, *NF1*, *PTEN*, *RBI*, *STK11*, *TP53* and *WT1* genes (Table 4). Several
29
30 somatic mutational hotspots were found for micro-deletions/micro-insertions in the *APC* gene,
31
32 the largest of which contained 33 mutations (positions 4303-4398) and forms part of a previously
33
34 reported mutation cluster region [Miyoshi et al., 1992]. Hotspots identified in different
35
36 mutational spectra were however unique to that spectrum. The only overlap noted between
37
38 mutational hotspots identified in germline and somatic micro-deletion/micro-insertion mutational
39
40 spectra was observed for the *APC* gene (the overlapping region comprising nucleotide positions
41
42 3919-3933). This micro-deletion/micro-insertion hotspot also includes codon 1309 (cDNA
43
44 positions 3925-3927) found to be frequently mutated in Greek and French patients with familial
45
46 adenomatous polyposis [Fostira et al. 2010; Lagarde et al. 2010].
47
48
49
50
51
52

53 Inspection of hotspot regions revealed that they are rich in repetitive elements, runs of identical
54
55 nucleotides and CpG/CpHpG oligonucleotides, offering immediate explanations for the elevated
56
57 mutability.
58
59
60

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, WKVNRRRRNVWK [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

1
2
3 and are characterised by higher mutability rates (both non-disease- and disease-associated).
4
5
6 Further, we found that shared mutations tend to occur in those codons that have been more
7
8 highly conserved evolutionarily, and are associated with more dramatic chemical differences
9
10 between the substituted (wild-type) and substituting amino acids. Taken together, it would thus
11
12 appear that shared mutations are influenced to a greater extent by the local nucleotide sequence
13
14 context than either germline or somatic non-recurrent missense mutations. Since this implies that
15
16 shared mutations (the mutation category most likely to harbour driver mutations) have a
17
18 tendency to arise through the action of similar endogenous mutational mechanisms, we may infer
19
20 that endogenous mechanisms of mutagenesis exert a disproportionate effect on tumorigenesis.
21
22
23

24
25 In an analysis of an unrelated dataset, we demonstrated that 17% of somatic missense
26
27 mutations previously identified as being probable drivers [Carter et al., 2009] were found to
28
29 possess the same features as shared (both recurrent and non-recurrent) mutations. A further 32%
30
31 of these probable driver mutations shared the features expected of recurrent somatic mutations.
32
33 Thus, we may conclude that ~50% of these somatic missense mutations possess features
34
35 consistent with their being either shared or recurrent, suggesting that a disproportionate number
36
37 of such lesions are likely to be drivers of tumorigenesis.
38
39
40

41
42 A sizeable proportion of shared (39%) and germline (30%) micro-lesions were found to be
43
44 located in runs of identical nucleotides ≥ 4 bp, making mononucleotide runs a hotspot for micro-
45
46 deletion and micro-insertions. The most likely underlying causative mechanism for these
47
48 mutations is slipped mispairing at DNA replication mediating duplications and 'de-duplications'
49
50 [Kondrashov & Rogozin, 2004]. With regard to missense mutations, CpG and CpHpG
51
52 oligonucleotides were found to be hotspots for shared recurrent and shared non-recurrent
53
54 missense mutations; 34% (10%) and 21% (9%) of respective mutations were found in CpG
55
56 (CpHpG) oligonucleotides. Further, 12% of the 568 probable driver mutations [derived from
57
58 Carter et al., 2009] were found to occur in CpG/CpHpG oligonucleotides. 41% of probable
59
60

1
2
3 driver mutations were found in repeats that were capable of non-B DNA structure formation (cf.
4
5 23% for potential mutations). Several hotspot regions were found in the mutational spectra of
6
7 various genes; one of these, in the *APC* gene, was a hotspot for both somatic and germline
8
9 micro-deletions/micro-insertions and corresponded to a previously recognized mutation hotspot
10
11 [Miyoshi et al., 1992].
12
13

14
15 Taken together, the results and analysis presented herein strongly suggest that algorithms that
16
17 attempt to predict the relative impact of tumour-associated micro-lesions on (tumour suppressor)
18
19 gene and protein function [Tavtigian et al., 2008; Couch et al., 2008; Thusberg and Vihinen,
20
21 2009], should take into consideration the origin (i.e. somatic, germline or shared) of the
22
23 mutations, their sequence context and repetitivity, as well as their frequency of occurrence.
24
25
26
27
28

29 **Acknowledgements**

30
31 We are most grateful to Eamonn Maher (Birmingham, UK), Gareth Evans (Manchester, UK) and
32
33 the late Michael Baser for their provision of somatic mutation data. We are also very grateful to
34
35 Larry Brody (NIH, Bethesda, MD, USA), Shiu-Ru Lin (Kaohsiung Medical University, Taiwan)
36
37 and Tone Lovig (University of Oslo, Norway) for providing us with further information on
38
39 specific somatic mutations and Rachel Karchin (Johns Hopkins University, Baltimore, USA) for
40
41 making available her dataset of probable driver mutations.
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

References

- 1
2
3
4
5
6 Ali IU, Schriml LM, Dean M. 1999. Mutational spectra of *PTEN/MMAC1* gene: a tumor
7
8 suppressor with lipid phosphatase activity. *J Natl Cancer Inst* 91:1922-1932.
9
10
11 Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova NA, Abeysinghe SS, O'Connell
12
13 CD, Cooper DN, Wells RD. 2004. Breakpoints of gross deletions coincide with non-B
14
15 DNA conformations. *Proc Natl Acad Sci USA* 101:14162-14167.
16
17
18 Bacolla A, Wells RD. 2004. Non-B DNA conformations, genomic rearrangements, and human
19
20 disease. *J Biol Chem* 279:47411-47414.
21
22
23 Bacolla A, Collins JR, Gold B, Chuzhanova N, Yi M, Stephens RM, Stefanov S, Olsh A,
24
25 Jakupciak JP, Dean M, Lempicki RA, Cooper DN, Wells RD. 2006. Long
26
27 homopurine•homopyrimidine sequences are characteristic of genes expressed in brain
28
29 and the pseudoautosomal region. *Nucleic Acids Res* 34: 2663-2675.
30
31
32 Ball EV, Stenson PD, Krawczak M, Cooper DN, Chuzhanova NA. 2005. Micro-deletions and
33
34 micro-insertions causing human genetic disease: common mechanisms of mutagenesis
35
36 and the role of local DNA sequence complexity. *Human Mutat* 26:205-213.
37
38
39 Baser ME, Contributors to the International NF2 Mutation Database. *Hum. Mutat.* 27:297-306.
40
41
42 Boland CR, Ricciardiello L. 1999. How many mutations does it take to make a tumor? *Proc Natl*
43
44 *Acad Sci USA* 96:14675-14677.
45
46
47 Buard J, Collick A, Brown J, Jeffreys AJ. 2000. Somatic versus germline mutation processes at
48
49 minisatellite CEB1 (D2S90) in humans and transgenic mice. *Genomics* 65:95-103.
50
51
52 Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R.
53
54 2009. Cancer-specific high-throughput annotation of somatic mutations: computational
55
56 prediction of driver missense mutations. *Cancer Res* 69:6660-6667.
57
58
59
60

- 1
2
3 Chuzhanova NA, Anassis EJ, Ball E, Krawczak M, Cooper DN. 2003. Meta-analysis of indels
4 causing human genetic disease: mechanisms of mutagenesis and the role of local DNA
5 sequence complexity. *Hum Mutat* 21:28-44.
6
7
8
9
10 Chuzhanova N, Chen JM, Bacolla A, Patrinos GP, Férec C, Wells RD, Cooper DN. 2009. Gene
11 conversion causing human inherited disease: the evidence for involvement of
12 recombination-associated motifs and non-B DNA-forming sequences in DNA breakage.
13 *Hum Mutat* 30:1189-1198.
14
15
16
17
18
19
20 Cole DN, Carlson JA, Wilson VL. 2008. Human germline and somatic cells have similar *TP53*
21 and *Kirsten-RAS* gene single base mutation frequencies. *Environ. Mol. Mutagen.* 2008
22 49:417-425.
23
24
25
26
27 Cooper DN, Mort M, Stenson PD, Ball EV, Chuzhanova NA. 2010. Methylation-mediated
28 deamination of 5-methylcytosine appears to give rise to mutations causing human
29 inherited disease in CpNpG trinucleotides as well as in CpG dinucleotides. *Hum*
30 *Genomics* 4:406-410..
31
32
33
34
35
36
37 Couch FJ, Rasmussen LJ, Hofstra R, Monteiro AN, Greenblatt MS, de Wind N; IARC
38 Unclassified Genetic Variants Working Group. 2008. Assessment of functional effects of
39 unclassified genetic variants. *Hum Mutat* 29:1314-1326.
40
41
42
43
44 Dallosso AR, Jones S, Azzopardi D, Moskvina V, Al-Tassan N, Williams GT, Idziaszczyk S,
45 Davies DR, Milewski P, Williams S, Beynon J, Sampson JR, Cheadle JP. 2009. The *APC*
46 variant p.Glu1317Gln predisposes to colorectal adenomas by a novel mechanism of
47 relaxing the target for tumorigenic somatic *APC* mutations. *Hum Mutat* 30:1412-1418.
48
49
50
51
52
53 Fostira F, Thodi G, Sandaltzopoulos R, Fountzilias G, Yannoukakos D. 2010. Mutational
54 spectrum of *APC* and genotype-phenotype correlations in Greek FAP patients. *BMC*
55 *Cancer* 10:389.
56
57
58
59
60

- 1
2
3 Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR.
4
5 2004. A census of human cancer genes. *Nat Rev Cancer* 4:177-183.
6
7
8 Gallou C, Joly D, Mejean A, Staroz F, Martin N, Tarlet G, Orfanelli MT, Bouvier R, Droz D,
9
10 Chretien Y, Maréchal JM, Richard S, Junien C, Bérout C. 1999. Mutations of the *VHL*
11
12 gene in sporadic renal cell carcinoma: definition of a risk factor for VHL patients to
13
14 develop an RCC. *Hum Mutat* 13:464-475.
15
16
17 Giovannone B, Sabbadini G, Di Maio L, Calabrese O, Castaldo I, Frontali M, Novelletto A,
18
19 Squitieri F. 1997. Analysis of (CAG)_n size heterogeneity in somatic and sperm cell DNA
20
21 from intermediate and expanded Huntington disease gene carriers. *Hum Mutat* 10:458-
22
23 464.
24
25
26
27 Glazko GV, Koonin EV, Rogozin IB. 2004. Mutation hotspots in the p53 gene in tumors of
28
29 different origin: correlation with evolutionary conservation and signs of positive
30
31 selection. *Biochim Biophys Acta* 1679:95-106.
32
33
34 Goode EL, Ulrich CM, Potter JD. 2002. Polymorphisms in DNA repair genes and associations
35
36 with cancer risk. *Cancer Epidemiol Biomarkers Prev* 11:1513-1530.
37
38
39 Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science*
40
41 185:862-864.
42
43
44 Greenblatt MS, Grollman AP, Harris CC. 1996. Deletions and insertions in the p53 tumor
45
46 suppressor gene in human cancers: confirmation of the DNA polymerase
47
48 slippage/misalignment model. *Cancer Res* 56:2130-2136.
49
50
51 Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP. 2003. Detailed
52
53 computational study of p53 and p16: using evolutionary sequence analysis and disease-
54
55 associated mutations to predict the functional consequences of allelic variants. *Oncogene*
56
57 22:1150-1163.
58
59
60

- 1
2
3 Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J,
4
5 Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S,
6
7 Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K,
8
9 Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko
10
11 T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T,
12
13 West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur
14
15 F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P,
16
17 Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung
18
19 SY, Wooster R, Futreal PA, Stratton MR. 2007. Patterns of somatic mutation in human
20
21 cancer genomes. *Nature* 446:153-158.
22
23
24
25
26
27 Groves C, Lamlum H, Crabtree M, Williamson J, Taylor C, Bass S, Cuthbert-Heavens D,
28
29 Hodgson S, Phillips R, Tomlinson I. 2002. Mutation cluster region, association between
30
31 germline and somatic mutations and genotype-phenotype correlation in upper
32
33 gastrointestinal familial adenomatous polyposis. *Am J Pathol* 160:2055-2061.
34
35
36
37 Haber D, Harlow E. 1997. Tumour-suppressor genes: evolving definitions in the genomic age.
38
39 *Nat Genet* 16:320-322.
40
41
42 Hess ST, Blake JD, Blake RD. 1994. Wide variations in neighbor-dependent substitution rates. *J*
43
44 *Mol Biol* 236:1022-1033.
45
46
47 Jago N, Thomas G, Hamelin R. 1993. Short direct repeats flanking deletions, and duplicating
48
49 insertions in p53 gene in human cancers. *Oncogene* 8:209-213.
50
51
52 Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon
53
54 R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S,
55
56 Davis DP, Carlton VE, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard
57
58 DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, de Sauvage FJ, Faham M,
59
60

- 1
2
3 Seshagiri S. 2010. Diverse somatic mutation patterns and pathway alterations in human
4 cancers. *Nature* 466:869-873.
5
6
7
8 Karlin S, Macken C. 1991. Some statistical problems in the assessment of inhomogeneity of
9 DNA sequence data. *J Am Statist Assoc* 86:27–35.
10
11
12 Knudson AG. 2001. Two genetic hits (more or less) to cancer. *Nat Rev Cancer* 1:157-162.
13
14
15 Kohavi R. 1996. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid.
16 Proceedings of the Second International Conference on Knowledge Discovery and Data
17 Mining. AAAI Press, pp 202-207.
18
19
20
21
22 Kolomietz E, Meyn MS, Pandita A, Squire JA. 2002. The role of *Alu* repeat clusters as mediators
23 of recurrent chromosomal aberrations in tumors. *Genes Chrom Cancer* 35:97-112.
24
25
26
27 Kondrashov AS, Rogozin IB. 2004. Context of deletions and insertions in human coding
28 sequences. *Hum Mutat* 23:177–185.
29
30
31
32 Kotsiantis S, Kanellopoulos D, Pintelas P. 2006. Handling imbalanced datasets: a review.
33 GESTS International Transactions on Computer Science and Engineering 30:25-36.
34
35
36
37 Krawczak M, Smith-Sorensen B, Schmidtke J, Kakkar VV, Cooper DN, Hovig E. 1995. Somatic
38 spectrum of cancer-associated single basepair substitutions in the *TP53* gene is
39 determined mainly by endogenous mechanisms of mutation and by selection. *Hum Mutat*
40 5:48-57.
41
42
43
44
45
46 Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-
47 line single-base-pair substitution in human genes. *Am J Hum Genet* 63:474-488.
48
49
50
51 Lagarde A, Rouleau E, Ferrari A, Noguchi T, Qiu J, Briaux A, Bourdon V, Rémy V, Gaildrat P,
52 Adélaïde J, Birnbaum D, Lidereau R, Sobol H, Olschwang S. 2010. Germline
53 *APC* mutation spectrum derived from 863 genomic variations identified through a
54 15-year medical genetics service to French patients with FAP. *J Med Genet*
55 [Epub ahead of print] PubMed PMID: 20685668.
56
57
58
59
60

- 1
2
3 Lamlum H, Ilyas M, Rowan A, Clark S, Johnson V, Bell J, Frayling I, Efstathiou J, Pack K,
4
5 Payne S, Roylance R, Gorman P, Sheer D, Neale K, Phillips R, Talbot I, Bodmer W,
6
7 Tomlinson I. 1999. The type of somatic mutation at *APC* in familial adenomatous
8
9 polyposis is determined by the site of the germline mutation: a new facet to Knudson's
10
11 'two-hit' hypothesis. *Nat Med* 5:1071-1075.
12
13
14
15 Latchford A Volikos E, Johnson V, Rogers P, Suraweera N, Tomlinson I, Phillips R, Silver A.
16
17 2007. *APC* mutations in FAP-associated desmoid tumours are non-random but not 'just
18
19 right'. *Hum Mol Genet* 16:78-82.
20
21
22
23 Leclercq S, Rivals E, Jarne P. 2010. DNA slippage occurs at microsatellite loci without minimal
24
25 threshold length in human: a comparative genomic approach. *Genome Biol Evol* 2:325-
26
27 335.
28
29
30
31 Leeflang EP, Tavare S, Marjoram P, Neal CO, Srinidhi J, MacFarlane H, MacDonald ME,
32
33 Gusella JF, de Young M, Wexler NS, Arnheim N. 1999. Analysis of germline mutation
34
35 spectra at the Huntington's disease locus supports a mitotic mutation mechanism. *Hum*
36
37 *Mol Genet* 8:173-183.
38
39
40
41 Lengauer C, Kinzler KW, Vogelstein B. 1998. Genetic instabilities in human cancers. *Nature*
42
43 396:643-649.
44
45
46
47 Lister R, Pelizzola M, Downen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z,
48
49 Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson
50
51 JA, Ren B, Ecker JR. 2009. Human DNA methylomes at base resolution show
52
53 widespread epigenomic differences. *Nature* 19:315-322.
54
55
56
57 Lobo GP, Waite KA, Planchon SM, Romigh T, Nassif NT, Eng C. 2009. Germline and somatic
58
59 cancer-associated mutations in the ATP-binding motifs of PTEN influence its subcellular
60
localization and tumor suppressive function. *Hum. Mol. Genet.* 18:2851-2862.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- 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.

- 1
2
3 Parmigiani G, Boca S, Lin J, Kinzler KW, Velculescu V, Vogelstein B. 2009. Design and
4
5 analysis issues in genome-wide somatic mutation studies of cancer. *Genomics* 93:17-21.
6
7
8 Pollard LM, Sharma R, Gomez M, Shah S, Delatycki MB, Pianese L, Monticelli A, Keats BJ,
9
10 Bidichandani SI. 2004. Replication-mediated instability of the GAA triplet repeat
11
12 mutation in Friedreich ataxia. *Nucleic Acids Res* 32:5962-5971.
13
14
15 Richter S, Vandezande K, Chen N, Zhang K, Sutherland J, Anderson J, Han L, Panton R, Branco
16
17 P, Gallie B. 2003. Sensitive and efficient detection of *RBI* gene mutations enhances care
18
19 for families with retinoblastoma. *Am J Hum Genet* 72:253-269.
20
21
22 Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. 2001. Somatic mutation hotspots
23
24 correlate with DNA polymerase eta error spectrum. *Nat Immunol* 2:530-536.
25
26
27 Rubin AF, Green P. 2009. Mutation patterns in cancer genomes. *Proc Natl Acad Sci USA* 106:
28
29 21766-21770.
30
31
32 Sharma R, Bhatti S, Gomez M, Clark RM, Murray C, Ashizawa T, Bidichandani SI. 2002. The
33
34 GAA triplet-repeat sequence in Friedreich ataxia shows a high level of somatic instability
35
36 *in vivo*, with a significant predilection for large contractions. *Hum Mol Genet* 11:2175-
37
38 2187.
39
40
41 Sherr CJ. 2004. Principles of tumor suppression. *Cell* 116:235-246.
42
43 Schmutte C, Jones PA. 1998. Involvement of DNA methylation in human carcinogenesis. *Biol*
44
45 *Chem* 379:377-388.
46
47
48 Shanks ME, May CA, Dubrova YE, Balaesque P, Rosser ZH, Adams SM, Jobling MA. 2008.
49
50 Complex germline and somatic mutation processes at a haploid human minisatellite
51
52 shown by single-molecule analysis. *Mutat. Res.* 648:46-53.
53
54
55 Simpson AJ. 2009. Sequence-based advances in the definition of cancer-associated gene
56
57 mutations. *Curr Opin Oncol* 21:47-52.
58
59
60 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

- 1
2
3 D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman
4
5 KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. 2006. The consensus
6
7 coding sequences of human breast and colorectal cancers. *Science* 314:268-274.
8
9
10 Stead JD, Jeffreys AJ. 2000. Allele diversity and germline mutation at the insulin minisatellite.
11
12 *Hum Mol Genet* 9:713-723.
13
14
15 Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The
16
17 Human Gene Mutation Database: 2008 update. *Genome Med.* 1:13.
18
19
20 Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. *Nature* 458:719-724.
21
22
23 Talavera D, Taylor MS, Thornton JM. 2010. The (non)malignancy of cancer amino acidic
24
25 substitutions. *Proteins* 78:518-529.
26
27
28 Tappino B, Chuzhanova NA, Regis S, Dardis A, Corsolini F, Stroppiano M, Tonoli E, Beccari T,
29
30 Rosano C, Mucha J, Blanco M, Szlago M, Di Rocco M, Cooper DN, Filocamo M. 2009.
31
32 Molecular characterization of 22 novel UDP-N-acetylglucosamine-1-phosphate
33
34 transferase alpha- and beta-subunit (GNPTAB) gene mutations causing mucopolidosis
35
36 types IIalpha/beta and IIIalpha/beta in 46 patients. *Hum Mutat* 30:E956-973.
37
38
39 Tartaglia M, Martinelli S, Stella L, Bocchinfuso G, Flex E, Cordeddu V, Zampino G, van der
40
41 Burgt I, Palleschi A, Petrucci TC, Sorcini M, Schoch C, Foà R, Emanuel PD, Gelb BD.
42
43 2006. Diversity and functional consequences of germline and somatic *PTPN11* mutations
44
45 in human disease. *Am J Hum Genet* 78:279-290.
46
47
48 Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB; IARC Unclassified Genetic Variants
49
50 Working Group. 2008. *In silico* analysis of missense substitutions using sequence-
51
52 alignment based methods. *Hum Mutat* 29:1327-1336.
53
54
55 Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F,
56
57 Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B,
58
59 McKay-Chopin S, Thomas A, Vallée MP, Voegelé C, Webb PM, Whiteman DC;
60

- 1
2
3 Australian Cancer Study; Breast Cancer Family Registries (BCFR); Kathleen Cuningham
4
5 Foundation Consortium for Research into Familial Aspects of Breast Cancer (kConFab),
6
7 Sangrajang S, Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G.
8
9 2009. *Am J Hum Genet* 85:427-446.
10
11
12 Thomas L, Kluwe L, Chuzhanova N, Mautner V, Upadhyaya M. 2010. Analysis of *NF1* somatic
13
14 mutations in cutaneous neurofibromas from patients with high tumor burden.
15
16 *Neurogenetics* 11:391-400.
17
18
19 Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of
20
21 missense mutations using bioinformatics methods. *Hum Mutat* 30:703-714.
22
23
24 Tornaletti S, Pfeifer GP. 1995. Complete and tissue-independent methylation of CpG sites in the
25
26 *p53* gene: implications for mutations in human cancers. *Oncogene* 10:1493-1499.
27
28
29 Upadhyaya M, Han S, Consoli C, Majounie E, Horan M, Thomas NS, Potts C, Griffiths S,
30
31 Ruggieri M, von Deimling A, Cooper DN. 2004. Characterization of the somatic
32
33 mutational spectrum of the neurofibromatosis type 1 (*NF1*) gene in neurofibromatosis
34
35 patients with benign and malignant tumors. *Hum Mutat* 23:134-136.
36
37
38 Upadhyaya M, Kluwe L, Spurlock G, Monem B, Majounie E, Mantripragada K, Ruggieri M,
39
40 Chuzhanova N, Evans DG, Ferner R, Thomas N, Guha A, Mautner V. 2008. Germline
41
42 and somatic *NF1* gene mutation spectrum in *NF1*-associated malignant peripheral nerve
43
44 sheath tumors (MPNSTs). *Hum Mutat* 29:74-82.
45
46
47
48 Vogelstein B, Kinzler KW. 2004. Cancer genes and the pathways they control. *Nat Med* 10:789-
49
50 799.
51
52
53 Walker DR, Bond JP, Tarone RE, Harris CC, Makalowski W, Boguski MS, Greenblatt MS.
54
55 1999. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation
56
57 with p53 protein structural and functional features. *Oncogene* 18:211-218.
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

For Peer Review

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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	total
<i>APC</i>	177	300	1	6	0	0	0	7	0	1	0	0	0	1
<i>ATM</i>	157	232	0	12	0	1	0	13	0	3	0	0	0	3
<i>BRCA1</i>	70	192	0	12	0	0	0	12	0	6	0	0	0	6
<i>BRCA2</i>	116	310	3	15	2	0	0	20	0	2	0	0	0	2
<i>CDHI</i>	135	116	1	5	0	0	0	6	0	0	0	0	0	0
<i>CDKN2A</i>	50	16	35	3	10	0	0	48	9	0	0	0	0	9
<i>NF1</i>	226	275	0	6	0	0	0	6	0	1	3	0	0	4
<i>NF2</i>	89	59	1	4	0	1	0	6	0	0	0	0	0	0
<i>PTCH1</i>	345	213	2	4	0	0	0	6	0	0	0	0	0	0
<i>PTEN</i>	14	33	1	1	0	5	4	11	2	0	0	0	0	2
<i>RB1</i>	80	81	4	3	2	0	0	9	1	1	1	1	0	3
<i>STK11</i>	137	60	4	3	2	2	0	11	0	0	0	0	0	0
<i>TP53</i>	15	22	8	0	0	35	28	71	10	0	0	23	8	41
<i>TSC1</i>	147	139	0	1	0	0	0	1	0	0	0	0	0	0
<i>TSC2</i>	454	238	0	19	1	0	1	21	0	7	1	0	1	8
<i>VHL</i>	78	24	7	2	4	0	5	18	0	2	4	0	2	6
<i>WT1</i>	143	70	0	9	0	0	0	9	0	4	0	0	0	4
TOTAL	2433	2380	67	105	21	44	38	275	22	27	9	24	11	62

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
<i>APC</i>			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)
<i>ATM</i>		8479-8494 (6)		
<i>BRCA1</i>		181-191 (6) 5085-5098 (8) 5201-5222 (9) 5236-5258 (8)		
<i>BRCA2</i>		8165-8182 (4)		6196-6203 (4) 6443-6450 (8)
<i>NF1</i>		2329-2352 (6) 2530-2543 (5) 4255-4274 (6)		6788-6798 (5)
<i>PTEN</i>	269-287 (18)	367-371 (4)		
<i>RBI</i>		1960-1970 (5)		202-220 (7)
<i>STK11</i>		526-545 (5)		150-197 (11) 737-757 (6)
<i>TP53</i>		832-848 (11)		
<i>TSC1</i>				2101-2112 (5)
<i>TSC2</i>				2059-2074 (5) 4247-4268 (5)
<i>WT1</i>		1174-1201 (13)		

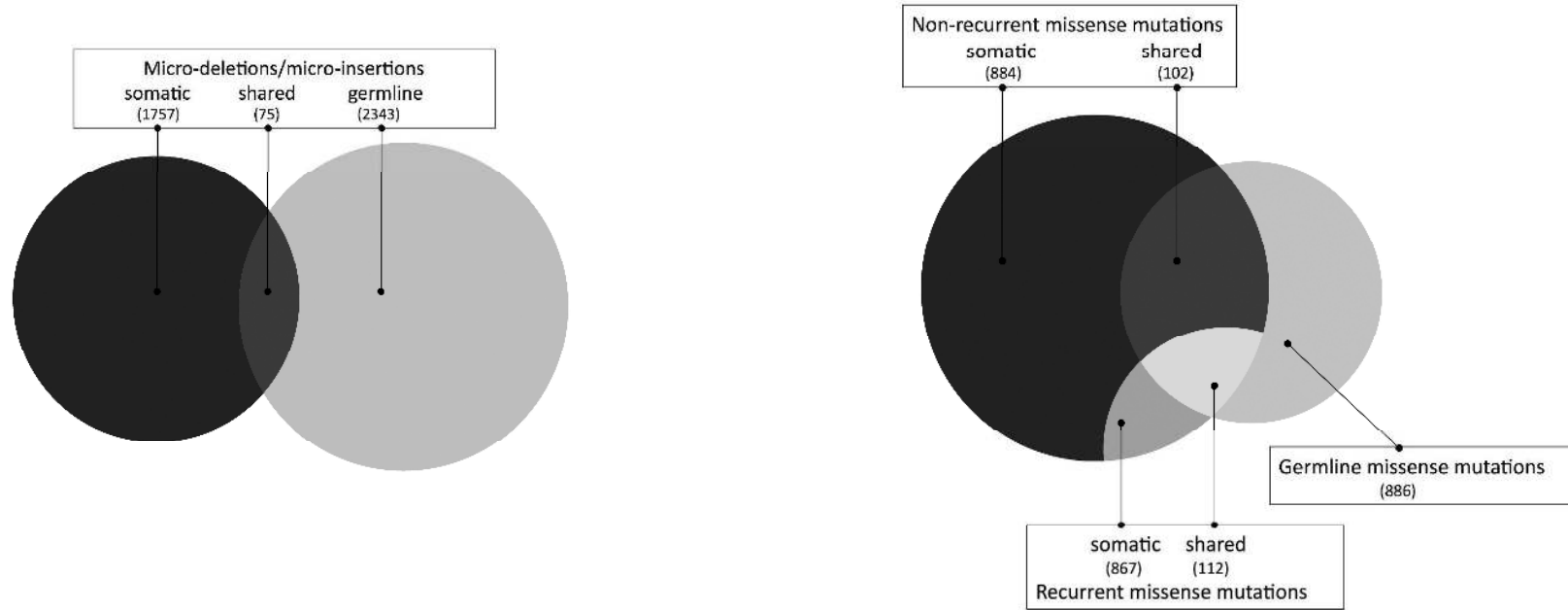


Figure 1. Diagrammatic representation of the number of various types of mutations analysed in the present study.

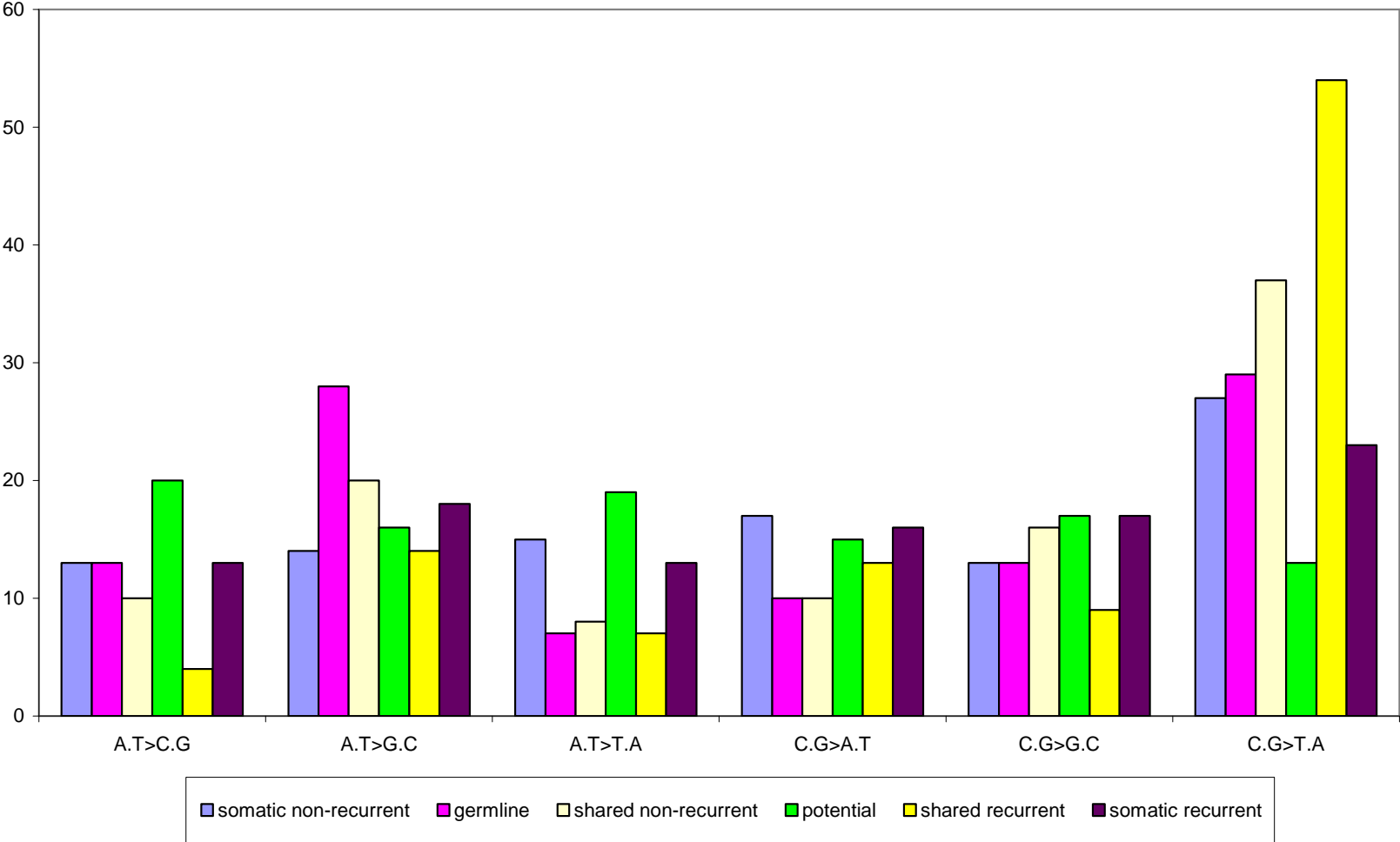


Figure 2. Nucleotide substitution patterns of missense mutations in 17 tumour suppressor genes.

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

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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.

Peer Review

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			<i>STK11</i>	1.66					<i>STK11</i>	25		
			<i>PTCH1</i>	1.06								
		8.4	<i>CDKN2A</i>	1.01	<i>CDKN2A</i>	0.38			<i>CDKN2A</i>	20	<i>CDKN2A</i>	5.2
		7.9	<i>APC</i>	0.83								
		5.6	<i>PTEN</i>	0.53								
		4.6	<i>TP53</i>	0.5	<i>TP53</i>	0.17			<i>RB1</i>	18	<i>TP53</i>	2.8
					<i>VHL</i>	0.14			<i>BRCA2</i>	16		
								<i>PTCH1</i>	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												
		7.2			<i>TSC2</i>	0			<i>BRCA1</i>	7	<i>BRCA1</i>	3.6
		7.3					<i>NF1</i>	98				
		7.6							<i>NF1</i>	7		
		7.9	<i>ATM</i>	0.79	<i>ATM</i>	0	<i>ATM</i>	98	<i>ATM</i>	15	<i>ATM</i>	3.8
		7.9	<i>BRCA1</i>	0.81	<i>VHL</i>	0	<i>VHL</i>	99	<i>BRCA1</i>	16		
		8.7	<i>BRCA2</i>	0.81					<i>NF1</i>	18		
			<i>PTEN</i>	0.92							<i>TSC2</i>	8.1
			<i>RB1</i>	0.99							<i>WT1</i>	10.8
		<i>NF1</i>	1.03									
		<i>TSC2</i>	1.03									

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

	<i>WT1</i>	10.1	<i>WT1</i>	1.22	<i>WT1</i>	0			<i>TSC2</i>	21		
			<i>CDH1</i>	1.27	<i>BRCA1</i>	0.14			<i>APC</i>	24		
					<i>CDKN2A</i>	0.29			<i>CDH1</i>	26		

For Peer Review

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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
<i>APC</i>	16	5	2	0	2	0	9	58	87	6	151
<i>ATM</i>	32	2	11	0	0	0	13	2	43	0	45
<i>BRCA1</i>	20	1	30	0	0	0	31	0	82	2	84
<i>BRCA2</i>	18	6	18	0	0	0	24	2	79	3	84
<i>CDH1</i>	24	4	0	0	0	0	4	5	8	0	13
<i>CDKN2A</i>	24	49	13	5	2	0	69	35	7	1	43
<i>NF1</i>	31	1	20	0	0	0	21	2	85	2	89
<i>NF2</i>	24	6	3	0	1	0	10	49	12	3	64
<i>PTCH1</i>	23	5	8	1	0	0	14	5	23	0	28
<i>PTEN</i>	44	27	3	1	9	0	40	42	13	1	56
<i>RB1</i>	48	3	10	1	0	0	14	4	41	1	46
<i>STK11</i>	33	3	6	0	2	0	11	1	20	1	22
<i>TP53</i>	30	60	2	1	132	23	218	147	1	0	148
<i>TSC1</i>	23	0	3	0	0	0	3	0	27	0	27
<i>TSC2</i>	23	0	13	0	0	0	13	1	29	0	30
<i>VHL</i>	17	3	9	2	0	2	16	25	7	2	34
<i>WT1</i>	26	0	6	0	0	0	6	3	4	0	7
TOTAL	25	175	157	11	148	25	516	381	568	22	971

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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

	Germline	37	76	117	53	11	37	257	0.14	
<i>STK11</i>	Somatic	20	10	5	1	1	20	17	1.18	0.17
	Germline	30	27	47	24	3	30	101	0.30	
<i>TP53</i>	Somatic	1229	96	512	238	0	1229	846	1.45	24.89
	Germline	94	10	16	5	3	94	34	2.76	
<i>TSC1</i>	Somatic	2	1	1	0	0	2	2	1.00	0.02
	Germline	7	37	53	25	4	7	119	0.06	
<i>TSC2</i>	Somatic	2	1	3	2	1	2	7	0.29	0.03
	Germline	89	74	110	46	3	89	233	0.38	
<i>VHL</i>	Somatic	88	15	180	44	1	88	240	0.37	1.82
	Germline	143	27	63	37	5	143	132	1.08	
<i>WT1</i>	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)

```


1
2
3
4
5 | | | | | | | | | | Hess_value > 4: (0.48) (0.28) (0.04) (0.16) (0.04)
6 | | | | | | | | | | Krawczak_value > 0.6155: (0.22) (0.5) (0.06) (0.17) (0.06)
7 | | | | | | | | | | Krawczak_value > 1.083: (0.17) (0.08) (0.08) (0.58) (0.08)
8 | | | | | | | | | | Grantham_score > 40.5
9 | | | | | | | | | | Hess_value <= 5.05
10 | | | | | | | | | | Grantham_score <= 194.5
11 | | | | | | | | | | Krawczak_value <= 0.365
12 | | | | | | | | | | Hess_value <= 3.95
13 | | | | | | | | | | Grantham_score <= 66.5
14 | | | | | | | | | | Hess_value <= 2.65: (0.21) (0.07) (0.38) (0.07) (0.28)
15 | | | | | | | | | | Hess_value > 2.65
16 | | | | | | | | | | Evol <= 0.275: (0.05) (0.05) (0.79) (0.05) (0.05)
17 | | | | | | | | | | Evol > 0.275: (0.32) (0.08) (0.36) (0.2) (0.04)
18 | | | | | | | | | | Grantham_score > 66.5
19 | | | | | | | | | | Grantham_score <= 159.5: (0.36) (0.37) (0.12) (0.15) (0.01)
20 | | | | | | | | | | Grantham_score > 159.5: (0.19) (0.04) (0.3) (0.07) (0.41)
21 | | | | | | | | | | Hess_value > 3.95
22 | | | | | | | | | | Krawczak_value <= 0.229: (0.26) (0.19) (0.04) (0.48) (0.04)
23 | | | | | | | | | | Krawczak_value > 0.229: (0.39) (0.07) (0.04) (0.04) (0.46)
24 | | | | | | | | | | Krawczak_value > 0.365
25 | | | | | | | | | | Hess_value <= 4.55
26 | | | | | | | | | | Hess_value <= 4.3
27 | | | | | | | | | | Grantham_score <= 105.5: (0.51) (0.14) (0.03) (0.29) (0.03)
28 | | | | | | | | | | Grantham_score > 105.5
29 | | | | | | | | | | Hess_value <= 3.3: (0.50) (0.33) (0.06) (0.06) (0.06)
30 | | | | | | | | | | Hess_value > 3.3: (0.36) (0.16) (0.04) (0.28) (0.16)
31 | | | | | | | | | | Hess_value > 4.3: (0.06) (0.24) (0.06) (0.29) (0.35)
32 | | | | | | | | | | Hess_value > 4.55: (0.39) (0.04) (0.04) (0.48) (0.04)
33 | | | | | | | | | | Grantham_score > 194.5: (0.09) (0.09) (0.73) (0.05) (0.05)
34 | | | | | | | | | | Hess_value > 5.05
35 | | | | | | | | | | Grantham_score <= 45.5: (0.04) (0.11) (0.54) (0.29) (0.04)
36 | | | | | | | | | | Grantham_score > 45.5
37 | | | | | | | | | | Evol <= 0.51
38 | | | | | | | | | | Hess_value <= 7.25
39 | | | | | | | | | | Evol <= 0.28: (0.07) (0.43) (0.07) (0.36) (0.07)
40 | | | | | | | | | | Evol > 0.28: (0.27) (0.27) (0.24) (0.18) (0.03)
41 | | | | | | | | | | Hess_value > 7.25
42 | | | | | | | | | | Hess_value <= 7.6: (0.09) (0.18) (0.09) (0.55) (0.09)
43 | | | | | | | | | | Hess_value > 7.6
44 | | | | | | | | | | Grantham_score <= 69: (0.57) (0.09) (0.04) (0.26) (0.04)
45
46
47


```

1
2
3
4
5 | | | | | Grantham_score > 91.5:      (0.33) (0.17) (0.03) (0.03) (0.43)
6 | | | | | Hess_value > 12.1
7 | | | | |   Evol <= 0.51
8 | | | | |     Repeats = 0
9 | | | | |       Grantham_score <= 44.5: (0.20) (0.07) (0.07) (0.60) (0.07)
10 | | | | |         Grantham_score > 44.5
11 | | | | |           Grantham_score <= 51: (0.09) (0.27) (0.09) (0.09) (0.45)
12 | | | | |             Grantham_score > 51: (0.24) (0.53) (0.03) (0.18) (0.03)
13 | | | | |               Repeats = 1:      (0.32) (0.45) (0.05) (0.14) (0.05)
14 | | | | |                 Evol > 0.51
15 | | | | |                   Hess_value <= 13.35: (0.27) (0.15) (0.04) (0.5) (0.04)
16 | | | | |                     Hess_value > 13.35: (0.28) (0.44) (0.04) (0.2) (0.04)
17 | | | | | Hess_value > 42.75
18 | | | | |   Repeats = 0
19 | | | | |     Evol <= 0.59
20 | | | | |       Evol <= 0.255:      (0.08) (0.12) (0.73) (0.04) (0.04)
21 | | | | |         Evol > 0.255
22 | | | | |           Evol <= 0.375: (0.18) (0.03) (0.03) (0.28) (0.49)
23 | | | | |             Evol > 0.375: (0.40) (0.13) (0.07) (0.33) (0.07)
24 | | | | |               Evol > 0.59
25 | | | | |                 Grantham_score <= 139: (0.02) (0.20) (0.75) (0.02) (0.02)
26 | | | | |                   Grantham_score > 139: (0.36) (0.43) (0.07) (0.07) (0.07)
27 | | | | | Repeats = 1
28 | | | | |   Hess_value <= 59.5
29 | | | | |     Hess_value <= 50.35: (0.40) (0.15) (0.05) (0.35) (0.05)
30 | | | | |       Hess_value > 50.35: (0.67) (0.13) (0.04) (0.13) (0.04)
31 | | | | |         Hess_value > 59.5: (0.19) (0.63) (0.06) (0.06) (0.06)

```

```

32 === Stratified cross-validation ===

```

```

33 === Summary ===

```

```

34 Correctly Classified Instances      2797          63.1377 %
35 Incorrectly Classified Instances  1633          36.8623 %
36 Kappa statistic                    0.5392
37 Mean absolute error                 0.1878
38 Root mean squared error             0.3177
39 Relative absolute error             58.6858 %

```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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

For Peer Review

1
2
3 Comparative Analysis of Germline and Somatic Micro-lesion Mutational Spectra in
4
5 17 Human Tumour Suppressor Genes
6
7

8 Dobril Ivanov^{1,2}, Stephen E. Hamby³, Peter D. Stenson¹, Andrew D. Phillips¹,
9 Hildegard Kehrer-Sawatzki⁴, David N. Cooper¹ and Nadia Chuzhanova³
10

11
12
13 ¹Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, CF14 4XN, UK
14

15 ²MRC Centre for Neuropsychiatric Genetics and Genomics, Department of Psychological Medicine
16 and Neurology, Biostatistics and Bioinformatics Unit, School of Medicine, Cardiff University,
17 Cardiff, CF14 4XN, UK
18
19

20
21 ³School of Science and Technology, Nottingham Trent University, Nottingham, NG11
22 8NS, UK
23
24

25 ⁴Institute of Human Genetics, University of Ulm, Albert-Einstein-Allee 11, 89081 Ulm,
26 Germany
27
28
29
30
31
32
33
34
35
36
37
38

39 *All correspondence to: Prof. Nadia Chuzhanova, School of Science and Technology
40 Nottingham Trent University, Clifton Lane, Nottingham NG11 8NS, UK
41
42
43 Tel: +44 (0) 0115 848 8304 E-mail: nadia.chuzhanova@ntu.ac.uk
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

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

Deleted: In this study

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

1
2 which serves to explain not only the phenomenon of genetic anticipation characteristic of these
3 disorders but also their inherent inter-individual clinical variability [Giovannone et al., 1997;
4 Leeflang et al., 1999; Martorell et al., 2000; Sharma et al., 2002; Pollard et al., 2004]. However,
5
6
7
8 by contrast, highly variable human minisatellites can display markedly different degrees of
9
10 instability between the soma and the germline [Buard et al., 2000; Stead and Jeffreys, 2000;
11
12 Shanks et al., 2008]. These studies notwithstanding, few attempts have so far been made to
13
14 compare the nature, location and relative frequency of germline and somatic mutations.
15

16 Human cancer genes usually harbour either somatic or germline mutations [Goode et al., 2002;
17
18 Futreal et al., 2004; Vogelstein and Kinzler, 2004]. There is, however, one category of cancer
19
20 gene, broadly termed tumour suppressors, that by virtue of their being mutated in both the
21
22 germline and the soma, provides us with an ideal model system to compare somatic vs. germline
23
24 mutational spectra [Futreal et al., 2004]. Tumour suppressor genes, defined as “genes that sustain
25
26 loss-of-function mutations in the development of cancer” [Haber and Harlow, 1997], are
27
28 involved in the regulation of a diverse array of different cellular functions including cell cycle
29
30 checkpoint control, detection and repair of DNA damage, protein ubiquitination and degradation,
31
32 mitogenic signalling, cell specification, differentiation and migration, and tumour angiogenesis
33
34 [Sherr, 2004]. They encode proteins with a regulatory role in cell cycle progression (e.g. Rb),
35
36 DNA-binding transcription factors (e.g. p53) and inhibitors of cyclin-dependent kinases required
37
38 for cell cycle progression (e.g. p16). In inherited cancer syndromes, the mutational inactivation
39
40 of both tumour suppressor alleles is required to change the phenotype of the cell. This ‘two hit
41
42 hypothesis’ provides the basis for our mechanistic understanding of tumour suppressor gene
43
44 mutagenesis: a first (inherited) mutation in one tumour suppressor allele is followed by the
45
46 somatic loss of the remaining wild-type allele via a number of different mutational mechanisms
47
48 [Knudson, 2001]. Whereas the inherited lesion is usually fairly subtle, the second (somatic) hit
49
50 may also involve the deletional loss of the entire gene or even a substantial portion of the
51
52
53
54
55
56
57
58
59
60

1
2 chromosome involved. Alternatively, both 'hits' may constitute somatic mutations: whatever the
3
4 actual mechanism, the end result is the same – the loss or inactivation of both gene copies. Some
5
6 interplay may however occur between the soma and the germline in that the location of the
7
8 germline mutation can in some instances influence the nature, frequency and location of the
9
10 subsequent somatic mutation [Lamlum et al., 1999; Groves et al., 2002; Latchford et al., 2007;
11
12 Dallosso et al., 2009; Dworkin et al., 2010].

13
14 Tumour suppressor genes are often somatically inactivated by mutational mechanisms that are
15
16 almost exclusively confined to the soma and which are found only infrequently in the germline
17
18 (e.g. gross mutations characterized by loss of heterozygosity, epi-mutations such as methylation-
19
20 mediated promoter inactivation, and micro-lesions within highly repetitive sequence elements
21
22 that are consequent to microsatellite instability). However, a typical spectrum of somatic
23
24 mutations associated with tumorigenesis may also include gross rearrangements, copy number
25
26 variation, and various types of micro-lesion (e.g. micro-deletions, micro-insertions and indels)
27
28 including single base-pair substitutions [Loeb and Harris, 2008; Stratton et al., 2009]. Although
29
30 the somatic micro-lesions are often quite similar to their germline counterparts, few studies of
31
32 tumour suppressor genes have so far attempted to compare and contrast germline and somatic
33
34 mutational spectra with respect to these relatively subtle types of mutation. However, several
35
36 such studies have indicated that germline and somatic micro-lesions can display remarkable
37
38 similarities in terms of mutation type, location and relative frequency of occurrence, and hence
39
40 by inference the putative underlying mechanisms of mutagenesis [Marshall et al., 1997; Ali et
41
42 al., 1999; Gallou et al., 1999; Richter et al., 2003; Upadhyaya et al., 2004; Glazko et al., 2004;
43
44 Tartaglia et al., 2006; Baser et al., 2006; Upadhyaya et al., 2008].

45
46 We attempt here a first formal comparison between germline and somatic micro-lesion
47
48 mutational spectra for a total of 17 different human tumour suppressor genes [*APC* (MIM#
49
50 611731), *ATM* (MIM# 607585), *BRCA1* (MIM# 113705), *BRCA2* (MIM# 600185), *CDHI*

(MIM# 192090), *CDKN2A* (MIM# 600160), *NF1* (MIM# 162200), *NF2* (MIM# 607379), *PTCH1* (MIM# 601309), *PTEN* (MIM# 601728), *RBI* (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>; *RBI* and *PTEN*), the *Breast Cancer Information Core* (<http://research.nhgri.nih.gov/bic>; *BRCA1*), the *RBI Gene Mutation Database* (<http://www.verandi.de/joomla>; *RBI*), 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

1
2 also to be absent from a non-tumour tissue (usually blood) from the same patient. Hence,
3
4 mutational data derived from 'sporadic' patients were not included unless a non-tumour tissue
5
6 had also been examined in order to exclude the possibility that the lesions detected were
7
8 constitutional in origin. Depending upon the number of independent occurrences, f , of a given
9
10 somatic or shared mutation described in the literature, these mutation types were further
11
12 subdivided into two categories: *recurrent mutations* ($f > 1$) and *non-recurrent mutations* ($f = 1$). At
13
14 the time this study was initiated (October 2006), the number of available germline and somatic
15
16 missense mutations for each of the 17 studied tumour suppressor genes were as listed in Table 1.
17

18 The analysis reported here focussed exclusively on missense mutations and micro-deletions/
19
20 micro-insertions. Nonsense mutations in tumour suppressor genes have already been addressed
21
22 in the context of a general meta-analysis of this type of lesion [Mort et al., 2008]. Indels
23
24 (complex lesions, representing combined micro-deletion/micro-insertions) were excluded from
25
26 the analysis owing to their paucity.

Deleted: s

Deleted: ed

Deleted: /

Deleted: s

30 *Control datasets of potential mutations*

31
32 For every tumour suppressor gene examined, all possible single base-pair substitutions in the
33
34 gene coding sequence that (i) could potentially have given rise to a missense mutation and (ii)
35
36 were not already included in either of the corresponding observed somatic and/or germline
37
38 mutational spectra, were generated. These 'potential missense mutations' were used as a control
39
40 dataset.
41

42 For each tumour suppressor gene, a matching control dataset of 'potential micro-deletions' was
43
44 also generated by randomly selecting a first breakpoint and then choosing the length of the
45
46 simulated micro-deletion (and hence the position of the second breakpoint) by reference to the
47
48 probability distribution calculated for micro-deletions (from 1 bp to 20 bp) observed in the
49
50 corresponding dataset of mutations. A matching dataset of micro-insertions was generated in
51
52
53
54
55
56
57
58
59
60

1
2 similar fashion, with the sites of insertion being randomly selected. Since some of the micro-
3 deletion/micro-insertion breakpoints occurred within an intron, extended cDNA sequences
4 comprising exons and additional flanking intronic sequence were used to generate corresponding
5 control datasets.
6
7
8
9

10 11 12 *Grantham scores*

13
14 The 'Grantham score' or 'Grantham difference' [Grantham, 1974] measures the chemical
15 difference between wild-type and mutated amino acid residues in terms of their side chain
16 composition (i.e. the weight ratio of non-carbon components in end-groups or rings to carbons in
17 side chains), polarity (i.e. basic, acidic or nonpolar depending upon side chain charge) and
18 molecular volume.
19
20
21
22
23

24 On average, the physicochemical differences manifested by orthologous amino acid
25 substitutions that have accumulated over evolutionary time will tend to be relatively small. By
26 contrast, disease-causing substitutions are expected to exhibit higher Grantham scores, indicative
27 of more dramatic physicochemical differences between the wild-type and mutated amino acid
28 residues [Krawczak et al., 1998]. The values tabulated by Grantham [1974] were used in this
29 study to calculate a median Grantham score for each set of missense mutations for each tumour
30 suppressor gene.
31
32
33
34
35
36
37
38
39
40

41 *Degree of evolutionary conservation*

42 Amino acid residues that are highly conserved in orthologous proteins frequently represent sites
43 of structural or functional importance. Hence, such highly conserved amino acid residues/protein
44 regions often constitute hotspots for observed pathological mutations as a consequence of
45 phenotype selection (rather than intrinsic mutability). To assess the degree of evolutionary
46 conservation of those codons affected by somatic/germline mutations, orthologous tumour
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 suppressor cDNA and protein sequences from different vertebrate species were retrieved from
3
4 NCBI's Entrez Gene database (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>). The species
5
6 used as a source of these cDNA and protein sequences are listed in Supp. Table 1 for each
7
8 tumour suppressor gene/protein. ClustalX (<http://www.clustal.org/>) was used to align the protein
9
10 sequences. A program was written to replace all amino acids in the protein alignments by cDNA-
11
12 derived codons, thereby avoiding the introduction of gaps within codons.
13

14 The evolutionary constraints acting upon the 17 human tumour suppressor genes at the codon
15
16 level were inferred by calculating the $\frac{K_a}{K_a + K_s}$ ratio for each codon where K_s and K_a are
17
18 respectively the relative numbers of synonymous and nonsynonymous substitutions between
19
20 codons in two aligned sequences [Walker et al., 1999]. If two aligned codons required more than
21
22 one substitution to be transformed into each other, then the minimum number of substitutions
23
24 was assumed, and the most parsimonious path was determined using a PAM100 matrix and the
25
26 Nei & Gojobori [1986] pathway method. Gaps inserted into the non-human vertebrate
27
28 orthologous cDNA sequences during alignment were treated as being equivalent to a non-
29
30 synonymous substitution. Codons that were not present in the human cDNA sequence were not
31
32 considered. A value representing the median level of evolutionary conservation across all codons
33
34 was then derived for each mutational spectrum; the higher values correspond to less conserved
35
36 genes whereas the lower values refer to more highly conserved ones.
37
38
39
40
41

42 *Relative mutability rates*

43
44 To assess the likelihood of observing a certain nucleotide change in a given position and in a
45
46 specific context, two tabulated measures of the nearest neighbour-dependent mutation rate were
47
48 employed. The first was derived from 20,200 single base-pair substitutions inferred from
49
50 alignments of paired human gene/pseudogene sequences [Hess et al., 1994]. This was termed the
51
52 *non-disease-associated mutability rate* and, since it approximates to the neutral mutation
53
54

1
2 frequency, it should reflect the intrinsic mutability of the underlying DNA sequence. One would
3
4 expect the non-disease-associated mutation rates associated with pathological mutations to be
5
6 low implying that these specific substitutions are much less likely to occur as neutral
7
8 substitutions.
9

10 The nearest neighbour-dependent mutation rates derived from germline single base-pair
11
12 substitutions [using data from the Human Gene Mutation Database (HGMD); Stenson et al.,
13
14 2009] by Krawczak et al. [1998] were used as an approximation of the *disease-associated*
15
16 *mutability rate*. This mutation rate is a function of selection for loss of biological function as
17
18 well as the underlying intrinsic mutability of the DNA sequence. This mutability rate varies
19
20 between 0.032 for the C(A>T)G mutation and 13.023 for the C(G>A)G mutation [Krawczak et
21
22 al., 1998].
23
24

25 26 27 *Repetitive sequence elements*

28 A variety of repetitive sequence elements have been reported in association with human gene
29
30 mutations causing both inherited disease and cancer. Direct and inverted repeats and symmetric
31
32 elements [see Chuzhanova et al. 2003 for definitions] of length ≥ 8 bp, and less than 21 bp apart,
33
34 capable of forming non-B DNA structures, were therefore sought within the extended cDNA
35
36 sequences (comprising exons and up to ± 85 bp of flanking sequence) using purposely designed
37
38 software. In addition, DNA sequences were screened for the presence of mononucleotide runs of
39
40 ≥ 4 bp.
41
42

43 44 45 *Mutation descriptors*

46 Each missense mutation was ascribed various descriptors indicating (a) the type of mutation [i.e.
47
48 shared mutation (i.e. found to occur both somatically and in the germline); exclusively somatic;
49
50 exclusively germline; shared recurrent mutation (i.e. found to occur not only in the germline but
51
52
53
54

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

14
15 All micro-lesions (*viz.* missense mutations, micro-deletions and micro-insertions) in each gene
16 were also labelled with respect to their occurrence within a region spanning a repetitive element
17 or mononucleotide run including ± 5 bp of flanking sequence. If a missense mutation (or at least
18 one micro-deletion/micro-insertion breakpoint) was found to occur within this extended region,
19 the micro-lesion was labelled as being found in association with the corresponding type of
20 repeat.

21 *Assessing the statistical significance of the results generated*

22 To assess the similarity (or dissimilarity) of the germline and somatic mutational spectra with
23 respect to (i) the frequency with which the missense mutations were located within CpG/non-
24 CpG dinucleotides or CpHpG/non-CpHpG trinucleotides and (ii) the frequency with which the
25 micro-deletions/micro-insertions were found within/outwith repeats, the various non-overlapping
26 mutation datasets (bearing specific descriptors) were compared by means of the χ^2 test. Since the
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 normality assumption did not hold for the datasets studied, the Wilcoxon rank-sum test was used
3
4 to compare and contrast missense mutational spectra with respect to the Grantham score, degree
5
6 of evolutionary conservation, and both the non-disease- and disease-associated mutability rates.
7

8 The permutation-based method [Olshen and Jain, 2002] was used to estimate the significance
9
10 of our findings and to allow for multiple testing wherever appropriate. For each comparison, the
11
12 null hypothesis [viz. no overall difference between two groups of mutations (e.g. somatic and
13
14 potential) with respect to the specific property in question (e.g. occurrence in CpG or non-CpG
15
16 nucleotides)], was tested for, either in the context of each gene or all genes combined. χ^2 or rank-
17
18 sum statistics were calculated for the observed germline and somatic mutations as well as for
19
20 10,000 control sets of mutations created from the original sets by random permutation of the
21
22 assigned mutational descriptors (e.g. randomly chosen mutations labelled as 'somatic' were re-
23
24 labelled as 'germline'; randomly chosen mutations labelled as 'shared' were re-labelled as
25
26 'somatic', etc.). The test statistic (χ^2 or rank-sum) for the original datasets that exceeded the 95th
27
28 percentile of χ^2 maxima for 10,000 control sets was deemed to be statistically significant; the
29
30 corresponding p-value was termed the 'gene-wise' p-value. To allow for multiple testing in those
31
32 cases where specific mutations in all genes were combined, a Bonferroni correction was applied;
33
34 the corresponding p-value was termed the 'experiment-wise' p-value.
35

36
37 Power calculations for the χ^2 tests were performed using the Pwr.Chisq.test package, part of
38
39 the R Statistical Language (<http://cran.r-project.org/>). A data based simulation method [Walters
40
41 2004] was used to perform power calculations for the Wilcoxon rank-sum tests. Only results
42
43 showing $\geq 80\%$ power to detect experiment- or gene-wise significance were reported.
44

45 46 47 *Naïve Bayes classifier*

48
49 A decision tree classifier known as a Naïve Bayes tree [NBTree; Kohavi, 1996], implemented in
50
51 the Weka machine learning package [Witten and Frank, 2005], was trained to discriminate
52
53
54

1
2 between somatic, germline, shared, recurrent somatic and recurrent shared missense mutations.
3
4 Each mutation was described by a total of six features including the degree of evolutionary
5
6 conservation, the non-disease-associated and disease-associated relative mutability rates,
7
8 Grantham score, and occurrence in CpG/CpHpG, non-CpG/non-CpHpG doublets/triplets or in
9
10 repeats/mononucleotide runs. Ten-fold cross-validation was used to assess the accuracy of
11
12 classification. The mutation datasets were balanced using random oversampling [Kotsiantis et
13
14 al., 2006] by replicating random instances from the minority classes until all classes were
15
16 represented by the same number of instances as the majority class.
17
18
19
20
21
22

23 Results and Discussion

24
25 The availability of both germline and somatic mutational spectra from tumour suppressor genes
26
27 provides us with an ideal opportunity to study the nature of mutation of the same gene sequences
28
29 in both the germline and the soma. The analysis reported here explores for the first time the
30
31 similarities and differences exhibited by the germline, somatic (and shared) micro-lesion
32
33 mutational spectra in 17 human tumour suppressor genes. The study presented here focussed
34
35 upon missense mutations and micro-deletions as well as micro-insertions. Nonsense mutations in
36
37 tumour suppressor genes have already been addressed elsewhere in the context of a general
38
39 meta-analysis of this type of lesion [Mort et al., 2008].
40
41
42

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

44
45 Taken together, the combined mutational spectra for all 17 tumour suppressor genes contained
46
47 twice as many somatic (61%) as germline (31%) mutations. [Further details are provided in the](#)
48
49 [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*, *RBI*, *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.

1
2 Shared mutations are of particular interest because identical mutational mechanisms operating
3 in the germline and the soma may be inferred for such lesions. The expected number of shared
4 mutations for each gene was calculated as $p_{\text{somatic}} \times p_{\text{germline}} \times (\text{total number of mutations})$,
5
6 where p denotes the relative frequencies of somatic and germline mutations. Although the
7
8 proportion of shared mutations varies markedly between genes (from 0% to 25% of the total),
9
10 only two genes (*TP53* and *VHL*) were found to have a higher than expected number of shared
11
12 mutations as calculated above.
13
14
15
16
17
18

19 *Patterns of germline and somatic missense mutations by mutation type*

20
21 Missense mutations were characterised by a predominance of transitions over transversions
22
23 (Figure 2). The transition:transversion ratio was at its highest for shared recurrent mutations (3.5)
24
25 and shared non-recurrent mutations (2.7). By contrast, the transition:transversion ratio for the
26
27 control group (i.e. potential mutations) was 0.85. Significant differences in the
28
29 transition:transversion ratio were observed between all mutation types ($p < 0.05$) with the
30
31 exception of germline vs. shared mutations (Figure 2).
32

33
34 Not surprisingly, a strong positive correlation was noted between somatic and shared
35
36 mutational spectra (Pearson's correlation $r = 0.986$, $p = 2.91 \times 10^{-4}$) with respect to the frequencies
37
38 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
39
40 C.G>T.A. Weaker negative correlations were found between somatic mutations and the control
41
42 dataset of mutations ($r = -0.887$, $p = 0.019$) and between the shared and control ($r = -0.837$,
43
44 $p = 0.038$) mutational spectra, indicative of the non-randomness of somatic mutation.

45
46 C.G>T.A transitions constituted the most frequent type of mutation in shared (46%), germline
47
48 (29%) and somatic (25%) mutational spectra, significantly higher proportions than noted in the
49
50 spectrum of mutations within our control dataset (13%, $p < 0.001$) (Figure 2). Intriguingly, the
51
52 number of A.T>G.C mutations was significantly higher (28%) in the germline as compared to
53
54
55
56
57
58
59
60

1
2 the somatic (16%), shared (17%) and control (16%) mutational spectra (Figure 2). A.T>C.G
3
4 mutations were significantly under-represented in the shared mutational spectrum (7%, $p<0.001$)
5
6 as compared to the other spectra whereas A.T>T.A mutations were under-represented (7%,
7
8 $p<0.001$) in both the germline and shared mutational spectra compared to both somatic and
9
10 potential mutations (Figure 2). Finally, C.G>A.T mutations were significantly underrepresented
11
12 in the germline mutational spectrum (10%) as compared to the somatic (16%, $p=1.2\times 10^{-5}$) and
13
14 potential (15%, $p=2.6\times 10^{-5}$) spectra. Thus, the main similarity between the somatic and germline
15
16 missense mutational spectra was in relation to C.G>T.A transitions whereas the main differences
17
18 between these spectra involved the A.T>G.C, A.T>T.A and C.G>A.T mutations. In passing, it
19
20 should be noted that the patterns of somatic nucleotide substitution exhibited by the 17 tumour
21
22 suppressor genes studied here were markedly different from the genome-wide patterns of
23
24 somatic nucleotide substitution Deleted: previously observed in various cancer genome sequencing studies [Sjöblom
25
26 et al., 2006; Greenman et al., 2007; Kan et al., 2010]; [these mutation datasets are likely to differ](#)
27
28 [quite dramatically with respect to their relative proportions of 'passenger' mutations.](#)
29
30

31 32 *CpG- and CpHpG-located missense mutations*

33
34 The CpG dinucleotide is a well known mutational hotspot in the human genome as a
35
36 consequence of the spontaneous (and endogenous) deamination of 5-methylcytosine. In addition,
37
38 Lister et al. [2009] reported abundant DNA methylation in CpHpG trinucleotides in the human
39
40 genome, where H is either A, C or T, raising the possibility that CpHpG might also be a
41
42 generalized mutation hotspot [Cooper et al., 2010].

43
44 The proportion of missense mutations that were either C>T or G>A within CpG or CpHpG
45
46 oligonucleotides in the 17 tumour suppressor genes was found to vary between 0% and 100%
47
48 (Table 2). This wide range in values may be attributed to the small size of some of the gene
49
50 mutation datasets under study. Importantly, the CpG and CpHpG oligonucleotides were found to
51
52

1
2 be disproportionately likely to harbour shared mutations; thus, 34% of shared recurrent
3
4 mutations and 21% of shared non-recurrent mutations were C>T and G>A mutations in CpG
5
6 dinucleotides with an additional 10% and 9% of mutations, respectively, occurring within
7
8 CpHpG trinucleotides. Since driver mutations tend to occur disproportionately frequently within
9
10 CpG dinucleotides [Talavera et al., 2010], we postulate that missense mutations identified as
11
12 being shared are highly likely to be driver mutations.

13
14 Significant differences were noted between the relative frequencies of CpG- and CpHpG-
15
16 located mutations for somatic, germline, shared, somatic recurrent and shared recurrent missense
17
18 mutations (Supp. Table 2).

19
20 We have previously shown that 18.2% and 9.9% of all missense/nonsense mutations recorded
21
22 in the HGMD are C>T and G>A transitions in CpG and CpHpG oligonucleotides respectively
23
24 [Cooper et al., 2010]. In the present study, we observed that the mutational spectra of shared and
25
26 shared recurrent missense mutations in tumour suppressor genes were both found to be
27
28 significantly enriched in CpG-located mutations (χ^2 -test; p-values, 0.028 and 1.1×10^{-9}
29
30 respectively). This implies that the CpG dinucleotide is a generalized mutation hotspot in both
31
32 the soma and the germline as a consequence of the endogenous mutational mechanism of
33
34 methylation-mediated deamination of 5-methylcytosine. By contrast, the number of CpG-located
35
36 mutations was significantly underrepresented (χ^2 -test; p-values $< 5 \times 10^{-14}$) in the other mutational
37
38 spectra (i.e. non-recurrent somatic, somatic recurrent and germline mutations) by comparison
39
40 with HGMD data. To perform these comparisons, missense mutations (Table 2) and nonsense
41
42 mutations [previously reported in Mort et al., 2008; see Table 6 therein] in all 17 tumour
43
44 suppressor genes were combined. The proportion of shared recurrent missense mutations in
45
46 tumour suppressor genes that were CpHpG-located was found to be significantly higher
47
48 (p=0.023) than for mutations recorded in the HGMD whereas CpHpG-located somatic and
49
50 recurrent somatic mutations were significantly under-represented (p $< 4 \times 10^{-10}$). Significant
51
52
53
54
55
56
57
58
59
60

1
2 enrichment in CpHpG-located mutations was observed for germline mutations as compared to
3 somatic mutations ($p < 3 \times 10^{-10}$) consistent with the reported decrease in CpHpG methylation in
4 differentiated cells [Lister et al., 2009]. In summary, germline and shared missense mutations
5 were found to be significantly enriched at CpG and CpHpG oligonucleotides.
6
7

8
9
10 The numbers of somatic and shared C>T and G>A transitions recorded within CpG
11 dinucleotides for each gene (Table 2) did not correlate with the numbers of CpG dinucleotides
12 found in these genes ($r < -0.5$, $p > 0.127$) and hence do not simply reflect intragenic CpG
13 frequency. A weak positive correlation between CpG-located mutations and the number of genic
14 CpG dinucleotides was however noted for germline mutations ($r = 0.489$, $p = 0.046$) indicating that
15 CpG methylation is not entirely unrelated to the number of CpG dinucleotides, at least with
16 respect to the germline; the relationship is however clearly more complex in the soma, possibly
17 due to inter-tissue differences in gene methylation patterns [Tornaletti and Pfeifer, 1995] or
18 transcription-coupled repair [Rubin and Green, 2009].
19
20
21
22
23
24
25
26
27

28
29 No correlation was found between the numbers of somatic, germline and shared mutations
30 recorded within CpHpG trinucleotides and the corresponding numbers of CpHpG trinucleotides
31 for these genes ($r = -0.316, 0.373, -0.414$; p -values 0.281, 0.216 and 0.098, respectively)
32 indicating that mutation within CpHpG trinucleotides is likely to be very much a gene-specific
33 phenomenon (presumably dependent on both the extent and the degree of spatial localization of
34 CpHpG methylation in the germline and/or soma).
35
36
37
38
39

40
41 Finally, the number of CpG dinucleotides in the various tumour suppressor genes studied
42 (Table 2) was not found to correlate with gene length ($r = 0.3$, p -value=0.241). By contrast, we
43 found a significant correlation ($r = 0.885$, p -value= 2.35×10^{-6}) between tumour suppressor gene
44 length and the number of CpHpG trinucleotides (excluding those with mutations), indicating that
45 the tumour suppressor genes under study possess a similar density of CpHpG trinucleotides per
46
47
48
49
50

1
2 unit length. We surmise that the factors that govern the establishment of the methylation pattern
3
4 of CpHpG trinucleotides are likely to be quite complex.
5
6

7
8 *Evolutionary conservation of tumour suppressor genes in relation to the sites of somatic and*
9
10 *germline missense mutations*

11 For all 17 tumour suppressor genes, the degree of evolutionary conservation, as measured by
12
13 Ka/Ks , was less than unity, indicating that these genes (and proteins) have been highly
14
15 conserved evolutionarily as a consequence of the action of purifying selection. Indeed, the
16
17 degree of evolutionary conservation displayed by most of the studied genes was markedly lower
18
19 than the average (~0.18) noted in a comparison of 1880 human, rat and mouse gene orthologues
20
21 [Makalowski and Boguski, 1998]. However, three genes (*CDKN2A*, *BRCA1* and *BRCA2*) were
22
23 found to exhibit a higher rate of evolutionary conservation than the average between human and
24
25 rodents.
26
27

28
29 The evolutionary conservation of each mutated codon was inferred by calculating the $\frac{Ka}{Ka + Ks}$
30
31 ratio; for each gene/spectrum, the mean value was then calculated across all mutations in the
32
33 corresponding gene/spectrum. Shared recurrent missense mutations were found to occur
34
35 disproportionately in highly conserved amino acid residues (mean degree of evolutionary
36
37 conservation, 0.072) followed by shared non-recurrent mutations (0.138), somatic recurrent
38
39 (0.169), germline (0.175), non-recurrent somatic (0.265), and control dataset mutations (0.255).
40
41 The observed differences in the degree of evolutionary conservation for the different mutational
42
43 spectra are shown in Supp. Table 2. These quite specific findings are consistent with the
44
45 previously reported general tendency for cancer-associated mutations to occur frequently at
46
47 evolutionarily conserved sites [Greenblatt et al., 2003; Tavtigian et al., 2009; Talavera et al.,
48
49 2010].
50
51
52
53
54
55
56
57
58
59
60

1
2 Somatic non-recurrent mutations were found to occur in codons characterized by the highest
3
4 mean value of $\frac{K_a}{K_a + K_s}$ ratios as compared not only to the shared recurrent and shared non-
5
6 recurrent mutations (see above) but also to the mutations within the control dataset. This is
7
8 consistent with the interpretation that a high proportion of non-recurrent somatic mutations, and
9
10 most notably those which are located in less evolutionarily conserved regions ([characterised by](#)
11
12 [higher values of the degree of evolutionary conservation](#)), are likely to be ‘passenger’ mutations.
13
14
15

16 17 *Missense mutations in relation to the disease- and non-disease-associated substitution rates*

18 Employing alignments of paired human gene/pseudogene sequences, Hess et al. [1994] derived
19
20 relative (non-disease-associated) nearest-neighbour-dependent mutability rates using the lowest
21
22 frequency substitution type, C(T>G)A/T(A>C)G, as a baseline. These mutability rates were
23
24 found to vary over a 52-fold range, with unity being assigned to the lowest frequency
25
26 substitution type. This *non-disease-associated* mutability rate approximates to the neutral
27
28 mutation frequency and hence reflects the intrinsic mutability of the underlying DNA sequence.
29
30 Depending upon the observed nearest-neighbour context, we retrieved the corresponding non-
31
32 disease-associated mutability rate (from the data of Hess et al. 1994) for each mutation (either
33
34 observed or from the control dataset) and calculated the median value for each mutational
35
36 spectrum. These median values are indicative of the relative mutability of each tumour
37
38 suppressor gene. [Further details are provided in the Supplementary Text online.](#)
39
40

41
42 When data from all 17 genes were combined, shared recurrent mutations were found to be
43
44 characterised by intrinsically low non-disease-associated mutability (median=11), followed by
45
46 even lower median mutability values for shared non-recurrent mutations (7.9), germline
47
48 mutations (7.2), somatic recurrent and non-recurrent (4.7) and control dataset mutations (4.1).
49
50 Such low median mutability values across all groups indicates that at least half of the mutations
51
52 within observed triplets are unlikely to be neutral in the sense defined by Hess et al. [1994] and
53
54

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 (*RBI*) and 11 (*P TEN*) 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.¶)

1
2 hence are not simply explicable in terms of intrinsic DNA mutability. The low median mutability
3
4 values for the control dataset of mutations within tumour suppressor genes reflect the high level
5
6 of evolutionary conservation manifested by tumour suppressor gene coding sequences across
7
8 different species, implying that any mutation within a triplet characterized by a low non-disease-
9
10 associated mutation rate is very likely to have pathological consequences and would thus be
11
12 subject to purifying selection.

13
14 In contrast to the non-disease-associated mutability rate (which is purely a reflection of the
15
16 intrinsic DNA mutability), the disease-associated mutability rate reflects (in addition to the
17
18 intrinsic DNA mutability) the increased likelihood of coming to clinical attention conferred by
19
20 the loss of biological function. The C(G>T)T mutation is one of the most frequent types of
21
22 mutation associated with the loss of biological function [disease-associated mutability rate
23
24 10.255; Krawczak et al., 1998] but occurs much less frequently among neutral mutations [non-
25
26 disease-associated mutability rate 4.4; Hess et al., 1994].

27
28 For each tumour suppressor gene and each mutational spectrum, the disease-associated median
29
30 mutability values were calculated using mutability rates derived from Krawczak et al. [1998].

31
32 The disease-associated median value was found to be 0.85 for the germline mutations. [Further](#)
33
34 [details are provided in the Supplementary Text online.](#) We found that shared recurrent and

35
36 shared non-recurrent mutational spectra were characterized by higher median values of the
37
38 disease-associated mutability rates (1.42 and 1.01 respectively) whereas somatic non-recurrent,
39
40 somatic recurrent and control dataset mutations exhibited lower median mutability rates (0.5, 0.5
41
42 and 0.4 respectively) as compared to germline mutations (0.85). The finding that the shared
43
44 mutations (which, by definition, occur in both the germline and the soma) are characterized by
45
46 higher disease-associated mutability rates is not surprising since mutations that occur with the
47
48 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).

1
2
3 We postulated that those mutations which occur both in the germline and the soma, and which
4 are characterised by higher disease-associated mutability rates are disproportionately likely to be
5 drivers of tumour development. Consistent with this postulate, somatic recurrent and non-
6 recurrent mutational spectra are characterized by lower median disease-associated mutability
7 rates as compared to the germline spectrum. However, given that higher disease-associated
8 mutability rates are a characteristic feature of driver mutations, a certain proportion of the
9 somatic mutations, namely those characterised by higher disease-associated mutability rates, may
10 correspond to functionally significant driver mutations.
11
12
13
14
15
16
17

18 In assessing the significance of our results, it was appropriate to consider the possibility that
19 somatic mutations might display quite different nearest-neighbour-dependent disease-associated
20 mutability rates from germline mutations. However, since a good correlation was observed
21 between the mutability rates derived from inherited disease data [Krawczak et al., 1998] and the
22 neighbour-dependent mutability rates calculated for the somatic mutations of the 17 tumour-
23 suppressor genes studied here (Pearson's correlation $r=0.703$, $p=6.6\times 10^{-30}$), this *caveat* appears
24 not to be an issue.
25
26
27
28
29
30
31
32
33

34 *Distribution of Grantham scores with respect to tumour suppressor gene mutations*

35 Shared recurrent mutations were found to exhibit the largest median chemical difference value
36 (Grantham scores) between the wild-type and mutated amino acid residues (100) followed by
37 shared non-recurrent mutations and germline mutations (both 93), somatic recurrent (85),
38 somatic non-recurrent (80) and potential mutations (78). Since there was an obvious trend for
39 shared recurrent and non-recurrent mutations to cause the most dramatic chemical changes of the
40 affected codon, we may infer that these types of lesion are also more likely to be driver
41 mutations. However, bearing in mind that the range of theoretically possible values varies
42 between 5 (Leu ↔ Ile) and 215 (Cys ↔ Trp), less elevated median values may simply indicate
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 that a proportion of the mutations in each mutational spectrum are likely to be chemically less
4
5 dramatic (Grantham scores <100).
6
7

8
9 *Missense mutations occurring within repeats and runs of identical nucleotides*

10
11 A number of studies have noted that single base-pair substitutions associated with inherited
12
13 disease occur disproportionately either within, or in close proximity to, repetitive sequences
14
15 [Jego et al., 1993; Greenblatt et al., 1996; Tappino et al., 2009; Thomas et al., 2010; Leclercq et
16
17 al., 2010]. Hence, we wished to assess whether either germline or somatic mutations occurred
18
19 disproportionately either within, or in the vicinity (see *Mutation descriptors*) of, direct, inverted
20
21 and symmetric repeats or mononucleotide runs in the 17 tumour suppressor genes under study
22
23 (Table 3, Supplementary Tables 4-6).
24

25 On average, direct repeats of length ≥ 8 bp were found to cover 5.6% of the cDNA lengths of
26
27 the 17 tumour suppressor genes. [Further details are provided in the Supplementary Text online](#)

28
29 On average, mononucleotide runs ≥ 4 bp spanned 19.9% of the cDNA lengths. Approximately
30
31 24% of non-recurrent somatic and 20% of germline missense mutations were found in

32
33 mononucleotide runs; these proportions were significantly higher than noted for shared non-

34
35 recurrent missense mutations (4.9%, $p \leq 1.6 \times 10^{-4}$). A greater proportion of non-recurrent somatic

36
37 missense mutations was found in direct repeats (7%) as compared to recurrent somatic missense

38
39 mutations (2%, $p = 8.8 \times 10^{-7}$), germline missense (4%, $p = 0.028$) and potential missense mutations

40
41 (3.7%, $p = 8.1 \times 10^{-7}$). This result may reflect the disproportionate number of CpG/CpHpG

42
43 mutations among shared and recurrent somatic missense mutations. Further, for all mutational

44
45 spectra examined (with the exception of the shared mutations), missense mutations were

46
47 preferentially found in association with inverted and symmetric repeats as compared to the

48
49 control dataset of mutations ($p < 0.05$). However, no statistically significant differences were

50
51 found between mutational spectra. [Further details are provided in the Supplementary Text online.](#)
52
53
54
55
56
57
58
59
60

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 *RBI* (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*). ¶

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

▼ -----

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. 4x2 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.

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. ¶

1
2 An additional non-overlapping dataset of 568 missense somatic mutations, identified in the 17
3
4 tumour suppressor genes under study, were extracted from a collection of 2,488 mutations
5
6 identified as being probable driver mutations [Carter et al., 2009]. Features such as the degree of
7
8 evolutionary conservation, Grantham score, mutability rates, CpG/CpHpG location, occurrence
9
10 within repeats/mononucleotide runs were again determined for each of these mutations.
11
12 Employing our classifier, 7% and 10% respectively of these 568 mutations were found to possess
13
14 features consistent with their being shared recurrent and shared non-recurrent mutations. In
15
16 addition, 32% of these probable driver mutations were found to bear features characteristic of
17
18 recurrent somatic mutations (i.e. mutations documented in different tumours). A further 25% of
19
20 the probable (somatic) driver mutations were classified as possessing features characteristic of
21
22 germline mutations and hence could conceivably be treated as shared mutations missing from the
23
24 original training dataset. The remaining 25% of mutations were classified as non-recurrent
25
26 somatic mutations. Using this classifier, which is based on a very modest number (6) of
27
28 predictive features, to analyse an independent dataset of probable driver mutations, we were able
29
30 to predict that ~50% of these somatic missense mutations exhibited features specific to either
31
32 shared or recurrent mutations, indicating that a disproportionate number of such lesions are likely
33
34 to be drivers of tumorigenesis. This percentage is certainly lower (79%) than that obtained by
35
36 Carter et al., [2009] through the application of a Random Forest Classifier based on 500 trees and
37
38 >50 predictive features (using an 'out-of-the-bag' error estimate similar to the cross-validation
39
40 procedure) to the set of putative 2,488 driver mutations. However, based on the results of this
41
42 study, we may conclude that, in general, the mutational spectrum of driver mutations is likely to
43
44 contain a disproportionate number of somatic mutations that have germline counterparts (~17%)
45
46 whilst an additional 32% of the driver mutations are likely to occur recurrently in the soma.
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 ($r = \sim 1$) 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

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 $< 2.20 \times 10^{-16}$). All other mutational spectra were characterized by the predominance of truncating mutations. ¶

1
2 micro-deletions and micro-insertions in both the soma and the germline. The preponderance of
3 such mutations in mononucleotide runs is unsurprising in the context of the shared mutations
4 since all mutations that occur with high frequency within mutation hotspots are more likely to be
5 shared between the germline and the soma (as previously noted for CpG and CpHpG mutations).
6
7 No other types of repeat were disproportionately associated (after correction for multiple testing)
8
9 with micro-deletions and micro-insertions.
10
11
12
13
14
15

16 *Regional hotspots in somatic and germline mutational spectra*

17
18 For the purposes of the following analysis, a regional mutation hotspot was defined as a stretch
19 of DNA of length ≤ 20 bp where four or more independent mutational events have been reported
20 and a significant degree ($p \leq 0.05$) of clustering of these mutations was evident for a given stretch
21 of DNA. In this definition of a regional hotspot, each recurrent mutation was considered only
22 once. The order statistics, r-scans, as described by Karlin and Macken [1991] and applied in
23 Bacolla et al. [2006], were used to detect significant clustering of mutations by comparison with
24 a Poisson distribution of mutations along the gene sequence. Overlapping hotspot regions were
25 considered as a single regional hotspot.
26
27
28
29
30
31
32
33

34 The only regional mutational hotspot for somatic missense mutations was observed in the
35 *PTEN* gene and comprised 18 mutations in the region between nucleotide positions 269 and 286.
36
37 Several germline regional mutational hotspots were however detected for missense mutational
38 spectra in the *ATM*, *BRCA1*, *BRCA2*, *NF1*, *PTEN*, *RBI*, *STK11*, *TP53* and *WT1* genes (Table 4).
39
40 Several somatic regional mutational hotspots were found for micro-deletions/micro-insertions in
41 the *APC* gene, the largest of which contained 33 mutations (positions 4303-4398) and forms part
42 of a previously reported mutation cluster region [Miyoshi et al., 1992]. Regional hotspots
43 identified in different mutational spectra were however unique to that spectrum. The only
44 overlap noted between regional mutational hotspots identified in germline and somatic micro-
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 deletion/micro-insertion mutational spectra was observed for the *APC* gene (the overlapping
3 region comprising nucleotide positions 3919-3933). This micro-deletion/micro-insertion hotspot
4 also includes codon 1309 (cDNA positions 3925-3927) found to be frequently mutated in Greek
5 and French patients with familial adenomatous polyposis [Fostira et al. 2010; Lagarde et al.
6 2010].
7
8
9
10
11

12 Inspection of regional hotspot sequences revealed that they are rich in repetitive elements, runs
13 of identical nucleotides and CpG/CpHpG oligonucleotides, offering immediate explanations for
14 the elevated mutability.
15
16
17
18

19 *Germline and somatic mutations located within specific hotspot motifs*

20 The cDNA sequences of 17 tumour suppressor genes were screened for the presence of nine
21 specific motifs (and their complements) previously reported as being hotspots for mutation.
22 These motifs included the putative somatic (cancer) mutation hotspot, WKVNRRRNVWK [the
23 ‘THEMIS motif’; Makridakis et al., 2009], the RGYW motif that correlates with the DNA
24 polymerase eta error spectrum [Rogozin et al., 2001] and several so-called ‘super hotspot’ motifs
25 originally found in germline micro-insertions and micro-deletions [Ball et al., 2005] and indels
26 [Chuzhanova et al., 2003]. For the purposes of this analysis, the shared mutations were added to
27 both the germline and somatic mutational spectra. Both germline and somatic micro-deletions
28 and micro-insertions were found to be significantly overrepresented ($p \leq 0.002$) in the ‘indel super
29 hotspot’ motif GTAAGT and its complement. Somatic micro-deletions and micro-insertions
30 were also significantly overrepresented ($p = 0.009$) with respect to the micro-deletion/micro-
31 insertion super hotspot AAATCT and its complement. The number of germline (but not somatic)
32 micro-deletions/micro-insertions in the THEMIS motif were significantly overrepresented
33 ($p = 0.003$) as compared to the controls. No significant difference was however observed in the
34 number of missense mutations occurring in any motifs analysed.
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 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.

Deleted: A number of important

Deleted: profound

Deleted: s

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

specific somatic mutations and Rachel Karchin (Johns Hopkins University, Baltimore, USA) for making available her dataset of probable driver mutations.

For Peer Review

References

- 1
2
3
4
5 Ali IU, Schriml LM, Dean M. 1999. Mutational spectra of *PTEN/MMAC1* gene: a tumor
6
7 suppressor with lipid phosphatase activity. *J Natl Cancer Inst* 91:1922-1932.
8
9 Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova NA, Abeysinghe SS, O'Connell
10
11 CD, Cooper DN, Wells RD. 2004. Breakpoints of gross deletions coincide with non-B
12
13 DNA conformations. *Proc Natl Acad Sci USA* 101:14162-14167.
14
15 Bacolla A, Wells RD. 2004. Non-B DNA conformations, genomic rearrangements, and human
16
17 disease. *J Biol Chem* 279:47411-47414.
18
19 Bacolla A, Collins JR, Gold B, Chuzhanova N, Yi M, Stephens RM, Stefanov S, Olsh A,
20
21 Jakupciak JP, Dean M, Lempicki RA, Cooper DN, Wells RD. 2006. Long
22
23 homopurine•homopyrimidine sequences are characteristic of genes expressed in brain
24
25 and the pseudoautosomal region. *Nucleic Acids Res* 34: 2663-2675.
26
27 Ball EV, Stenson PD, Krawczak M, Cooper DN, Chuzhanova NA. 2005. Micro-deletions and
28
29 micro-insertions causing human genetic disease: common mechanisms of mutagenesis
30
31 and the role of local DNA sequence complexity. *Human Mutat* 26:205-213.
32
33 Baser ME, Contributors to the International NF2 Mutation Database. *Hum. Mutat.* 27:297-306.
34
35 Boland CR, Ricciardiello L. 1999. How many mutations does it take to make a tumor? *Proc Natl*
36
37 *Acad Sci USA* 96:14675-14677.
38
39 Buard J, Collick A, Brown J, Jeffreys AJ. 2000. Somatic versus germline mutation processes at
40
41 minisatellite CEB1 (D2S90) in humans and transgenic mice. *Genomics* 65:95-103.
42
43 Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R.
44
45 2009. Cancer-specific high-throughput annotation of somatic mutations: computational
46
47 prediction of driver missense mutations. *Cancer Res* 69:6660-6667.
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Chuzhanova NA, Anassis EJ, Ball E, Krawczak M, Cooper DN. 2003. Meta-analysis of indels
4 causing human genetic disease: mechanisms of mutagenesis and the role of local DNA
5 sequence complexity. *Hum Mutat* 21:28-44.
6
7
8 Chuzhanova N, Chen JM, Bacolla A, Patrinos GP, Férec C, Wells RD, Cooper DN. 2009. Gene
9 conversion causing human inherited disease: the evidence for involvement of
10 recombination-associated motifs and non-B DNA-forming sequences in DNA breakage.
11 *Hum Mutat* 30:1189-1198.
12
13
14 Cole DN, Carlson JA, Wilson VL. 2008. Human germline and somatic cells have similar *TP53*
15 and Kirsten-RAS gene single base mutation frequencies. *Environ. Mol. Mutagen.* 2008
16 49:417-425.
17
18
19 Cooper DN, Mort M, Stenson PD, Ball EV, Chuzhanova NA. 2010. Methylation-mediated
20 deamination of 5-methylcytosine appears to give rise to mutations causing human
21 inherited disease in CpNpG trinucleotides as well as in CpG dinucleotides. *Hum*
22 *Genomics* 4:406-410.
23
24
25 Couch FJ, Rasmussen LJ, Hofstra R, Monteiro AN, Greenblatt MS, de Wind N; IARC
26 Unclassified Genetic Variants Working Group. 2008. Assessment of functional effects of
27 unclassified genetic variants. *Hum Mutat* 29:1314-1326.
28
29
30 Dallosso AR, Jones S, Azzopardi D, Moskvina V, Al-Tassan N, Williams GT, Idziaszczyk S,
31 Davies DR, Milewski P, Williams S, Beynon J, Sampson JR, Cheadle JP. 2009. The *APC*
32 variant p.Glu1317Gln predisposes to colorectal adenomas by a novel mechanism of
33 relaxing the target for tumorigenic somatic *APC* mutations. *Hum Mutat* 30:1412-1418.
34
35
36
37
38
39
40
41
42
43
44
45 [Dworkin AM, Ridd K, Bautista D, Allain DC, Iwenofu OH, Roy R, Bastian BC, Toland AE.](#)
46 [2010. Germline variation controls the architecture of somatic alterations in tumors. *PLoS*](#)
47 [Genet](#) 6: e1001136.
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Fostira F, Thodi G, Sandaltzopoulos R, Fountzilias G, Yannoukakos D. 2010. Mutational
4 spectrum of *APC* and genotype-phenotype correlations in Greek FAP patients. *BMC*
5 *Cancer* 10:389.
6
7
8 Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR.
9
10 2004. A census of human cancer genes. *Nat Rev Cancer* 4:177-183.
11
12 Gallou C, Joly D, Mejean A, Staroz F, Martin N, Tarlet G, Orfanelli MT, Bouvier R, Droz D,
13
14 Chretien Y, Maréchal JM, Richard S, Junien C, Bérout C. 1999. Mutations of the *VHL*
15 gene in sporadic renal cell carcinoma: definition of a risk factor for VHL patients to
16 develop an RCC. *Hum Mutat* 13:464-475.
17
18
19
20 Giovannone B, Sabbadini G, Di Maio L, Calabrese O, Castaldo I, Frontali M, Novelleto A,
21
22 Squitieri F. 1997. Analysis of (CAG)_n size heterogeneity in somatic and sperm cell DNA
23 from intermediate and expanded Huntington disease gene carriers. *Hum Mutat* 10:458-
24 464.
25
26
27
28 Glazko GV, Koonin EV, Rogozin IB. 2004. Mutation hotspots in the p53 gene in tumors of
29 different origin: correlation with evolutionary conservation and signs of positive
30 selection. *Biochim Biophys Acta* 1679:95-106.
31
32
33
34 Goode EL, Ulrich CM, Potter JD. 2002. Polymorphisms in DNA repair genes and associations
35 with cancer risk. *Cancer Epidemiol Biomarkers Prev* 11:1513-1530.
36
37
38
39 Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science*
40 185:862-864.
41
42
43 Greenblatt MS, Grollman AP, Harris CC. 1996. Deletions and insertions in the p53 tumor
44 suppressor gene in human cancers: confirmation of the DNA polymerase
45 slippage/misalignment model. *Cancer Res* 56:2130-2136.
46
47
48
49 Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP. 2003. Detailed
50 computational study of p53 and p16: using evolutionary sequence analysis and disease-
51
52
53
54
55
56
57
58
59
60

- 1
2 associated mutations to predict the functional consequences of allelic variants. *Oncogene*
3
4 22:1150-1163.
5
6 Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J,
7
8 Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S,
9
10 Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K,
11
12 Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko
13
14 T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T,
15
16 West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur
17
18 F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P,
19
20 Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung
21
22 SY, Wooster R, Futreal PA, Stratton MR. 2007. Patterns of somatic mutation in human
23
24 cancer genomes. *Nature* 446:153-158.
25
26 Groves C, Lamlum H, Crabtree M, Williamson J, Taylor C, Bass S, Cuthbert-Heavens D,
27
28 Hodgson S, Phillips R, Tomlinson I. 2002. Mutation cluster region, association between
29
30 germline and somatic mutations and genotype-phenotype correlation in upper
31
32 gastrointestinal familial adenomatous polyposis. *Am J Pathol* 160:2055-2061.
33
34 Haber D, Harlow E. 1997. Tumour-suppressor genes: evolving definitions in the genomic age.
35
36 *Nat Genet* 16:320-322.
37
38 Hess ST, Blake JD, Blake RD. 1994. Wide variations in neighbor-dependent substitution rates. *J*
39
40 *Mol Biol* 236:1022-1033.
41
42 Jegu N, Thomas G, Hamelin R. 1993. Short direct repeats flanking deletions, and duplicating
43
44 insertions in p53 gene in human cancers. *Oncogene* 8:209-213.
45
46 Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon
47
48 R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S,
49
50 Davis DP, Carlton VE, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard
51
52
53
54
55
56
57
58
59
60

- 1
2
3 DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, de Sauvage FJ, Faham M,
4 Seshagiri S. 2010. Diverse somatic mutation patterns and pathway alterations in human
5 cancers. *Nature* 466:869-873.
6
7
8 Karlin S, Macken C. 1991. Some statistical problems in the assessment of inhomogeneity of
9 DNA sequence data. *J Am Statist Assoc* 86:27–35.
10
11
12 Knudson AG. 2001. Two genetic hits (more or less) to cancer. *Nat Rev Cancer* 1:157-162.
13
14
15 Kohavi R. 1996. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hHybrid.
16
17 Proceedings of the Second International Conference on Knowledge Discovery and Data
18 Mining. AAAI Press, pp 202-207.
19
20
21 Kolomietz E, Meyn MS, Pandita A, Squire JA. 2002. The role of *Alu* repeat clusters as mediators
22 of recurrent chromosomal aberrations in tumors. *Genes Chrom Cancer* 35:97-112.
23
24
25 Kondrashov AS, Rogozin IB. 2004. Context of deletions and insertions in human coding
26 sequences. *Hum Mutat* 23:177–185.
27
28
29 Kotsiantis S, Kanellopoulos D, Pintelas P. 2006. Handling imbalanced datasets: a review.
30
31 GESTS International Transactions on Computer Science and Engineering 30:25-36.
32
33
34 Krawczak M, Smith-Sorensen B, Schmidtke J, Kakkar VV, Cooper DN, Hovig E. 1995. Somatic
35 spectrum of cancer-associated single basepair substitutions in the *TP53* gene is
36 determined mainly by endogenous mechanisms of mutation and by selection. *Hum Mutat*
37 5:48-57.
38
39
40
41 Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-
42 line single-base-pair substitution in human genes. *Am J Hum Genet* 63:474-488.
43
44
45 Lagarde A, Rouleau E, Ferrari A, Noguchi T, Qiu J, Briaux A, Bourdon V, Rémy V, Gaildrat P,
46 Adélaïde J, Birnbaum D, Lidereau R, Sobol H, Olschwang S. 2010. Germline
47
48 *APC* mutation spectrum derived from 863 genomic variations identified through a
49
50
51
52
53
54
55
56
57
58
59
60

1
2 15-year medical genetics service to French patients with FAP. *J Med Genet*
3 47:721-722.

4
5
6 Lamlum H, Ilyas M, Rowan A, Clark S, Johnson V, Bell J, Frayling I, Efstathiou J, Pack K,
7
8 Payne S, Roylance R, Gorman P, Sheer D, Neale K, Phillips R, Talbot I, Bodmer W,
9
10 Tomlinson I. 1999. The type of somatic mutation at *APC* in familial adenomatous
11
12 polyposis is determined by the site of the germline mutation: a new facet to Knudson's
13
14 'two-hit' hypothesis. *Nat Med* 5:1071-1075.

15
16 Latchford A, Volikos E, Johnson V, Rogers P, Suraweera N, Tomlinson I, Phillips R, Silver A.
17
18 2007. *APC* mutations in FAP-associated desmoid tumours are non-random but not 'just
19
20 right'. *Hum Mol Genet* 16:78-82.

21
22 Leclercq S, Rivals E, Jarne P. 2010. DNA slippage occurs at microsatellite loci without minimal
23
24 threshold length in human: a comparative genomic approach. *Genome Biol Evol* 2:325-
25
26 335.

27
28 Leeflang EP, Tavare S, Marjoram P, Neal CO, Srinidhi J, MacFarlane H, MacDonald ME,
29
30 Gusella JF, de Young M, Wexler NS, Arnheim N. 1999. Analysis of germline mutation
31
32 spectra at the Huntington's disease locus supports a mitotic mutation mechanism. *Hum*
33
34 *Mol Genet* 8:173-183.

35
36 Lengauer C, Kinzler KW, Vogelstein B. 1998. Genetic instabilities in human cancers. *Nature*
37
38 396:643-649.

39
40 Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z,
41
42 Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson
43
44 JA, Ren B, Ecker JR. 2009. Human DNA methylomes at base resolution show
45
46 widespread epigenomic differences. *Nature* 462:315-322.

- 1
2 Lobo GP, Waite KA, Planchon SM, Romigh T, Nassif NT, Eng C. 2009. Germline and somatic
3 cancer-associated mutations in the ATP-binding motifs of PTEN influence its subcellular
4 localization and tumor suppressive function. *Hum. Mol. Genet.* 18:2851-2862.
5
6
7
8
9 Loeb LA, Harris CC. 2008. Advances in chemical carcinogenesis: a historical review and
10 prospective. *Cancer Res* 68: 6863-6872.
11
12 Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian
13 genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad*
14 *Sci USA* 95:9407-9412.
15
16
17
18 Makridakis NM, Caldas Ferraz LF, Reichardt JK. 2009. Genomic analysis of cancer tissue
19 reveals that somatic mutations commonly occur in a specific motif. *Hum Mutat* 30:39-48.
20
21
22 Marshall B, Isidro G, Carvalhas R, Boavida M. 1997. Germline versus somatic mutations of the
23 *APC* gene: evidence for mechanistic differences. *Hum Mutat* 9:286-288.
24
25
26
27 Martorell L, Monckton DG, Gamez J, Baiget M. 2000. Complex patterns of male germline
28 instability and somatic mosaicism in myotonic dystrophy type 1. *Eur J Hum Genet* 8:
29 423-430.
30
31
32
33 Miyoshi Y, Nagase H, Ando H, Horii A, Ichii S, Nakatsuru S, Aoki T, Miki Y, Mori T,
34 Nakamura Y. 1992. Somatic mutations of the *APC* gene in colorectal tumors: mutation
35 cluster region in the *APC* gene. *Hum Mol Genet* 1:229-33.
36
37
38
39 Mort M, Ivanov D, Cooper DN, Chuzhanova NA. 2008. A meta-analysis of nonsense mutations
40 causing human genetic disease. *Hum Mutat* 29:1037-1047.
41
42
43
44 Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and
45 nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418-426.
46
47
48
49 Oldenburg J, Rost S, El-Maarri O, Leuer M, Olek K, Muller CR, Schwaab R. 2000. *De novo*
50 factor VIII gene intron 22 inversion in a female carrier presents as a somatic mosaicism.
51 *Blood* 96: 2905-2906.
52
53
54
55
56
57
58
59
60

- 1
2
3 Olshen AB, Jain AN. 2002. Deriving quantitative conclusions from microarray expression data.
4
5 Bioinformatics 18:961-970.
- 6
7 Parmigiani G, Boca S, Lin J, Kinzler KW, Velculescu V, Vogelstein B. 2009. Design and
8
9 analysis issues in genome-wide somatic mutation studies of cancer. Genomics 93:17-21.
- 10
11 Pollard LM, Sharma R, Gomez M, Shah S, Delatycki MB, Pianese L, Monticelli A, Keats BJ,
12
13 Bidichandani SI. 2004. Replication-mediated instability of the GAA triplet repeat
14
15 mutation in Friedreich ataxia. Nucleic Acids Res 32:5962-5971.
- 16
17 Richter S, Vandezande K, Chen N, Zhang K, Sutherland J, Anderson J, Han L, Panton R, Branco
18
19 P, Gallie B. 2003. Sensitive and efficient detection of *RBI* gene mutations enhances care
20
21 for families with retinoblastoma. Am J Hum Genet 72:253-269.
- 22
23 Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. 2001. Somatic mutation hotspots
24
25 correlate with DNA polymerase eta error spectrum. Nat Immunol 2:530-536.
- 26
27 Rubin AF, Green P. 2009. Mutation patterns in cancer genomes. Proc Natl Acad Sci USA 106:
28
29 21766-21770.
- 30
31 Sharma R, Bhatti S, Gomez M, Clark RM, Murray C, Ashizawa T, Bidichandani SI. 2002. The
32
33 GAA triplet-repeat sequence in Friedreich ataxia shows a high level of somatic instability
34
35 *in vivo*, with a significant predilection for large contractions. Hum Mol Genet 11:2175-
36
37 2187.
- 38
39 Sherr CJ. 2004. Principles of tumor suppression. Cell 116:235-246.
- 40
41 Schmutte C, Jones PA. 1998. Involvement of DNA methylation in human carcinogenesis. Biol
42
43 Chem 379:377-388.
- 44
45 Shanks ME, May CA, Dubrova YE, Balaesque P, Rosser ZH, Adams SM, Jobling MA. 2008.
46
47 Complex germline and somatic mutation processes at a haploid human minisatellite
48
49 shown by single-molecule analysis. Mutat Res 648:46-53.
- 50
51 Simpson AJ. 2009. Sequence-based advances in the definition of cancer-associated gene
52
53 mutations. Curr Opin Oncol 21:47-52.
- 54
55
56
57
58
59
60

- 1
2
3 Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J,
4 Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson
5 D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman
6 KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. 2006. The consensus
7 coding sequences of human breast and colorectal cancers. *Science* 314:268-274.
8
9
10
11
12 Stead JD, Jeffreys AJ. 2000. Allele diversity and germline mutation at the insulin minisatellite.
13
14 *Hum Mol Genet* 9:713-723.
15
16
17 Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The
18
19 Human Gene Mutation Database: 2008 update. *Genome Med.* 1:13.
20
21 Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. *Nature* 458:719-724.
22
23 Talavera D, Taylor MS, Thornton JM. 2010. The (non)malignancy of cancer amino acidic
24
25 substitutions. *Proteins* 78:518-529.
26
27 Tappino B, Chuzhanova NA, Regis S, Dardis A, Corsolini F, Stroppiano M, Tonoli E, Beccari T,
28
29 Rosano C, Mucha J, Blanco M, Szlago M, Di Rocco M, Cooper DN, Filocamo M. 2009.
30
31 Molecular characterization of 22 novel UDP-N-acetylglucosamine-1-phosphate
32
33 transferase alpha- and beta-subunit (*GNPTAB*) gene mutations causing mucopolidosis
34
35 types IIalpha/beta and IIIalpha/beta in 46 patients. *Hum Mutat* 30:E956-973.
36
37 Tartaglia M, Martinelli S, Stella L, Bocchinfuso G, Flex E, Cordeddu V, Zampino G, van der
38
39 Burgt I, Palleschi A, Petrucci TC, Sorcini M, Schoch C, Foà R, Emanuel PD, Gelb BD.
40
41 2006. Diversity and functional consequences of germline and somatic *PTPN11* mutations
42
43 in human disease. *Am J Hum Genet* 78:279-290.
44
45 Tavigian SV, Greenblatt MS, Lesueur F, Byrnes GB; IARC Unclassified Genetic Variants
46
47 Working Group. 2008. *In silico* analysis of missense substitutions using sequence-
48
49 alignment based methods. *Hum Mutat* 29:1327-1336.
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 Tavgigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F,
4 Byrnes GB, Chuang SC, Forey N, Feuchtinger C, Gioia L, Hall J, Hashibe M, Herte B,
5 McKay-Chopin S, Thomas A, Vallée MP, Voegelé C, Webb PM, Whiteman DC;
6 Australian Cancer Study; Breast Cancer Family Registries (BCFR); Kathleen Cuningham
7 Foundation Consortium for Research into Familial Aspects of Breast Cancer (kConFab),
8 Sangrajrang S, Hopper JL, Southey MC, Andrulis IL, John EM, Chenevix-Trench G.
9
10 2009. *Am J Hum Genet* 85:427-446.
11
12
13
14
15
16 Thomas L, Kluwe L, Chuzhanova N, Mautner V, Upadhyaya M. 2010. Analysis of *NF1* somatic
17 mutations in cutaneous neurofibromas from patients with high tumor burden.
18
19 *Neurogenetics* 11:391-400.
20
21
22
23 Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of
24 missense mutations using bioinformatics methods. *Hum Mutat* 30:703-714.
25
26
27 Tornaletti S, Pfeifer GP. 1995. Complete and tissue-independent methylation of CpG sites in the
28 *p53* gene: implications for mutations in human cancers. *Oncogene* 10:1493-1499.
29
30
31 Upadhyaya M, Han S, Consoli C, Majounie E, Horan M, Thomas NS, Potts C, Griffiths S,
32 Ruggieri M, von Deimling A, Cooper DN. 2004. Characterization of the somatic
33 mutational spectrum of the neurofibromatosis type 1 (*NF1*) gene in neurofibromatosis
34 patients with benign and malignant tumors. *Hum Mutat* 23:134-136.
35
36
37
38 Upadhyaya M, Kluwe L, Spurlock G, Monem B, Majounie E, Mantripragada K, Ruggieri M,
39 Chuzhanova N, Evans DG, Ferner R, Thomas N, Guha A, Mautner V. 2008. Germline
40 and somatic *NF1* gene mutation spectrum in *NF1*-associated malignant peripheral nerve
41 sheath tumors (MPNSTs). *Hum Mutat* 29:74-82.
42
43
44
45
46
47 Vogelstein B, Kinzler KW. 2004. Cancer genes and the pathways they control. *Nat Med* 10:789-
48
49 799.
50
51
52
53
54
55
56
57
58
59
60

1
2 Walker DR, Bond JP, Tarone RE, Harris CC, Makalowski W, Boguski MS. Greenblatt MS.
3

4 1999. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation
5 with p53 protein structural and functional features. *Oncogene* 18:211-218.
6
7

8 [Walters, S. J. 2004. Sample size and power estimation for studies with health related quality of](#)
9 [life outcomes: a comparison of four methods using the SF-36. *Health Qual Life*](#)
10 [*Outcomes* 2, p. 26.](#)
11
12

13
14 Witten IH, Frank E. 2005. *Data mining: practical machine learning tools and techniques*, 2nd
15 ed. Morgan Kaufmann, San Francisco, pp. 365-483.
16
17

18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

1
2
3 **Figure Legends**

4 **Figure 1.** Diagrammatic representation of the number of various types of mutations analysed in
5 the present study.
6

7
8 **Figure 2.** Nucleotide substitution patterns of missense mutations in 17 tumour
9 suppressor genes.
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

Dobril Ivanov^{1,2}, Stephen E. Hamby³, Peter D. Stenson¹, Andrew D. Phillips¹,
Hildegard Kehrer-Sawatzki⁴, David N. Cooper¹ and Nadia Chuzhanova³

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

²MRC 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

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

⁴Institute 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

1
2
3
4
5
6 *Gene-wise characteristics of germline and somatic missense mutations with respect to mutation*
7
8 *type*
9

10 Taken together, the combined mutational spectra for all 17 tumour suppressor genes contained
11 twice as many somatic (61%) as germline (31%) mutations. For five genes (*APC*, *CDKN2A*,
12 *NF2*, *PTEN* and *TP53*), a predominance of somatic over germline mutations was noted, with the
13 *TP53* gene having the highest proportion of somatic mutations (92%). For the majority of genes,
14 however (namely *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *PTCH1*, *RBI*, *STK11*, *TSC1*, *TSC2*, *VHL*
15 and *WT1*), the analysed dataset included more germline than somatic mutations, with >97% of all
16 mutations in the *BRCA1*, *NF1*, *TSC2* and *WT1* genes being germline in origin.
17
18
19
20
21
22
23
24
25
26
27
28
29

30 *Gene-wise characteristics of missense mutations in relation to the disease- and non-disease-*
31 *associated substitution rates*
32
33

34 The median values were found to vary between 4 (*NF2*) and 8.9 (*STK11*) for somatic mutations,
35 4.1 (*TP53*) and 10.1 (*WT1*) for germline mutations, and 7.2 (*RBI*) and 11 (*PTEN*) for shared
36 mutations (values given only for genes with more than three mutations in the corresponding
37 category; see Supp. Table 3, indicating that many of the median values are quite low and hence
38 the corresponding mutations are unlikely to be neutral.
39
40
41
42
43
44
45

46 The highest and lowest disease-associated median values for the mutation rates were noted for
47 somatic mutations in the *STK11* gene (1.7; Supp. Table 3) and for germline mutations in the
48 *TP53* (0.42) gene (values given only for genes with more than three mutations in the
49 corresponding category).
50
51
52
53
54
55
56
57

58 *Gene-wise occurrence of missense mutations within repeats and runs of identical nucleotides*
59
60

1
2
3 On average, the coverage of the respective gene sequences by direct repeats of length ≥ 8 bp
4 was found to vary between 2.5% (*BRCA2*) and 17% (*PTEN*). The corresponding proportion of
5 the cDNA lengths for inverted repeats ≥ 8 bp was found to vary between 4.5% (*PTCH1*) and *RBI*
6 15.7% (*RBI*) while symmetric elements ≥ 8 bp were found to vary between 15.5% for *APC* and
7 44% for *PTEN* genes.
8
9

10
11 On average, mononucleotide runs ≥ 4 bp spanned 19.9% of the cDNA lengths, varying between
12 9.5% (*VHL*) and 29% (*TP53*).
13
14

15
16 No correlation was observed between the number of mutations located within repeats and the
17 fractional length of the cDNA covered by repeats, indicating that not every repeat sequence is
18 mutation-prone. However, a strong correlation between the fractional length of the cDNA
19 covered by repeats and cDNA length of genes ($r > 0.87$ and $p < 10^{-6}$) served to demonstrate that
20 repeat density per unit length was approximately the same for all tumour suppressor genes
21 studied.
22
23
24
25
26
27
28
29
30
31
32
33

34 *Truncating vs non-truncating mutations in the germline and soma*

35
36 Somatic mutational spectra from the *BRCA2*, *CDKN2A*, *STK11*, *TP53* and *TSC1* genes were
37 characterized by the predominance of non-truncating (i.e. missense) lesions over truncating
38 lesions (i.e. nonsense mutations, frameshift micro-deletions, micro-insertions and indels) when
39 nonsense mutations [reported in Mort et al. (2008)] and micro-indels (excluded from previous
40 analyses) were also considered (Supp. Table 8). A similar predominance of non-truncating over
41 truncating lesions was observed for the germline mutational spectra of the *CDKN2A*, *TP53*, *VHL*
42 and *WT1* genes. In general, the ratio of non-truncating to truncating lesions was found to be
43 significantly higher in the soma (0.85) than in the germline (0.30; p -value $< 2.20 \times 10^{-16}$). All other
44 mutational spectra were characterized by the predominance of truncating mutations.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

For Peer Review

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)

```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

```
      | 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)
```


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

													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)				

```

1
2
3
4
5 | | | | | | | Krawczak_value > 0.341
6 | | | | | | | | Krawczak_value <= 0.463: (0.38) (0.46) (0.04) (0.08) (0.04)
7 | | | | | | | | Krawczak_value > 0.463: (0.21) (0.31) (0.03) (0.41) (0.03)
8 | | | | | | | | Hess_value > 3.3: (0.20) (0.27) (0.01) (0.51) (0.01)
9 | | | | | | | | Evol > 0.065: (0.36) (0.5) (0.05) (0.05) (0.05)
10 | | | | | | | | Evol > 0.18: (0.10) (0.05) (0.05) (0.76) (0.05)
11 | | | | | | | | Hess_value > 4.35
12 | | | | | | | | Evol <= 0.045
13 | | | | | | | | Grantham_score <= 30.5
14 | | | | | | | | | Hess_value <= 5.55: (0.43) (0.18) (0.04) (0.32) (0.04)
15 | | | | | | | | | Hess_value > 5.55
16 | | | | | | | | | Grantham_score <= 26.5: (0.18) (0.44) (0.03) (0.32) (0.03)
17 | | | | | | | | | Grantham_score > 26.5: (0.11) (0.11) (0.05) (0.68) (0.05)
18 | | | | | | | | | Grantham_score > 30.5
19 | | | | | | | | | Grantham_score <= 118.5
20 | | | | | | | | | Grantham_score <= 95.5
21 | | | | | | | | | | Hess_value <= 10.6
22 | | | | | | | | | | Grantham_score <= 75.5
23 | | | | | | | | | | Grantham_score <= 69.5
24 | | | | | | | | | | | Hess_value <= 7.05
25 | | | | | | | | | | | Hess_value <= 4.65
26 | | | | | | | | | | | | Hess_value <= 4.55: (0.07) (0.23) (0.03) (0.13) (0.53)
27 | | | | | | | | | | | | Hess_value > 4.55: (0.30) (0.30) (0.30) (0.05) (0.05)
28 | | | | | | | | | | | | Hess_value > 4.65: (0.07) (0.21) (0.03) (0.31) (0.38)
29 | | | | | | | | | | | | Hess_value > 7.05: (0.23) (0.02) (0.02) (0.32) (0.41)
30 | | | | | | | | | | | | Grantham_score > 69.5: (0.10) (0.10) (0.33) (0.02) (0.45)
31 | | | | | | | | | | | | Grantham_score > 75.5
32 | | | | | | | | | | | | | Grantham_score <= 92.5: (0.13) (0.29) (0.04) (0.5) (0.04)
33 | | | | | | | | | | | | | Grantham_score > 92.5: (0.18) (0.32) (0.41) (0.05) (0.05)
34 | | | | | | | | | | | | | Hess_value > 10.6: (0.26) (0.23) (0.03) (0.46) (0.03)
35 | | | | | | | | | | | | | Grantham_score > 95.5
36 | | | | | | | | | | | | | | Hess_value <= 5.55
37 | | | | | | | | | | | | | | Hess_value <= 4.65: (0.27) (0.45) (0.09) (0.09) (0.09)
38 | | | | | | | | | | | | | | Hess_value > 4.65: (0.03) (0.06) (0.03) (0.2) (0.69)
39 | | | | | | | | | | | | | | Hess_value > 5.55
40 | | | | | | | | | | | | | | Grantham_score <= 102.5: (0.18) (0.56) (0.02) (0.13) (0.11)
41 | | | | | | | | | | | | | | Grantham_score > 102.5: (0.06) (0.2) (0.03) (0.37) (0.34)
42 | | | | | | | | | | | | | | Grantham_score > 118.5
43 | | | | | | | | | | | | | | Grantham_score <= 149.5: (0.08) (0.13) (0.18) (0.04) (0.57)
44 | | | | | | | | | | | | | | Grantham_score > 149.5
45
46
47

```

```

1
2
3
4
5 | | | | | | | | | Hess_value <= 10.45
6 | | | | | | | | | | Krawczak_value <= 0.428: (0.36) (0.09) (0.09) (0.36) (0.09)
7 | | | | | | | | | | Krawczak_value > 0.428: (0.07) (0.26) (0.56) (0.09) (0.02)
8 | | | | | | | | | Hess_value > 10.45: (0.04) (0.16) (0.24) (0.06) (0.50)
9 | | | | | | | | | Evol > 0.045: (0.33) (0.37) (0.04) (0.22) (0.04)
10 | | | | | | | | | CpG/CHG = 1
11 | | | | | | | | | Grantham_score <= 99.5
12 | | | | | | | | | | Hess_value <= 10.05
13 | | | | | | | | | | Grantham_score <= 86: (0.07) (0.14) (0.07) (0.21) (0.5)
14 | | | | | | | | | | Grantham_score > 86: (0.03) (0.03) (0.88) (0.03) (0.03)
15 | | | | | | | | | | Hess_value > 10.05
16 | | | | | | | | | | Evol <= 0.07
17 | | | | | | | | | | Krawczak_value <= 12.275
18 | | | | | | | | | | Krawczak_value <= 9.211
19 | | | | | | | | | | Krawczak_value <= 8.5135
20 | | | | | | | | | | | Krawczak_value <= 7.551: (0.45) (0.27) (0.09) (0.09) (0.09)
21 | | | | | | | | | | | Krawczak_value > 7.551: (0.03) (0.14) (0.03) (0.03) (0.76)
22 | | | | | | | | | | | Krawczak_value > 8.5135: (0.47) (0.35) (0.06) (0.06) (0.06)
23 | | | | | | | | | | | Krawczak_value > 9.211
24 | | | | | | | | | | | Hess_value <= 46.4: (0.26) (0.11) (0.05) (0.05) (0.53)
25 | | | | | | | | | | | Hess_value > 46.4: (0.02) (0.02) (0.22) (0.06) (0.68)
26 | | | | | | | | | | | Krawczak_value > 12.275: (0.08) (0.03) (0.72) (0.03) (0.14)
27 | | | | | | | | | | | Evol > 0.07: (0.07) (0.03) (0.03) (0.03) (0.83)
28 | | | | | | | | | | | Grantham_score > 99.5
29 | | | | | | | | | | | Krawczak_value <= 7.519: (0.03) (0.03) (0.03) (0.1) (0.82)
30 | | | | | | | | | | | Krawczak_value > 7.519
31 | | | | | | | | | | | Grantham_score <= 113: (0.02) (0.19) (0.02) (0.06) (0.70)
32 | | | | | | | | | | | Grantham_score > 113: (0.13) (0.57) (0.04) (0.22) (0.04)
33 | | | | | | | | | | | Evol > 0.205
34 | | | | | | | | | | | Hess_value <= 9.65
35 | | | | | | | | | | | Repeats = 0
36 | | | | | | | | | | | Hess_value <= 8.8
37 | | | | | | | | | | | Grantham_score <= 40.5
38 | | | | | | | | | | | Hess_value <= 2.65: (0.60) (0.07) (0.07) (0.20) (0.07)
39 | | | | | | | | | | | Hess_value > 2.65
40 | | | | | | | | | | | Krawczak_value <= 1.083
41 | | | | | | | | | | | Krawczak_value <= 0.269: (0.11) (0.39) (0.06) (0.39) (0.06)
42 | | | | | | | | | | | Krawczak_value > 0.269
43 | | | | | | | | | | | Krawczak_value <= 0.6155
44 | | | | | | | | | | | Hess_value <= 4: (0.68) (0.05) (0.05) (0.16) (0.05)
45
46
47

```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

							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)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

											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)

```

1
2
3
4
5 | | | | | Grantham_score > 91.5:      (0.33) (0.17) (0.03) (0.03) (0.43)
6 | | | | | Hess_value > 12.1
7 | | | | |   Evol <= 0.51
8 | | | | |     Repeats = 0
9 | | | | |       Grantham_score <= 44.5: (0.20) (0.07) (0.07) (0.60) (0.07)
10 | | | | |         Grantham_score > 44.5
11 | | | | |           Grantham_score <= 51: (0.09) (0.27) (0.09) (0.09) (0.45)
12 | | | | |             Grantham_score > 51: (0.24) (0.53) (0.03) (0.18) (0.03)
13 | | | | |               Repeats = 1:      (0.32) (0.45) (0.05) (0.14) (0.05)
14 | | | | |                 Evol > 0.51
15 | | | | |                   Hess_value <= 13.35: (0.27) (0.15) (0.04) (0.5) (0.04)
16 | | | | |                     Hess_value > 13.35: (0.28) (0.44) (0.04) (0.2) (0.04)
17 | | | | | Hess_value > 42.75
18 | | | | |   Repeats = 0
19 | | | | |     Evol <= 0.59
20 | | | | |       Evol <= 0.255:      (0.08) (0.12) (0.73) (0.04) (0.04)
21 | | | | |         Evol > 0.255
22 | | | | |           Evol <= 0.375: (0.18) (0.03) (0.03) (0.28) (0.49)
23 | | | | |             Evol > 0.375: (0.40) (0.13) (0.07) (0.33) (0.07)
24 | | | | |               Evol > 0.59
25 | | | | |                 Grantham_score <= 139: (0.02) (0.20) (0.75) (0.02) (0.02)
26 | | | | |                   Grantham_score > 139: (0.36) (0.43) (0.07) (0.07) (0.07)
27 | | | | | Repeats = 1
28 | | | | |   Hess_value <= 59.5
29 | | | | |     Hess_value <= 50.35: (0.40) (0.15) (0.05) (0.35) (0.05)
30 | | | | |       Hess_value > 50.35: (0.67) (0.13) (0.04) (0.13) (0.04)
31 | | | | |         Hess_value > 59.5: (0.19) (0.63) (0.06) (0.06) (0.06)

```

```

32 === Stratified cross-validation ===

```

```

33 === Summary ===

```

```

34 Correctly Classified Instances      2797          63.1377 %
35 Incorrectly Classified Instances   1633          36.8623 %
36 Kappa statistic                    0.5392
37 Mean absolute error                 0.1878
38 Root mean squared error             0.3177
39 Relative absolute error             58.6858 %

```

1
2
3
4
5 Root relative squared error 79.4156 %
6 Total Number of Instances 4430

7
8 === 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	

15
16 === Confusion Matrix ===

17
18 a b c d e <-- classified as
19 447 125 63 207 44 | a = 1
20 170 377 89 153 97 | b = 2
21 12 9 792 9 64 | c = 3
22 181 144 85 421 55 | d = 4
23 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	<i>Xenopus laevis</i>	U64442.1	AAB41671.1
	<i>Bos taurus</i>	XM_865627.1	XP_870720.1
	<i>Rattus norvegicus</i>	NM_012499.1	NP_036631.1
	<i>Mus musculus</i>	NM_007462.1	NP_031488.1
ATM	<i>Gallus gallus</i>	XM_417160.1	XP_417160.1
	<i>Xenopus laevis</i>	AY668954.1	AAT72929.1
	<i>Rattus norvegicus</i>	XM_236275.3	XP_236275.3
	<i>Sus scrofa</i>	AY587061	AAT01608.1
	<i>Canis familiaris</i>	XM_845871.1	XP_850964.1
	<i>Mus musculus</i>	NM_007499	NP_031525.1
BRCA1	<i>Gallus gallus</i>	NM_204169.1	NP_989500.1
	<i>Xenopus laevis</i>	AF416868.1	AAL13037.1
	<i>Bos taurus</i>	NM_178573.1	NP_848668.1
	<i>Rattus norvegicus</i>	NM_012514.1	NP_036646.1
	<i>Canis familiaris</i>	NM_001013416.1	NP_001013434.1
	<i>Mus musculus</i>	NM_009764.2	NP_033894.2
BRCA2	<i>Gallus gallus</i>	NM_204276.1	NP_989607.1
	<i>Danio rerio</i>	XM_690042.1	XP_695134.1
	<i>Bos taurus</i>	XM_583622.2	XP_583622.2
	<i>Rattus norvegicus</i>	NM_031542.1	NP_113730.1
	<i>Canis familiaris</i>	NM_001006653.4	NP_001006654.2
	<i>Mus musculus</i>	NM_009765.1	NP_033895.1
CDH1	<i>Xenopus laevis</i>	BC068940.1	AAH68940.1
	<i>Danio rerio</i>	NM_131820.1	NP_571895.1
	<i>Bos taurus</i>	NM_001002763.1	NP_001002763.1
	<i>Rattus norvegicus</i>	NM_031334.1	NP_112624.1
	<i>Canis familiaris</i>	XM_536807.2	XP_536807.2
	<i>Mus musculus</i>	NM_009864.1	NP_033994.1
CDKN2A	<i>Gallus gallus</i>	NM_204433.1	NP_989764.1
	<i>Takifugu rubripes</i>	AJ250231.1	CAC12808.1
	<i>Bos taurus</i>	XM_868375.1	XP_873468.1
	<i>Rattus norvegicus</i>	NM_031550.1	NP_113738.1
	<i>Canis familiaris</i>	XM_538685.2	XP_538685.2
	<i>Mus musculus</i>	AF044336.1	AAC08963.1
NF1	<i>Gallus gallus</i>	XM_415914.1	XP_415914.1
	<i>Takifugu rubripes</i>	AF064564.2	AAD15839.1
	<i>Rattus norvegicus</i>	NM_012609.1	NP_036741.1
	<i>Canis familiaris</i>	XM_537738.2	XP_537738.2
	<i>Mus musculus</i>	NM_010897.1	NP_035027.1
NF2	<i>Gallus gallus</i>	NM_204497.2	NP_989828.2
	<i>Danio rerio</i>	NM_212951.1	NP_998116.1
	<i>Bos taurus</i>	XM_611643.2	XP_611643.2
	<i>Rattus norvegicus</i>	XM_341248.2	XP_341249.2
	<i>Canis familiaris</i>	XM_534729.2	XP_534729.2
	<i>Mus musculus</i>	NM_010898.2	NP_035028.2
PTCH1	<i>Xenopus laevis</i>	AF302765.1	AAK15463.1
	<i>Gallus gallus</i>	NM_204960.1	NP_990291.1
	<i>Danio rerio</i>	NM_130988.1	NP_571063.1
	<i>Meriones unguiculatus</i>	AB188226.1	BAE78534.1
	<i>Rattus norvegicus</i>	NM_053566.1	NP_446018.1
	<i>Mus musculus</i>	NM_008957.1	NP_032983.1
PTEN	<i>Xenopus laevis</i>	AF144732.1	AAD46165.1

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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			<i>STK11</i>	1.66					<i>STK11</i>	25		
			<i>PTCH1</i>	1.06								
		8.4	<i>CDKN2A</i>	1.01	<i>CDKN2A</i>	0.38			<i>CDKN2A</i>	20	<i>CDKN2A</i>	5.2
		7.9	<i>APC</i>	0.83								
		5.6	<i>PTEN</i>	0.53								
		4.6	<i>TP53</i>	0.5	<i>TP53</i>	0.17			<i>RB1</i>	18	<i>TP53</i>	2.8
					<i>VHL</i>	0.14			<i>BRCA2</i>	16		
								<i>PTCH1</i>	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												
		7.2			<i>TSC2</i>	0			<i>BRCA1</i>	7	<i>BRCA1</i>	3.6
		7.3					<i>NF1</i>	98				
		7.6							<i>NF1</i>	7		
		7.9	<i>ATM</i>	0.79	<i>ATM</i>	0	<i>ATM</i>	98	<i>ATM</i>	15	<i>ATM</i>	3.8
		7.9	<i>BRCA1</i>	0.81	<i>VHL</i>	0	<i>VHL</i>	99	<i>BRCA1</i>	16		
		8.7	<i>BRCA2</i>	0.81					<i>NF1</i>	18		
			<i>PTEN</i>	0.92							<i>TSC2</i>	8.1
			<i>RB1</i>	0.99							<i>WT1</i>	10.8
		<i>NF1</i>	1.03									
		<i>TSC2</i>	1.03									

	<i>WT1</i>	10.1	<i>WT1</i>	1.22	<i>WT1</i>	0			<i>TSC2</i>	21		
			<i>CDH1</i>	1.27	<i>BRCA1</i>	0.14			<i>APC</i>	24		
					<i>CDKN2A</i>	0.29			<i>CDH1</i>	26		

For Peer Review

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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

Supplementary 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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

	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	