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

A thesis submitted for the Degree of Doctor of Philosophy at Cardiff University

by

Dobril Kirilov Ivanov

2009

Department of Medical Genetics
School of Medicine
Cardiff University
Heath Park
Cardiff CF14 4XN

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration and Statements

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed...(candidate)
Date.........24/02/10...

Statement 1

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references. A bibliography is appended.

Signed...(candidate)
Date.........24/02/10...

Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and the summary to be made available to outside organisations.

Signed...(candidate)
Date.........24/02/10...

Acknowledgements

I would like to express my gratitude to my PhD supervisors, Prof. David N Cooper and Dr. Nadia Chuzhanova, for their devotion to this PhD project, but also for their hard work on reading the individual chapters and for their invaluable guidance throughout.
I am grateful to Dr. Andrew Philips, Dr. Edward Ball, Mr. Matthew Mort and Dr. Nick Thomas (*HGMD*) for their initial help with getting the germline mutations together.

A huge thank you goes to Valentina for her invaluable expertise and advice on the multiple hypotheses testing. I am also grateful to George and Denise for their comments on few of the chapters.

I would also like to thank Ivan for his moral support and having the nerves to listen to my experiences as a PhD student.
An enormous thank you also goes to my parents and immediate family and friends for their support.

Last, but not least, an enormous thank you goes to Didi for having to endure being my wife during my PhD, but also for her precious and critical scientific advice throughout this PhD project.

Publications

Mort M, Ivanov D, Cooper DN, Chuzhanova NA (2008). A meta-analysis of nonsense mutations causing human genetic disease. *Hum Mutat* 29 (8), pp. 1037-1047

Abbreviations

5mC	5-Methyl cytosine
A	Alanine
A-GVGD	Align-Grantham Variation Grantham Deviation
ATP	Adenosine tri-phosphatase
BIC	Breast Cancer Information Core Database
BLAST	Basic local sequence alignment search tool
BLOSUM	Blocks of amino acid substitution matrix
bp	Base-pair/s
C	Cysteine
CCR5	Cellular chemokine receptor 5
CDK	Cyclic dependant kinase
cDNA	Complementary deoxyribonucleic acid
CIN	Chromosomal instability
COSMIC	The Catalogue of Somatic Mutations In Cancer
D	Aspartic acid
df	Degree of freedom
DNA	Deoxyribonucleic acid
DSB	Double-strand break
E	Glutamic acid
e.g.	Exempli gratia
EEJ	Exon-exon junction
EJC	Exon junction complex
ESTR	Expanded simple tandem repeats
et al.	Et alia
F	Phenylalanine
FAP	Familial Adenomatosis Polyposis
FASTA	FAST-All
G	Glycine
GB	Giga byte
GHz	Giga hertz
GTP	Guanosine tri-phosphatase
H	Histidine
HGMD	The Human Gene Mutation Database
HIV	Human immunodeficiency virus
HPLC	High pressure liquid chromatography
HR	Homologous recombination
I	Isoleucine
i.e.	id est
IARC	International Agency for Research on Cancer
Indel	Insertion deletion
K	Lysine
L	Leucine
LINE	Long Interspersed Nuclear Element
LOH	Loss of heterozygosity
m	Mean
M	Methionine
Max	Maximum

MCR	Mutation cluster region (*APC* gene)
MIN	Microsatellite instability
mRNA	Messenger ribonucleic acid
N	Asparagine
NCBI	National Center for Biotechnology Information
NHEJ	Non-homologous end joining
NMD	Nonsense-mediated mRNA decay
nt	Nucleotide
ORF	Open reading frame
OS	Operating system
P	Proline
PAM	Point accepted mutation or Percent accepted mutation
PC	Personal computer
Perl	Practical extraction and report language
PolyPhen	Polymorphism phenotyping
PTC	Premature termination codon
Q	Glutamine
R	Arginine
RAM	Random access memory
RINS	Runs of identical nucleotides
s	Standard deviation
S	Serine
SIFT	Sorting intolerant from tolerant
SINE	Short interspersed nuclear element
SSB	Single-strand break
T	Threonine
TSG	Tumour suppressor gene
UTR	Untranslated region
UV	Ultraviolet
V	Valine
viz.	Videlicet
vs.	Versus
W	Tryptophan
Y	Tyrosine
χ^2	Chi-square

Internet resources used

Resource	Internet link
The Human Gene Mutation Database (HGMD)	http://www.hgmd.org
National Center for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/
National Center for Biotechnology Information (NCBI), PubMed	http://www.ncbi.nlm.nih.gov/sites/entrez
National Center for Biotechnology Information (NCBI), Gene Database	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
National Center for Biotechnology Information (NCBI), GenBank	http://www.ncbi.nlm.nih.gov/sites/entrez?db=Nucleotide
The AAindex Database	http://www.genome.jp/dbget-bin/www_bfind?aaindex
The *p53* database	http://p53.free.fr/
Catalogue of Somatic Mutations in Cancer (COSMIC)	http://www.sanger.ac.uk/genetics/CGP/cosmic/
Breast Cancer Information Core (BIC)	http://research.nhgri.nih.gov/bic/
RB1 Gene Mutation Database	http://rb1-lsdb.d-lohmann.de
International *NF2* Mutation Database	http://neurosurgery.mgh.harvard.edu/NFclinic/NFresearch.htm
VHL Mutation Database	http://www.umd.be/VHL/
CDKN2A Database	https://biodesktop.uvm.edu/perl/p16
International Agency for Research on Cancer (IARC) *TP53* Mutation Database	http://www-p53.iarc.fr/
Spidey program	http://www.ncbi.nlm.nih.gov/spidey/spideyexec.html
R statistical language	http://cran.r-project.org/
Comprehensive Perl Archive Network (CPAN)	http://cpan.org
Comprehensive Perl Archive Network (CPAN), CPAN Modules	http://search.cpan.org
Comprehensive Perl Archive Network (CPAN), Statistics::Distributions Package	http://search.cpan.org/~mikek/Statistics-Distributions-1.02/Distributions.pm
The Epidermal Growth Factor Receptor Mutation Database (EGFR)	http://www.cityofhope.org/mdl/egfr/Pages/default.aspx
Basic Local sequence Alignment SearchTool (BLAST)	http://blast.ncbi.nlm.nih.gov/Blast.cgi

Table of contents

List of Tables

List of Figures

List of equations

Summary

The known somatic (N>4000) and germline (N>4000) cancer-associated mutational spectra (*viz.* missense and nonsense mutations; micro-deletions, micro-insertions and micro-indels ≤20bp) of 17 human tumour suppressor genes (*viz. APC, ATM, BRCA1, BRCA2, CDH1, CDKN2A, NF1, NF2, PTCH, PTEN, RB1, STK11, TP53, TSC1, TSC2, VHL* and *WT1*) were compared in order to identify similarities and differences. Analysed parameters included the recurrence status of mutations, CpG mutability; Grantham difference; evolutionary conservation of affected codons; role of nonsense-mediated mRNA decay and co-location with repetitive sequence elements.

Only a small proportion of the mutations (~5%) were found to be shared between the germline and soma, although the proportions varied between different types of mutation (from 11% for missense mutations to ~1% for micro-indels). Shared mutations are unlikely to be coincidental and are probably indicative of underlying shared (and endogenous) mutational mechanisms. Shared missense mutations were found to be more likely to be drivers of tumorigenesis than either exclusively somatic or exclusively germline missense mutations.

Shared micro-lesions combined for all genes occurred disproportionately within repetitive elements by comparison with both somatic or germline micro-lesions, consistent with an endogenous mutational mechanism. For some genes (e.g. TP53), shared CpG-dinucleotide mutations evidenced the action of an endogenous mutational mechanism (viz. methylation-mediated deamination of 5-methylcytosine) in both the soma and the germline.

Differences between mutational spectra were also noted. Germline missense mutations were found to be more likely to bear relatively more drastic functional consequences by comparison with somatic missense mutations, but also more likely to be truncating mutations. Germline micro-lesions (combined for all genes) were also found to be more likely to be co-located with repetitive elements than somatic micro-lesions. This could be due to the germline being relatively more protected from the action of exogenous mutagens by comparison to the soma.

This study of 17 human tumour suppressor genes has therefore provided a first glimpse of the similarities and differences between germline and somatic mutational spectra.

1. General introduction

1.1. Cancer genes

'Cancer is, in essence, a genetic disease' (Vogelstein and Kinzler 2004), in which DNA sequence and epigenetic changes are considered causative of neoplasms. Generally, there are three types of 'cancer' genes that when mutated can promote or substantially contribute to tumorigenesis: oncogenes, DNA repair/genome stability genes and tumour-suppressor genes (Vogelstein and Kinzler 2004).

Oncogenes can promote tumour development via mutations that render the genes continually active. For example, the *BRAF* oncogene encodes a serine/threonine protein kinase that when activated, e.g. via phosphorylation of amino-acids at positions 598 and 601, phosphorylates downstream targets, such as the extracellular signal-regulated kinase (Robinson and Cobb 1997). Thus, missense mutations, mostly affecting the kinase domain (Wan et al. 2004) render the *BRAF* oncogene constitutively active in the absence of wild-type activating signals. Thus, mutations in one allele of the oncogenes are generally sufficient to promote cell proliferation and confer a growth advantage.

DNA repair/genome stability genes on the other hand are responsible for the repair of subtle DNA sequence changes. These genes are involved in DNA repair mechanisms, such as mismatch repair, nucleotide-excision repair and base-excision repair (Vogelstein and Kinzler 2004). Mutations in the coding sequences of these genes result in cells being deficient in certain repair mechanisms. For example, skin cells from Xeroderma pigmentosum patients exhibit increased mutation frequency, due to defective nucleotide-excision repair (Friedberg 2003; Masutani et al. 2000). Mutations in the DNA repair genes give rise to an increased mutation frequency in all genes and therefore indirectly promote or contribute to tumour development. In addition, bi-allelic inactivation of both alleles of the DNA repair genes is required for a 'physiological effect' (Vogelstein and Kinzler 2004).

Tumour suppressor genes (TSGs) are responsible for a variety of cell functions that, as suggested by their name, suppress tumour development. These include pivotal functions, such as regulation of cell proliferation, maintenance and surveillance of the human genome, cellular response to DNA damage and ubiquitination of proteins (Sherr 2004). TSGs are defined as 'genes that sustain loss-of-function mutations in the development of cancer' (Haber and Harlow 1997). Following Knudson's 'two-hit' hypothesis (Knudson 1971, 1978), generally a bi-allelic inactivation is required for a phenotypic effect. Thus, one inherited (germline) hit and subsequent somatic inactivation of the hitherto unaffected allele, or alternatively somatic bi-allelic inactivation is required for tumour initiation and/or development. However, a number of studies have suggested that some tumour suppressor genes may not conform to Knudson's two-hit hypothesis. Reports, suggest that gene dosage (Fodde and Smits 2002), in heterozygous carriers of mutations, could confer increased cancer susceptibility (e.g. *BRCA2* gene, Howlett et al. 2002; *TP53* gene Venkatachalam et al. 1998). Nevertheless, the commonly accepted model of TSGs is that, in contrast to oncogenes, bi-allelic inactivation of tumour suppressor genes is sufficient to promote tumour development and/or progression.

1.2. Importance of the mutational spectrum in the development of cancer

Cancer, being a disorder of the soma, arises as a consequence of somatic DNA sequence and/or epigenetic changes (Stratton et al. 2009). Cancer cells also frequently exhibit genetic and genomic instability and abnormalities (e.g. chromosome (CIN) and microsatellite instability (MSI); Charames and Bapat 2003). As a consequence of genetic instability, somatic cells can acquire anywhere between <1000 to >100,000 DNA sequence changes during tumorigenesis (Stratton et al. 2009). However, it is generally considered that between 5 and 7 and up to 20 mutations are responsible for tumour initiation and development (Beerenwinkel et al. 2007). Thus, it can be seen that not all sequence changes contribute equally towards cancer development. Some of the changes do not confer a selective (growth) advantage upon the cells harbouring them and these changes have been termed ‘passenger’ mutations, whereas other mutations are more likely to be ‘drivers’ of tumorigenesis (Greenman et al. 2007; Thomas et al. 2007). Therefore, a great deal of effort has been put forward, to distinguish mutations that drive tumorigenesis from those that are mere ‘passengers’.

Mutations can be placed into numerous categories, based on the number of affected nucleotides, the functional consequences of the mutant product, the transmission ability to descendants, the recurrence status, the location in the genome, etc. With respect to inheritance, germline mutations can be transmitted to descendents, whereas somatic mutations are confined exclusively to the soma. In addition, the genome of a cancer cell exhibits a diverse range of mutations, from single base-pair substitutions to gross genomic rearrangements (Stratton et al. 2009), in terms of the size of mutated or affected DNA sequence. Germline intra-genic single base-pair substitutions along with micro-deletions and micro-insertions of ≤20bp, are by far the most frequent types of mutations logged in *HGMD* (*The Human Gene Mutation Database*; http://www.hgmd.org; Stenson et al. 2003). Furthermore, small intra-genic mutations (missense and nonsense mutations and micro-lesions) represent the bulk of mutations in the COSMIC database (*The Catalogue of Somatic Mutations in Cancer*; http://www.sanger.ac.uk/genetics/CGP/cosmic/; Bamford et al. 2004; Forbes et al. 2006; Forbes et al. 2008).

Missense mutations are defined as intra-genic single base-pair substitutions (e.g. CA**C**->CA**A**; histidine->glutamine) that lead to a non-synonymous change of the wild-type amino-acid encoded by a specific codon. Based on their functional consequences (in terms of an increase in susceptibility to developing or directly causing disease), missense mutations could be categorised as being neutral, deleterious, and beneficial or of unknown clinical importance (Chan et al. 2007; Strachan and Read 2004). Following this classification, the functional importance of some of the missense mutations during the development of cancer is somewhat unclear. Such missense mutations are of unknown functional and clinical importance. Because, functional assays are difficult to perform or do not exist for every single missense mutation that turns up during clinical diagnostic procedures, numerous *in silico* algorithms have been proposed to try to assess the functional significance of missense mutations (Chan et al. 2007; Miller and Kumar 2001; Vitkup et al. 2003). Most of these algorithms rely on datasets of mutations with already known functional consequences. However, by definition, these training datasets do not comprise all observed missense mutations. Thus, chance variation is likely to influence the outcome of these *in silico* algorithms. Indeed, it has been estimated that the overall predictive accuracy of these algorithms ranges from ~70% to ~90% (Chan et al. 2007). Despite difficulties in quantifying the relative functional/clinical importance of some missense mutations, studies have shown that some missense variants are very likely to significantly contribute towards tumour development. For example, a number of mutations in the *BRCA1* gene have been shown to display a significantly negative effect (as compared to wild-type product) on the function of the protein (reviewed in Carvalho et al. 2009). More detailed description and analysis of the functional consequences of missense mutations are presented in Chapter 4.

Nonsense mutations are defined as single base-pair substitutions that lead to the introduction of nonsense or premature termination (stop) codons (e.g. CA**G**->TA**G**; glutamine->stop codon) within the coding regions of genes. Generally, the functional consequence of the majority of nonsense mutations is abrupt translational termination.

Nonsense mutations are important because in the majority of cases they are 'equivalent to nonsense sequences' (Kuzmiak and Maquat 2006), and if translated (assuming stable and functional mRNA to support translation), it is very likely that such mutant products could have either very limited or no function at all. Such an assertion is supported by an estimation that inherited nonsense mutations are twice as likely to come to clinical attention as compared to the most extreme missense mutations (extreme in terms of the chemical difference between the substituted wild-type and substituting amino acid residues; Krawczak et al. 1998). Thus, nonsense mutations are frequently selected for their likely 'loss-of-function' effect during tumour development or clonal expansion (e.g. *APC* gene in colorectal cancer; Beroud and Soussi 1996; Fearnhead et al. 2001). A more detailed description and analysis of some of the functional consequences (i.e. the role of the nonsense-mediated mRNA decay) of nonsense mutations are presented in Chapter 5.

Micro-lesions, *viz*. micro-deletions, micro-insertions and micro-indels in this PhD project were defined as intra-genic (although some micro-lesions extended into the intronic parts of the genes) deleted and/or inserted nucleotides (nt) of length ≤20nt. When the length (in base-pairs or nucleotides) of the affected DNA coding sequence is not divisible by three, a frameshift occurs and premature termination of translation is to be expected, due to the triplet nature of the genetic code (Crick et al. 1961; Yanofsky 2007). Such micro-lesions are very likely to result in similar functional consequences as the abovementioned nonsense mutations. On the other hand, micro-lesions with length (in base-pairs) of affected bases divisible by three (i.e. in-frame), would be expected to have relatively less severe functional consequences, as only a few amino-acids would be lost. Indeed, a comprehensive meta-analysis of micro-insertions and micro-deletions in inherited human genetic disease (Ball et al. 2005) has revealed that such in-frame micro-lesions (i.e. length of affected nucleotides-3bp and 6bp) exhibit a markedly decreased frequency. Therefore, these micro-lesions are less likely to come to clinical attention, most likely because of less severe functional consequences. A more detailed description and analysis of micro-lesions are presented in Chapter 6.

The functional importance of single base-pair substitutions is exemplified by the mutational spectrum in the *TP53* gene. The *TP53* gene has been referred to as 'the guardian of human genome' (Lane 1992). The latter assertion is supported by the fact that mutations in the *TP53* gene are observed in >50% of all human cancers (Soussi and Beroud 2001; Toledo and Wahl 2006). Furthermore, the majority of mutations found in *TP53* are indeed single base-pair substitutions, and in particular missense mutations (>80%; the *p53* database, http://p53.free.fr/ Soussi and Beroud 2001). Some missense mutations, part of the mutational spectrum in the *TP53* gene, are likely to have additional functional consequences, such as a 'loss-of-function' effect. Others are selected for their likely 'gain-of-function' effect during clonal expansion (Blagosklonny 2000; Glazko et al. 2006; Glazko et al. 2004; Koonin et al. 2005).

1.3. Somatic and germline mutations

Cancer predisposition genes can exhibit either somatic or germline mutations (Futreal et al. 2004; Kinzler KW 2002; Vogelstein and Kinzler 2004). A major distinction to be made between somatic and germline mutations is that the former occur during meiosis, whereas the latter are generally meiotic in nature. Both germline and somatic cells divide mitotically, whereas meiotic division is exclusively confined to the germline cells. However, a hallmark of meiotic division, recombination, is not exclusively confined to germline cells. Somatic mitotic recombination has also been reported (LaFave and Sekelsky 2009), albeit a rare event per cell division (Dong and Fasullo 2003).

Despite the relative scarcity of reports on the comparative analysis of germline and somatic mutations, some similarities and differences have been observed. Out of the ~22,000 protein-coding genes in the human genome, ~350 have been found to contribute significantly to oncogenesis (Futreal et al. 2004; Stratton et al. 2009). The majority of mutations found in these cancer genes (~90%) are somatic mutations as compared to ~10% germline mutations (Futreal et al. 2004; Stratton et al. 2009). These observations clearly demonstrate that human cancer is a disorder of the soma. Despite their relatively lower frequency of occurrence, germline mutations have been shown to play an important role in the process of tumorigenesis by conferring cancer susceptibility. For example, the lifetime risk (by the age of 70) of breast cancer carriers of inherited (germline) mutations in the *BRCA1* gene is estimated to be ~50% and for germline carriers in the *BRCA2* gene ~35% (Antoniou et al. 2002; Ford et al. 1998). Even although these estimations suggest that several common low penetrance genes other than *BRCA1* and *BRCA2* may account for the residual risk (Antoniou et al. 2002), germline mutations in these two genes are responsible for the majority of familial cases of breast and ovarian cancers (Ramus et al. 2007). Despite the relatively high risk associated with germline *BRCA1* and *BRCA2* mutation carriers, the underlying mechanisms responsible for the increased risk are still unclear. To account for the increased risk, research has suggested a potential disruption of hormone-signalling pathways (Mote et al. 2004) and deficiency in DNA repair through homologous recombination (Barwell et al. 2007; Stefansson et al. 2009). In addition, 'radiosensitivity' (repair deficiency following ionizing radiation) has been shown in heterozygous carriers in *BRCA1* and *BRCA2* genes (Buchholz et al. 2002). Therefore, potential mechanisms that could account for the increased risk include a gene-dosage effect and haploinsufficiency, whereby one functional allele of the genes is insufficient to suppress tumour development (Buchholz et al. 2002; Meric-Bernstam 2007).

Germline mutations have also been shown to play an important role in shaping the somatic mutational spectrum. Thus, germline and somatic mutations in some genes are not just two separate mutational events, but intricate germline-soma interplay is evident. The position of germline mutations in the *APC* gene has been shown to have the potential to influence the position and type of the second (somatic) hit in familial adenomatous polyposis coli (Albuquerque et al. 2002; Fearnhead et al. 2001; Lamlum et al. 1999; Latchford et al. 2007). Similarly, inherited variation (i.e. haplotype block) in the *JAK2* gene has been proposed to either confer a somatic hypermutability at the *JAK2* locus, or a stronger selective advantage over the somatic cells in myeloproliferative neoplasms (Campbell 2009; Jones et al. 2009; Kilpivaara et al. 2009; Olcaydu et al. 2009). On the other hand, germline mutations in the *CHEK2* gene have been associated with a decreased risk of lung and upper aero-digestive cancers (Cybulski et al. 2008), although the mechanism to account for the decreased risk remains elusive.

Some remarkable differences and similarities have been shown between the germline and the soma. DNA mismatch-repair-deficient *C. elegans* mutants have been found to display similar germline and somatic repeat instability (Tijsterman et al. 2002). The frequency of somatic micro-indels in mice have been shown to be similar to the human germline micro-indels (i.e. *TP53* micro-indels; Gonzalez et al. 2007). A similar age-related shift has been noted in the frequencies of human somatic and germline mutations (Evans et al. 2005), although the frequency of somatic micro-indels in mice has not been shown to display any age-related difference (Gonzalez et al. 2007).

However, mutation rates in the soma and the germline may also display some differences. Thus, the germline mutation rates are suggested to be lower as compared to the soma (Azad and Woodruff 2006; Drake et al. 1998; Neel 1983; Walter et al. 1998). The highly variable minisatellites show extreme germline instability, whereas somatic mutants have been shown to be rare (Buard et al. 2000; Stead and Jeffreys 2000).

Research has shown that the mutational spectrum of tumour cells is influenced by the action of both endogenous mutational mechanisms and exogenous mutagens. The positional occurrence of micro-lesions (*viz.* micro-deletions, micro-insertions and micro-indels) is likely to be influenced mainly by endogenous mutational mechanisms, such as 'slipped-mispairing' and 'strand-switching' mechanisms (Cooper and Krawczak 1993; Efstratiadis et al. 1980; Krawczak and Cooper 1991; Ripley 1982). These mutational mechanisms have been shown to be often promoted by numerous repetitive elements, such as direct, inverted and mirror repeats; runs of mononucleotides; various non-B DNA secondary structures (e.g. C/G-quartets); sequence motifs, etc. (Bacolla et al. 2004; Bacolla and Wells 2009; Chuzhanova et al. 2003; Cooper and Krawczak 1993; Efstratiadis et al. 1980; Krawczak and Cooper 1991; Ripley 1982; Wells 2007), although Cheung et al. (2007) have noted no overrepresentation of repetitive elements around the breakpoints of deletions and insertions in the *BRCA1* and *BRCA2* genes. In addition, environmental mutagens (e.g. mitomycin C, cyclophosphomide and radiation) could induce slippage in repetitive elements (i.e. tetranucleotide repeats; Lyons-Darden and Topal 1999; Niwa 2006; Pineiro et al. 2003). Then again, the mutagenic effect of low doses of radiation in mice has been shown to be very similar in both the soma and the germline (Vilenchik and Knudson 2000). In addition, environmental agents can directly cause DNA damage, such as double and single DNA-strand breaks, abasic sites, oxidised bases, etc. (Breen and Murphy 1995; Sankaranarayanan and Wassom 2005).

Spontaneous deamination of 5-methyl cytosine in the context of CpG-dinucleotides resulting in C->T (on the coding DNA strand) and G->A (on the non-coding strand) transitions, is largely responsible for the increased mutation rate (estimated transition rates - 5 times the base mutation rate; Krawczak et al. 1998) at CpG-dinucleotides (Pfeifer 2006). Consequently, differences or similarities in the frequency and positional occurrence of CpG-located mutations are dependent of the methylation status of CpG-dinucleotides. Furthermore, epigenetic silencing of gene expression, via hypermethylation of promoter regions, has been shown in a number of human cancers (Nagarajan and Costello 2009; Schulz and Hoffmann 2009; Tost 2009) and CpG-dinucleotides have been shown as a mutational hotspot for a number of tumour suppressor genes (e.g. *RB1*, *APC*, *BRCA1*, *BRCA2*, *TP53*; Cheung et al. 2007; Farrell and Clayton 2003; Radpour et al. 2009; Soussi and Beroud 2003). However, mutations at CpG-dinucleotides (*viz.* C->T and G->A) are also thought not only to result from endogenous mutagenesis (i.e. spontaneous deamination of 5mC) but also to the action of carcinogens (e.g. benzo[α]pyrene and cyclic aromatic hydrocarbon) at least in the *TP53* gene (Pfeifer and Besaratinia 2009). For a more detailed description and analysis of mutations within CpG-dinucleotides, see Chapter 4 (Missense mutations) and Chapter 5 (Nonsense mutations).

Therefore, it is evident, from past research, that somatic and germline mutations arise from the action of endogenous mechanisms and/or the influence of exogenous mutagens. However, the relative contribution of exogenous mutagens and exogenous mutational mechanisms is often difficult to quantify.

1.4. How could the comparative analysis of somatic and germline mutational spectrum help us to better understand tumorigenesis?

As pointed out above, somatic and germline mutations, amongst other genetic and epigenetic changes, contribute or play an important role in tumour development and/or initiation. Understanding the mutational mechanisms that predispose or directly contribute to the process of tumorigenesis is pivotal in trying to assess the clinical significance of DNA sequence changes, such that a better understanding of these mutational mechanisms is likely to lead to a better risk assessment, cancer treatment and prevention therapies. A key component is the relative contribution of endogenous mutational mechanisms and exogenous or environmental mutagens (e.g. carcinogens).

It is surprising that relatively few studies have attempted a formal comparison of germline and somatic mutations. Tumour suppressor genes, being subject to bi-allelic inactivation, could potentially provide an appropriate model system to study not only the relative contribution of somatic and germline mutations, but also the relative contribution of endogenous mutational mechanisms and environmental mutagens in both the soma and the germline, in the process of tumorigenesis. Furthermore, a potential elucidation of the relative contribution of exogenous mutagens and endogenous mutational mechanisms is likely to help the understanding of mutagenesis in other types of genetic disorders (Elespuru and Sankaranarayanan 2007).

1.5. Objectives of this PhD project

The current PhD project is a formal attempt to try to shed some light upon the mutational mechanisms that operate to influence the known mutational spectra in both the soma and the germline in 17 human tumour suppressor genes. Several key objectives were defined at the beginning of this PhD project (the end of 2005). These objectives comprised several general key questions:

How do germline and somatic mutational spectra for each of the studied human tumour suppressor gene compare with respect to the relative proportions of each type of mutation (i.e. missense and nonsense mutations, micro-deletions, micro-insertions and micro-indels)?

What proportion of the observed mutations (*viz.* missense and nonsense mutations and micro-lesions) is found in both the soma and the germline (i.e. shared)?

Are shared mutations merely coincidental and what is their relative functional importance with respect both to exclusively somatic and exclusively germline mutations?

Do specific DNA sequence features account for both single base-pair substitutions (i.e. missense and nonsense mutations within CpG-dinucleotides) and micro-lesions (e.g. co-localisation of repetitive elements and micro-deletions, micro-insertions and micro-indels) for their mutability and in particular recurrent somatic mutations?

In addition, specific questions were also asked with respect to different types of mutations.

How do somatic and germline missense mutations compare to each other (for each tumour suppressor gene and the combined mutations for all 17 genes) with respect to nucleotide substitution rates derived from non-disease and disease-associated substitution rates; physicochemical difference between wild-type and mutant amino-acids; degree of evolutionary conservation; co-localisation within CpG-dinucleotides?

How do somatic and germline nonsense mutations compare to each other (for each tumour suppressor gene and the combined mutations for all 17 genes) with respect to the potential involvement of nonsense-mediated mRNA decay?

How do somatic, germline and shared micro-lesions (*viz.* micro-deletions, micro-insertion and micro-indels) compare to each other with respect to their occurrence in the vicinity of repetitive elements?

2. General materials

2.1. Sources of mutation data

Germline mutations were obtained from the *Human Gene Mutation Database* (HGMD; http://www.hgmd.org; Stenson et al. 2003), a collection of >90,000 germ-line mutations in >3500 nuclear genes underlying or associated with human inherited disease. The HGMD data were privately communicated with Peter D. Stenson and Andrew D. Philips. Only one example of each reported mutation is present in HGMD, a policy designed so as to avoid confusion between recurrent and identical-by-descent lesions.

Sources of somatic mutation data included various somatic mutational databases, PubMed-based literature searches and data privately communicated by Gareth Evans (*NF2* gene) and Eamon Maher (*VHL* gene). The sources of somatic mutation data, used to extract mutations at the beginning of the PhD project (October 2005) are summarised in Table 1.

Table 1 Sources of somatic and germline mutational data

Name	Source	Data obtained
Human Gene Mutation Database (HGMD)	http://www.hgmd.org	Germline mutation data
Catalogue of Somatic Mutations in Cancer (COSMIC)	http://www.sanger.ac.uk/cosmic	Somatic mutation data for *RB1* and *PTEN*
Breast Cancer Information Core (BIC)	http://research.nhgri.nih.gov/bic/	Somatic mutation data for *BRCA1*
RB1 Gene Mutation Database	http://rb1-lsdb.d-lohmann.de	Somatic mutation data for *RB1*
International *NF2* Mutation Database	http://neurosurgery.mgh.harvard.edu/NFclinic/NFresearch.htm*	Somatic mutation data for *NF2*
Gareth Evans	Privately communicated; University Department of Medical Genetics, St. Mary's Hospital, Manchester M13 OJH, UK	Somatic mutation data for *NF2*
VHL Mutation Database	http://www.umd.be/VHL	Somatic mutation data for *VHL*
Eamonn Maher	Privately communicated; Section of Medical and Molecular Genetics, University of Birmingham, School of Medicine, B15 2TT, UK	Somatic mutation data for *VHL*
CDKN2A Database	https://biodesktop.uvm.edu/perl/p16	Somatic mutation data for *CDKN2A*
International Agency for Research on Cancer (IARC) *TP53* Mutation Database	http://www-p53.iarc.fr/index.html	Somatic mutation data for *TP53*
PubMed	http://www.ncbi.nlm.nih.gov/sites/entrez	Somatic mutation data for *APC*, *ATM*, *BRCA1*, *BRCA2*,

		CDH1, *NF1*, *PTCH*, *STK11*, *TSC1*, *TSC2* and *WT1*

* no longer existing link

Obviously, incorrect, incomplete or ambiguous data were disregarded irrespective of whether they were derived from databases or the original literature.

In order to avoid the repetition of analyses on multiple regularly updated mutational datasets, the collection of mutations was deemed to be complete by October 2005. Six different categories of germline and somatic micro-lesions were collated for 17 different human tumour suppressor genes (Table 2).

Table 2 The 17 human tumour suppressor genes studied

Gene symbol*	Gene ID*	Chromosome	Official name*
APC	324	5	Adenomatous polyposis coli
ATM	472	11	Ataxia telangiectasia mutated
BRCA1	672	17	Breast cancer 1, early onset
BRCA2	675	13	Breast cancer 2, early onset
CDH1	999	16	Cadherin 1, type 1, E-cadherin (epithelial)
CDKN2A	1029	9	Cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
NF1	4763	17	Neurofibromin 1
NF2	4771	22	Neurofibromin 2 (merlin)
PTCH1	5727	9	Patched homolog 1
PTEN	5728	10	Phosphatase and tensin homolog
RB1	5925	13	Retinoblastoma 1
STK11	6794	19	Serine/threonine kinase 11
TP53	7157	9	Tumour protein p53
TSC1	7248	6	Tuberous sclerosis 1
TSC2	7249	17	Tuberous sclerosis 2
VHL	7428	3	Von Hippel-Lindau tumour suppressor
WT1	7490	11	Wilms tumour 1

*Gene symbol, Gene ID and Official name were derived from Entrez, NCBI's gene database, available at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene.

The categories of mutations comprised: single base-pair substitutions that introduced missense and nonsense mutations in the coding regions of the 17 tumour suppressor genes; intra-genic micro-deletions, micro-insertions and micro-indels involving ≤20bp either deleted and/or inserted.

2.2. Collection policy

To allow ready comparison with the HGMD data, the somatic single base-pair substitutions (i.e. missense and nonsense mutations) were collated as triplet changes with an additional flanking base (shown in lower case) included when the mutated bases occurred in either the first or third positions in the triplet (e.g. gATG-TTG, ATGt-ATC).

Micro-deletions, micro-insertions and micro-indels of ≤20 bp were augmented with 10bp genomic DNA sequence flanking both sides of the lesion (e.g. CCAAGA^AAACagGGGCCCGAAA). The '^' symbol indicates the start of an amino-acid codon, such that it is not part of the deleted or inserted sequences and the deleted and/or inserted nucleotides are indicated in lower case. In addition, where deleted/inserted nucleotides or the 10bp flanking sequences extended into an intron of a gene, the position of the intron/exon boundary was also recorded (e.g. GAAG_I25E26_G^TTTTTccTTGATATAGC, CCAAA^TCACAgttatttcttaa_E19bI19b_gtaaattTCAGTCACCA).

The clinical phenotype (histological clinical phenotype of the tumours of associated mutational data), mutation sequence (triplet changes for nonsense and missense mutations and deleted/inserted nucleotides with corresponding 10bp flanking sequence for micro-insertions, micro-deletions and micro-indels), amino-acid position (referring to the amino acid immediately following the symbol '^' for micro-deletions, micro-insertions and micro-indels), reference (author, journal, volume, page and year) were also collected. Examples of a logged missense mutation and a micro-deletion are provided in Table 3 and Table 4.

Table 3 An example of a logged somatic missense mutation in the *NF1* gene

Gene	Clinical phenotype	Mutation sequence	Amino-acid substitution	Amino-acid position	Reference				
					Author	Journal	Volume	Page	Year
NF1	Neurofibroma	CAA-CCA	Gln-Pro	519	Upadhyaya	Hum Mutation	23	134	2004

Table 4 An example of a logged somatic micro-deletion in the *NF1* gene

Gene	Clinical phenotype	Mutation sequence	Amino-acid position (^)	Reference

				Author	Journal	Volume	Page	Year
NF1	Neurofibroma	GTGGTTCTTTatttatAG^GCATTTTG	219	Serra	Hum Genet	108	416	2001

To be regarded as *bona fide* somatic mutations, and therefore suitable for inclusion in the analysis, reported lesions had to have been found in tumour tissue but also to have been shown to be absent from a non-tumorous tissue from the same patient. Thus, mutation data derived from "sporadic" patients were not included, unless non-tumorous tissue had also been examined in order to exclude the possibility that they were constitutional mutations.
Only one example of each somatic lesion was selected, although individual examples of independently recurring somatic lesions in each dataset were noted and marked accordingly. Thus, mutations with more than one, but fewer than ten, independent examples were marked with a symbol '*', whereas examples of mutations that recurred more than ten times were marked with the symbol '**'.

Mutational data were stored in the form of plain text files (tab-delimited format), separately for each gene and for each mutational type (*viz.* germline and somatic). A summary of the numbers of collected mutations is presented in Table 5. In addition, a graphical overview of the collected somatic and germline mutations is presented in Figure 1 and Figure 2.

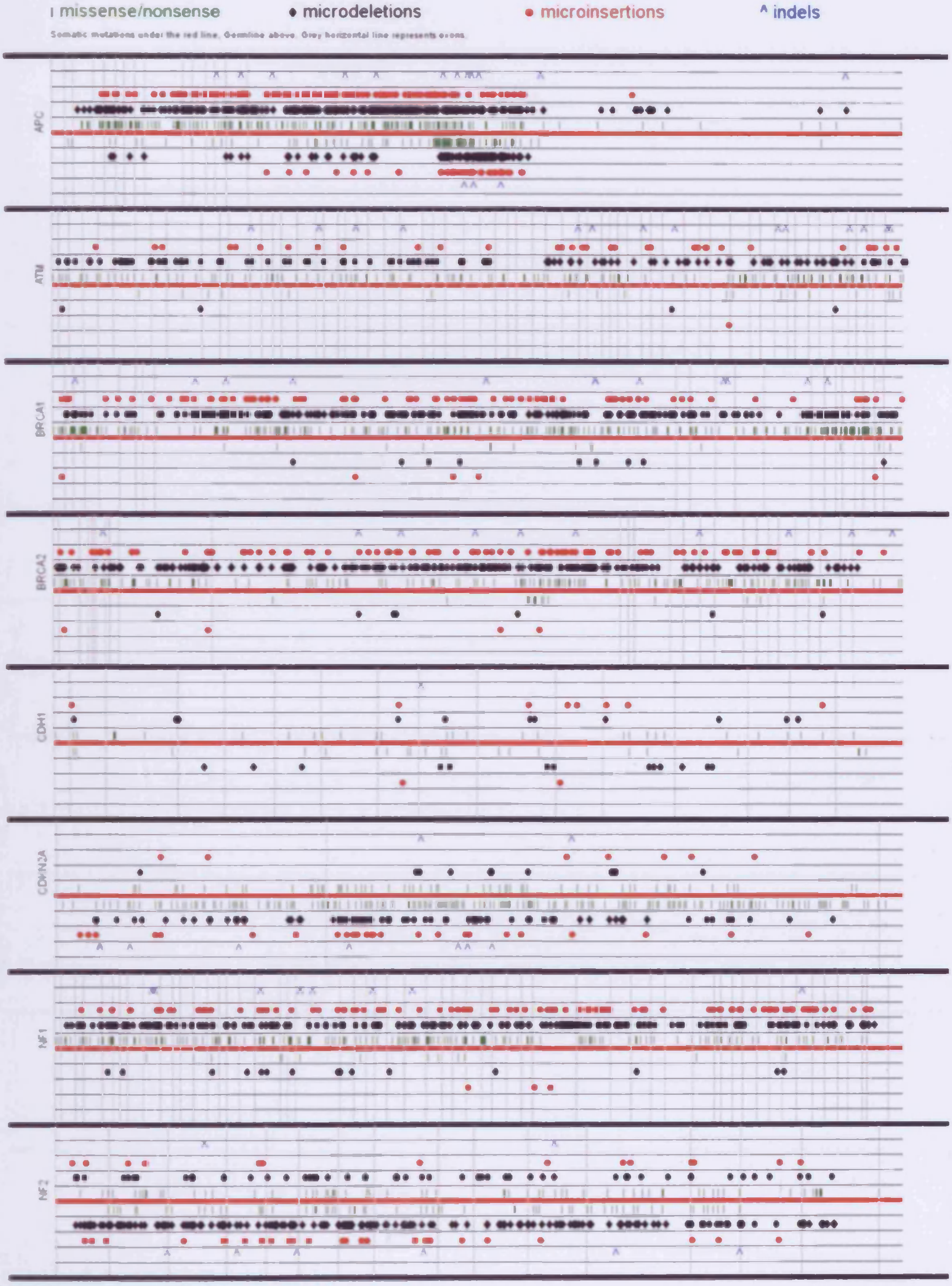
Table 5 Summary of numbers of collected mutations in the 17 human tumour suppressor genes studied

		Mutations					
		Missense	Nonsense	Micro-deletions	Micro-insertions	Micro-indels	Total
APC	Somatic	38	3	137	44	3	225
	Germline	22	64	284	115	36	521
	Shared	1	4	15	0	0	20
	Recurrent	4	2	33	4	0	43
	Total	61	71	436	159	39	766
ATM	Somatic	11	4	4	1	0	20
	Germline	75	69	122	35	14	315
	Shared	0	2	0	0	0	2
	Recurrent	1	0	0	0	0	1
	Total	86	75	126	36	14	337
BRCA1	Somatic	5	3	6	3	0	17
	Germline	169	109	255	83	12	628
	Shared	1	4	3	2	0	10
	Recurrent	0	0	2	0	0	2
	Total	175	116	264	88	12	655

BRCA2	Somatic	20	1	7	2	0	30
	Germline	85	75	244	88	10	502
	Shared	1	0	1	2	0	4
	Recurrent	2	0	1	0	0	3
	Total	106	76	252	92	10	536
CDH1	Somatic	14	5	13	2	0	34
	Germline	18	9	12	8	1	48
	Shared	1	2	0	0	0	3
	Recurrent	0	0	0	0	0	0
	Total	33	16	25	10	1	85
CDKN2A	Somatic	170	13	76	24	8	291
	Germline	34	1	10	6	2	53
	Shared	28	5	1	1	0	35
	Recurrent	6	3	9	3	0	21
	Total	232	19	87	31	10	379
NF1	Somatic	2	4	13	3	0	22
	Germline	83	105	218	105	8	519
	Shared	0	10	3	0	0	13
	Recurrent	0	0	1	0	0	1
	Total	85	119	234	108	8	554
NF2	Somatic	23	24	176	28	6	257
	Germline	20	25	50	16	2	113
	Shared	0	18	5	0	0	23
	Recurrent	3	18	8	2	0	31
	Total	43	67	231	44	8	393
PTCH	Somatic	13	7	14	6	1	41
	Germline	23	25	42	32	8	130
	Shared	1	2	0	0	0	3
	Recurrent	0	0	0	0	0	0
	Total	37	34	56	38	9	174
PTEN	Somatic	201	39	145	47	4	436
	Germline	23	15	23	18	3	82
	Shared	22	12	6	4	0	44
	Recurrent	47	17	45	16	0	125
	Total	246	66	174	69	7	562
RB1	Somatic	22	12	30	12	2	78
	Germline	34	61	112	53	10	270
	Shared	3	15	4	0	1	23
	Recurrent	1	9	2	0	0	12
	Total	59	88	146	65	13	371
STK11	Somatic	17	7	3	1	0	28
	Germline	27	23	45	24	2	121
	Shared	3	3	2	0	1	9
	Recurrent	2	1	1	0	0	4
	Total	47	33	50	25	3	158
TP53	Somatic	1138	87	504	234	0	1963
	Germline	6	1	8	3	3	21
	Shared	88	9	8	4	0	109
	Recurrent	781	85	162	57	0	1085
	Total	1232	97	520	241	3	2093
TSC1	Somatic	2	1	1	0	0	4
	Germline	7	37	53	25	4	126

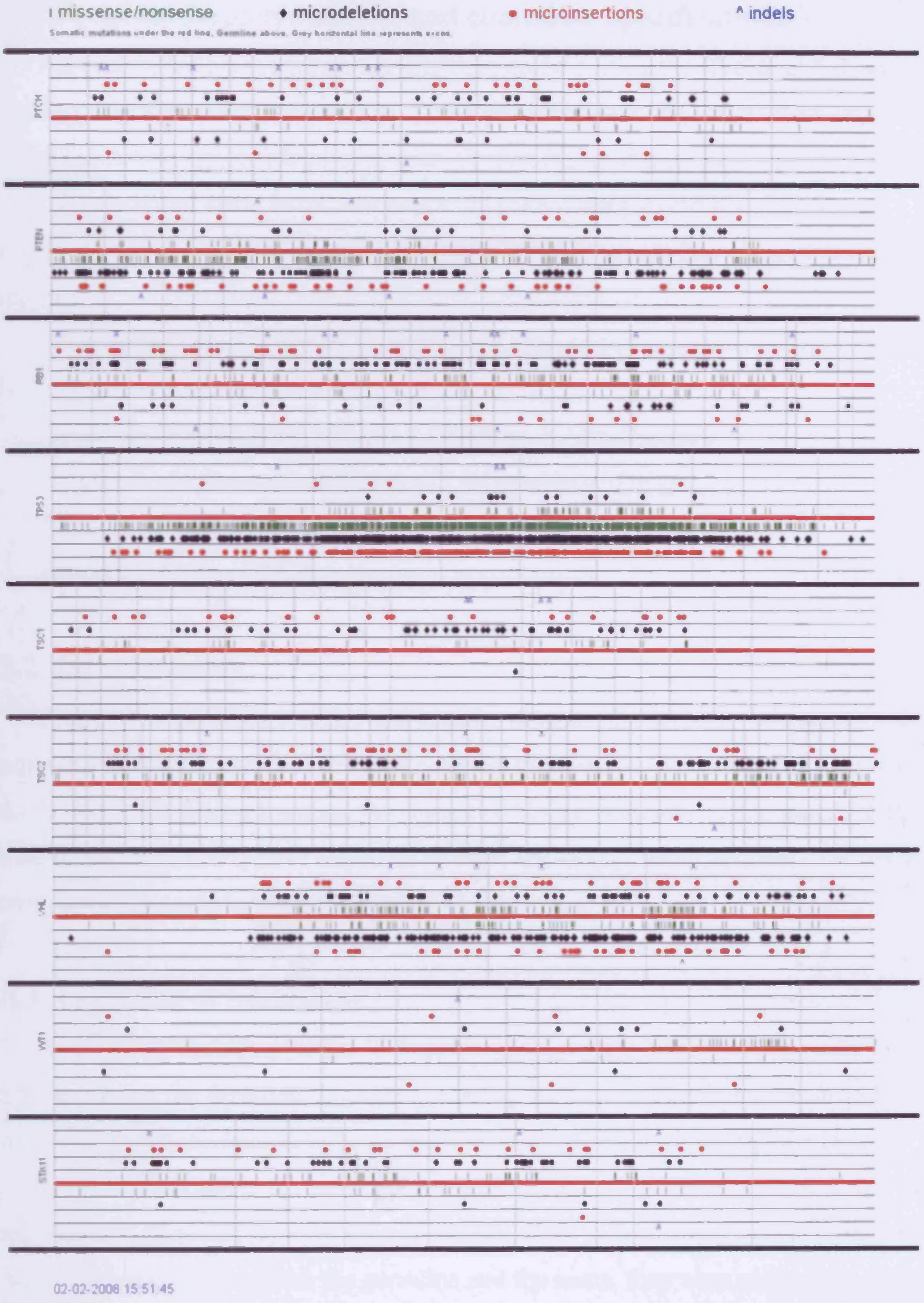
	Shared	0	0	0	0	0	0
	Recurrent	0	0	0	0	0	0
	Total	9	38	54	25	4	130
TSC2	Somatic	0	0	3	2	1	6
	Germline	87	72	110	46	3	318
	Shared	2	1	0	0	0	3
	Recurrent	0	0	0	0	0	0
	Total	89	73	113	48	4	327
VHL	Somatic	43	3	171	38	1	256
	Germline	98	8	55	31	5	197
	Shared	45	6	8	6	0	65
	Recurrent	5	2	14	2	0	23
	Total	186	17	234	75	6	518
WT1	Somatic	1	0	4	3	0	8
	Germline	39	11	8	4	1	63
	Shared	0	3	0	0	0	3
	Recurrent	0	0	0	0	0	0
	Total	40	14	12	7	1	74
ALL	Somatic	1720	213	1307	450	26	3716
	Germline	850	710	1651	692	124	4027
	Shared	196	96	56	19	2	369
	Recurrent	852	137	278	84	0	1351
	Total	2766	1019	3014	1161	152	8112

Figure 1 A graphical representation of the somatic and germline mutations in 8 of the 17 human tumour suppressor genes studied (*APC*, *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *CDKN2A*, *NF1* and *NF2*)



02-02-2006 15:53:17

Figure 2 A graphical representation of the somatic and germline mutations in 9 of the 17 human tumour suppressor genes studied (*PTCH*, *PTEN*, *RB1*, *TP53*, *TSC1*, *TSC2*, *VHL*, *WT1* and STK11)



3. General methods

3.1. Program implementation and computer specifications

All custom-built computer programs were created using the 'Practical Extraction and Reporting Language' (Perl; version 5.8). Those programs were the result of my own design and implementation, unless stated otherwise. Certain Perl packages were used to perform specific functions and those have been explicitly acknowledged, either as a literature reference or internet link. All algorithmic steps of these programs have been described, where appropriate.

All programs used in this PhD thesis were executed on a personal computer (PC) with the following specifications:

Operating System (OS): Microsoft Windows XP Professional
Processor: single-core Intel Xeon at 2.8 Giga Hertz (GHz)
Random Access Memory (RAM): 1 Giga Byte (GB)
Hard-Disk Drive: 150GB

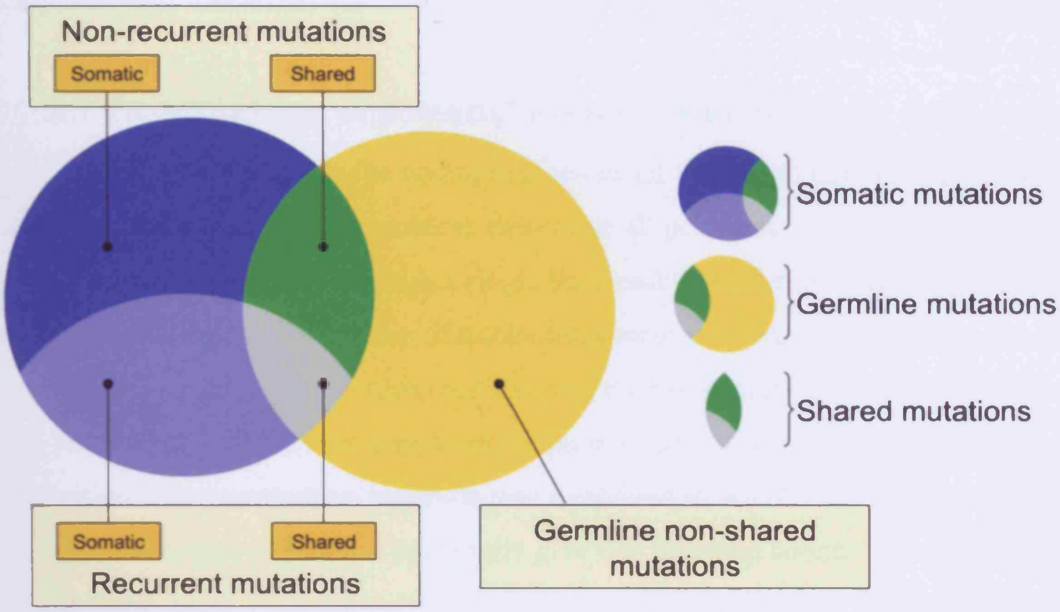
3.2. Data integrity

A computer program was devised in order to check the accuracy of the manually extracted/curated mutational data. Where errors in these data were discovered, they were manually corrected. This minimised the chance of having inaccurately annotated (in terms of position of mutation and type of sequence change, flanking genomic sequence for the micro-lesions) mutations included in the analysis.

3.3. Labelling of mutations

Some of the mutations in the 17 tumour suppressor genes under study were found in both the soma and the germline. In order to identify such mutations that were shared (i.e. mutations that were reported in both the soma and the germline), I devised a computer program that made possible the automatic recognition and labelling of mutations that were shared. For each gene and for each type of mutation (*viz.* missense and nonsense mutations and micro-lesions) found in both the germline and the soma, they were removed from the list of mutations and a new single entry with the label 'shared' was created. Thus, mutations were labelled somatic (when they were found exclusively in the soma), germline (i.e. when they were found exclusively in the germline) or shared (when they were found in both the soma and the germline). A graphical representation of the labelling of mutations is given in Figure 3.

Figure 3 A graphical representation of labelling of mutations



3.4. Extended cDNA sequences

Some of the analyses performed in this PhD project required extended cDNA sequences for all 17 tumour suppressor genes under study. In order to acquire such extended cDNA sequences, genomic DNA sequences (sequence contigs) for each gene were collected. A sequence contig encompassing the gene sequence was identified via a link in NCBI's GenBank http://www.ncbi.nlm.nih.gov/sites/entrez?db=Nucleotide. For every gene, the genomic sequence was stored in a text file (FASTA format). In order to find the positions where the cDNA mapped to the genomic sequence, the Spidey program (Wheelan et al. 2001) was used. It allows alignment of spliced sequences (i.e. cDNA) to genomic sequences. The PC executable form of the Spidey program available at http://www.ncbi.nlm.nih.gov/spidey/spideyexec.html was used locally to obtain the exact positions where the cDNA mapped to the genomic sequence with the following parameters:

-i (file name, containing genomic sequence in FASTA format)

-m (file name, containing cDNA sequence in FASTA format)

I devised a computer program that took Spidey's output (text file) and extracted the extended cDNA sequence (i.e. 85bp around every exon) as well as the positions of splice junctions. For every gene, the program outputted a text file with the extended cDNA sequence (FASTA format) as well as a mapping file that mapped the exact positions of the beginning and end of every exon and intron to the extended cDNA. The cDNA and extended cDNA sequences for all 17 tumour suppressor genes studied could be found in the supplementary materials.

3.5. Identification of potential missense and nonsense mutations

Every codon within the coding regions of all 17 human tumour suppressor genes under study was examined in order to determine all potential missense and nonsense mutations that could arise through a single base-pair substitution. Thus, at position 1 of the codon, all possible combinations of nucleotides were introduced (3 possible combinations excluding the existing nucleotide) and keeping the nucleotides at positions 2 and 3 unchanged (i.e. simulating a single base-pair substitution in position 1). After each change of nucleotide in position 1, the new codon sequence was examined to see if it changed the wild-type amino acid (i.e. missense) or it could potentially give rise to a stop codon. This process was repeated for positions 2 and 3, where nucleotides 1 and 3 respectively were kept unchanged (i.e. change in position 2) and positions 1 and 2 were kept unchanged (i.e. change in position 3). An example is shown in Table 6.

Table 6 Identification of potential missense and nonsense mutations through a single base-pair substitution

Position	**1**	**2**	**3**	**Codon**	**Amino acid**
Codon	C	G	A	CGA	Arginine (R)
Possible change	T	—	—	TGA	Stop
	A	—	—	AGA	Arginine (R)
	G	—	—	CGA	Arginine (R)
	—	T	—	CTA	Leucine (L)
	—	A	—	CAA	Glutamine (Q)
	—	C	—	CCA	Proline (P)
	—	—	T	CGT	Arginine (R)
	—	—	G	CGG	Arginine (R)
	—	—	C	CGC	Arginine (R)

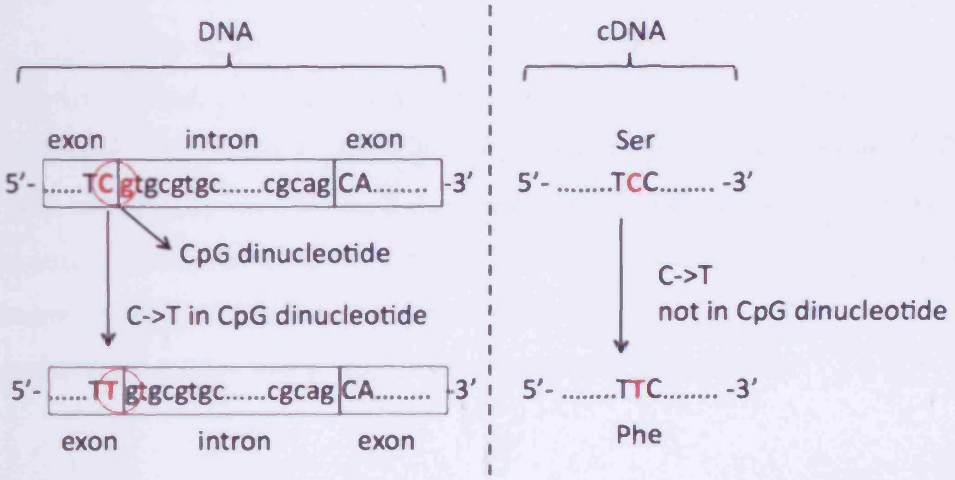
Symbol "—" indicates that corresponding position within the codon is unchanged

All calculations of potential missense and nonsense mutations were performed according to the canonical open reading frame (ORF). I devised a stand-alone computer program in order to define, codon-by-codon, all potential single base-pair substitutions leading to a non-synonymous change (missense and nonsense mutations) of the wild-type amino acid within the coding sequences of the genes, thereby minimising manual intervention and maximising error-free definitions of potential missense and nonsense mutations.

3.5.1. Identification of potential missense and nonsense mutations in CpG-dinucleotides

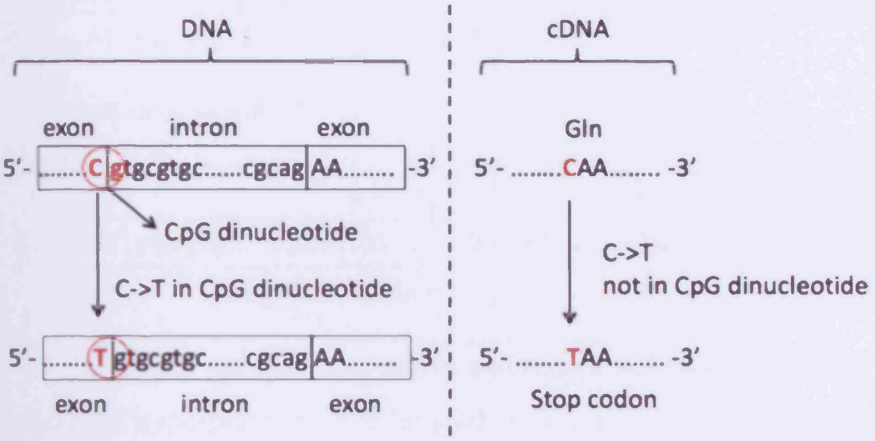
In order to recognise all missense and nonsense mutations that could arise in CpG dinucleotides, a slight modification to the algorithm, described in 3.5, had to be applied in order to include exon-intron junction-spanning CpG dinucleotides. Generally, the coding sequences of human genes are split by introns, the exception being a few intronless genes. Let us take an example of a TCC codon (encoding the amino acid serine) that is split by an intron after the first C nucleotide. Since the nucleotide sequence for the splice donor site of the intervening intron invariably starts with a G nucleotide, a C->T transition in the CpG dinucleotide in the last base of the exon in the gene's genomic sequence would generate a missense mutation (i.e. TTC) after exon-exon splicing. The above example is shown in Figure 4.

Figure 4 A single base-pair substitution (missense mutation) in a CpG dinucleotide at exon-intron boundary


DNA
cDNA
exon
intron
exon
5'- ......TC gtgcgtgc......cgcag CA........ -3'
CpG dinucleotide
C->T in CpG dinucleotide
5'- ......TT gtgcgtgc......cgcag CA........ -3'
exon
intron
exon
Ser
5'- ....TCC........ -3'
C->T
not in CpG dinucleotide
5'- ....TTC........ -3'
Phe

Similarly, if we take an example of a CAA codon (encoding the amino acid glutamine) that is split by an intron after the C nucleotide. Since the nucleotide sequence for the splice donor site of the intervening intron invariably starts with a G nucleotide, a C->T transition in the CpG dinucleotide in the last base of the exon in the gene's genomic sequence would generate a nonsense mutation (i.e. TAA) after exon-exon splicing. The above example is shown in Figure 5.

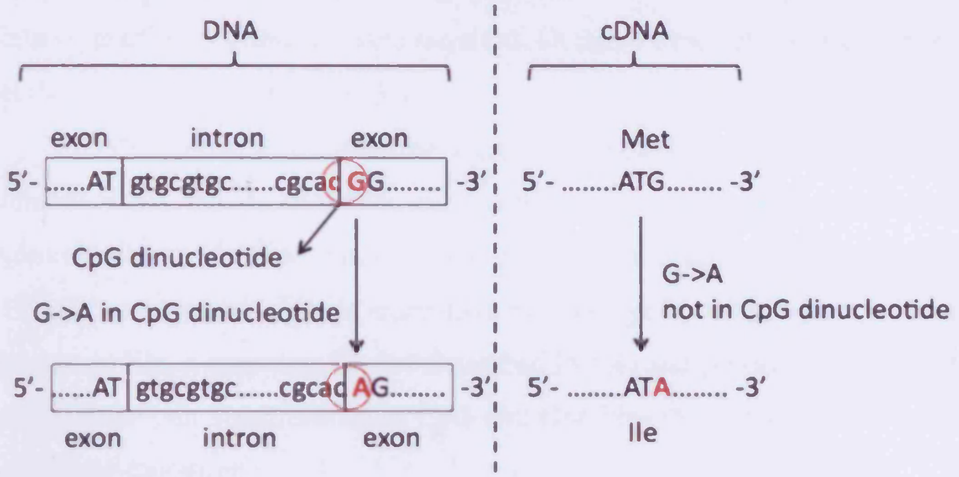
Figure 5 A single base-pair substitution (nonsense mutation) in a CpG dinucleotide (exon-intron boundary)


DNA
cDNA
exon
intron
exon
5'- ....C gtgcgtgc......cgcag AA........ -3'
CpG dinucleotide
C->T in CpG dinucleotide
5'- ....T gtgcgtgc......cgcag AA........ -3'
exon
intron
exon
Gln
5'- ....CAA........ -3'
C->T
not in CpG dinucleotide
5'- ....TAA........ -3'
Stop codon

These special cases (arguably could be classified as splice site mutations, although the mutations are exonic in location) would however have been missed, if the analysis had employed cDNA (rather than genomic DNA) sequence to identify single base-pair substitutions leading to missense and nonsense mutations in CpG dinucleotides (i.e. if the analysis had been based on cDNA sequence, the TCC->TTC and CAA->TAA substitutions would not have been counted as a missense and nonsense mutation in CpG dinucleotides, because they would not obviously have occurred in a CpG).

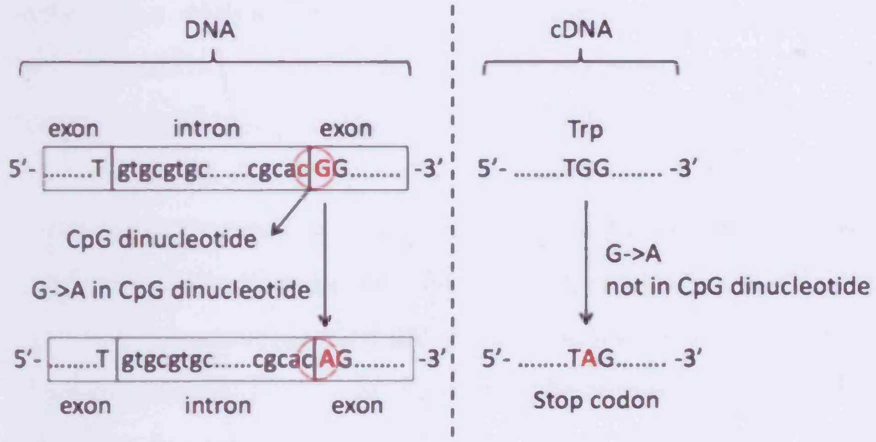
In addition, not all splice sites contain the canonical GT-AG consensus splice site sequence. Indeed, a very small proportion of introns possess AT-AC splice sequences instead of the usual GT-AG consensus splice site sequence (reviewed in Mount 2000). Therefore, in a very similar way (as described above and shown in Figure 4), if an ATG codon, encoding the amino acid methionine, were to be split by an intron after the T nucleotide and the consensus acceptor splice site ends with the nucleotide C, a potential G->A substitution would create a ATA missense mutation (as shown in Figure 6).

Figure 6 A single base-pair substitution (missense mutation) in a CpG dinucleotide at intron-exon boundary


DNA
cDNA
exon
intron
exon
5'- ......AT gtgcgtgc......cgcac GG........ -3'
Met
5'- ....ATG........ -3'
CpG dinucleotide
G->A in CpG dinucleotide
G->A
not in CpG dinucleotide
5'- ......AT gtgcgtgc......cgcac AG........ -3'
5'- ....ATA........ -3'
exon
intron
exon
Ile

Similarly (as described above and shown in Figure 5), a TGG codon, encoding the amino acid tryptophan, were to be split by an intron after the T nucleotide and the consensus acceptor splice site ends with the nucleotide C, a potential G->A substitution would create a TAG nonsense mutation (as shown in Figure 7).

Figure 7 A single base-pair substitution (nonsense mutation) in a CpG dinucleotide (intron-exon boundary)


DNA
cDNA
exon
intron
exon
5'- ....T gtgcgtgc......cgcac GG........ -3'
Trp
5'- ....TGG........ -3'
CpG dinucleotide
G->A in CpG dinucleotide
G->A
not in CpG dinucleotide
5'- ....T gtgcgtgc......cgcac AG........ -3'
5'- ....TAG........ -3'
exon
intron
exon
Stop codon

Again, as described above, these special cases (arguably they could be classified as splice site mutations, although the mutations are exonic in nature) would be missed, if the analysis were to have been solely based on cDNA sequence to identify single base-pair substitutions leading to missense and nonsense mutations in CpG dinucleotides (i.e. the G->A substitutions giving rise to TGG->TAG and ATG->ATA are not in a CpG dinucleotide when the cDNA is considered).

Taking these special cases into consideration, in order to examine whether such cases exist in the 17 tumour suppressor genes, 'extended cDNA sequences' that include exon-intron junction sequences, were required. Detailed description of the generation of extended cDNA sequences is given in 3.4.

I developed a computer program to determine if such special cases of CpG dinucleotides, discussed above, exist in the 17 tumour suppressor genes and could be converted into missense and nonsense mutations through a single base-pair substitution in the 17 tumour suppressor genes examined. For each gene, the program took an extended cDNA sequence file, a mapping file (as described in 3.4) and produced any of these special cases of single base-pair substitutions in CpG dinucleotides that could lead to a missense and nonsense mutation.

After running the program sequentially for each gene, such special cases (i.e. missense mutations) of CpG dinucleotides were indeed noted. These special cases included, TC|g...|C->TTC, AC|g...|A->ATA and AT|...c|G->ATA (|g...| denotes the first nucleotide of an intron and |...c| denotes the last nucleotide of an intron) in the *ATM*, *CDH1* and *STK11* genes respectively. Having found such special cases of substitutions in CpG dinucleotides leading to missense mutations, extended cDNA sequences were used in the subsequent analyses (i.e. analyses in Chapter 4; Missense mutations) for the abovementioned genes.

In addition, such special cases of CpG dinucleotides were also noted for the potential nonsense mutations, although none of the potential single base-pair substitutions (i.e. C->T or G->A) led to the introduction of a stop codon. Having checked in this way that the use of cDNA rather than genomic sequence would not cause errors, cDNA sequence from each gene was used in order to make the subsequent algorithms more efficient (i.e. analyses in Chapter 5; Nonsense mutations). Certainly, if this analysis were ever to be extended to a wider range of genes, genomic DNA sequence should be employed because non-canonical splice sequences do occasionally occur.

3.6. Statistical methods used

3.6.1. Hypothesis testing

3.6.1.1. Association testing

When analysing a binary or categorical variable, the distribution of the variable can be represented as a contingency table (Table 7).

Table 7 Distribution of a binary variable, represented as a contingency table

	In repeats	**Not in repeats**	**Marginal totals**
Number somatic mutations	a	b	a+b
Number germline mutations	c	d	c+d
Marginal totals	a+c	b+d	n=(a+b+c+d)

In order to assess the association or relationship between different type of mutations (e.g. somatic and germline in the example given in Table 7) and the binary variable in question (e.g. status of the mutations with respect to its occurrence in repeats), i.e. the non-random distribution of the status of mutations with respect to its occurrence in repeats in the two samples of mutations, a Pearson's χ^2 test with 1 degree of freedom (df) can be calculated (Altman 1991). The test compares observed frequencies with the expected frequencies (see Equation 1), under the assumption of independence (i.e. repeats are not associated with the positional occurrence of mutations and the distribution of the repeats is similar in both types of mutations).

Equation 1 Pearson's χ^2 statistic (after Field 2005)

$$\chi^2 = \sum_{i=1}^{4} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$, where

$$Expected_i = \frac{Row\_total_i * Column\_total_i}{n}$$, where n is the total number of observations (i.e. $n = a + b + c + d$ in Table 7); i is the number of the cell in the contingency table; m is the total number of cells within the contingency table

The χ^2 statistic is an appropriate statistical measure, when 80% of the cells in the contingency table have expected counts ≥5, but also when all of the cells show expected counts >1 (Altman 1991). In the case of small expected frequencies, Fisher's exact test or simulation-based tests should be used. The small expected frequencies would inflate the χ^2 statistic. Furthermore, when the observed counts are small, then the χ^2 statistic tends to be overestimated, because the assumption of a continuous χ^2 distribution introduces some bias.

The calculation of the χ^2 statistic was performed using the formulae given in Equation 1. Those calculations were mainly incorporated into Perl programs. Some of the calculations were performed using the chisq.test function in the R statistical language (http://cran.r-project.org/). The chisq.test function in R was only used for the calculation of the χ^2 statistic in Chapter 6 for the combination of micro-lesions.

The associated p-values of the χ^2 statistic were calculated either within a Perl program or within R. Within a Perl program, the p-values were calculated using the chisqrprob function, which is a part of the Statistics::Distributions package, available at http://search.cpan.org/~mikek/Statistics-Distributions-1.02/Distributions.pm. Within R, the chisq.test function (see above) outputs include both the χ^2 statistic and the associated p-value.

3.6.1.2. Non-parametric tests

Most of the statistical tests rely on parametric assumptions about the data, most notably- normally distributed data (Field 2005). In addition, for non-normally distributed data, it is not always possible to correct for an unknown distribution. Such was the case with the analysis of a number of parameters detailed in Chapter 4. Therefore, a non-parametric test was used, i.e. Wilcoxon rank-sum test (Wilcoxon 1945).

Non-parametric tests are usually known as assumption-free tests, because they make fewer assumption as compared to parametric tests (Field 2005). The Wilcoxon rank-sum test compares whether two independent samples have come from the same distribution, but is also known to be used for testing differences between medians (Field 2005).

I created a computer program that, when supplied with two sets of data (e.g. somatic and germline substitution rates), calculates the Wilcoxon rank-sum statistic (i.e. W_S statistic), the mean of the test statistic ($\overline{W_S}$) and the standard error (SE_{W_s}). The calculations of the Wilcoxon rank-sum statistic comprised the following algorithm: the values in two datasets under investigation were combined into one dataset, but keeping a record to which dataset they belong. Those values were sorted in ascending order and assigned a potential rank, ignoring the dataset to which they belong. If two values have the same number (i.e. tied ranks), they were assigned ranks that were the average of the potential ranks. The W statistic is calculated by adding up all the ranks in each dataset and choosing the lowest of the two sums to be the test statistic (Field 2005). Based on these values, the W statistic could be easily converted into a z-score using the following formula:

Equation 2 Converting Wilcoxon rank-sum statistic into z-score (Field 2005)

$$z = \frac{X - \overline{X}}{s} = \frac{W_S - \overline{W}_S}{SE_{W_s}}$$

$$SE_{W_s} = \frac{\sqrt{n_1 n_2 (n_1 + n_2 + 1)}}{12}$$

$$\overline{W}_S = \frac{n_1(n_1 + n_2 + 1)}{2}$$

Then, by using the properties of the normal distribution with mean 0 and variance 1 ($\mu = 0, \sigma = 1$), the z-score is easily converted into a p-value in order to assess the statistical significance of the results at the chosen alpha level of significance (i.e. $\alpha = 0.05$).

3.6.2. Multiple hypothesis testing

The difference between two groups is considered as statistically significant if the corresponding p-value is smaller than the significance level alpha (α) chosen for the particular experiment. The commonly accepted significance level is $\alpha = 0.05$ (Fisher 1990). The significance level α indicates the probability of observing a difference between two groups by chance under the null hypothesis of no difference. Thus, when more than one ($N > 1$) hypothesis is tested, each hypothesis has a probability α of being falsely determined as being significant and therefore the expected number of (false) significant findings, assuming the null hypothesis in each test, is equal to $N\alpha$ and the probability of finding at least one significant difference by chance if the tests are independent is $p = 1 - (1 - \alpha) * N$ (e.g. if $N = 200$ this probability equals to 0.99996).

Correction for multiple hypothesis testing attempts to maintain the probability p at the chosen significance level α. The most widely used method of multiple hypotheses correction is the Bonferroni correction, where the α is simply divided by the number of tests performed, and the overall chance of finding any false positive remains the same as in a single hypothesis experiment. The Bonferroni correction assumes that the tests are independent, and is considered to be a conservative adjustment when tests are dependent (Sokal and Rohlf 1995).

The way the data were split, many comparisons were found to have strong correlations with one another within a gene, and the use of the Bonferroni adjustment would therefore be likely to be a conservative correction. I implemented a permutation-based correction, which computes p-values that are adjusted for the number of tests undertaken but in a way that is less conservative than the Bonferroni method. The permutation-based methods are often (and successfully) used in the context of microarray expression data (Olshen and Jain 2002). At each permutation, the label (i.e. somatic, germline, shared and potential, where applicable) is randomly reshuffled, ensuring that the number of each type of mutation (*viz.* somatic, germline, shared and potential, where applicable) is the same and that the differences between groups occur purely by chance while preserving the correlation structure between the tests.

The algorithm of the permutation method used is as follows:

1) Compute χ^2 statistic (original statistic) for every possible comparison in a particular gene or the combination of mutations in all genes;
2) Randomly permute the label (i.e. somatic, germline, shared and potential), thereby breaking the relationship between the studied variables and the observed mutations;
3) Compute the same χ^2 statistic as in (1) using the permuted labels. Save the maximum statistic;
4) Compare the maximum statistic with every original value of statistic (see 1) and record a success, if the maximum statistic is greater or equal than original value of statistic
5) Repeat (2) and (3) 10,000 times*;
6) For each test the permuted p-value is derived by dividing the number of successes recorded in (4) by the number of permutations performed (i.e. 10,000).

* It is not feasible to use all possible permutations, as the number of combinations is computationally expensive or time-consuming.

The permutation-based method used here, only tries to maintain the probability of falsely finding any significant hypothesis at the α value for each gene or the combination of mutations in all genes, but does not account for the tests performed in different genes. It is computationally expensive to use permutation-based method to account for the tests performed for all genes. If it were to be performed, the permutation method (described above) used for each gene had to be performed 10,000 times. Instead, Bonferroni correction for multiple testing was applied which is always valid as a conservative estimate. The new α value for each gene was obtained after dividing the permuted critical value by the number of genes tested (i.e. 17).

Due to the nature of the permutation method applied to correct for multiple hypothesis testing, if the original statistics is very large, the statistics obtained at a particular permutation sometimes never reached the observed one. In this case the corrected p-values were reported as $<1/M$, where M is the number of permutations performed (in all cases this was 10,000).

All p-values reported in the Results section of the individual chapters are corrected for multiple testing as described above.

3.6.3. Calculation of statistical power

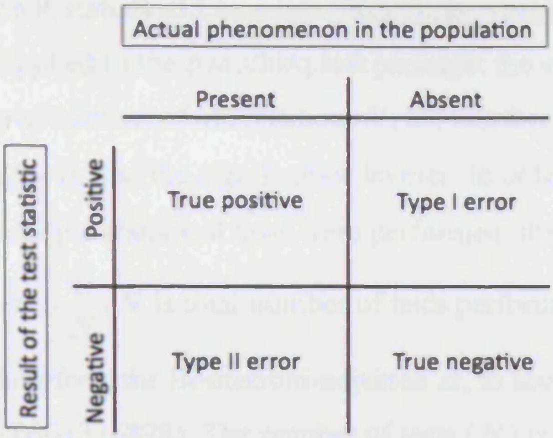
3.6.3.1. Type I and type II errors

Generally, statistics deals with a subset or a subsample of the population of interest. Thus, statistics uses statistical tests to determine if an effect or a phenomenon (e.g. difference in proportions, difference in the distributions, etc.) exists in the studied population. Normally, the true state of the population is unknown, i.e. it is unknown whether an effect exists or not. Therefore, a test statistic and associated probabilities could indicate which is more likely. In this process, one could commit two types of errors.

Type I error, also called the false positive error (Figure 8), is the probability of falsely rejecting the null hypothesis, when there is no true effect in the population. The most commonly accepted Type I error rate is $\alpha = 0.05$ (Fisher 1990). Therefore, there is only a small, i.e. 5% chance, of the result occurring by chance alone.

Type II error, also called the false negative error (Figure 8), is the probability of falsely accepting the null hypothesis, when in fact there is a true effect in the population. Cohen (1988) has suggested that the maximum acceptable probability of the Type II error should be $\beta = 0.2$ or 20% (Cohen 1988; Field 2005). Thus, there would be a 20% chance that an existing genuine or true effect/phenomenon in the population would not be detected by the statistical test.

Figure 8 Type I and Type II errors



		Actual phenomenon in the population	
		Present	Absent
Result of the test statistic	Positive	True positive	Type I error
	Negative	Type II error	True negative

3.6.3.2. Effect size

The effect size or an observed effect size is the strength of the relationship between two variables being measured. For the test of χ^2 statistic, the effect size was calculated by using the *ES.w2* function in the *PWR* package, part of the R statistical language (http://cran.r-project.org/) that utilizes the following formula (Cohen 1988):

Equation 3 Effect size for χ^2 statistic

$$w = \sqrt{\sum_{i=1}^{4} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}},$$

$$P_{0i} = P_{ir}.P_{ic}$$

where P_{0i} is the proportion in cell i posited by the null hypothesis (e.g. a/n in Table 7), and P_{ir} and P_{ic} are the proportions of the marginal totals in the contingency table (i.e. $P_{ir} = (a+b)/n$ and $P_{ic} = (a+c)/n$ in Table 7); P_{1i} is the proportion in cell i posited by the alternative hypothesis.

It is generally considered that an effect size of 0.10 represents a small effect, 0.30 a medium effect and 0.50 a large effect.

3.6.3.3. Power

The power of a statistical test represents the probability to detect an effect size of a particular magnitude (w or r) with a specified Type I error rate (α) and a particular sample size, $power = 1 - \beta$ (where β is Type II error rate). The power analysis requires an assumption that a true effect exists in the population under study.

Power calculations for the χ^2 tests were performed using the pwr.chisq.test package, part of the R statistical Language (http://cran.r-project.org/). The following parameters were supplied to the pwr.chisq.test package: the effect size w (calculated using Equation 3), the total number of observations N, the number of degrees of freedom df (for all tests performed, $df = 1$), and the significance level α. In order to keep the overall α at the 0.05 level, as multiple statistical tests were performed, the value of α used in the power calculations was set to $\frac{\alpha}{N}$ (N is total number of tests performed, e.g. 374 tests performed in Chapter 6, therefore, the Bonferroni-adjusted α, to account for multiple testing, was $\alpha = 0.05/374$, or 0.0001336898). The number of tests (N) is explicitly given in each of the results chapters (i.e. Chapters 4, 5 and 6). Since Bonferroni correction is considered to be a conservative correction for multiple testing, the power calculated for the χ^2 statistical tests is a conservative estimate.

For power calculations, with respect to Wilcoxon rank-sum tests used, a data-based simulation method for statistical inference was used (Walters 2004). The simulation method involved repeatedly drawing random sub-samples from the original data, with replacement, thereby generating the non-standard distribution of the observed data. For power analysis, the following algorithm was used, based on Walters' Method 4 (Walters 2004):

For each test, calculate observed difference of means $\delta = \bar{x} - \bar{y}$, where $\bar{x}$ is the mean in the first sample (S_1) and $\bar{y}$ is the mean in the second sample (S_2)

1) Draw two random samples (S_1 and S_2) from the combined mutational data, with N_1 and N_2, where N_1 and N_2 are the observed number of mutations in the two original datasets.
2) Add δ to each of the samples in S_1.
3) Calculate the Wilcoxon rank-sum test statistic and associated significance (p-value).
4) A success is recorded, if $p \le \frac{\alpha}{2N}$ (two-sided test, thus $\alpha / 2 = 0.025$)
5) Steps 1-4 are repeated 10,000 times and power is calculated by the proportion of successes among the 10,000 simulations.

3.7. Supplementary Data

Owing to the immense volume of data generated during this project, only the most interesting (i.e. the statistically significant) results are presented in paper format.

Nevertheless, comprehensive results from the analyses are presented in the form of supplementary tables, and are supplied on a CD at the back of the thesis. The supplementary CD comprises: cDNA and extended cDNA sequences, comprehensive results for the missense, nonsense and micro-lesions mutational analyses, somatic and germline mutations, repetitive elements, properties of corresponding types of mutations (i.e. disease and non-disease nucleotide substitution rates, Grantham difference values, CpG-located mutations, for missense mutations; predicted NMD status and CpG-located mutations for nonsense mutations; location within repetitive elements for micro-deletions, micro-insertions and micro-indels), for each of the 17 human tumour suppressor genes studied.

4. Missense mutations

4.1. Introduction

4.1.1. The importance of missense mutations

Missense mutations (i.e. nonsynonymous mutations) are defined as single base-pair substitutions in the coding regions of genes that lead to a nonsynonymous change of the wild-type amino acid encoded by a specific codon. For instance, a single base-pair substitution C->A in the third position of the codon CAC (i.e. CAC->CAA), would change the encoded (wild-type) amino acid histidine to glutamine. Missense mutations could be classified into several different categories with respect to their functional importance. These include neutral, deleterious and beneficial missense variants and missense variants of unknown clinical importance (Chan et al. 2007; Strachan and Read 2004).

4.1.1.1. Deleterious effect of missense mutations

Proteins often contain domains of relatively high functional importance. These domains themselves contain key amino acid residues, responsible for DNA-binding, transactivation, oligomerization, promotion or suppression of cell division, etc. Thus, substitution of these amino acids is likely to alter or abrogate the function of the domain/domains affected. Generally, deleterious missense mutations are defined as missense variants that have a significantly negative impact on the function of a protein as compared to the wild-type product (Carvalho et al. 2009), hence the term deleterious. Various estimations have shown that ~20% of all *de novo* missense substitutions are likely to be strongly detrimental - indeed, these mutations may predispose an individual to a disease state (Kryukov et al. 2007; Yampolsky et al. 2005).

Numerous research groups have performed functional assays to determine if certain missense variants impair either critical regions or the overall function of an affected gene product. For example, certain missense variants in the *BRCA1* gene (e.g. V1833M) have been shown to reduce transactivation activity to ~30%, as compared to the activity of the wild-type protein (Carvalho et al. 2009). Further, a number of missense variants have been shown to negatively affect the function of the *BRCA1* protein (reviewed in Carvalho et al. 2009).

Missense mutations in other well studied human tumour suppressor genes have also been shown to abrogate crucial functions of the affected proteins. For example, these affect the kinase activity of the *ATM* gene (Mitui et al. 2009), the DNA-binding domain of *BRCA2* (Farrugia et al. 2008), the calcium-binding domain responsible for cell-cell adhesion of *CDH1* (Corso et al. 2007), the *CDK*-interacting domains of *CDKN2A* (Ruas et al. 1999), the GTPase-activating protein related domain of *NF1* (Upadhyaya et al. 1997), the phosphatase domain of *PTEN* (Han et al. 2000), the DNA-binding domain of *TP53* (Khromova et al. 2008), the tuberin-binding domain of *TSC1* (Mak et al. 2005), the β domain of *VHL* (Li et al. 2007), and the DNA-binding domain of *WT1* (Little et al. 1995).

Additionally, genes/proteins with several functionally important domains could exhibit deleterious (as defined above) missense mutations in different domains. Such is the case with the human mismatch repair gene *MSH2*, where mutations in the amino-terminal and lever domains affect protein stability whereas mutations in the ATPase domain affect mismatch binding or repair (Ollila et al. 2008).

On the other hand, missense variants could indirectly exert their negative effects on the function of a protein by altering the splicing phenotype. Examples are a germline missense mutation (R141S) which results in the skipping of exon 4 in the *APC* gene (Aretz et al. 2004) and a germline missense mutation (D153Y) in the *CDKN2A* gene which results in an alternatively spliced product, comprising either a 75bp deletion or the complete skipping of exon 2 (Rutter et al. 2003). Both missense variants lead to a nonsynonymous change of the wild-type amino acid, but also affect normal splicing. In the case of *APC* R141S, the observed clinical phenotype and the segregation patterns within families indicate that the altered splicing is disease-causing (Aretz et al. 2004). In the case of D153Y, exon 2 has been reported to be required for nucleolar localisation; therefore exon skipping could potentially have a negative effect on the function of the protein (Rutter et al. 2003; Zhang and Xiong 1999).

4.1.1.2. Neutral effect of missense mutations

Generally, missense variants that have been shown to have no negative impact on the function of proteins are classified as neutral and are likely to be of little clinical importance (Abkevich et al. 2004; Carvalho et al. 2009). These variants are termed neutral because they are associated with little or no disease risk, although the definition of 'no negative impact' varies from study to study. Thus, Carvalho et al. (2009) suggests using a tentative criterion of >50% intact product activity in comparison to the activity of the wild-type protein, in order for a variant to be classified as neutral. By contrast, Mitui et al. (2009) regard a variant as being neutral or 'operationally neutral' if >36% of the activity of the wild-type protein remains intact. Nevertheless, missense variants have been invariably shown to have no or very little effect on the function of a mutant protein and estimations show that ~27% of all *de novo* missense substitutions are effectively neutral (Kryukov et al. 2007; Yampolsky et al. 2005). Analysis of a germline missense mutation (S1613G) in the *BRCA1* gene indicated no change in the activity (i.e. quantitative transcription assay Carvalho et al. 2009) of the mutant product in comparison with the wild-type activity. This substitution has been confirmed to be a neutral polymorphic variant (Friedman et al. 1994; Tavtigian et al. 2006). Similarly, R841W (*BRCA1* gene), Y42C and P655R (*BRCA2* gene) have also been shown to be probably neutral missense variants (Goldgar et al. 2004).

4.1.1.3. Unknown effect of missense mutations

Despite numerous classification procedures, classifying the functional effect of some missense variants remains elusive. For example, Alter et al. (2007) reported 5 different missense variants in the *BRCA2* gene which are of unknown clinical significance (Alter et al. 2007). In addition, other gene products harbouring missense mutations show intermediate activity (*viz.* somewhere between neutral and deleterious), with respect to wild-type activity with the functional consequences not being readily determined (e.g. K1487R in the *BRCA1* gene Carvalho et al. 2009). Moreover, some 50% of all unique variants in the *BRCA1* and *BRCA2* genes are classified as being of unknown effect (Breast Cancer Information Core (BIC) database, http://research.nhgri.nih.gov/bic/; Goldgar et al. 2004) and 13% of the variants detected in 7461 individuals sequenced for the *BRCA1* and *BRCA2* genes are also classified as being of uncertain clinical significance (Frank et al. 2002).

Kryukov et al. (2007) have estimated that ~50% of all *de novo* missense substitutions are mildly deleterious. Mildly deleterious mutations are defined as mutations that are neither neutral nor strongly deleterious (i.e. mutations subject to purifying selection). Therefore, the majority of *de novo* missense mutations described in human genes are effectively of unknown status. The effect of a given missense mutation may be unclear for a number of reasons. In general, there might be insufficient evidence to determine the functional consequences. Some variants could be benign polymorphisms, whereas others might simply co-segregate with known deleterious variants within families.

4.1.1.4. Beneficial effect of missense mutations

Interestingly, some missense mutations in certain genes appear to display a protective effect, or selective advantage, on the individual carrying those mutations. Cellular chemokine receptors act as co-receptors for the entry of pathogens (i.e. *CCR5* for M-tropic strain of *HIV* and *DARC* for the malarial parasite *Plasmodium vivax*; Tamasauskas et al. 2001). A germline missense mutation (R89C) in the *DARC* gene results in a reduced level of the associated protein (i.e. *DARC*-negative) and confers resistance to infection by the malarial parasite (Miller et al. 1976; Pogo and Chaudhuri 2000; Tamasauskas et al. 2001). Similarly, the R60S germline missense change in the *CCR5* gene reduces the ability of *HIV-1* entry (72% entry efficiency compared to that of a wild-type product Tamasauskas et al. 2001) and has been shown in an unaffected yet *HIV-1*-exposed individual (Carrington et al. 1997). Low frequency germline missense mutations in the *CHEK2* gene (e.g. I157T) have been shown to be associated with a significantly lower incidence of lung cancer (Brennan et al. 2007; Cybulski et al. 2008), although the underlying mechanism remains unknown. Likewise, a germline missense variant (S408N) in the *CSNK1E* gene, which plays an important role in the regulation of circadian clock rhythms, has been found with reduced frequency in cases of Delayed Sleep Phase Disorder, as compared with healthy controls, but also has much higher functional activity than the wild-type protein (Takano et al. 2004). Takano et al. (2004) have speculated that the aforementioned allele may play a protective role in the development of Delayed Sleep Phase Disorder.

4.1.2. The challenge of classifying the functional consequences of missense mutations

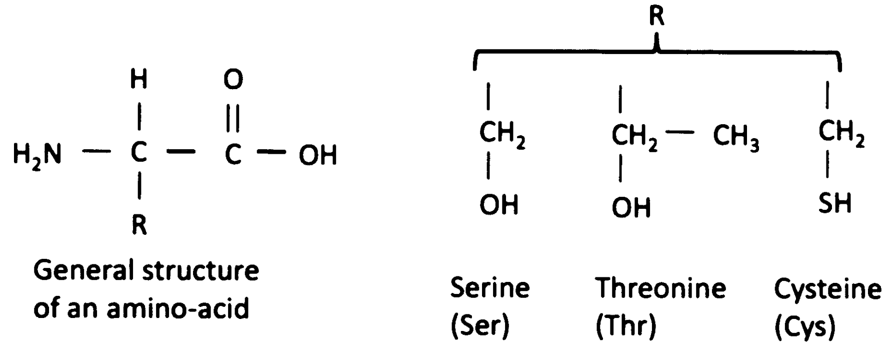
The optimal means of identifying the functional consequences of missense mutations is a reliable *in vitro* functional assay that would measure not only the activity, but also the properties of the mutant product (i.e. the protein harbouring the missense mutation), long-term effects and interaction with other gene products. For a limited number of disease-related genes, such functional assays exist, but for many others they are costly to construct, unreliable or difficult to perform. Even for a relatively small gene, such as *CDKN2A* (156 codons in total), there are 2945 possible missense variants, but for only 100 (<3%) of these have functional assays been performed (Chan et al. 2007). Therefore, in the absence of *in vitro* assays, a variety of *in silico* algorithms have been developed to aid in the classification of missense mutations. Computational methods are relatively cheap and allow for an unlimited number of variants to be tested. These methods rely on various sources of data including the biochemical and physicochemical properties of amino acids; known secondary and tertiary structure of affected proteins; evolutionary conservation data and mutation rates.

4.1.2.1. Biochemical and physicochemical properties of amino acids

Some amino acids are very similar to each other with respect to their chemical and/or physicochemical composition, whereas others are extremely different. One of the most widely used scores for measuring the chemical and physical differences between amino acids, is the so called 'Grantham score' or 'Grantham difference' (Grantham 1974). This measure describes the difference between the side chain composition (i.e. weight ratio of noncarbon components in end groups or rings to carbons in side chains), polarity (i.e. basic, acidic or nonpolar depending on the side chain charge) and molecular volume of two amino acids. Even although numerous other measures of amino acid difference have been devised (Clarke 1970; Epstein 1967; Miyata et al. 1979), the Grantham difference is useful because it is a continuous measure and most importantly helps to quantify the 'severity of amino acid changes' (Miller and Kumar 2001).

Some amino acid residues play a crucial role in proteins and therefore could not be easily substituted by others without drastically altering protein structure and/or function. For instance, cysteine could form disulphide bridges and plays an important role in formation of the secondary structure of proteins (Grantham 1974). In addition, cysteine is also a unique amino acid, as it is the only one with a sulfhydryl group in its side-chain. By contrast, isoleucine and valine, or serine and threonine, have very similar side chains (shown as R in Figure 9); therefore the physicochemical difference between them is one of the lowest (Grantham 1974). The chemical composition of the above-mentioned amino acids is shown in Figure 9.

Figure 9 Chemical composition of some amino acids (adapted from Strachan and Read 2004)



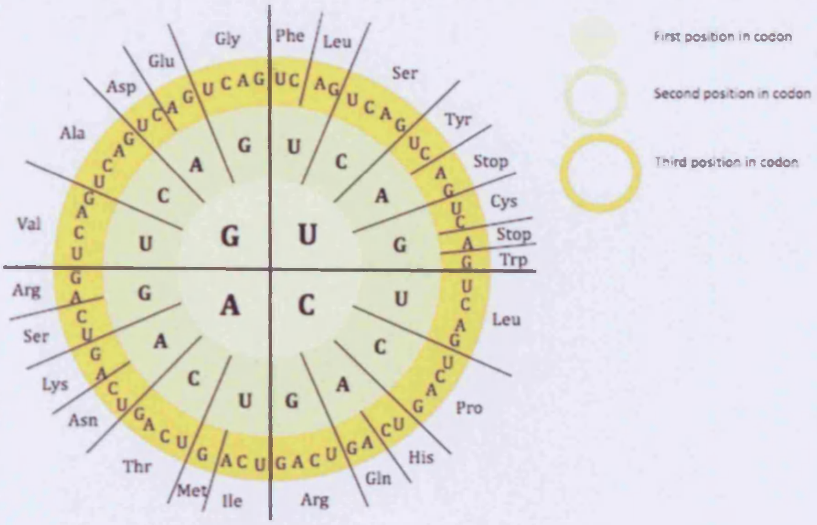
Although Grantham differences range from 5 to 215, Tavtigian et al. (2008) have suggested that Grantham differences of 5-60 are to be considered 'conservative', 60-100 'non-conservative' and >100 'radical'. Disease-associated germline missense mutations in 7 human genes (*CFTR*, *TSC2*, *G6PD*, *L1CAM*, *PAH*, *RS1*, *PAX6*) exhibit a greater average chemical difference than the average difference observed in interspecific comparisons of missense changes in orthologous proteins (Miller and Kumar 2001). In addition, Miller and Kumar (2001) have also noticed fewer 'radical' disease-associated amino acid changes (i.e. Grantham difference >100), than 'non-conservative' (Grantham difference 60-100). A similar result has been reported by Notaro et al. (2000) for the *G6PD* gene. These authors suggested that 'radical' mutations are likely to be lethal, hence relatively fewer radical changes are observed. Thus, there is strong, negative purifying selection pressure acting on those 'radical' amino acid changes. Furthermore, Krawczak et al. (1998) have shown that the Grantham difference is positively correlated with a measure which they termed the 'relative clinical observation likelihood'; thus, germline missense changes that give rise to a greater difference in terms of chemical composition are more likely to come to clinical attention.

These results indicate that non-conservative (with respect to the amino acid difference between the wild-type and mutant amino acids) disease-associated mutations are more likely to lead to an observed disease phenotype in patients, in comparison to less radical changes.

4.1.2.2. Relative mutability rates

DNA in living organisms comprises 4 basic nucleotides (viz. A, C, G and T). At the mRNA level, a set of three nucleotides forms a codon, which encodes a specific amino acid. Strings of amino acids (i.e. amino acid sequences or polypeptides) define proteins. As a consequence, there are 4^3 (i.e. 4 possible nucleotides in each of the three positions in a codon), or 64 different codons (i.e. different set of trinucleotides), but only 20 standard amino acids. Therefore, different codons can encode the same amino acid, known as codon degeneracy (shown in Figure 10).

Figure 10 Codon degeneracy (adapted from Ellington and Cherry 2001)



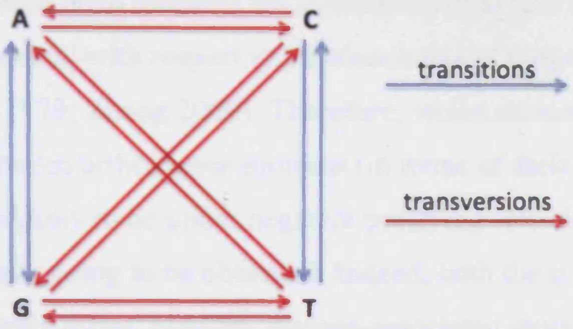
For example, the codons CGA, CGC, CGG and GCU all encode the same amino acid, arginine. In addition, the last position of the CG- codon is said to be fourfold degenerate, as all possible nucleotides (viz. A, C, G and T) at this position encode the same amino acid (i.e. synonymous substitutions). There are also twofold (i.e. when 1 out of the 3 possible substitutions is a synonymous change) and non-degenerate sites (i.e. when all 3 possible substitutions are non-synonymous). Interestingly, for some amino acids that are encoded by more than one codon, there is a bias in the usage of synonymous codons (Irwin et al. 1995; Tats et al. 2008). As a result, some codons are preferred over others. Studies have suggested that preferences in the use of the genetic code may be due to different rates of translation efficiency and accuracy (Bossi and Ruth 1980; Irwin et al. 1995; Stormo et al. 1986).

The physicochemical differences, the design of the genetic code, and codon usage differentially affect amino acid mutability (defined as relative rates of amino acid substitutions). Thus, some substitutions will be observed more frequently than others. The relative rate of mutability in an evolutionary context has been defined as the rate of change of an amino acid in a pair of aligned sequences (i.e. the number of changes divided by the total number of occurrences of a particular amino acid (Collins and Jukes 1994; Dayhoff et al. 1978). These relative rates of substitution are usually portrayed as substitution matrices, such as PAM (Dayhoff et al. 1978) and BLOSUM (Henikoff and Henikoff 1992). The term PAM matrix stands for Point Accepted Mutation and measures the probability or rate of substitution of one amino acid by another over time. It is calculated from the sequence alignments of closely related protein families. Thus, PAM1 is 1 substitution per 100 amino acids or 1% and is usually used for closely related sequences. For othologues from more distantly related species, PAM matrices are extrapolated from PAM1 matrix, by multiplying PAM1 by itself. On the other hand, BLOSUM matrix (Henikoff and Henikoff 1992) stands for BLOcks of amino acid SUbstitution Matrix and is mainly used for relatively divergent sequences. In contrast to PAM matrices, BLOSUM is derived from local sequence alignments (i.e. blocks of protein alignments without gaps), without extrapolation. Numerous BLOSUM matrices have been devised, according to the relatedness of the sequences used, e.g. BLOSUM80 for relatively closely related sequences and BLOSUM45 for more divergent sequences and the number represents clustering of the blocks at certain percentage level. Others have used a different approach to calculate substitution rates. For example, Hess et al. (1994) have shown a strong neighbour-dependent bias of the substitution rates using ~20,000 point substitutions in aligned human gene/pseudogene sequences. In addition, the substitutions rates estimated by Hess et al. (1994) have been derived from sequences that are no longer under evolutionary pressure.

Despite the differences in the methods used in calculating substitution rates, it is clear that relative substitution rates are not uniformly distributed between different amino acid changes. Estimates show that, in an evolutionary context, one of the least mutable amino acids is cysteine, whereas serine and threonine are among the most mutable ones (Collins and Jukes 1994). The latter findings are not surprising bearing in mind that cysteine, as described above, is the only amino acid with a sulfhydryl group and serine and threonine are quite similar, with respect to their chemical composition.

At the nucleotide level, transitions are single base-pair substitutions of a pyrimidine for another pyrimidine (T⇔C) or a purine by a purine (G⇔A), while transversions are substitutions of a pyrimidine by a purine and *vice versa* (depicted in Figure 11).

Figure 11 Transitions and transversions (after Strachan and Read 2004)



As shown in Figure 11, there are twice as many possible transversions than transitions. Therefore, based purely on their frequencies, transversions are expected to be twice as frequent compared with transitions. However, in an evolutionary context, studies have shown quite the opposite, with transitions being found more frequently observed than transversions. Collins and Jukes (1994) have calculated that the ratio of transversions to transitions for nonsynonymous changes is 1.2 as compared to 2.4 expected from the genetic code. Thus, there is an excess of transitions over transversions (almost 2 times), even with a correction for the use of genetic code. To a large extent, this transitional bias is due to the spontaneous deamination of 5mC in the context of CpG dinucleotides, resulting in C->T (coding DNA strand) and G->A transitions (non-coding DNA strand) in a CpG dinucleotide context (Coulondre et al. 1978; Grippo et al. 1968). As is evident from numerous studies, this spontaneous deamination of 5mC is also largely responsible for a highly increased mutation rate at CpG dinucleotides (Cooper and Youssoufian 1988; Gaffney and Keightley 2008; Krawczak et al. 1998).

Other authors have derived mutability rates from disease-associated single-base pair substitutions. Thus, Krawczak et al. (1998), employed single-base-pair substitutions associated with inherited disease that were logged at that time in the Human Gene Mutation Database (*HGMD,* Stenson et al. 2003). These substitution rates therefore represent mutability rates associated with inherited disease.

Comparison of relative mutability rates derived from disease-associated mutations with those derived from the interspecific comparison of orthologous protein sequences, could indicate patterns of specific amino acid exchanges or the severity of the amino acid exchanges associated with disease. It is expected that on average the physicochemical difference of amino acid exchanges, over evolutionary time, will have been relatively small. By contrast, disease-associated mutations are expected to have been much more drastic with respect to the physicochemical differences of the amino acids involved. Indeed, a number of studies have revealed the similarities of amino acid substitutions between orthologous proteins, with respect to physicochemical properties (Clarke 1970; Epstein 1967; Miyata et al. 1979; Zhang 2000). Therefore, whilst amino acid substitutions will be very similar between orthologous proteins (in terms of their physicochemical properties), drastic changes are likely to be under negative purifying selection (Miller and Kumar 2001) and hence are rarely going to be observed. Indeed, both the type and frequency of amino acid exchanges greatly differ, between disease-associated mutations and the exchanges observed between orthologous proteins (Miller and Kumar 2001). Most importantly, disease-associated changes are more radical overall than changes observed in orthologous sequences (Miller and Kumar 2001; Vitkup et al. 2003). In contrast to the strong nearest neighbour-dependent substitution rates reported by Hess et al. (1994) for evolutionary substitutions, disease-associated substitution rates exhibit a very limited nearest neighbour effect (Krawczak et al. 1998).

4.1.2.3. Evolutionary data

DNA sequence is said to be evolutionarily conserved if the orthologous sequence is similar or nearly identical in multiple organisms. Under natural selection (genetic variations that confer an advantage or disadvantage upon the organism in terms of its ability to survive and reproduce), changes in the DNA sequence (e.g. mutations) would be neutral or nearly neutral (i.e. silent mutations), deleterious or advantageous. Mutations are said to be neutral if they neither confer an advantage nor a disadvantage to an organism. Some amino acid residues frequently vary between orthologous proteins, indicating that they are tolerated by natural selection and might be under less stringent selection pressure (Miller and Kumar 2001). On the other hand, some amino acid residues are virtually invariant (i.e. they exhibit a high degree of evolutionary conservation), when orthologous sequences are compared in different species. Many authors have suggested that some amino acid residues could play a relatively more important role than others, with respect to protein function. Assuming that such sites are susceptible to mutation, the fact that they are found to be virtually invariant among different species suggests that mutations at these sites might exert a detrimental effect on the function of the protein product. Therefore, these sites will have been under negative selection pressure via natural selection. Nevertheless, evolutionarily conserved sites are not necessarily under strong purifying selection pressure. Indirect evidence comes from a phenomenon, termed 'pseudogeneralization' (Wang et al. 2006), which represents the loss of a gene (through deactivating mutation and independently from other species) during human evolution, since divergence from the chimpanzee lineage. This gene loss has been the basis for the 'less is more' hypothesis (Olson and Varki 2003). It suggests that the loss of specific genes in humans, since divergence from the chimpanzee lineage, may have allowed brain size expansion (Stedman et al. 2004). Thus, Stedman et al. (2004) have suggested that loss of masticatory muscle strength may have relaxed the evolutionary constraints on encephalisation. In addition, a nonsynonymous mutation that leads to sickle cell anaemia confers resistance to malaria in heterozygotes (described in more detail in 4.1.1.4). Thus, even though some mutations at evolutionarily conserved sites confer a negative effect on the function of the protein, they may nevertheless be tolerated by natural selection.

Nevertheless, it is commonly accepted that drastic amino acid exchanges (in terms of the physicochemical difference between substituted and substituting amino acids) during the evolution of species, would be subject to strong negative selection. In addition, studies suggest that the majority of changes are neutral or nearly neutral with respect to selection pressure (Kimura 1991; Kimura and Ota 1974). There is evidence to show that drastic amino acid changes are depleted in the genomes of higher organisms (Kimura 1991) and that the majority of changes are physicochemically similar or effectively neutral. On the other hand, disease-associated missense changes are much more drastic. Significantly more disease-associated mutations are observed at invariant or highly conserved amino acid positions than would be expected by chance alone (Abkevich et al. 2004; Miller and Kumar 2001; Walker et al. 1999).

During tumour development, pathological amino acid changes occurring in the soma are generally considered to abrogate the function of tumour suppressor genes (Tavtigian et al. 2008), whereas gain-of-function mutations are associated with oncogenes, such as the *KRAS* and *HRAS* genes (Schubbert et al. 2007). Tumour suppressor genes are responsible for key processes, such as response to DNA damage (e.g. *ATM*), inhibit cell proliferation (e.g. *TP53*), responsible for DNA repair (e.g. *BRCA1, BRCA2*), etc. (Sherr 2004). Therefore, tumour development and progression require the elimination of key tumour suppressor genes. Nevertheless, some tumour suppressor genes have been suggested to 'evolve' via positive selection, during tumorigenesis (Glazko et al. 2006; Glazko et al. 2004). Since, the majority (>80%) of tumour-associated *TP53* sequence changes are missense mutations (The p53 database, http://p53.free.fr/ Soussi and Beroud 2001) and a significant excess of non-synonymous mutations is observed as compared to neutral expectations (Glazko et al. 2006), the mutant *TP53* gene could acquire new functions during tumour development. Indeed, various studies have suggested that, unusually, *TP53* is not a simple tumour suppressor gene, but may also possess some properties of an oncogene (Blagosklonny 2000; Pugacheva et al. 2002). In other words, during tumour development, at least for the *TP53* gene, there is a preferential fixation of missense mutations over nonsense or silent mutations. Thus, nonsense mutations, which are generally considered to abrogate gene function, are likely to be eliminated, possibly via negative selection, whereas missense mutations are preferentially acquired and may result in gains of function (i.e. oncogenic properties). Despite nonsense mutations being generally considered to abrogate gene function, several studies have shown that mutant *TP53* proteins retain specific wild-type functions, such as an ability to induce apoptosis (Rutherford et al. 2002) or are as abundant as the wild-type protein (Anczukow et al. 2008). Thus, considerable variation could exist between different nonsense mutations, with respect to functional consequences. As mentioned above, in general the majority of pathological missense mutations are found at evolutionarily conserved positions. The *TP53* gene is no exception to this rule. These key positions are located in several functionally important domains responsible for DNA binding, conformation, transactivation and tetramerization (Glazko et al. 2004; Joerger and Fersht 2008). Thus, these observations strongly suggest that missense changes in the *TP53* gene at key amino acid positions, could promote tumour development or progression, through the acquisition of new gene functions. Similar findings, although with a relatively smaller effect, has been reported for other tumour suppressor genes, such as the *BRCA1, BRCA2* and *CDKN2A* (Glazko et al. 2006).

4.1.2.4. Hotspot analysis

The mutational spectra of both the soma (i.e. tumours) and the germline contain examples of mutations, which frequently re-occur at particular positions (i.e. hotspots) in a number of different tumour suppressor genes. This observation suggests that mutations are not randomly distributed along the gene sequence, but rather can occur at hotspots due to the action of both exogenous mutagens and endogenous mutational mechanisms. Carcinogens are responsible for some of the mutational hotspots (Besaratinia and Pfeifer 2006). For example, sunlight (UV light) and aflatoxin B(1) exposure are both associated with specific mutational spectra in the *TP53* gene in skin and liver cancer respectively (Pfeifer et al. 2005). UV irradiation is usually characterized by C->T or CC->TT transitions at dipyrimidine sites (Drobetsky et al. 1994; Sage et al. 1996), whereas aflatoxin exposure is strongly associated with G->C and T->A transversions, predominantly in CpG dinucleotides (Besaratinia et al.

2009; Hussain and Harris 1999). It has however also been shown that endogenous mechanisms also play an important part in shaping the mutational spectra associated with tumour development. These endogenous mechanisms include methylation-mediated deamination of 5-methylcytosine in CpG dinucleotides, slippage of the complementary DNA strands, post-replicative mismatch repair, exonucleolytic proofreading mechanisms, etc (Krawczak et al. 1998; Wells et al. 2005).

Furthermore, only 19% (73 hotspots) of *TP53* codons account for 88% of all reported *TP53* mutations whilst just 6 codons account for 25% of all somatic point mutations in this gene (Walker et al. 1999). Walker et al. (1999) have reported that these hotspots are situated at evolutionarily conserved codons thereby indicating a relationship to functionally important amino acid residues.

4.1.3. Accuracy of existing methods to classify the functional consequences of missense variants

In the absence of reliable functional assays to determine the functional importance of missense variants, numerous *in silico* methods and algorithms have been proposed. These methods take into consideration important factors that contribute to, or play an integral part in mutagenesis. In order to predict the pathogenicity of missense mutations, these methods rely on evolutionary conservation, based on multiple sequence alignments, amino acid physical and chemical composition, structural properties of wild-type and mutant proteins, amino acid substitutions matrices, nucleotide mutability rates, etc (Tavtigian et al. 2008). The prediction accuracy of these algorithms is relatively high. For three algorithms (SIFT, PolyPhen and A-GVGD), Chan et al. (2007) have reported an overall prediction accuracy of 73-82% when missense variants are scored with each of the programs. Similarly, several other tools reach an accuracy of prediction which ranges from 75% to 95% (Tavtigian et al. 2008). Further, the combination of the three algorithms tested by Chan et al. (2007) increases the overall predictive ability to ~88% and up to ~96% for mutations at invariant amino acids with respect to evolutionary conservation.

Comparison of these methods indicates that all algorithms, which use evolutionary conservation, are superior to those methods, which only use structural information (Chan et al. 2007). In addition, predictions are less accurate when the degree of evolutionary conservation is not used (Goldgar et al. 2004).

Even although these methods achieve a relatively high degree of prediction accuracy, the parameters used are only proxies for the quantity of interest, i.e. pathogenicity (Kryukov et al. 2007; Tavtigian et al. 2008). Most of these algorithms rely on training sets, usually based on recurring mutations (i.e. hotspots). However, there are some indications that the majority of rare missense alleles (usually not included in the training sets) are in fact deleterious (Kryukov et al. 2007). Furthermore, even although these rare missense alleles may be subject to purifying selection (i.e. since they are likely to be deleterious), some could predispose to disease. Indeed, if the 'common disorder, rare allele' hypothesis (Kryukov et al. 2007) turns out to be correct, some of these deleterious alleles may play an important role in the development and/or initiation of a disease phenotype. Nevertheless, there is no single test that is capable of achieving 100% accuracy and unequivocally determining the pathogenicity and functional importance of missense variants. Thus, further effort is required to improve the predictive accuracy of these algorithms.

4.1.4. Could the comparison of somatic and germline mutational spectra help to improve the accuracy of pathogenicity prediction?

It is already known that the mutational spectra of both the soma and germline exhibit similarities, but also differences. For example, in response to ionizing radiation dose, both the germline and the soma show similar damage rates (Vilenchik and Knudson 2000). On the other hand, it has been shown that the germline exhibits extreme minisatellite instability, whereas this is rarely observed in the soma (Buard et al. 2000).

Research suggests that an interaction between germline and somatic mutations could play an important role in the aetiology of familial adenomatous polyposis (Latchford et al. 2007). This interaction is evident from the observation that germline mutations can predict or even direct the type and position of subsequent somatic mutations with respect to tumour development. In addition, germline inherited susceptibility not only has the potential to influence the somatic mutation rate directly, but could also confer a stronger selective advantage upon the cells. Hence, these cells might be more likely to undergo clonal expansion and subsequent tumour development. Evidence for such interplay between the soma and the germline comes from the analysis of myeloproliferative neoplasms (Campbell 2009). These neoplasms are associated with somatic mutations in the *JAK2* gene. Three independent studies have shown that these somatic mutations are preferentially acquired within a particular inherited haplotype within the *JAK2* gene (Jones et al. 2009; Kilpivaara et al. 2009; Olcaydu et al. 2009). These authors proposed two hypotheses to account for the preferentially acquired somatic mutations. Firstly, inherited variants could confer a selection advantage and secondly, the inherited variants could promote an increased somatic mutation rate. Additional research is required to ascertain which hypothesis is valid (i.e. selective advantage or increased mutation rate), but these findings clearly put the differential selection advantage of somatically acquired mutations with respect to inherited variants in perspective. To speculate further, if the stronger selection advantage hypothesis is valid, a somatic mutation might or might not promote tumour development, based on inherited (i.e. germline) mutation or variation (e.g. a specific haplotype). Therefore, interplay between germline and somatic variants could well be very important when assessing the functional importance of missense variants.

Based on the importance of somatic and germline mutational spectra with respect to tumour development, it is quite surprising that relatively few studies have attempted to compare and contrast mutational spectra in the soma and the germline. Potential differences in the mutational mechanisms operating in the germline and the soma could influence the overall accuracy of any prediction algorithm, with respect to the pathogenicity of missense variants. On the other hand, different selection constraints could also distort the overall prediction accuracy. Therefore, studying mutational mechanisms, with regard to the germline and soma, should not only serve to contribute substantially to improving our understanding of tumour development, but could also help to improve the accuracy of *in silico* algorithms to predict the pathogenicity of missense variants.

4.1.5. Aims of the analysis

The main objectives of the analysis here were to explore any similarities or differences that somatic and germline missense mutations might exhibit with respect to nucleotide substitution rates derived from disease-associated mutations and nucleotide substitution rates derived from non-disease-associated mutations. The main objectives also included exploring similarities and differences between the soma and germline with respect to amino acid physical and physicochemical differences and the degree of evolutionary conservation. To accomplish these objectives, a number of tasks were performed and a number of parameters or properties were addressed in this analysis.

- **Definition and calculation of potential missense mutations**

For every tumour suppressor gene examined, all possible single base-pair substitutions were calculated (described in 4.2.2.5). Those mutations that were not part of the observed somatic and germline mutational spectra were termed 'potential missense mutations'. As the potential missense mutations in these particular tumour suppressor genes have not so far been associated with any disease phenotype, they were used as a "control set" to draw inferences about the non-randomness of various parameters observed in somatic or/and germline missense mutations. The following parameters were assessed:

- **Non-disease-associated single base-pair substitution rates**

For every missense mutation, observed or potential, non-disease-associated nucleotide substitution rates were derived from Hess et al. (1994).

- **Disease-associated single-base pair substitution rates**

For every missense mutation, observed or potential, disease-associated relative nucleotide substitution rates were derived from Krawczak et al. (1998).

- **Degree of evolutionary conservation**

In order to estimate the degree of evolutionary conservation at every codon in each gene, orthologous gene sequences were derived from a number of vertebrate species. They were used to produce codon-by-codon multiple sequence alignments. These alignments, allowed the estimation of the degree of evolutionary conservation at each amino acid position.

- **Degree of physical and physicochemical difference of amino acid substitutions**

For every amino acid substitution (observed or potential), the physicochemical difference between wild-type and mutant amino acids, calculated by Grantham (Grantham 1974), were used.

These tasks were performed to provide meaningful answers to the following questions:

Are there any differences/similarities between somatic, germline, shared and potential missense mutations for **each tumour suppressor gene and all genes** combined in terms of disease and non-disease-associated nucleotide substitutions rates, evolutionary conservation or physicochemical difference?

Is there any difference/similarity between the combination of somatic, germline and shared missense mutations (henceforth called observed mutations) and potential missense mutations for **each tumour suppressor gene and all genes** combined in terms of disease and non-disease associated nucleotide substitutions rates, evolutionary conservation or physicochemical difference?

Are there any differences/similarities between CpG- and non-CpG located missense mutations between somatic, germline, shared and potential missense mutations for **each tumour suppressor gene and all genes** combined?

Are there any differences/similarities between CpG- and non-CpG located missense mutations between observed and potential missense mutations for **each tumour suppressor gene and all genes** combined?

4.2. Materials and methods

4.2.1. Materials

4.2.1.1. General definition of missense mutations

A missense mutation is defined as a single base-pair substitution responsible for the non-synonymous change of the wild-type amino acid encoded by a specific codon. This definition excludes single base-pair substitutions that lead to the introduction of a stop codon within the coding region of a gene (these are termed nonsense mutations). Employing this definition, there are 392 different possible single-base pair substitutions in 61 codons that could lead to a missense mutation. These single base-pair substitutions are listed in Table 8.

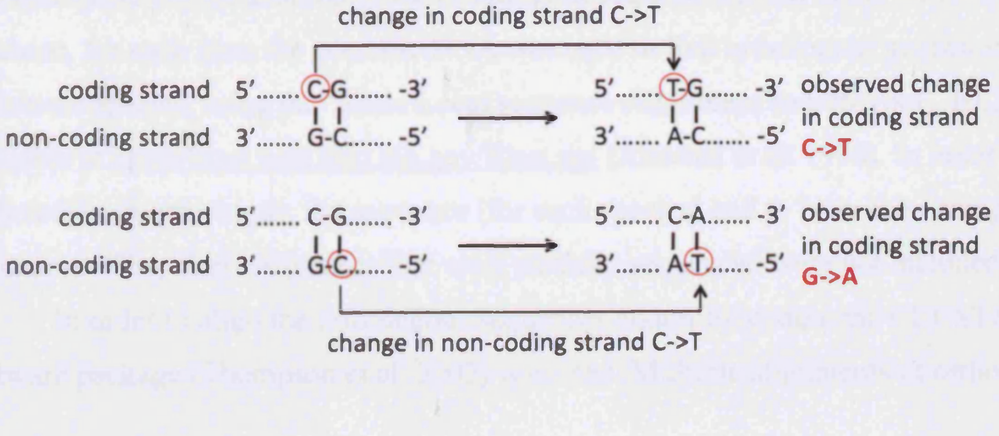
4.2.1.2. Labelling of somatic, germline and shared missense mutations

For detailed description of labelling of mutations, see 3.3. A summary of the studied missense mutations is given in Table 9 and Table 11.

4.2.1.3. General definition of single base-pair substitutions generating missense mutations in CpG dinucleotides

CpG missense mutations were defined as C->T transitions found in the context of CpG dinucleotides. In addition, all observed missense mutations were logged according to the coding strand of DNA. Hence, single base-pair substitutions in CpG dinucleotides on the non-coding DNA strand would appear as G->A transitions (as shown in Figure 12).

Figure 12 Single base-pair substitutions in CpG dinucleotides


change in coding strand C->T
coding strand 5'.......C-G....... -3'
non-coding strand 3'....... G-C....... -5'
5'.......T-G....... -3'
3'....... A-C....... -5'
observed change in coding strand C->T
coding strand 5'....... C-G....... -3'
non-coding strand 3'....... G-C....... -5'
5'....... C-A....... -3'
3'....... A-T...... -5'
observed change in coding strand G->A
change in non-coding strand C->T

4.2.2. Methods

4.2.2.1. Identification of potential missense mutations

The identification of potential missense mutations and potential missense mutations in CpG-dinucleotides was accomplished as described in 3.5.

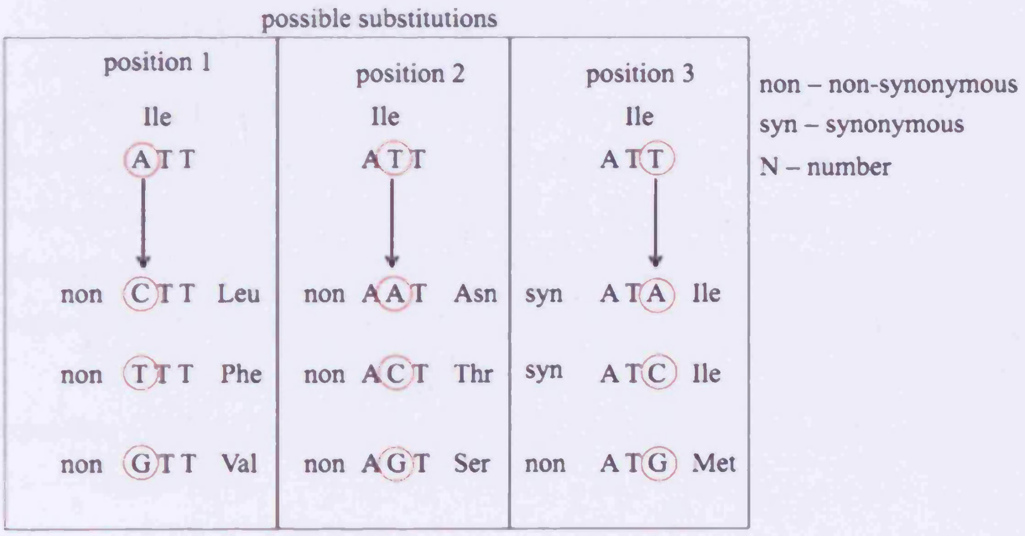
4.2.2.2. Calculation of degree of evolutionary conservation

In order to assess the degree of evolutionary conservation at the codon level for each of the 17 tumour suppressor genes under study, multiple sequence alignments were required. For each gene, several sequence orthologues from vertebrate species were obtained (cDNA and protein sequence). Detailed information on species and sequences used is given in Table 10. Research suggests that the disease-associated mutations tend to occur at evolutionarily conserved sites (Abkevich et al. 2004; Miller and Kumar 2001; Tavtigian et al. 2008; Walker et al. 1999). In the majority of cases when a disease-associated mutation is found at an evolutionarily variable position, substitutions have occurred in phylogenetic lineages least related to humans (Miller and Kumar 2001). While inclusion of a wide variety of species could protect from chance variation (i.e. an amino acid position may be evolutionarily invariant due to the limited number of sequences used Tavtigian et al. 2008), amino acid variation may be correlated with functional differences of the associated products (Miller and Kumar 2001). Therefore, only orthologous sequences from vertebrate species were used. Orthologous sequences were retrieved from NCBI's Entrez Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene Maglott et al. 2005). The aim was to include as many orthologous sequences from vertebrate species as possible. At the time of conducting the analysis (beginning of 2006), there were on average 6-7 vertebrate species available for each of the 17 studied human tumour suppressor genes (detailed information is given in Table 10). In addition to the orthologous sequences listed in the NCBI's Entrez Gene database, for each gene the genomic DNA was used to find orthologous sequence from vertebrate species, using the 'Basic Local sequence Alignment Search Tool', BLAST, available at http://blast.ncbi.nlm.nih.gov/Blast.cgi (Altschul et al. 1990). In order to be included as an orthologue, the sequence (for each species) had to have been completely sequenced. Therefore, sequences that were partially sequenced were not included.

In order to align the orthologous sequences codon-by-codon, the CLUSTALX software package (Thompson et al. 2002) was used. Multiple alignments of orthologous protein sequences, for each gene, were generated using the default parameters in CLUSTALX. In order to generate cDNA codon-by-codon multiple sequence alignments, I created a computer program that took the output of CLUSTALX (i.e. multiple protein orthologous sequence alignments) and consecutively, for each of the codons and for each species, generated multiple orthologous cDNA sequence alignments. These multiple cDNA sequence alignments were then used to assess the evolutionary constraints at codon level.

In order to estimate the evolutionary constraints acting upon each of the 17 human tumour suppressor genes and to estimate how fast each gene has been evolving, 4 different algorithms were employed. The most common measure used to estimate sequence diversity is the *Ka*/*Ks* ratio (Yang and Bielawski 2000). *Ka* represents the number of non-synonymous substitutions per non-synonymous site and *Ks* is the number of synonymous substitutions per synonymous site. Each site (i.e. position 1, 2 or 3 within a codon) could thus be synonymous, non-synonymous, or partially non-synonymous. A comprehensive example is given in Figure 13.

Figure 13 Example of synonymous and non-synonymous sites

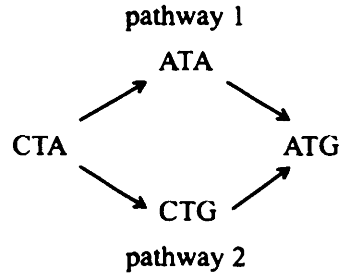


position	1	2	3		
synonymous site	0	0	2/3	⟶	total N synonymous sites 2/3
non-synonymous site	1	1	1/3	⟶	total N non-synonymous sites 2⅓

Generally, the algorithms for calculating the *Ka*/*Ks* ratio contain the following steps. The numbers of synonymous and non-synonymous sites are counted, then for each pair of aligned (codon-by-codon) sequences, the number of synonymous and non-synonymous changes is determined. If two aligned codons differ by more than 1 substitution, depending on the number of different substitutions (i.e. 1, 2 or 3), there are 2 or 6 different pathways to account for the substitutions. An example is given in Figure 14.

Figure 14 Example of number of synonymous and non-synonymous differences between two aligned codons

sequence 1	CTA
sequence 2	ATG



	Leu	Ile	Met	
pathway 1	CTA →	ATA →	ATG →	2 non-synonymous changes
pathway 2	CTA →	CTG →	ATG →	1 synonymous,1 non-synonymous changes
	Leu	Leu	Met	

Number synonymous differences (s_d) 1/4 x 2 = 1/2

Number non-synonymous differences (s_n) 3/4 x 2 = 3/2

The proportions of synonymous and non-synonymous differences are then calculated from the total number of synonymous and non-synonymous differences. The numbers of synonymous substitutions per synonymous site, *Ka* and the number of non-synonymous substitutions per non-synonymous site *Ks* are then estimated, by using for example the Jukes-Cantor formula (Jukes and Cantor 1969).

Equation 4 Jukes-Cantor formula

$$d = -\frac{3}{4}\log_e(1 - \frac{4p}{3})$$

$$p = \frac{n_d}{n}$$

where n_d - number of nucleotide differences
n - total number of nucleotides compared

The algorithm, described above is the Nei and Gojobori (1986) unweighted pathway method for estimating synonymous substitutions. A plethora of methods and algorithms exist to account for multiple substitutions at two aligned codons, which take into account the transition/transversion bias, unequal base frequencies, varying substitution rates among sites, substitution rates between sites and substitution patterns among lineages (Kumar et al. 2004). As mentioned above, four of the most common methods were used to estimate, for each gene, the rate of synonymous and non-synonymous substitutions per site. These included the Li-Wu-Luo (Li et al. 1985) and Pamilo-Bianchi (Pamilo and Bianchi 1993) codon models, both part of the MEGA software package (Kumar et al. 2004). In addition, the Nei and Gojobori (1986) codon model and the Goldman and Yang (1994) maximum likelihood method were also used, both part of the PAML (CODEML program) software package (Yang 1997). For each of the methods, both software packages produce pair-wise Ka/Ks ratios between the orthologous sequences from two species. In order to calculate the overall Ka/Ks ratio for all pair-wise comparisons (i.e. for each gene, and for a pair of aligned codon-by-codon orthologous sequences), the average Ka/Ks ratio was taken for all pair-wise comparisons.

In order to estimate the evolutionary constraints acting at the codon level, a method described in Walker et al. (1999) was used. The program SUBROLL part of the SEALS software package (Walker and Koonin 1997) that was used in Walker et al. (1999), was not available. Therefore I created a program according to the method described in Walker et al. (1999) to estimate evolutionary constraints at the codon level. The program utilizes orthologous cDNA sequences, aligned codon-by-codon as described above. In a pair-wise fashion (i.e. comparing only two sequences at a time), for all combinations of pairs of aligned orthologous cDNA sequences and consecutively for each pair of codons, the numbers of synonymous (Ks) and non-synonymous (Ka) differences were counted. The pathway method of Nei and Gojobori (1986) was used to create all possible pathways between two codons. If two aligned codons differed by more than one substitution, the minimum number of substitutions was assumed and the most favourable path was determined using a PAM100 matrix. Gaps in the cDNA sequences of non-human species were treated as being equivalent to a non-synonymous substitution. Codons that were not present in human cDNA sequence were not taken into consideration. As a result, the evolutionary constraints acting upon the 17 human tumour suppressor genes at codon level were inferred by calculating $Ka/(Ka+Ks)$. The measure corrects for the fact that some substitutions occur at degenerate codons and do not change the amino acid.

4.2.2.3. Nucleotide substitution rates

Nucleotide substitution rates were taken from Hess et al. (1994) and Krawczak and Cooper (1991). Both substitution rates take into account the influence of nucleotide context (e.g. frequency of nucleotide triplets) on mutational bias. Thus, these nucleotide substitution rates take into consideration the effect of adjacent nucleotides, either side of the reported single base-pair substitution.

4.2.2.4. Amino acid difference between wild-type and mutant amino acids

For each amino acid change in all datasets (i.e. somatic, germline, shared and potential) and for every gene, a value corresponding to the amino acid difference between the wild-type and mutant residues, was assigned according to Grantham (Grantham 1974).

4.2.2.5. Comparisons and calculation of statistical significance

In order to answer the questions posed in the Aims of this analysis (Section 4.1.5), the following tests for each gene were performed:

Soma vs. potential (simulated spectra) missense mutations
Germline vs. potential (simulated spectra) missense mutations
Shared vs. potential (simulated spectra) missense mutations
Observed (the combination of numbers of somatic, germline and shared mutations) vs. potential (simulated spectra) missense mutations
Soma vs. germline missense mutations
Soma vs. shared missense mutations
Germline vs. shared missense mutations
Recurrent somatic vs. non-recurrent somatic missense mutations
Recurrent somatic shared vs. recurrent somatic non-shared missense mutations
Non-recurrent somatic shared vs. non-recurrent somatic non-shared missense mutations
Recurrent somatic shared vs. non-recurrent somatic non-shared missense mutations

In addition, the numbers of mutations in all genes were combined only if missense mutations possessed the same label (i.e. somatic, germline or shared) to represent the combination of mutations in all genes. Each of these comparisons were performed with respect to nucleotide substitution rates (values derived from Hess et al. (1994) and Krawczak et al. (1998)), degree of evolutionary conservation at the codon level (as described in 4.2.2.2), Grantham amino acid difference (as described in 4.2.2.4) and CpG dinucleotide context (described in 4.2.2.1). All mutations were categorized into two groups, i.e. 'within CpG dinucleotide' or not, and the tests with respect to CpG dinucleotide context were performed using a χ^2 statistic (as described in 3.6.1.1). The rest of the comparisons (i.e. involving nucleotide substitution rates, degree of evolutionary conservation and Grantham amino acid difference) used continuous measures. Therefore, in order to determine what was the most appropriate test statistic, a normality test (to assess whether the data were drawn from a normally distributed population or not) was performed for each of the datasets (e.g. somatic, germline, shared, etc.). The normality tests were performed using the Shapiro-Wilk test for normality, part of the R software package (http://www.r-project.org/). All of the datasets (e.g. somatic, germline, shared, etc.) for each gene were found to deviate significantly from a normal distribution (an example is given in Figure 15). As a result, the rest of the test was performed using the Wilcoxon rank-sum test (Wilcoxon 1945). Detailed description of Wilcoxon rank-sum test is given in 3.6.1.2

To allow for multiple hypotheses testing for the tests performed for each gene (listed at the beginning of this section), 10,000 resampling permutations were performed and the resulting statistic was termed 'permuted'. To allow for multiple hypotheses testing for the 2 separate tests performed in each gene (i.e. CpG-dinucleotide analysis and Wilcoxon rank-sum tests), a Bonferroni correction was applied and the resulting statistic was termed 'gene-wise'. Therefore, each permuted p-value was multiplied by 2 (gene-wise $\alpha = 0.05/2$ or 0.025), to account for the different tests. To allow for multiple hypothesis testing, for the tests performed in all genes, a Bonferroni correction was also applied. Therefore, each gene-wise p-value was multiplied by 17 (overall experiment-wise $\alpha = \frac{0.05}{2*17}$ or 0.0015).

I designed a computer program that automatically performs the χ^2 and Wilcoxon rank-sum statistics for each test along with the re-sampling permutations.

4.2.2.6. Calculation of power and effect size

Calculations of the power and associated effect sizes are described in 3.6.3. In order to keep the overall α at the 0.05 level, as multiple statistical tests were performed, the value of α used in the power calculations was set to 0.0001336898 (total number of tests performed 374; therefore, the Bonferroni-adjusted α, to account for multiple testing, was $\alpha = 0.05/374$, or 0.0001336898 for χ^2 test statistic and $\alpha = \frac{0.05}{374*2}$ (two-sided test) for the Wilcoxon rank-sum test statistic). As Bonferroni correction is considered to be a conservative correction for multiple testing, the power calculated for the χ^2 statistical tests is a conservative estimate.

4.3. Results

The results are presented in comparison-wise fashion. Due to the overwhelming quantity of results that were generated during the work described in this chapter, only summaries of statistically significant results are discussed and presented in the form of tables (Table 9, Table 11, Table 12, Table 13, Table 14 and Table 15). These tables are to be found at the end of this chapter.

Nevertheless, the presented tables capture the results obtained for all comparisons performed, along with the directionality of the statistically significant results observed and power calculations. For further information, complete results for all comparisons performed are to be found in the Supplementary Tables.

In order to facilitate readability, whenever a comparison was statistically significant, it was substituted with the words "significant" or "significantly"; gene-wise statistically significant p-values were substituted with the symbol p_G and experiment-wise statistically significant p-values were substituted with the symbol p_E. In addition, whenever a comparison exhibited gene-wise and experiment-wise statistical significance, only the experiment-wise p-values were given and whenever a comparison exhibited only gene-wise, but not experiment-wise statistical significance, only the gene-wise p-value was listed. Additionally, all gene-wise and experiment-wise statistically significant results are graphically summarized in Table 13, along with the direction of the result and power calculations, but it was not referenced throughout the Results section, in order to reduce repetition.

4.3.1. Degree of evolutionary conservation

In order to estimate how fast each gene has been evolving and the degree of the evolutionary constraints acting upon the 17 human tumour suppressor genes, four different algorithms (detailed description is given in 4.2.2.2) for estimating the rate of evolution, namely Ka/Ks ratio, were performed. A summary graph of the results is presented in Figure 16.

For all genes, the Ka/Ks ratio was smaller than 1. This indicates that over time natural selection has tended to eliminate deleterious mutations in these genes, thereby yielding highly evolutionarily conserved gene and protein sequences. The evolutionary divergence for most of the genes was well below the average rate of sequence divergence between human and rodent, Ka/Ks~0.180 derived from 1880 human, rat and mouse gene orthologues (Makalowski and Boguski 1998). The *TP53* gene exhibited a rate of evolution that was similar to the average rate of gene evolution between human and rodent. In addition, the *CDKN2A, BRCA1* and *BRCA2* genes exhibited on average comparatively higher rates of evolution than the average rate of evolution between human and rodent. All of the genes showed *Ka<Ks*, indicative of the relative functional importance of these genes and that they might be under strong negative purifying selection.

4.3.2. Somatic vs. potential missense mutations

For a number of genes, it was evident that nucleotide context, measured in terms of both disease-associated and non-disease-associated mutability rates, significantly influences the occurrence of somatic, when compared to potential missense mutations. This was certainly the case for the *APC, CDKN2A, PTEN* and *TP53* genes (p_E<0.0034), with medians ranging from 4.6 to 8.4 and from 4.1 to 4.5 for the somatic and potential mutations respectively for non-disease-associated mutability rates and ranging from 0.5 to 1.1 and from 0.38 to 0.5 for the somatic and potential mutations for disease-associated mutability rates. In addition, the *STK11* gene showed experiment-wise significantly higher median relative disease-associated mutability rate (p_E<0.0034; median 1.66 and 0.44 for the somatic and potential mutations respectively). On the other hand, two genes (i.e. *BRCA2, RB1*) did not exhibit significant results, but nevertheless showed ≥80% power to detect an experiment-wise significant difference. Thus, it is very likely that nucleotide context does not play a significant part in shaping the somatic missense mutational spectrum in these two genes, i.e. *BRCA2* and *RB1*. These results indicate, that the somatic mutational spectrum in some genes is strongly influenced by the nucleotide sequence context (i.e. *APC, CDKN2A, PTEN* and *TP53*), whereas for other genes (i.e. *BRCA2* and *RB1*) nucleotide context plays little or no role.

The *RB1* and *CDKN2A* genes exhibited significantly more somatic missense mutations located in CpG dinucleotides (13% and 15% for *RB1* and *CDKN2A* respectively), when compared to potential missense mutations (1% and 4%, p_E<0.0034). In addition, several other genes (i.e. *ATM, BRCA2* and *STK11*) exhibited only gene-wise statistical significance (p_G ranging from 0.01 to 0.03), indicating more somatic missense mutations located in CpG-dinucleotides (proportions ranging from 9% to 24%) than potential mutations (proportions ranging from ~0% to 3%), but did not reach experiment-wise statistical significance (p_E ranging from 0.170 to 0.510). These results indicate that for the *RB1* and

CDKN2A genes, there is likely to be heavy intra-genic methylation in the soma whereas other genes are likely to be methylated to a relatively lesser degree (i.e. *ATM*, *BRCA2* and *STK11*).

No individual gene showed significantly different median Grantham physicochemical difference, when somatic mutations were compared to potential missense mutations. Therefore, it is likely that the mutant amino acids that comprise part of the somatic missense mutational spectrum are not associated with a higher median Grantham difference, as compared to wild-type amino acids, although no individual comparison showed ≥80% statistical power. Thus, either no true difference between wild-type and mutant amino acids, with respect to Grantham difference, and/or a paucity of mutations, may have contributed to these results.

The somatic mutations of the *TP53* and *VHL* genes were found to preferentially target evolutionarily conserved sites (p_E<0.0034; medians 0.14 and 0.17 for *TP53* and *VHL* respectively), when compared to potential missense mutations (medians 0.29, 0.43 for the *TP53* and *VHL* genes respectively). In addition, there was a trend in the *CDKN2A* gene (p_G=0.006) for somatic missense mutations to target evolutionarily conserved codons (medians 0.38 and 0.46 for somatic and potential mutations respectively), but this result did not reach experiment-wise statistical significance (p_E=0.102). On the other hand, the *APC* and *PTEN* genes, showed ≥80% power to detect an experiment-wise significance, but did not reach a statistically significant threshold (experiment-wise $\alpha \leq 0.05$). Therefore, for these two genes, it is very likely that somatic missense mutations do not preferentially target evolutionarily conserved codons. Thus, for the *TP53*, *VHL* and to some extent the *CDKN2A* genes, somatic missense mutations are more likely to be found in functionally important sites, whereas for the *APC* and *PTEN* genes, evolutionary conservation does not seem to play an important role.

4.3.3. Germline vs. potential missense mutations

In a similar pattern to the somatic mutational spectrum, it was evident that nucleotide context (measured by both disease and non-disease associated mutability rates) plays an important part in shaping the germline missense mutational spectrum in some genes, but not others, when compared to potential missense mutations. Seven genes (i.e. *ATM*, *BRCA1*, *BRCA2*, *NF1*, *RB1*, *TSC2* and *WT1*), showed significantly higher disease and non-disease-associated mutability rates (p_E<0.0034) with non-disease-associated medians ranging from 7.2 to 10.1 and 4.1 to 4.5 and disease-associated medians ranging from 0.79 to 1.27 and 0.38

to 0.43 for the germline and potential mutations respectively. Additionally, the *CDH1* and *PTEN* genes showed experiment-wise significantly higher disease-associated mutability rates (p_E<0.0034) for both genes; medians 1.27/0.41 and 0.92/0.38 for the germline/potential mutations in the *CDH1* and *PTEN* genes. Some genes (i.e. *APC*, *NF2* and *PTCH*) showed enough statistical power (i.e. ≥80%), but did not exhibit statistically significant results, for both disease and non-disease-associated mutability rates. Hence their germline missense mutational spectra are very likely not influenced by nucleotide context, with some other mechanism(s) influencing the occurrence of mutations. Therefore, for a number of genes (i.e. *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *PTEN*, *RB1*, *TSC2* and *WT1*), the germline mutational spectrum was characterized by a significantly higher median, with respect to disease and non-disease-associated nucleotide substitution rates, but not for others, such as the *APC*, *NF2* and *PTCH*. This is to be expected, since the disease-associated nucleotide substitution rates have been derived from germline missense mutations.

The *ATM* and *TSC2* genes showed significant co-localisation of germline missense mutations within CpG-dinucleotides (p_E<0.0034; 8% and 15% within CpG-dinucleotides for *ATM* and *TSC2* respectively), as compared to potential missense mutations (1% and 3% within CpG-dinucleotides for *ATM* and *TSC2* genes respectively). In addition, a number of genes (i.e. *APC*, *BRCA1*, *BRCA2*, *CDH1*, *NF1* and *NF2*) showed a trend (p_G ranging from 0.008 to 0.038 and p_E ranging from 0.136 to 0.646) for germline missense mutations being preferentially found in CpG-dinucleotides (proportions ranging from 5% to 22% and ~0% to 2% for the germline and potential mutations respectively). Thus, the germline missense mutations for the *ATM* and *TSC2* genes were much more likely to be found in CpG dinucleotides than potential missense mutations, whereas for other genes (i.e. *APC*, *BRCA1*, *BRCA2*, *CDH1*, *NF1* and *NF2*) only a trend was observed.

The wild-type amino acids in the *NF1* and *VHL* genes were much more likely to be substituted by mutant amino acids, characterized by a significantly higher Grantham difference as a result of germline missense mutations (p_E<0.0034 for both genes), with medians ranging from 98 to 99 and 71 to 76 for the germline and potential mutations respectively. In addition, the *ATM* gene exhibited a trend in the same direction, i.e. significantly higher median Grantham difference (p_G=0.04 and p_E=0.68) for the germline as compared to potential missense mutations. For the rest of the genes, no conclusions could be made, as there was not enough statistical power and none of the genes showed statistically significant results.

The germline mutations for the *ATM*, *BRCA1* and *VHL* genes preferentially targeted evolutionarily conserved codons (p_E<0.0034), with medians ranging from 0 to 0.14 and 0.17 to 0.54 for the germline and potential mutations respectively. In addition, the *CDKN2A*, *TSC2* and *WT1* genes showed only a modest association of germline missense mutations and evolutionarily conserved sites (p_G ranging from 0.008 to 0.032), when compared to potential mutations (germline medians ranging from 0 to 0.29 and potential medians ranging from 0 to 0.46). Conversely, the *NF1* and *PTEN* genes showed enough statistical power, but did not exhibit statistically significant results. Therefore, in the *ATM*, *BRCA1*, *VHL* and possibly *CDKN2A*, *TSC2* and *WT1* genes, germline missense mutations are much more likely to be found in evolutionarily conserved codons as compared to potential mutations, whereas for the *NF1* and *PTEN* genes no such finding was evident.

4.3.4. Shared vs. potential missense mutations

The *CDKN2A*, *PTEN* and *TP53* genes exhibited experiment-wise significantly higher nucleotide substitution rates, for both disease (medians ranging from 1.08 to 1.28) and non-disease substitution rates (medians ranging from 8.9 to 11), when shared mutations were compared to potential missense mutations (p_E<0.0034; medians ranging from 0.38 to 0.5 and 4.1 to 4.5 for disease and non-disease-associated substitution rates). In addition, the *VHL* gene exhibited only a significantly higher median, with respect to disease-associated mutations (p_E<0.0034; medians 1 vs. 0.44 for the shared and potential mutations), but not with respect to non-disease-associated substitution rates.

The *BRCA2*, *CDKN2A*, *PTEN*, *RB1*, *TP53* and *TSC2* genes exhibited a significantly higher proportion of CpG-located shared mutations (ranging from 14% to 100% of the mutations found within CpG-dinucleotides), when compared to potential missense mutations (p_E≤0.0034), with proportions ranging from ~0% to 4% of the mutations found within CpG-dinucleotides.

Only the *TP53* gene exhibited a significantly higher difference between wild-type and mutant amino acids with respect to Grantham difference (p_G=0.014; medians 98 and 74 for the shared and potential mutations respectively), but did not reach experiment-wise significance (p_E=0.238).

Three genes (*viz. CDKN2A*, *TP53* and *VHL*) showed significantly many more shared missense mutations found in evolutionarily conserved sites as compared to potential mutations ($p_E<0.0034$, medians 0 for the shared mutations and medians ranging from 0.29 to 0.46 for the potential missense mutations).

From the results presented for the comparisons between shared and potential missense mutations, three genes (*viz. TP53*, *CDKN2A* and *VHL*) clearly stood out. The *TP53* gene exhibited statistically significant results for all of the comparisons performed. Thus, it is evident that the shared mutational spectrum associated with the *TP53* gene is strongly influenced by nucleotide context, measured in terms of disease and non-disease-associated substitution rates; mutant amino acids exhibited a greater physicochemical difference as compared to wild-type amino acids; shared missense mutations preferentially targeted evolutionarily conserved sites and CpG-dinucleotides.

Similarly, the *CDKN2A* gene showed exactly the same pattern, with one exception: shared missense mutations did not exhibit significantly higher Grantham difference.

The shared missense mutations in the *VHL* gene were found to be characterised by a significantly higher median nucleotide substitution rate (disease associated) and preferential location within evolutionarily conserved sites, as compared with potential mutations.

4.3.5. Somatic vs. germline missense mutations

Interestingly, not a single gene showed a statistically significant result for any of the comparisons performed, although it should be noted that only the *CDKN2A* and *STK11* genes had enough statistical power to detect an experiment-wise significant difference, with respect to disease and/or non-disease-associated mutability rates. Thus, one could conclude for these genes that no difference exists between somatic and germline missense mutations, with respect to nucleotide context measured by both disease and non-disease associated substitution rates. It is clear that nucleotide context for these genes (i.e. *CDKN2A* and *STK11*) influences both the somatic and germline missense mutational spectrum in a very similar way.

In addition, the *CDH1* and *PTEN* genes showed ≥80% statistical power for detecting an experiment-wise difference, with respect to disease associated substitution rates, but no significant differences were detected (both gene-wise and experiment-wise). Therefore, nucleotide context, measured in terms of disease-associated substitution rates, influences both the somatic and germline missense mutational spectrum for the *PTEN* and *CDH1* genes in a very similar way.

One could conclude that nucleotide context strongly influences both the germline and somatic spectrum in a very similar way for the *CDH1*, *CDKN2A*, *PTEN* and *STK11* genes.

When shared mutations were compared to potential missense mutations, with respect to evolutionarily conserved positions in the *PTEN* gene, there was enough statistical power, but the comparison did not reach a statistically significant threshold (both gene and experiment-wise). At least for the *PTEN* gene, the location of somatic and germline missense mutations did not differ with respect to evolutionarily conserved codons. Thus, it is very likely that somatic and germline missense mutations within the *PTEN* gene did not specifically target evolutionarily conserved sites.

4.3.6. Somatic vs. shared missense mutations

Clearly, only 4 genes (i.e. *CDKN2A*, *PTEN*, *TP53* and *VHL*) had enough shared and somatic missense mutations that resulted in enough statistical power to derive any meaningful conclusions (Table 12).

The *TP53* gene exhibited a significantly higher median substitution rate (p_E<0.0034 for both disease and non-disease-associated substitution rates) of shared mutations (medians 4.6 and 0.5), when compared to somatic missense mutations (medians 8.9 and 1.6). The proportion of CpG-located shared missense mutations was also found to be significantly higher than its somatic counterpart (p_E<0.0034; 3% and 23% missense mutations were found in CpG-dinucleotides for the somatic and shared mutations respectively). In addition, somatic missense mutations showed significantly higher median evolutionary variation, when compared to shared missense mutations (p_E<0.0034; somatic median 0.17 and germline median 0) for the *TP53* gene. These results indicate that shared missense mutations in the *TP53* gene are much more likely to be influenced by nucleotide context and to preferentially target evolutionarily conserved codons and CpG-dinucleotides, when compared to somatic missense mutations.

The *PTEN* gene showed only a modest statistical significance of disease-associated nucleotide substitution rates (p_G=0.04, and p_E=0.68; somatic median 0.53 and shared median 1.23) and non-disease-associated substitution rates (p_G=0.022 and p_E=0.374; somatic median 5.6 and shared median 11). Thus, shared missense mutations within the *PTEN* gene were found to be associated with higher median substitution rates as compared to somatic missense mutations. Conversely, there was enough statistical power for the comparison of shared and somatic missense mutations, with respect to evolutionary variation, but the results were not found to be significant (p_G~1 and p_E~1). Therefore, it is likely that both shared and somatic missense mutations do not preferentially target evolutionarily conserved codons.

Contrary to the *TP53* and *PTEN* genes, there was enough statistical power to detect an experiment-wise difference between the shared and somatic missense mutations for the *CDKN2A* gene, with respect to both disease and non-disease associated substitution rates, but no statistically significant difference was found. Thus, both somatic and shared missense mutations in the *CDKN2A* gene are equally strongly influenced by nucleotide context. The *CDKN2A* gene exhibited a modest significantly higher median evolutionary variation of somatic missense mutations (median 0.38) when compared to shared missense mutations (p_G=0.034 and p_E=0.578) with median 0. Therefore, to some degree, it is likely that shared missense mutations at least for the *CDKN2A* gene, targeted relatively more evolutionarily conserved codons than the somatic missense mutations.

4.3.7. Germline vs. shared missense mutations

As with the somatic vs. shared missense mutations comparison, four genes had enough germline and shared missense mutations to yield sufficient statistical power to detect an experiment-wise significant difference. The *CDKN2A*, *PTEN*, *TP53* and *VHL* genes did not exhibit statistically significant results, with respect to nucleotide substitution rates, disease and/or non-disease associated mutability rates. Therefore, it is very likely that both the germline and shared missense mutations are equally strongly influenced by nucleotide context, as both the germline and shared mutations were separately found to be characterized by significantly higher medians, when compared to potential missense mutations (for details see sections 4.3.3 and 4.3.4).

For the rest of the tests, there was not enough statistical power and none of the comparisons reached statistical significance.

4.3.8. Recurrent somatic missense mutations

Only two genes (i.e. *PTEN* and *TP53*) had enough mutations to permit further conclusions to be drawn. In fact, the number of recurrent somatic missense mutations in both genes represented 90% of the recurrent mutations observed in the 17 genes studied (Table 9).

Furthermore, for these two genes, all (100%) shared missense mutations that were found in CpG-dinucleotides were also found to be recurrent (66% for all genes combined). In addition, ~27% of all recurrent somatic missense mutations (the recurrence status of the germline missense mutations could not be determined, since this information is not recorded in *HGMD*) that were found in CpG dinucleotides were also found in the germline (27%% and 25% for the *PTEN* and *TP53* genes respectively).

The results obtained for the *TP53* gene suggest that recurrent somatic missense mutations are significantly more strongly influenced by nucleotide context (medians 0.53 and 4.6 for disease and non-disease associated substitution rates respectively), than non-recurrent mutations (p_E<0.0034 and medians 0.42/4.1 for the disease/non-disease-associated substitution rates respectively). Furthermore, the recurrent somatic missense mutations were also found to preferentially target evolutionarily conserved codons as compared to non-recurrent mutations (p_E<0.0034), with medians 0 and 0.36. Thus, somatic missense mutations that recur are influenced by nucleotide context, but selection towards functionally important domains is also likely to play an important part. Similarly, recurrent and shared somatic missense mutations were much more likely to be associated with a significantly higher median mutability rate, than both recurrent non-shared (p_E<0.0034; recurrent and shared medians 1.19 and 9 for the disease and non-disease associated mutability rates respectively; recurrent non-shared medians 0.53 and 4.6) and non-recurrent and non-shared somatic missense mutations (medians 4.1 and 0.42 for disease and non-disease associated mutability rates respectively). In addition, recurrent and shared somatic missense mutations were disproportionately more likely to be found in CpG dinucleotides (25%), than both recurrent non-shared (p_E=0.0034; 3%) and non-recurrent non-shared mutations (p_E<0.0034; 2%). Therefore, these results imply that recurrence status is heavily influenced by nucleotide context. In addition, recurrent somatic mutations that are also found in the germline preferentially target CpG dinucleotides, as compared to recurrent somatic mutations not found in the germline or somatic mutations that do not recur and were also not found in the germline. Thus, CpG dinucleotides are mutational hotspots for both recurrent somatic and germline missense mutations. This is likely to result from heavy intra-genic CpG methylation in both the germline and the soma for the *TP53* gene.

In contrast to *TP53*, the *PTEN* gene possessed enough statistical power, but did not exhibit statistically significant results for either the recurrent vs. non-recurrent or recurrent and shared vs. recurrent and non-shared somatic missense mutations. As a result, and in contrast to *TP53*, recurrent somatic mutations that are also found in the germline were unlikely to be influenced by nucleotide context. Nevertheless, recurrent somatic missense mutations that are also found in the germline are more likely to be associated with higher median nucleotide substitution rates (p_G=0.0360/0.002 and p_E=0.612/0.034 for disease/non-disease associated substitution rates respectively; medians 1.4, 12.7 and 0.43, 4.7 for recurrent and shared vs. non-recurrent non-shared mutations respectively) and were disproportionately more likely to be found within CpG dinucleotides (27% vs. 1% respectively), than somatic missense mutations that did not recur and were also not found in the germline (p_G=0.004 and p_E=0.068).

4.3.9. Combination of somatic, germline and shared vs. potential missense mutations for individual genes

The results for the combination of missense mutations in the individual 17 genes were very much dependent on the number of somatic, germline and shared missense mutations. Thus, the mutational spectra in the genes could be separated into several groups: predominantly somatic missense mutations (*APC*); predominantly germline missense mutations (*ATM, BRCA1, BRCA2, NF1, PTCH, TSC1, TSC2* and *WT1*); similar proportions of somatic and germline missense mutations (*NF2* and *CDH1*); predominantly somatic with a sizeable proportion of shared mutations (*CDKN2A, PTEN* and *TP53*); predominantly germline with a sizeable proportion of shared mutations (*RB1, STK11* and *VHL*). The results exhibited very similar patterns (e.g. direction of results and statistical significance) to the comparisons of the largest proportion of mutations in the individual genes.

4.3.10. Summary of results

4.3.10.1. Disease and non-disease-associated substitution rates

The *APC, CDKN2A, PTEN* and *TP53* genes exhibited significantly (gene-wise and/or experiment-wise) higher relative median values for both disease and non-disease- associated substitution rates for **somatic mutations, when compared to potential mutations.**

The *ATM, BRCA1, BRCA2, NF1, RB1, TSC2* and *WT1* genes exhibited significantly (gene-wise and/or experiment-wise) higher relative median values for both disease and non-disease-associated substitution rates for **germline mutations, when compared to potential mutations.**

The *CDKN2A*, *PTEN* and *TP53* genes exhibited significantly (gene-wise and/or experiment-wise) higher relative median values for both disease and non-disease-associated substitution rates for **shared mutations, when compared to potential mutations.**

None of the individual genes studied exhibited significantly (gene-wise and/or experiment-wise) different relative median values for both disease and non-disease-associated substitution rates for **somatic mutations, when compared to germline mutations**, even though there was enough power to detect an experiment-wise significant difference for the *CDKN2A* and *STK11* genes.

The *PTEN* and *TP53* genes exhibited significantly (gene-wise and/or experiment-wise) higher relative median values for both disease and non-disease-associated substitution rates for **shared missense mutations, when compared to somatic mutations.**

Only the *TP53* gene exhibited significantly (gene-wise and/or experiment-wise) higher relative median values for both disease and non-disease-associated substitution rates for **recurrent somatic missense mutations as compared to non-recurrent somatic mutations; recurrent and shared somatic mutations as compared to recurrent non-shared somatic mutations; recurrent and shared somatic as compared to non-recurrent and non-shared somatic missense mutations.**

It is interesting to note that all comparisons that showed a significant result with non-disease-associated substitution rates, also showed a significant result with disease-associated substitution rates while a number of comparisons showed only significant results with disease-associated mutability rates. Therefore, it may be concluded that the majority of the spectrum of missense mutations in most genes are likely to be associated with disease-associated substitution rates, hence are more likely to be 'drivers' of tumour development.

4.3.10.2. CpG dinucleotides

The *ATM*, *BRCA2*, *CDKN2A*, *RB1* and *STK11* genes exhibited a significantly (gene-wise and/or experiment-wise) higher proportion of **somatic mutations, when compared to potential mutations**, with respect to mutations found in CpG dinucleotides.

The *APC*, *ATM*, *BRCA1*, *BRCA2*, *CDH1*, *NF1*, *NF2* and *WT1* genes exhibited a significantly (gene-wise and/or experiment-wise) higher proportion of **germline mutations, when compared to potential mutations**, with respect to mutations found in CpG dinucleotides.

The *BRCA2*, *CDKN2A*, *PTEN*, *RB1*, *TP53* and *TSC2* genes exhibited a significantly (gene-wise and/or experiment-wise) higher proportion of **shared mutations, when compared to potential mutations**, with respect to mutations found in CpG dinucleotides. None of the genes exhibited a significantly (gene-wise and/or experiment-wise) different proportion of **somatic mutations, when compared to germline mutations**, with respect to mutations found in CpG dinucleotides.

Only the *TP53* gene exhibited a significantly (gene-wise and/or experiment-wise) higher proportion of **somatic vs. shared mutations; recurrent and shared vs. recurrent and non-shared somatic mutations; recurrent and shared vs. non-recurrent and non-shared somatic mutations**, with respect to CpG-located missense mutations.

4.3.10.3. Grantham difference

None of the genes exhibited a significantly higher median Grantham difference between wild-type and mutant **somatic missense mutations, when compared to potential mutations**.

The *ATM*, *NF1* and *VHL* genes exhibited a significantly (gene-wise and/or experiment-wise) higher median Grantham difference between wild-type and mutant **germline missense mutations, when compared to potential mutations.**

Only the *TP53* gene exhibited a significantly higher median Grantham difference between wild-type and mutant **shared missense mutations, when compared to potential mutations.**

4.3.10.4. Evolutionary conservation

The **somatic mutations** in the *CDKN2A*, *TP53* and *VHL* genes preferentially targeted evolutionarily conserved codons, when compared to **potential mutations**.

The **germline mutations** in the *ATM*, *BRCA1*, *CDKN2A*, *TSC2*, *VHL* and *WT1* genes preferentially targeted evolutionarily conserved codons, when compared to **potential mutations**.

The **shared mutations** in the *CDKN2A*, *TP53* and *VHL* genes preferentially targeted evolutionarily conserved codons, when compared to **potential mutations**.

The *TP53* gene exhibited significantly (gene-wise and/or experiment-wise) lower median evolutionary variation for **somatic as compared to shared missense mutations;**

recurrent as compared to non-recurrent somatic missense mutations; recurrent and shared as compared to non-recurrent and non-shared missense mutations.

The *CDKN2A* gene exhibited significantly (gene-wise) lower median evolutionary variation for **somatic as compared to shared mutations and non-recurrent shared as compared to non-recurrent non-shared somatic missense mutations.**

4.3.11. Combination of missense mutations in all genes

Around 88% of the somatic mutational spectra for all genes combined were represented by 3 genes (i.e. *CDKN2A*- 9.88%, *PTEN*- 11.69% and *TP53*- 66.16%). In addition, ~70% of the germline mutational spectrum for all genes was represented by 6 genes (i.e. *ATM*- 8.82%, *BRCA1*- 19.88%, *BRCA2*- 10.00%, *NF1*- 9.76%, *TSC2*- 10.24% and *VHL*- 11.53%) and >93% of the shared mutational spectrum for all genes was represented by 4 genes (i.e. *CDKN2A*- 14.29%, *PTEN*- 11.22%, *TP53*- 44.90% and *VHL*- 22.96%). Therefore, the results for the combination of mutations in all genes were very much influenced by these genes. Detailed proportions are presented in Table 12.

Nevertheless, when each type of mutations, namely somatic, germline and shared missense mutations for all genes, were compared to the corresponding potential missense mutations combined for all genes, they were found to exhibit significantly higher median mutability rates (both disease and non-disease associated), preferential location in CpG dinucleotides, higher median Grantham difference between wild-type and mutant amino acids and higher affinity towards evolutionarily conserved sites (summary of results is given in Table 15). Clearly, the somatic, germline and shared missense mutations are influenced by nucleotide context, but nevertheless selection pressure in the form of the physicochemical difference of affected amino acids and functionally important codons/domains must also play an important role. Thus, codons susceptible to mutations (i.e. hotspots) are also selected on the basis of damage to the protein.

Clear differences and similarities between the somatic, germline and shared missense mutations for all genes combined were evident. Indeed, when somatic were compared to germline or shared missense mutations, they were significantly less likely to be influenced by nucleotide context (measured by both disease and non-disease-associated mutability rates), less likely to be located in CpG-dinucleotides, showed smaller Grantham differences and targeted relatively less evolutionary conserved sites. In addition, when germline were compared to shared missense mutations, the shared

missense mutations exhibited a significantly higher median disease-associated substitution rate and were significantly more likely to be found in CpG dinucleotides. Furthermore, shared and germline missense mutations were equally likely to be found in evolutionarily conserved codons.

Therefore, these results suggest that a fine 'ranking' in the pathogenicity of missense mutations exists, with shared missense mutations being more likely to be 'drivers' of tumorigenesis, than pure somatic or pure germline missense mutations. Similarly, pure germline missense mutations are more likely to be associated with tumorigenesis, than purely somatic mutations.

4.4. Discussion

Cancer is regarded as a genetic disorder on the basis that genetic and epigenetic modifications often contribute to neoplasia. The underlying mutations include DNA sequence changes, such as copy number variation, gross rearrangements, micro-lesions (e.g. deletions, insertions and indels), single base-pair substitutions, etc. (Loeb and Harris 2008; Stratton et al. 2009). Missense mutations are an important part of the mutational spectrum associated with tumour development. For some missense mutations, research has unambiguously shown their functional importance, but for others the degree of pathogenicity remains largely unknown. Therefore, numerous classification procedures have been developed and employed to try to predict the pathogenicity of missense mutations. These classification procedures involve sometimes costly and labour-intensive functional assays, but which nevertheless could be helpful in assigning functional significance (Chan et al. 2007). In addition, it may not be very practical to perform functional assays on every single variant that turns up during routine clinical screening, because for some variants, functional assays may not exist or may be very difficult to perform. As a result, a plethora of *in silico* algorithms, methods and procedures has been developed to facilitate the classification of functional importance of genetic variants, and missense mutations in particular (Miller and Kumar 2001; Tavtigian et al. 2008). Normally, variants such as deletions, insertions and nonsense mutations are readily classified as functionally important because they generally disrupt gene function and/or structure. Most of the unclassified sequence variants are missense mutations (Tavtigian et al. 2008). Thus, *in silico* algorithms for the classification of sequence variants are usually focussed on missense mutations. These procedures utilize numerous measures and parameters. Some of these measures include properties of amino acids, derived from substitution matrices, such as PAM (Dayhoff et al. 1978) and BLOSUM (Henikoff and Henikoff 1992); physicochemical differences, e.g. Grantham (Grantham 1974); changes in protein structure (Goldgar et al. 2004), evolutionary conservation, disease and non-disease-associated nucleotide substitution rates (Hess et al. 1994; Krawczak et al. 1998); and many others. Usually, these procedures and methods rely heavily on training datasets. These datasets comprise classified, e.g. validated neutral/polymorphic, pathogenic/functional missense variants (Tavtigian et al. 2008). Because unclassified missense variants exist, these training sets would not encompass every single missense mutation; therefore, they are subject to chance variation.

Furthermore, while cancer or tumour development is considered to be a disorder of the soma (i.e. phenotypic manifestation as somatic tumours), there is growing evidence that inherited (i.e. germline) variants may play an important role in the development of a disease, including cancer. Three recent papers (Jones et al. 2009; Kilpivaara et al. 2009; Olcaydu et al. 2009) have shown very strong association of a particular inherited haplotype in the *JAK2* gene and the preferential acquisition of a particular somatic mutation (V617F), considered causative of myeloproliferative neoplasms (Campbell 2009). This activating point mutation, has been found in >95% of the individuals with polycythemia vera and 50-60% of the individuals with essential thrombocythemia (Campbell 2009; Levine et al. 2007). Campbell (2009) has proposed two competing hypotheses that could potentially account for the inherited predisposition. The somatic missense mutation occurs with a rate independent of the inherited haplotype, but inherited variation confers a stronger selective advantage over the cells; thus, such cells are more likely to undergo clonal expansion. The second hypothesis considers the differential mutation rate of the causative variant with respect to the germline haplotype. Thus, cells that inherit the germline variant exhibit hypermutability of the gene locus and these cells more frequently acquire mutations and hence are more likely to acquire the causative variant and undergo clonal expansion.

Knudson's 'two-hit' hypothesis (Knudson 1971, 1978) provides a framework for understanding tumour suppressor genetics and the development of cancer. This hypothesis states that in addition to an inherited hit (e.g. a sequence alteration or modification), a second, somatic hit affecting the hitherto unaffected allele of the gene is necessary for the initiation and development of cancer. In addition, research on the genetics of tumour suppressor genes has indicated that there might be a complex interaction between the two hits. The *APC* gene is an example of such a tumour suppressor gene that has been shown to exhibit somatic-germline interplay. It is a tumour suppressor gene consistent with Knudson's 'two-hit' hypothesis, in the sense that both inherited (i.e. germline) and acquired (i.e. somatic) sequence changes are required for tumour development (Miyoshi et al. 1992; Powell et al. 1992). In their seminal paper, Latchford et al. (2007) have shown a positional non-random occurrence of somatic mutations, also shown by others (Groves et al. 2002; Lamlum et al. 1999), but most importantly the position of the germline hit could direct the frequency and type of the second hit (i.e. somatic mutation). This 'first-second hit' relationship is held to be sufficient to maintain a 'just right' level of β-catenin protein (Latchford et al. 2007), in order to manifest a selective advantage of the mutant *APC* protein. Even although the lesions found in the *APC* gene are predominantly truncating mutations (nonsense mutations and frameshifts), these examples indicate the important role of the relationship between germline and somatic genetic changes.

Therefore, it is clear that interplay, similarities and differences between the soma and germline in terms of genetic changes, have to be taken into consideration, when inferring the functional consequences of DNA sequence changes and in particular missense mutations. In addition, the anecdotal nature of reports of the relationship between the germline and the somatic mutation status does not rule out the possibility that this phenomenon could be relatively widespread. Usually, germline mutations are considered to increase the risk of developing cancer, but the mechanisms are still unclear. Thus, we may see an increasing number of reports suggesting a possible intricate relationship between the germline and the soma in the context of mutagenesis.

In the light of the relatively few studies that have compared and contrasted somatic and germline missense mutations, the analysis presented in this chapter represents an attempt to shed some light on differences and similarities in the germline and soma, with respect to missense mutations.

4.4.1. *ATM* gene

The presented results for the *ATM* gene are very much in agreement with other studies on the mutational spectrum of the *ATM* gene. A number of studies have reported the absence of somatic missense mutations in breast cancer (Feng et al. 2003; Vorechovsky et al. 1996) and T-cell prolymphocytic leukaemia (Luo et al. 1998). The data presented in this chapter indicates that the missense mutational spectrum in the *ATM* gene comprised predominantly germline missense mutations (~87% germline vs. ~13% somatic missense mutations). In addition, studies have failed to show any association between germline missense mutations and loss of heterozygosity in breast cancer and ataxia-talagiectasia (Feng et al. 2003; Liberzon et al. 2004); hence it is likely that germline missense mutations contribute significantly more to tumorigenesis than somatic missense mutations. Furthermore, to account for this, Feng et al. (2003) have suggested that germline missense mutations in the *ATM* gene could exert a dominant effect, e.g. by inactivating a multiprotein complex. Indeed, a study by Scott et al. (2002) has identified a number of missense mutations (S2592C, V2716A, R2849P and G2867R) in breast cancer, using *in vitro* mutagenesis of full length *ATM* cDNA, that display dominant negative activity over the wild-type protein. Clearly, the mutational data from the *ATM* gene strongly suggest that the germline missense mutational spectrum disrupts the function of the mutant proteins, predicting a major effect of germline missense mutations in driving tumorigenesis.

On the other hand, analysis of the mutational spectrum of the *ATM* gene has revealed that ~80-85% of all sequence alterations are truncating mutations (Lavin et al. 2004). The data presented here showed that 52% of the somatic sequence alterations are truncating as compared to 76% in the germline (data are presented in Table 16). These results suggest that germline mutations (truncating and non-truncating- i.e. missense mutations) in the *ATM* gene predispose to tumour development to a relatively stronger degree as compared to somatic mutations. Based on the results presented in this chapter, one could conclude that somatic missense mutations present in the *ATM* gene are more likely to be result of genomic instability, thereby constituting passenger mutations.

4.4.2. *BRCA1* and *BRCA2* genes

Germline sequence variants in the *BRCA1* and *BRCA2* genes have been shown to contribute significantly to the development of breast and ovarian cancers (Easton et al. 1993; Ramus et al. 2007; Szabo et al. 1996). About a third of all reported (Breast Cancer Information Core (*BIC*) database: http://research.nhgri.nih.gov/bic/index.shtml) sequence changes in *BRCA1* are missense mutations and >50% of those are reported only once (Szabo et al. 2004). Furthermore, ~50% of all reported changes in *BRCA1* (~1200, *BIC*) have been reported only once. Likewise, ~11,000 mutations have been reported for the *BRCA2* gene and ~50% of these are classified as being of unknown functional consequence. These data indicate the highly variable nature of the mutational spectrum in the *BRCA1* and *BRCA2* genes. Some authors have suggested that the relatively poor evolutionary conservation of *BRCA1* across different species (Szabo et al. 1996) could indicate that most of the gene sequence might have relaxed functional and/or structural constraints. Therefore, numerous sequence variations could be relatively well tolerated, with respect to functional and/or structural characteristics. Indeed, the evolutionary conservation analysis performed in this chapter suggested that both the *BRCA1* and *BRCA2* genes are the two most variable genes from the 17 tumour suppressor genes studied, with respect to evolutionary conservation (Figure 16). Nevertheless, even although mutations in some regions of the *BRCA1* gene could be of little functional importance to the function of the gene, germline missense mutations were very likely to have functional importance. This was indicated by the fact that *BRCA1* germline missense mutations were rather more likely to be found in evolutionarily conserved codons than potential missense mutations. Conversely, the paucity of the somatic and germline missense mutations in the *BRCA2* gene precluded the possibility of inferring the evolutionary conservation of residues subject to mutations.

The *BRCA2* gene showed a greater proportion of germline (~80%) as compared to somatic (~19%) missense mutations (Table 12). Over and above that, the *BRCA1* gene showed an almost complete absence of somatic missense mutations (~3% of all missense mutations, Table 12). Due to the fact that both *BRCA1* and *BRCA2* genes are reported to be consistent with Knudsons's two hit hypothesis (Dworkin et al. 2009), different mechanisms could explain the dearth of somatic missense mutations, such as somatic abolition of protein expression through hypermethylation of the promoter region and protein truncation. Indeed, studies on methylation patterns in the *BRCA1* promoter region have shown hypermethylation in breast and ovarian cancer tissues as compared to normal tissues from the same individual (Baldwin et al. 2000; Esteller et al. 2000; Radpour et al. 2009). The germline missense mutations in the *BRCA1* gene showed, to some extent, a preferential location within CpG dinucleotides, implying hypermethylation of intra-genic CpG dinucleotides in the germline. On the contrary, promoter hypermethylation has been shown be an infrequent event in the *BRCA2* gene (Dworkin et al. 2009; Hilton et al. 2002). Nevertheless, germline, somatic and shared missense mutations all showed preferential localization within CpG dinucleotides. Thus, one could infer that heavy intra-genic CpG methylation may be present in both the soma and germline in the *BRCA2* gene.

Due to the lack of data on promoter hypermethylation, one could only speculate the possibility that the *BRCA1* and *BRCA2* genes could exhibit a reduction of protein expression in both the soma and germline through hypermethylation of the promoter region.

Alternatively, the lack of somatic missense mutations could potentially be explained by the predominance of truncating mutations. Around 79% of all somatic mutations in the *BRCA1* gene are indeed truncating aberrations (nonsense mutations, micro-deletions, micro-insertions and micro-indels; Table 16), comparable with the ~70% value reported in the literature (Szabo et al. 2004). It is interesting to note that the proportion of truncating germline mutations was very similar (~73%, Table 16). Therefore, protein truncation in both the soma and the germline appears to be a common mutational mechanism in the *BRCA1* gene.

On the contrary, ~61% of all somatic mutations within the *BRCA2* gene are non-truncating (i.e. missense mutations), in contrast to ~17% germline missense mutations. It has been suggested that *BRCA2* may not be a classic tumour suppressor gene with respect to

Knudson's two-hit hypothesis (Meric-Bernstam 2007), although heterozygous knock-out of *BRCA2* in mice has not shown a "strong tumour predisposition phenotype" (Evers and Jonkers 2006; Meric-Bernstam 2007).

Therefore, it would seem that missense mutations in both the *BRCA1* and *BRCA2* genes might not be severe enough to disrupt the function of the protein to such a degree to cause tumour development, but nevertheless could lead to a predisposition at least with germline missense mutations in the *BRCA1* gene. This is further supported by the observation that germline missense mutations in both genes did not show significantly higher Grantham differences as compared to wild-type amino acids. It is likely that somatic and germline missense mutations in the *BRCA2* and somatic missense mutations in the *BRCA1* genes are disproportionately "passenger" mutations, whereas germline *BRCA1* missense mutations give rise to predisposition to tumour development.

4.4.3. *CDKN2A* gene

The mutational spectrum of the *CDKN2A* gene is predominantly described by whole-gene deletions and point mutations being uncommon in most common cancers, such as colorectal, breast and gynaecological cancers (The *CDKN2A* database: https://biodesktop.uvm.edu/perl/p16, Murphy et al. 2004). Interestingly, the observed somatic and germline mutational spectra in the *CDKN2A* gene were described predominantly by non-truncating mutations (~61% and ~70% of all mutations for the somatic and germline mutations respectively, Table 12). Therefore, apart from the common whole-gene deletions, missense mutations are a relatively common event.

The majority of the missense mutational spectrum comprised somatic (~73%), followed by equal proportions of germline and shared missense mutations (~15% and ~12% for germline and shared missense mutations). It has to be said that the *CDKN2A* is the smallest gene, with respect to number of nucleotides, as compared to the rest of the genes. Thus, one could argue that it is relatively more likely that mutations in the soma could be also observed in the germline just by chance alone, as compared to larger genes. Nevertheless, the results presented herein indicated that shared missense mutations within *CDKN2A* are not merely accidental. Shared missense mutations were indeed found to preferentially target evolutionarily conserved sites, despite the fact that the *CDKN2A* gene might have relaxed functional and/or structural constraints. Similar to *BRCA1* and *BRCA2*, the *CDKN2A* gene showed relatively more evolutionary divergence as compared to the rest of the genes (Figure

16). In addition, both the germline and somatic missense mutations were also found to some degree within evolutionarily conserved sites. Thus, shared missense mutations showed a stronger association with evolutionarily conserved codons than both the soma and germline and were more likely to be found in evolutionarily conserved sites than somatic missense mutations.

Furthermore, the somatic and germline missense mutations did not show preferential location within CpG dinucleotides, where the shared mutations were preferentially located in CpG dinucleotides. This is a potential indication of intra-genic hypermethylation in both the soma and the germline. Indeed, promoter hypermethylation has been shown to be a strong cancer predictor in the *CDKN2A* gene, for breast cancer and oesophageal adenocarcinoma (Radpour et al. 2009; Wang et al. 2009).

These results suggest that missense mutations in both the soma and the germline are very likely to be caused by the same mechanisms. This was further supported by the fact that direct comparison between somatic and germline missense mutations did not exhibit significant differences, with respect to any of the parameters studied. Both the somatic and shared missense mutations exhibited higher relative median values of both disease and non-disease mutability rates, when each was compared to potential missense mutations. Therefore, these mechanisms are very likely to be DNA-sequence dependent, i.e. endogenous mutagenesis, such as methylation-mediated deamination of 5-methylcytosine in CpG dinucleotides, post-replicative mismatch repair and exonucleolytic proof-reading (Cooper and Krawczak 1993; Krawczak et al. 1998).

4.4.4. *NF1* gene

The *NF1* gene is regarded as being a classic tumour suppressor gene; thus, bi-allelic inactivation is required for tumour development (Glover et al. 1991; Rasmussen et al. 2000; Upadhyaya et al. 2008). Most frequent somatic inactivation reported has been large deletions, frequently encompassing numerous genes (Upadhyaya et al. 2008). By contrast, the inherited hit comprises a more complex spectrum of sequence changes. These comprise frameshifts, splice-site mutations, nonsense mutations, large deletions and infrequent missense mutations (Upadhyaya et al. 2008). It was not surprising that the majority of somatic (~94%) and germline (~84%) mutations in the *NF1* gene were truncating mutations (Table 16). Moreover, a preponderance of germline missense mutations was evident, when compared to somatic missense mutations (~98% germline vs. ~2% somatic missense mutations).

Therefore, at least with respect to inactivating mutations (e.g. micro-lesions ≤20bp), both the germline and the soma exhibit very similar frequencies. Nevertheless, it is clear that the germline mutational spectrum is more likely to comprise missense mutations, than the somatic mutational spectrum. These germline missense mutations are also likely to contribute significantly towards tumour development, as indicated by the higher median Grantham difference between mutant and wild-type amino acids, when germline were compared to potential mutations.

In contrast to the *BRCA1* and *BRCA2* genes, the *NF1* gene was found to be the most evolutionarily conserved of the 17 genes studied. This is an indication that in the majority of sequence changes, selection eliminates deleterious mutations. Thus, it was not surprising that germline missense mutations were not preferentially found within evolutionarily conserved codons as compared to potential mutations, due to the fact that codons in the *NF1* gene are evolutionarily conserved throughout.

It is interesting to note that the inherited lesions comprise a variety of sequence changes with a sizeable proportion of missense mutations that are also very likely to significantly impair the function of the protein, where such mutations are virtually absent within the soma. This is an indication of slightly different mutational mechanisms that operate in the germline and soma. Thus, it has been speculated before that germline allelic loss could confer a "significant selective disadvantage on many cells" (Upadhyaya et al. 2008).

4.4.5. *PTEN* gene

The *PTEN* gene is the second most frequently mutated gene, after *TP53*, in human cancers (Simpson and Parsons 2001). It's mutational spectrum comprises frameshifts, nonsense and missense mutations (Yin and Shen 2008). Thus, not surprisingly, missense mutations were a relatively frequent event (~46% and ~35% for somatic and germline sequence changes respectively, Table 16), but those were predominantly somatic missense mutations (~82% somatic vs. ~9% germline and shared missense mutations).

Along with *NF1*, the *PTEN* gene is one of the most evolutionarily conserved gene out of the 17 tumour suppressor genes studied; hence most missense mutations are to be found in evolutionarily conserved sites. Thus, it is not surprising that neither germline or somatic, nor shared missense mutations showed preferential location within evolutionarily conserved codons, when compared to potential mutations. This notwithstanding, most of the missense mutations are likely to be found in evolutionarily conserved sites, just because *PTEN* is highly conserved for most parts of the gene.

On the other hand, both the somatic and germline missense mutations exhibited similar nearest neighbour-dependent mutability rates; thus, similar mechanisms contribute towards the mutational spectrum in both the soma and the germline. Furthermore, mutations found in both the soma and the germline were found to have significantly higher median mutability rate than somatic missense mutations. Hence, one could speculate that similar mechanisms in the germline and soma, quite possibly due to sequence-dependent mechanisms, result in missense mutations in specific codons that are shared between the germline and the soma. Such hotspots that were found to be significantly overrepresented, were CpG dinucleotides. As a result, it is very likely that intra-genic hypermethylation of the *PTEN* gene is present in both the soma and the germline. Somatic promoter hypermethylation has indeed been confirmed in familial cerebral cavernous malformations and acute lymphoblastic leukaemia (Montiel-Duarte et al. 2008; Zhu et al. 2009). Moreover, recurrent somatic missense mutations that were also found in the germline were more likely to be found in CpG dinucleotides, than non-recurrent somatic mutations found exclusively in the soma. This is likely to be a result of heavy intra-genic methylation in the soma, but crucially it is also likely to be the case within the germline.

4.4.6. *TP53* gene

The *TP53* gene is regarded as the "guardian of the genome" (Lane 1992), as it is the most frequently (>50% of all human cancers; Toledo and Wahl 2006) mutated gene in human cancers (Levine et al. 2004; Vogelstein et al. 2000) and plays a vital role in crucial functions, such as activating cell cycle inhibitors and triggering apoptosis in response to DNA damage (Efeyan and Serrano 2007). Therefore, it was not surprising that ~66% of all somatic missense mutations for the 17 genes studied, were *TP53* mutations. Even more, ~45% of the combined spectrum of shared missense mutations in the 17 genes, were also *TP53* mutations. The *TP53* gene was also described by predominantly somatic missense mutations (~92%) as compared to virtually absent germline missense mutations (<1%).

Not surprisingly, the somatic missense mutations were found to be very likely to impair critical domains of the *TP53*, shown by the preferential location in evolutionarily conserved codons. They were also found to be influenced by sequence context, measured by disease and non-disease mutability rates. Therefore, the somatic missense mutational spectrum in the *TP53* gene is influenced by sequence-context, but it is also subject to negative selection. Thus, somatic missense mutations were likely to be selected for their likely negative impact on the function of the protein. Even although germline missense mutations were virtually absent from the missense mutational spectrum, a sizeable proportion of shared mutations was observed (~7%). Furthermore, ~94% of all germline missense mutations were also found in the soma. One could argue that due to the relatively small size of *TP53* (394 codons) and the large number of somatic mutations (~47% of all possible missense mutations, Table 11), shared missense mutations could be due to chance occurrence. The results obtained in this chapter have shown that missense mutations found in both the soma and the germline were not merely coincidental. In fact, they were extremely likely to impair the function of the protein, as compared to potential mutations. They were found in evolutionarily conserved sites and mutant amino acids were selected for their greater median physicochemical difference, with respect to wild-type amino acids. Additionally, shared missense mutations were more likely to be found in evolutionarily conserved sites than somatic missense mutations. **This suggests that shared missense mutations may be more detrimental to protein function than somatic missense mutations.** As a result, the majority of germline missense mutations are also found in the soma, but more importantly these shared lesions also seem to contribute significantly towards tumour development and/or progression.

It is well known that *TP53* germline mutations are associated with rare dominant inherited predisposition syndrome, i.e. Li-Fraumen (Malkin et al. 1990; Varley 2003). Thus, germline mutations are known to significantly contribute to cancer development. A closer look at the origin of the germline missense mutations in the *TP53* gene revealed that ~34% are derived from Li-Fraumeni syndrome patients (data not presented). Therefore, a logical conclusion would be that germline missense mutations are likely to be significantly associated not only with Li-Fraumeni syndrome, but may quite possibly play an important role in other cancers. A few studies have indicated that this could be so in the case of choroid plexus papiloma (Rutherford et al. 2002), glioblastoma and colon cancer (Yamada et al. 2009).

It has been recognized that the occurrence, frequency and distribution of mutations are shaped by mutational bias and/or selection. These two processes shape the mutational spectrum of *TP53* to such an extent that 11 hotspot mutations are found >100 times (Soussi et al. 2005). The relative contribution of these processes that shape the occurrence of these highly recurrent mutations is relatively unknown. The analysis of recurrent and non-recurrent somatic missense mutations presented herein, indicates that selection on the basis of functional impact on the protein might play a significant role. The recurrent somatic missense mutations were much more likely to be found in evolutionarily conserved codons, than non-recurrent ones. It is likely that evolutionarily conserved sites bear greater functional importance than relatively less conserved sites. Therefore, the recurrence status of somatic missense mutations could be linked to their functional importance; hence recurrent mutations are very likely to be drivers of tumorigenesis. Nevertheless, these recurrent somatic mutations were also found to be influenced to a much greater extent by nucleotide mutability rates than non-recurrent mutations. Thus, mutability could also play an important role. Furthermore, the great majority of recurrent somatic mutations were found in non-CpG dinucleotides (~95%). As a result, the significantly different nucleotide mutability rates between recurrent and non-recurrent mutations, could not be attributed to CpG dinucleotides. Moreover, the proportions of mutations found in CpG/non-CpG dinucleotides did not differ significantly between recurrent and non-recurrent somatic missense mutations; hence the mutability differences could not be attributed solely to mutations found in CpG dinucleotides. Thus, both nucleotide context and selection play important roles in the recurrence status of somatic missense mutations in the *TP53* gene. Additionally, it has been suggested that the selection criterion is towards gain-of-function mutations, i.e. positive selection (Glazko et al. 2006). Conversely, the proportion of recurrent somatic mutations that were also found in the germline was significantly greater, than the proportion of recurrent non-shared mutations co-localised within CpG-dinucleotides. In addition, recurrent and shared mutations were influenced to a significantly greater extent by nucleotide context than recurrent non-shared mutations. Thus, one could argue that recurrent shared missense mutations were more likely to be influenced by endogenous mutagenesis (i.e. spontaneous deamination of 5-methylcytosine within CpG dinucleotides) than mutations that recur, but are also not found in the germline. This is certainly a possibility, although the paucity of mutations and/or no true difference, precluded any conclusions for the comparison of recurrent and shared vs. recurrent non-shared missense mutations, with respect to both Grantham difference and evolutionary conservation.

The fact that shared CpG-located missense mutations were 100% recurrent, would suggest that intra-genic CpG methylation of the *TP53* gene is a relatively frequent event in both the soma and the germline. Indeed, promoter hypermethylation has been observed in breast cancer for both invasive and non-invasive lesions (Kang et al. 2001), but also tissue-independent complete intra-genic methylation (Tornaletti and Pfeifer 1995). Unfortunately, the scarcity of published reports on the intra-genic or promoter methylation status of the *TP53* gene in the germline precludes the possibility of support the prediction of possible germline intra-genic hypermethylation. Nevertheless, Magdinier et al. (2002) have shown a relatively high level of methylation of exon 4 of the *TP53* gene in blastocysts that supports the results and conclusions presented here to some degree.

4.4.7. *VHL* gene

Germline sequence changes in the *VHL* gene have been shown in the majority of patients with von Hippel-Lindau disease, associated with frequent tumours (Maher and Kaelin 1997). It is a tumour suppressor gene that has been shown to be consistent with Knudson's two hit hypothesis as point-mutations and deletions have been observed as first-second hits (Gnarra et al. 1997).

The somatic mutational spectrum comprises predominantly truncating mutations (~73%) as compared to equal proportions of truncating and non-truncating germline mutations (52% and 48% non-truncating and truncating mutations respectively). Thus, germline missense mutations were found to be a frequent event (~53%) as compared to somatic missense mutations (~23%).

Both the germline and the somatic missense mutations were found likely to be detrimental to the function of the protein. The germline and somatic missense mutations were more likely to be found in evolutionarily conserved codons than potential mutations. Moreover, germline missense mutations were also found to exhibit significantly higher Grantham differences between mutant and wild-type amino acids. Thus, it would seem that both germline and somatic missense mutations are very likely to play an important role in tumour development.

Furthermore, the mutational spectrum of the *VHL* gene was also described by a sizeable proportion of shared missense mutations (~24%). These were also found to preferentially target evolutionarily conserved codons, as compared to potential mutations. What is more, they exhibited a significantly higher median nucleotide disease-associated mutability rate when compared to potential mutations. Therefore, one may conclude that mutability and selection shape the distribution of missense mutations found in both the soma and the germline. Interestingly, both the somatic and germline mutational spectra are very likely to have been shaped in a similar way. Evidence comes from the observation that neither the somatic nor the germline missense mutations exhibited significant differences when compared to shared mutations, with respect to disease-associated mutability rates. Thus, we may infer that very similar mechanisms operate to influence both the germline and somatic missense mutations spectra in the *VHL* gene.

Reports have indicated an intricate first-hit second-hit relationship in the *VHL* gene. It has been reported that almost exclusively, whenever the inherited hit has been a deletion, the second hit is very likely to be a point mutation (Vortmeyer et al. 2002). Vortmeyer et al. (2002) have termed the relationship "mutation-deletion sequence". Furthermore, homozygous inactivation of the *VHL* gene has been reported to lead to embryonic lethality in mice (Gnarra et al. 1997). Hence it is likely that homozygous deletions as first and second hits are likely to be incompatible with cell survival. It is quite likely that this "mutation-deletion sequence" could potentially occur in reverse order. Thus, the germline mutation could be a missense mutation and the somatic hit a deletion. This is supported by reports that the second-hit usually comprises deletions with variable sizes (Glasker et al. 2006), but more importantly both the germline and somatic missense mutations are equally likely to be detrimental to the protein function and are influenced by similar mechanisms. In addition, such a scenario has been previously described (Wait et al. 2004).

Therefore, an intricate relationship between the germline and somatic mutations could play an important role in von Hippel-Lindau disease. In addition, both the germline and somatic missense mutations are quite likely to be drivers of tumorigenesis, especially those mutations found in both the soma and the germline.

4.4.8. Final conclusions

Despite the absence of other studies to support some of the findings presented in this chapter, a number of important conclusions could be drawn. Missense mutations falling into different categories can exhibit clear differences in terms of pathogenicity. It would appear that germline and missense mutations found in both the soma and the germline may exhibit profound effects on tumour development and/or progression as compared to pure somatic missense mutations, at least for the *TP53* gene and possibly also for *CDKN2A*. It is quite possible that other types of somatic mutation (other than missense) could have a greater impact on tumour development in addition to inherited predisposition (e.g. germline missense mutations). These include mutations that completely abolish gene structure and/or function, such as gross deletions, insertions, indels, gene rearrangements, etc. The data on micro-lesions (≤20bp) in these 17 human tumour suppressor genes partially support such a hypothesis. For a number of genes, namely *APC, ATM, BRCA2, CDH1, PTCH, PTEN, RB1, STK11* and *TSC1,* the ratio of truncating lesions (i.e. nonsense mutations, micro-deletions, micro-insertions and micro-indels) to non-truncating mutations (i.e. missense mutations) indirectly indicates that truncating lesions are found relatively more frequently in the soma than in the germline. This is further supported by the fact that the combination of lesions in the soma for all genes as compared to the combination of lesions in the germline, showed significantly more (p-value<2.20E-16, Table 17) truncating lesions than non-truncating ones (0.46 and 0.23 for the soma and germline respectively, Table 17). In addition, cancer is usually perceived as a disorder of post-reproductive age. Hence, an age-related shift in DNA repair mechanisms could potentially indicate the more deleterious impact of somatic lesions as compared to germline ones. However, studies on somatic micro-lesions indicate that gender, age or tissue-specificity might not play an important role in determining the frequency of micro-lesions (Gonzalez et al. 2007). Further analysis could help to resolve the relative impact of somatic and germline lesions. One possible way of answering those questions is to have matched mutational data (i.e. germline and somatic lesions) from tumours (i.e. soma) and normal cells (i.e. germline) from the same individual. This would allow a direct comparison between the germline and the soma, with respect to relative impact on gene/protein structure and/or function. A slightly different approach is to have mutational data of matched DNA from tumour and normal cells from the same individual, for bi-allelic inactivation of relevant genes, where there is the presence or absence of a germline hit.

Germline and missense mutations found in both the germline and the soma are influenced to a greater extent by the nucleotide context than pure somatic missense mutations. Therefore, if missense mutations found in both the soma and germline are the result of similar mechanisms, this would indicate that endogenous mutagenesis could have a significant impact on tumour development, at least for missense mutations.

Taken together, the results and analysis presented herein strongly suggest that algorithms and methods that attempt to predict the relative impact on the function of genes and proteins with respect to disease-associated missense mutations, have to take into consideration the different mutational categories that these mutations fall into (i.e. somatic, germline, shared and recurrent). Thus, some categories of missense mutations are more likely to result in a disease phenotype than others.

Table 8 Possible single base-pair substitutions leading to the non-synonymous change of the wild-type amino acid encoded by a specific codon

	Position in the codon			
	1	2	3	All positions in the codon
Codon	Number of possible missense changes			
CTT	3	3	0	6
GCC	3	3	0	6
GGA	2	3	0	5
GTC	3	3	0	6
TGC	3	3	1	7
AGT	3	3	2	8
TGT	3	3	1	7
TCA	3	1	0	4
CGA	1	3	0	4
ATT	3	3	1	7
TAT	3	3	0	6
ATC	3	3	1	7
AAC	3	3	2	8
AGC	3	3	2	8
TAC	3	3	0	6
AAT	3	3	2	8
ACT	3	3	0	6
ACA	3	3	0	6
TCG	3	2	0	5
GAC	3	3	2	8
CAA	2	3	2	7
CCG	3	3	0	6
CTG	2	3	0	5
GGT	3	3	0	6
GCA	3	3	0	6
AAG	2	3	2	7
GTG	3	3	0	6
TCC	3	3	0	6
TTT	3	3	2	8
AGG	2	3	2	7
CAC	3	3	2	8
GTT	3	3	0	6
CGT	3	3	0	6
CGG	2	3	0	5
CAT	3	3	2	8
ATA	3	3	1	7
AGA	1	3	2	6
GGG	3	3	0	6
CCC	3	3	0	6
ACC	3	3	0	6
GAG	2	3	2	7
TTA	2	1	2	5
CCA	3	3	0	6
GAT	3	3	2	8
CTA	2	3	0	5
TCT	3	3	0	6
TGG	3	2	2	7
TTC	3	3	2	8
CGC	3	3	0	6
CTC	3	3	0	6
GCG	3	3	0	6
TTG	2	2	2	6
GGC	3	3	0	6
GAA	2	3	2	7
GCT	3	3	0	6
CAG	2	3	2	7
CCT	3	3	0	6
ACG	3	3	0	6
AAA	2	3	2	7
ATG	3	3	3	9
GTA	3	3	0	6
Total	166	176	50	392

Table 9 Shared recurrent missense mutations and shared missense mutations found in CpG dinucleotides

Mutations	Shared	CpG-located		Recurrent		Recurrent and CpG-located			
Gene	F_{SH}	F_{CpG}	F_{CpG}/F_{SH}	F_{REC}	F_{REC}/F_{SH}	F_{REC_CpG}	F_{REC_CpG}/F_{SH}	F_{REC_CpG}/F_{CpG}	F_{REC_CpG}/F_{REC}
APC	1	0	0.00	0	0.00	0	0.00	N/A	N/A
ATM	0	0	N/A	0	N/A	0	N/A	N/A	N/A
BRCA1	1	0	0.00	1	1.00	0	0.00	N/A	0.00
BRCA2	1	1	1.00	0	0.00	0	0.00	0.00	N/A
CDH1	1	0	0.00	0	0.00	0	0.00	N/A	N/A
CDKN2A	28	7	0.25	1	0.04	0	0.00	0.00	0.00
NF1	0	0	N/A	0	N/A	0	N/A	N/A	N/A
NF2	0	0	N/A	0	N/A	0	N/A	N/A	N/A
PTCH	1	0	0.00	0	0.00	0	0.00	N/A	N/A
PTEN	22	3	0.14	11	0.50	3	0.14	1.00	0.27
RB1	3	1	0.33	1	0.33	0	0.00	0.00	0.00
STK11	3	1	0.33	0	0.00	0	0.00	0.00	N/A
TP53	88	20	0.23	79	0.90	20	0.23	1.00	0.25
TSC1	0	0	N/A	0	N/A	0	N/A	N/A	N/A
TSC2	2	2	1.00	1	0.50	1	0.50	0.50	1.00
VHL	45	6	0.13	7	0.16	3	0.07	0.50	0.43
WT1	0	0	N/A	0	N/A	0	N/A	N/A	N/A
Total	196	41	0.21	101	0.52	27	0.14	0.66	0.27

F_i is the number of mutations, where $i \in \{SH, CpG, REC, REC\_CpG\}$

SH-shared, CpG-CpG-located , REC-recurrent , REC_CpG-recurrent and CpG-located

Values marked in red denote genes that made a relatively large contribution to the corresponding mutational spectrum

Table 10 Species and sequences used to estimate evolutionary conservation

Gene	***Species***	**cDNA sequence identifier**	**Protein sequence identifier**
APC	*Xenopus laevis*	U64442.1	AAB41671.1
	Bos taurus	XM_865627.1	XP_870720.1
	Rattus norvegicus	NM_012499.1	NP_036631.1
	Mus musculus	NM_007462.1	NP_031488.1
ATM	*Gallus gallus*	XM_417160.1	XP_417160.1
	Xenopus laevis	AY668954.1	AAT72929.1
	Rattus norvegicus	XM_236275.3	XP_236275.3
	Sus scrofa	AY587061	AAT01608.1
	Canis familiaris	XM_845871.1	XP_850964.1
	Mus musculus	NM_007499	NP_031525.1
BRCA1	*Gallus gallus*	*NM_204169.1*	*NP_989500.1*
	Xenopus laevis	*AF416868.1*	*AAL13037.1*
	Bos taurus	*NM_178573.1*	*NP_848668.1*
	Rattus norvegicus	*NM_012514.1*	*NP_036646.1*
	Canis familiaris	*NM_001013416.1*	*NP_001013434.1*
	Mus musculus	*NM_009764.2*	*NP_033894.2*
BRCA2	*Gallus gallus*	*NM_204276.1*	*NP_989607.1*
	Danio rerio	*XM_690042.1*	*XP_695134.1*
	Bos taurus	*XM_583622.2*	*XP_583622.2*
	Rattus norvegicus	*NM_031542.1*	*NP_113730.1*
	Canis familiaris	*NM_001006653.4*	*NP_001006654.2*
	Mus musculus	*NM_009765.1*	*NP_033895.1*
CDH1	*Xenopus laevis*	*BC068940.1*	*AAH68940.1*
	Danio rerio	*NM_131820.1*	*NP_571895.1*
	Bos taurus	*NM_001002763.1*	*NP_001002763.1*
	Rattus norvegicus	*NM_031334.1*	*NP_112624.1*
	Canis familiaris	*XM_536807.2*	*XP_536807.2*
	Mus musculus	*NM_009864.1*	*NP_033994.1*
CDKN2A	*Gallus galus*	*NM_204433.1*	*NP_989764.1*
	Takifugu rubripes	*AJ250231.1*	*CAC12808.1*
	Bos taurus	*XM_868375.1*	*XP_873468.1*
	Rattus norvegicus	*NM_031550.1*	*NP_113738.1*
	Canis familiaris	*XM_538685.2*	*XP_538685.2*
	Mus musculus	*AF044336.1*	*AAC08963.1*
NF1	*Gallus gallus*	XM_415914.1	XP_415914.1
	Takifugu rubripes	AF064564.2	AAD15839.1
	Rattus norvegicus	NM_012609.1	NP_036741.1
	Canis familiaris	XM_537738.2	XP_537738.2
	Mus musculus	NM_010897.1	NP_035027.1
NF2	*Gallus gallus*	NM_204497.2	NP_989828.2
	Danio rerio	NM_212951.1	NP_998116.1
	Bos taurus	XM_611643.2	XP_611643.2
	Rattus norvegicus	XM_341248.2	XP_341249.2
	Canis familiaris	XM_534729.2	XP_534729.2
	Mus musculus	NM_010898.2	NP_035028.2
PTCH	*Xenopus laevis*	AF302765.1	AAK15463.1
	Gallus gallus	NM_204960.1	NP_990291.1
	Danio rerio	NM_130988.1	NP_571063.1
	Meriones unguiculatus	AB188226.1	BAE78534.1
	Rattus norvegicus	NM_053566.1	NP_446018.1
	Mus musculus	NM_008957.1	NP_032983.1
PTCH	*Xenopus laevis*	AF144732.1	AAD46165.1
	Gallus gallus	XM_421555.1	XP_421555.1
	Bos taurus	XM_613125.2	XP_613125.2
	Canis familiaris	NM_001003192.1	NP_001003192.1
	Rattus norvegicus	NM_031606.1	NP_113794.1

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

Figure 15 Distribution of nucleotide substitution rates derived from Hess et al. (1994) for the somatic missense mutations in the *TP53* gene

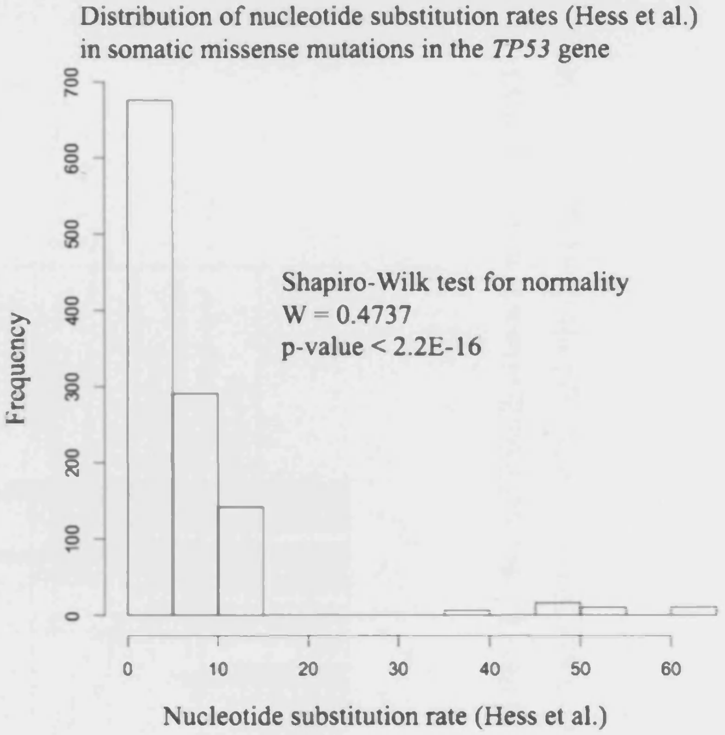
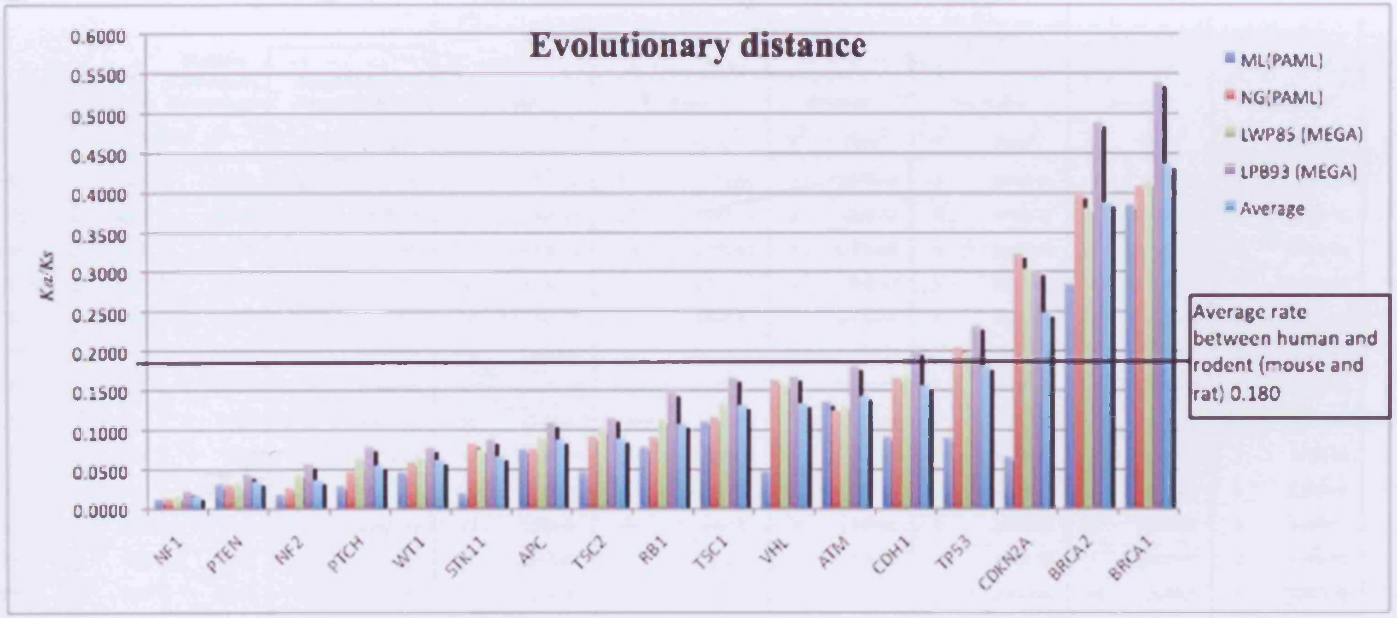


Figure 16 Rate of evolution in the studied 17 human tumour suppressor genes



ML- Maximum likelihood method; NG- Nei and Gojobori (1986) method; LWP85- Li, Wu and Pamillo-Bianchi method; LPB93- Li and Pamilo-Bianchi method; ML and NG are part of the PAML software package (Yang 1997); LWP85 and LPB93 are part of the MEGA (Kumar et al. 2004) software package

Table 11 Distribution of somatic and germline missense mutations in the 17 tumour suppressor genes studied

	Codons	Possible missense mutations	Possible CpG mutations		Missense mutations				CpG mutations of missense				CpG mutations of potential				Recurrent mutations		Shared mutations			
					Somatic		Germline		Somatic		Germline		Somatic		Germline		Somatic		Somatic		Germline	
Gene	N[1]	N[1]	N[1]	Freq[2]	N[1]	Freq[2]	N[1]	Freq[2]	N[1]	Freq[2]	N[1]	Freq[2]	N[1]	Freq[2]	N[1]	Freq[2]	N[1]	Freq[2]	N[1]	Freq[2]	N[1]	Freq[2]
APC	2844	18727	131	7.00E-03	38	2.03E-03	22	1.17E-03	1	1.67E-02	3	5.00E-02	1	7.63E-03	3	2.29E-02	4	1.05E-01	1	2.63E-02	1	4.55E-02
ATM	3057	20309	118	5.81E-03	11	5.42E-04	75	3.69E-03	1	1.16E-02	6	6.98E-02	1	8.47E-03	6	5.08E-02	1	9.09E-02	0	0.00E+00	0	0.00E+00
BRCA1	1864	12497	57	4.56E-03	5	4.00E-04	169	1.35E-02	0	0.00E+00	9	5.17E-02	0	0.00E+00	9	1.58E-01	0	0.00E+00	1	2.00E-01	1	5.92E-03
BRCA2	3419	22814	94	4.12E-03	20	8.77E-04	85	3.73E-03	2	1.90E-02	9	8.57E-02	2	2.13E-02	9	9.57E-02	2	1.00E-01	1	5.00E-02	1	1.18E-02
CDH1	883	5840	105	1.80E-02	14	2.40E-03	18	3.08E-03	1	3.13E-02	4	1.25E-01	1	9.52E-03	4	3.81E-02	0	0.00E+00	1	7.14E-02	1	5.56E-02
CDKN2A	157	1023	67	6.55E-02	170	1.66E-01	34	3.32E-02	26	1.27E-01	2	9.80E-03	26	3.88E-01	2	2.99E-02	6	3.53E-02	28	1.65E-01	28	8.24E-01
NF1	2819	18723	164	8.76E-03	2	1.07E-04	83	4.43E-03	0	0.00E+00	4	4.71E-02	0	0.00E+00	4	2.44E-02	0	0.00E+00	0	0.00E+00	0	0.00E+00
NF2	596	4004	70	1.75E-02	23	5.74E-03	20	5.00E-03	2	4.65E-02	2	4.65E-02	2	2.86E-02	2	2.86E-02	3	1.30E-01	0	0.00E+00	0	0.00E+00
PTCH	1448	9556	259	2.71E-02	13	1.36E-03	23	2.41E-03	2	5.56E-02	2	5.56E-02	2	7.72E-03	2	7.72E-03	0	0.00E+00	1	7.69E-02	1	4.35E-02
PTEN	404	2729	19	6.96E-03	201	7.37E-02	23	8.43E-03	4	1.79E-02	1	4.46E-03	4	2.11E-01	1	5.26E-02	47	2.34E-01	22	1.09E-01	22	9.57E-01
RB1	929	6136	65	1.06E-02	22	3.59E-03	34	5.54E-03	3	5.36E-02	2	3.57E-02	3	4.62E-02	2	3.08E-02	1	4.55E-02	3	1.36E-01	3	8.82E-02
STK11	434	2914	102	3.50E-02	17	5.83E-03	27	9.27E-03	4	9.09E-02	2	4.55E-02	4	3.92E-02	2	1.96E-02	2	1.18E-01	3	1.76E-01	3	1.11E-01
TP53	394	2604	60	2.30E-02	1138	4.37E-01	6	2.30E-03	30	2.62E-02	0	0.00E+00	30	5.00E-01	0	0.00E+00	781	6.86E-01	88	7.73E-02	88	1.47E+01
TSC1	1165	7709	104	1.35E-02	2	2.59E-04	7	9.08E-04	0	0.00E+00	1	1.11E-01	0	0.00E+00	1	9.62E-03	0	0.00E+00	0	0.00E+00	0	0.00E+00
TSC2	1808	11880	334	2.81E-02	0	0.00E+00	87	7.32E-03	0	0.00E+00	13	1.49E-01	0	0.00E+00	13	3.89E-02	0	0.00E+00	2	0.00E+00	2	2.30E-02
VHL	214	1406	69	4.91E-02	43	3.06E-02	98	6.97E-02	4	2.84E-02	2	1.42E-02	4	5.80E-02	2	2.90E-02	5	1.16E-01	45	1.05E+00	45	4.59E-01
WT1	450	3003	101	3.36E-02	1	3.33E-04	39	1.30E-02	0	0.00E+00	7	1.75E-01	0	0.00E+00	7	6.93E-02	0	0.00E+00	0	0.00E+00	0	0.00E+00
total	22885	151874	1919	1.26E-02	1720	1.13E-02	850	5.60E-03	80	3.11E-02	69	2.68E-02	80	4.17E-02	69	3.60E-02	852	4.95E-01	196	1.14E-01	196	2.31E-01

[1]-Number of missense mutations

[2]-Frequency

Table 12 Distribution of somatic, germline and shared mutations

Mutations	Missense	Somatic				Germline				Shared			
Gene	F_M	F_S	F_S/F_M	F_S/F_{ST}	F_S/F_{MT}	F_G	F_G/F_M	F_G/F_{GT}	F_G/F_{MT}	F_{SH}	F_{SH}/F_M	F_{SH}/F_{SHT}	F_{SH}/F_{MT}
APC	61	38	0.62	0.02	0.01	22	0.36	0.03	0.01	1	0.02	0.01	0.00
ATM	86	11	0.13	0.01	0.00	75	0.87	0.09	0.03	0	0.00	0.00	0.00
BRCA1	175	5	0.03	0.00	0.00	169	0.97	0.20	0.06	1	0.01	0.01	0.00
BRCA2	106	20	0.19	0.01	0.01	85	0.80	0.10	0.03	1	0.01	0.01	0.00
CDH1	33	14	0.42	0.01	0.01	18	0.55	0.02	0.01	1	0.03	0.01	0.00
CDKN2A	232	170	0.73	0.10	0.06	34	0.15	0.04	0.01	28	0.12	0.14	0.01
NF1	85	2	0.02	0.00	0.00	83	0.98	0.10	0.03	0	0.00	0.00	0.00
NF2	43	23	0.53	0.01	0.01	20	0.47	0.02	0.01	0	0.00	0.00	0.00
PTCH	37	13	0.35	0.01	0.00	23	0.62	0.03	0.01	1	0.03	0.01	0.00
PTEN	246	201	0.82	0.12	0.07	23	0.09	0.03	0.01	22	0.09	0.11	0.01
RB1	59	22	0.37	0.01	0.01	34	0.58	0.04	0.01	3	0.05	0.02	0.00
STK11	47	17	0.36	0.01	0.01	27	0.57	0.03	0.01	3	0.06	0.02	0.00
TP53	1232	1138	0.92	0.66	0.41	6	0.00	0.01	0.00	88	0.07	0.45	0.03
TSC1	9	2	0.22	0.00	0.00	7	0.78	0.01	0.00	0	0.00	0.00	0.00
TSC2	89	0	0.00	0.00	0.00	87	0.98	0.10	0.03	2	0.02	0.01	0.00
VHL	186	43	0.23	0.03	0.02	98	0.53	0.12	0.04	45	0.24	0.23	0.02
WT1	40	1	0.03	0.00	0.00	39	0.98	0.05	0.01	0	0.00	0.00	0.00
	(F_{MT})	(F_{ST})				(F_{GT})				(F_{SHT})			
total	2766	1720	0.62	1.00	0.62	850	0.31	1.00	0.31	196	0.07	1.00	0.07

F_i is the number of mutations, where $i \in \{M, S, G, SH, MT, GT, ST, SHT\}$

M-missense, S-somatic , G-germline , SH-shared, MT-missense total, GT-germline total, ST-somatic total, SHT-shared total

Marked in red are genes that contribute to a relatively greater extent to the somatic, germline or shared mutational spectrum for all genes combined

Table 13 Summary of statistically significant results in the studied 17 tumour suppressor genes

	Parameter	ALL	*APC*	*ATM*	*BRCA1*	*BRCA2*	*CDH1*	*CDKN2A*	*NF1*	*NF2*	*PTCH*	*PTEN*	*RB1*	*STK11*	*TP53*	*TSC1*	*TSC2*	*VHL*	*WT1*
Soma vs. pot	Hess	🟪↑	🟪↑			🟩		🟪↑			🟩	🟪↑	🟩	🟩	🟪↑				
	Krawczak	🟪↑	🟪↑			🟩		🟪↑		🟩	🟩↑	🟪↑	🟩	🟪↑	🟪↑				
	CpG	🟪↑		↑		🟩↑	🟩	🟪↑					🟪↑	↑					
	Grantham	🟩↑																	
	Evolution	🟪↓	🟩					↓				🟩			🟪↓			🟨↓	
Germ vs. pot	Hess	🟪↑	🟩	🟪↑	🟪↑	🟪↑	🟩		🟪↑	🟩	🟩	🟩	🟪↑				🟪↑		🟪↑
	Krawczak	🟪↑	🟩	🟪↑	🟪↑	🟪↑	🟪↑		🟪↑	🟩	🟩	🟪↑	🟪↑	🟩			🟪↑		🟪↑
	CpG	🟪↑	🟩↑	🟪↑	🟩↑	🟩↑	🟩↑		↑	↑							🟪↑		🟩
	Grantham	🟪↑		↑					🟨↑									🟪↑	
	Evolution	🟪↓		🟨↓	🟪↓			↓	🟩			🟩					🟩↓	🟪↓	🟩↓
Shared vs. pot	Hess	🟪↑						🟪↑				🟪↑		↑	🟪↑			🟩	
	Krawczak	🟪↑						🟪↑				🟪↑			🟪↑			🟪↑	
	CpG	🟪↑				🟪↑		🟪↑				🟪↑	🟪↑		🟪↑		🟪↑		
	Grantham	🟪↑													↑				
	Evolution	🟪↓						🟨↓				🟩			🟪↓			🟪↓	
Obs vs. pot	Hess	🟪↑	🟪↑	🟪↑	🟪↑	🟪↑	🟩↑	🟪↑	🟪↑	🟩	🟩	🟪↑	🟪↑	🟩	🟪↑		🟪↑	🟩	🟪↑
	Krawczak	🟪↑	🟪↑	🟪↑	🟪↑	🟪↑	🟪↑	🟪↑	🟪↑	🟩	🟪↑	🟪↑	🟪↑	🟪↑	🟪↑	🟩	🟪↑	🟩↑	🟪↑
	CpG	🟪↑	🟩↑	🟪↑	🟩↑	🟪↑	🟩↑	🟪↑	↑	↑		🟩↑	🟪↑	↑	🟩↑		🟪↑		🟩
	Grantham	🟪↑							🟨↑			🟨↑			🟩↑			↑	
	Evolution	🟪↓	🟩	🟪↓	🟪↓			🟪↓	🟩	🟩		🟩		🟩	🟪↓		🟩↓	🟪↓	🟩↓
Soma vs. germ	Hess	🟪↓						🟩						🟩					
	Krawczak	🟪↓					🟩	🟩				🟩		🟩					
	CpG	🟩↑																	
	Grantham	🟩↓																	
	Evolution	🟩↑										🟩							
Soma vs. shared	Hess	🟪↓						🟩				🟩↓			🟪↓				
	Krawczak	🟪↓						🟩				🟩↓			🟪↓			🟩	
	CpG	🟪↓													🟪↓				
	Grantham																		
	Evolution	🟪↑						↑				🟩			🟨↑				
Germ vs. shared	Hess	🟩						🟩							🟩			🟩	
	Krawczak	🟩↓						🟩				🟩			🟩			🟩	
	CpG	🟪↓																	
	Grantham																		
	Evolution	🟩																	
Rec vs. non-rec	Hess	🟩										🟩			🟪↑				
	Krawczak	🟩										🟩			🟪↑				
	CpG			🟩															
	Grantham																		
	Evolution	🟪↓										🟩			🟪↓				
Rec shared vs. rec non-shared	Hess	🟪↑										🟩			🟪↑				
	Krawczak	🟪↑										🟩			🟪↑				
	CpG	🟪↑													🟪↑				
	Grantham																		
	Evolution																		
Non-rec shared vs. non-rec non-shared	Hess	🟩↑																	
	Krawczak	🟩↑						🟩											
	CpG	🟩↑																	
	Grantham																		
	Evolution	🟩↓						↓											
Rec shared vs. non-rec non-shared	Hess	🟪↑										🟪↑			🟪↑				
	Krawczak	🟪↑										🟩↑			🟪↑				
	CpG	🟪↑										🟩↑			🟪↑				
	Grantham																		
	Evolution	🟪↓													🟪↓				

Legend: ↑ or ↓ shows the direction of gene- or experiment-wise statistically significant results. The direction is with respect to the first group in the comparison. 🟪 Grey shaded box represents an experiment-wise statistically significant result, non-shaded arrow (i.e. ↑ or ↓) represents a gene-wise statistically significant result, 🟩 Green shaded box represents ≥80% power to detect a statistically significant result for the comparison and associated effect size, 🟨 Yellow shaded box represents ≤80% power and experiment-wise statistically significant result; Soma- Somatic, Germ- Germline, Obs.- Observed (somatic, germline and shared), Pot.- Potential, Rec.- Recurrent, Non-rec.- Non-recurrent;

Table 14 Gene-wise somatic and germline missense mutations combined for all genes

	Mutation rate		Disease-associated mutation rate		Evolutionary variation rate		Grantham score		CpG-located missense mutations	
	Gene symbol	Median	Gene symbol	Median	Gene symbol	Median	Gene symbol	Median	Gene symbol	%
Somatic mutations			*STK11*	1.66					*STK11*	24
			PTCH	1.06						
	APC	8.4	*CDKN2A*	1.01	*CDKN2A*	0.38			*CDKN2A*	15
	CDKN2A	7.9	*APC*	0.83						
	PTEN	5.6	*PTEN*	0.53						
	TP53	4.6	*TP53*	0.5	*TP53*	0.17			*RB1*	14
					VHL	0.14			*BRCA2*	10
									ATM	9
for all 17 genes combined	somatic	4.7	somatic	0.53	somatic	0	somatic	78	somatic	5
	potential	4.1	potential	0.4	potential	0.2	potential	74	potential	1
	germline	7.2	germline	0.81	germline	0	germline	94	germline	8
Germline mutations	*TSC2*	7.3			*TSC2*	0				
	BRCA1	8.7							*BRCA1*	5
	NF1	7.3					*NF1*	98	*NF1*	5
	ATM	7.9	*ATM*	0.79	*ATM*	0	*ATM*	98	*ATM*	8
	RB1	7.6	*BRCA1*	0.81	*VHL*	0	*VHL*	99	*NF2*	10
	BRCA2	8.7	*BRCA2*	0.81					*BRCA2*	11
			PTEN	0.92						
			RB1	0.99						
			NF1	1.03						
			TSC2	1.03					*APC*	14
	WT1	10.1	*WT1*	1.22	*WT1*	0			*TSC2*	15
			CDH1	1.27	*BRCA1*	0.14			*CDH1*	22
					CDKN2A	0.29				

Legend: Genes exhibiting gene-wise or experiment-wise (shaded in grey) statistically significant results for the somatic vs. potential and germline vs. potential missense mutations

Table 15 Inequalities between somatic, germline and shared missense mutations for all genes combined

Parameter	Observed trend (p<0.05)
Median mutability rate with respect to Hess et al. (1994)	**shared>germline>>somatic** [8.7] [7.2] [4.7]
Median disease-associated mutability rate with respect to Krawczak et al. (1998)	**shared>germline>>somatic** [1.06] [0.81] [0.53]
Mean and median evolutionary variability	**shared << somatic** [0.10;0] [0.22;0] **somatic >> germline** [0.22;0] [0.18;0]
Median Grantham score	**somatic < germline** [78] [94]
Proportion of CpG-located mutations	**shared>>germline>somatic** [0.21] [0.08] [0.05]

Legend: >> and << indicate experiment-wise statistical significance; > indicates only gene-wise statistical significance; median or median and mean values are presented in square brackets

Table 16 Truncating vs. non-truncating lesions

Gene		Missense	Nonsense	Micro-deletions	Micro-insertions	Micro-indels	Non-truncating	Truncating	Non-truncating/Truncating
APC	Somatic	39	79	152	44	3	39	278	0.14
	Germline	23	180	299	115	12	23	606	0.04
ATM	Somatic	11	7	4	1	0	11	12	0.92
	Germline	76	75	122	35	14	76	246	0.31
BRCA1	Somatic	6	9	9	5	0	6	23	0.26
	Germline	170	121	259	85	12	170	477	0.36
BRCA2	Somatic	21	1	8	4	0	21	13	1.62
	Germline	86	76	247	90	11	86	424	0.20
CDH1	Somatic	15	7	13	2	0	15	22	0.68
	Germline	19	11	12	8	1	19	32	0.59
CDKN2A	Somatic	198	18	77	25	8	198	128	1.55
	Germline	62	7	11	7	2	62	27	2.30
NF1	Somatic	2	11	16	3	0	2	30	0.07
	Germline	83	115	221	105	8	83	449	0.18
NF2	Somatic	23	42	182	28	6	23	258	0.09
	Germline	20	43	55	16	2	20	116	0.17
PTCH	Somatic	14	9	14	6	1	14	30	0.47
	Germline	24	27	42	32	8	24	109	0.22
PTEN	Somatic	226	56	152	51	4	226	263	0.86
	Germline	45	28	29	22	3	45	82	0.55
RB1	Somatic	25	27	34	12	3	25	76	0.33
	Germline	37	76	117	53	11	37	257	0.14
STK11	Somatic	20	10	5	1	1	20	17	1.18
	Germline	30	27	47	24	3	30	101	0.30
TP53	Somatic	1229	96	512	238	0	1229	846	1.45
	Germline	94	10	16	5	3	94	34	2.76
TSC1	Somatic	2	1	1	0	0	2	2	1.00
	Germline	7	37	53	25	4	7	119	0.06
TSC2	Somatic	2	1	3	2	1	2	7	0.29
	Germline	89	74	110	46	3	89	233	0.38
VHL	Somatic	88	15	180	44	1	88	240	0.37
	Germline	143	27	63	37	5	143	132	1.08
WT1	Somatic	1	3	4	3	0	1	10	0.10
	Germline	40	14	8	4	1	40	27	1.48
Total	Somatic	1922	392	1366	469	28	1922	2255	0.85
	Germline	1048	948	1711	709	103	1048	3471	0.30

Table 17 Truncating vs. non-truncating somatic and germline mutations for all genes combined

Somatic				Germline					
Truncating		Non-truncating		Truncating		Non-truncating			
Number	Frequency	Number	Frequency	Number	Frequency	Number	Frequency	χ^2	p-value
1922	0.46	2255	0.54	1048	0.23	3471	0.77	1389.2130	<2.20E-16

5. Nonsense mutations

5.1. Introduction

5.1.1. Gene expression in eukaryotes

The gene expression pathway in eukaryotes comprises a number of interconnected steps. A key player is the mRNA, an intermediate between the genetic information stored in DNA and the process of translation into a protein. The mRNA precursor is transcribed from DNA and is transformed into a mature mRNA by removal of all or certain introns (alternative splicing) and addition of the terminal m(7)GpppN cap (which facilitates ribosome binding) and the poly (A) tail. The mature mRNA is then exported to the cytoplasm where it is generally translated into protein before ultimately being degraded.

Even although the complexity of the gene expression machinery allows control at many different levels, gene expression is susceptible to errors. These errors include mutations within the coding regions of the genes via incorrect processing (i.e. splice site mutations), frameshifts (e.g. deletions, insertions and indels) or nonsense mutations (Baker and Parker 2004; Hilleren and Parker 1999).

5.1.2. Importance of nonsense mutations and their functional consequences

An open translational reading frame (ORF) contains a sequence of bases within the mRNA that could potentially encode a protein. Generally, ORFs begin with an AUG initiation codon and end with a termination (or stop) codon (*viz.* UAA, UAG, UGA). Apart from the 'naturally-occurring termination codons' (Mort et al. 2008), various aberrations could introduce a premature termination codon (PTC) within the gene coding region thereby interrupting the ORF. These include single base-pair substitutions that directly introduce termination codons, and intra-genic (i.e. within the gene coding region) frameshift mutations (i.e. insertions, deletions and indels, where the size of the deleted or inserted bases is not divisible by three) or mutations that give rise to inefficient or inaccurate intron removal (e.g. intron retention) from the pre-mRNA or alternatively spliced mRNAs (Mendell et al. 2004) that result in the use of a non-natural termination codon. In addition, a small proportion

(0.05% to 0.5%) of human mRNA transcripts are estimated to acquire a PTC through various transcription errors (Muhlemann et al. 2008).

One commonly occurring mutation is the single base-pair substitution that introduces a stop codon (nonsense mutation or premature termination codon) within the coding region of a gene that, in a majority of cases, leads to premature translational termination. Nonsense mutations (as a result of a single base-pair substitution) are an integral and important part of the mutational spectrum associated with inherited disease. In this context, it has been estimated that nonsense mutations are twice as likely to come to clinical attention as the most extreme missense mutations (extreme in terms of the chemical difference between the substituted wild-type and substituting amino acid residues), and three times more likely to come to clinical attention than the average amino acid change (Krawczak et al. 1998). Numerous inherited diseases have been associated with premature termination codons, including cystic fibrosis, Duchenne muscular dystrophy, Hurler syndrome, several types of cancer, β-thalassemia, Marfan syndrome, etc (Culbertson 1999; Frischmeyer and Dietz 1999; Holbrook et al. 2004). Indeed, they account for ~20% of all disease-associated exonic single base-pair substitutions logged in the Human Gene Mutation Database (*HGMD*; http://www.hgmd.org; Mort et al. 2008; Stenson et al. 2003). Furthermore, others have estimated that 35% of all human alternatively spliced mRNAs harbour a premature termination codon (Lewis et al. 2003).

Nonsense mutations are important because they are 'equivalent to nonsense sequences' (Kuzmiak and Maquat 2006) and generally their occurrence precludes the synthesis of a full length protein. If translated (assuming of course that the mRNA is stable enough to support translation), these mutant forms (i.e. containing a PTC in the ORF) could have very limited or even no function at all (e.g. loss of key domains or amino acids). Thus, premature terminational translation could be regarded as a waste of energy expended on non-functional proteins (Seligmann and Pollock 2004). By contrast, some mutations would lead to a 'gain-of-function' if they were to give rise to a dominant negative form over the wild-type. Dominant negative forms usually arise in dimeric or multimeric proteins, such as that encoded by the *TP53* gene. Therefore, mutant forms of the protein could heterotetramerize (as in the case of *TP53*) with the wild-type product and exert a dominant negative effect over the naturally occurring product. In the given example of *TP53*, this effect is expressed in reduced DNA binding and transactivation of its target genes, *CDKN1A*, *MDM2* and *PIG2* (Willis et al. 2004). It has to be noted that the great majority of tumour suppressor mutations, whether inherited or somatic, are loss-of-function mutations (Sherr 2004). Nevertheless, numerous studies on tumour suppressor genes have identified mutations that present dominant negative effects over the wild-type product. Certain mutant forms of the *APC* gene are shown to exhibit such an effect and the mutant proteins encoded promote proliferation and chromosome instability (Dihlmann et al. 1999; Tighe et al. 2004). Other well studied tumour suppressor genes, such as the *ATM* (Chenevix-Trench et al. 2002; Oguchi et al. 2003; Scott et al. 2002), *CDH1* (Crane et al. 2004; Pfleger and Kirschner 2000), *BRCA1* (Deans et al. 2004; Hohenstein and Fodde 2003; Kim et al. 2003), *NF2* (Johnson et al. 2002), *PTCH* (Uchikawa et al. 2006), *PTEN* (Steelman et al. 2008), *RB1* (Li et al. 2000), *TSC2* (Rosner et al. 2003) and *WT1* (Han et al. 2007) genes, have also been shown to exhibit dominant negative forms. Some mutant forms of the 'guardian of the genome' (Vousden 2000), the *TP53* gene, lead to a 'loss-of-function' whereas others give rise to a potent dominant negative over the wild-type (Chan and Poon 2007; Hassan et al. 2008; Junk et al. 2008).

5.1.3. CpG dinucleotides and nonsense mutations

Cytosine (C) is subject to a post-replicative covalent modification in the form of methylation, converting cytosine into 5-methylcytosine (5mC), mainly in the context of CpG dinucleotides (Coulondre et al. 1978; Grippo et al. 1968). Spontaneous deamination of 5mC yields the DNA base thymidine (T) (Wang et al. 1982), whereas deamination of unmethylated cytosine generates uracil (U), an RNA base, which can be processed and removed by uracil DNA glycosylase (Lindahl 1974; Visnes et al. 2008). As is evident from numerous studies, this spontaneous deamination of 5mC is largely responsible for a greatly increased mutation rate at CpG dinucleotides (Cooper and Youssoufian 1988; Gaffney and Keightley 2008; Krawczak et al. 1998). Importantly, key tumour suppressor genes, such as *TP53* (Greenblatt et al. 1994) and *CDKN2A* (Pollock et al. 1996), are frequently found to exhibit mutations at CpG dinucleotides. Moreover, mutations at CpG sites that are compatible with a model of methylation-mediated deamination of 5mC (C>T on the coding strand and G>A on the non-coding strand) are found to account for 20-25% of all inherited mutations causing human disease (Cooper and Youssoufian 1988). The impact of this endogenous mechanism will largely depend, albeit indirectly, on the methylation patterns of the DNA sequences in question in the germline or the soma. Some reports suggest that the methylation status may differ between the germline and the soma. Indeed, it has been suggested that sperm cells may be hypermethylated as compared to oocytes, but both are hypomethylated with respect to somatic tissues (Allegrucci et al. 2005). In addition, there might be variation in the methylation pattern both within and between individuals (Millar et al. 1998) which would make it very difficult to extrapolate from methylation status or patterns to mutation rates in specific CpG dinucleotides.

The above notwithstanding, out of the 23 possible single base-pair substitutions that could lead to the introduction of a stop codon (as shown in Figure 19), the most frequent change is CGA->TGA, converting the arginine codon to a termination codon. Indeed, some 21% of all nonsense mutations causing human inherited disease logged in *HGMD* are located in CpG dinucleotides (Mort et al. 2008).

5.1.4. Quality control mechanisms and nonsense mutations

Although errors arise in the coding regions of genes and subsequently in the mRNA transcripts, it is unwise to assume that such transcripts invariably accumulate to give rise to defective or aberrant proteins. Indeed, eukaryotic cells have evolved quality control mechanisms to detect and eliminate abnormal mRNA transcripts (Fasken and Corbett 2005). Some of these abnormal mRNA transcripts are detected and degraded in the nucleus whereas others are degraded in the cytoplasm (Fasken and Corbett 2005). The surveillance mechanisms recognise transcripts that lack natural stop codons (Maquat 2002; Vasudevan et al. 2002), mRNAs that harbour premature termination codons and other imperfections (Gonzalez et al. 2001; Maquat 2004). These imperfections can include incomplete splicing, extended 3' untranslated regions (UTRs) and upstream open reading frames within the 5' UTRs of the genes (Mitrovich and Anderson 2000; Muhlrad and Parker 1999; Ruiz-Echevarria and Peltz 2000; Welch and Jacobson 1999).

5.1.5. Nonsense-mediated mRNA decay

One of the most studied quality control mechanisms is nonsense-mediated mRNA decay (NMD), which is also known as mRNA surveillance. It generally targets for degradation mRNAs that harbour premature termination codons (Baker and Parker 2004; Conti and Izaurralde 2005; Fasken and Corbett 2005; Holbrook et al. 2004; Lejeune and Maquat 2005; Maquat 2004; Wilkinson 2005; Yamashita et al. 2005).

Even although it may differ between species and is considered non-essential in lower eukaryotes such as worms and yeast (Medghalchi et al. 2001), the process and the proteins involved generally appear to be evolutionarily conserved and considerable evidence exists to suggest that this mechanism is essential for life (Amrani et al. 2004; Behm-Ansmant et al. 2007; Culbertson 1999; Gatfield et al. 2003; Longman et al. 2007; Maquat 2004).

5.1.5.1. Mechanism of NMD

The 'unified NMD model' (Muhlemann et al. 2008) proposes that an initial round of translation is required for nonsense codon recognition in which the UAA, UAG or UGA codons direct translational termination (Carter et al. 1995; Menon and Neufeld 1994; Qian et al. 1993), although PTC recognition entirely by the translating ribosome is insufficient for NMD (Zhang and Maquat 1997) and subsequent mRNA degradation. Thus, several lines of evidence suggest that the signal which distinguishes premature stop codons from *bona fide* naturally occurring stop codons is the exon junction complex (EJC; Frischmeyer and Dietz 1999; Hilleren and Parker 1999; Li and Wilkinson 1998). The EJC is a complex of proteins that is deposited as a consequence of pre-mRNA splicing (i.e. intron removal) upstream (i.e. 5') of the splicing-generated exon–exon junctions. The EJC is deposited 20-24 nucleotides upstream of exon-exon junctions after RNA splicing (Tange et al. 2004) in a sequence-independent manner (Le Hir et al. 2000; Muhlemann et al. 2008). As such, the EJC maintains the position of excised introns in the newly spliced mRNA (Chang et al. 2007). In normal mRNA transcripts, the advancing ribosome displaces the exon-exon junction complexes (Dostie and Dreyfuss 2002; Lejeune et al. 2002) as they are all deposited upstream of the stop codon. By contrast, there is generally at least one EJC downstream (or 3') of a premature termination codon, thereby triggering NMD. There is still an ongoing debate as to whether an EJC downstream of a PTC is actually required for PTC recognition or whether it simply functions as an NMD enhancer (Buhler et al. 2006; Muhlemann et al. 2008).

5.1.5.2. General rule for triggering NMD

According to the established rule, PTCs that are followed by an intron and are located more than 50–55 nucleotides (nt) upstream of the last exon-exon junction (EEJ), would generally elicit NMD (Nagy and Maquat 1998). Stated in terms of the spliced mRNA, PTCs that are followed by an exon–exon junction (measured after splicing) and which are located more than 50–55nt upstream (5') of the EEJ, would generally elicit NMD (the process is illustrated in Figure 17 and Figure 18).

Figure 17 General mechanism of nonsense-mediated mRNA decay

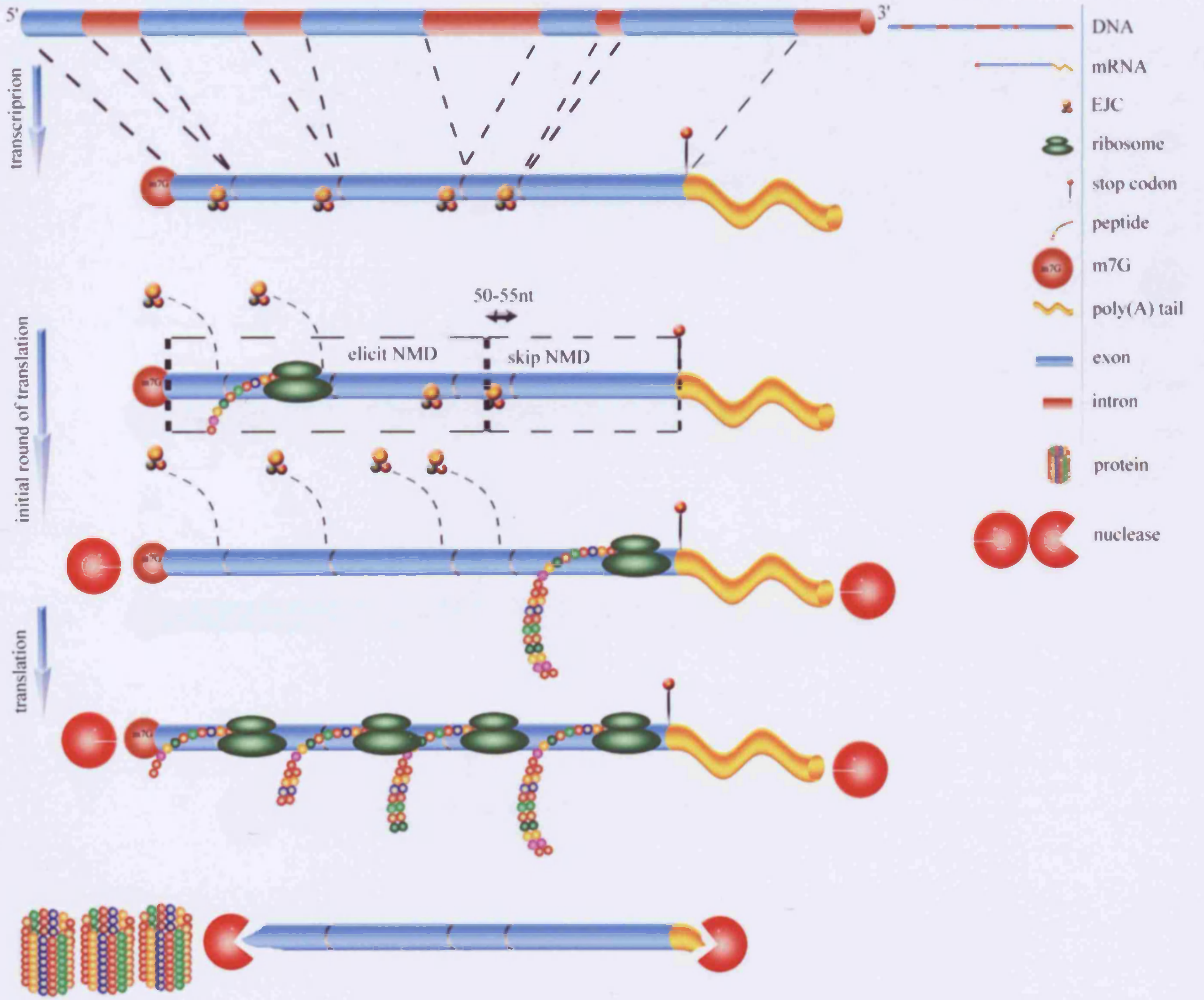
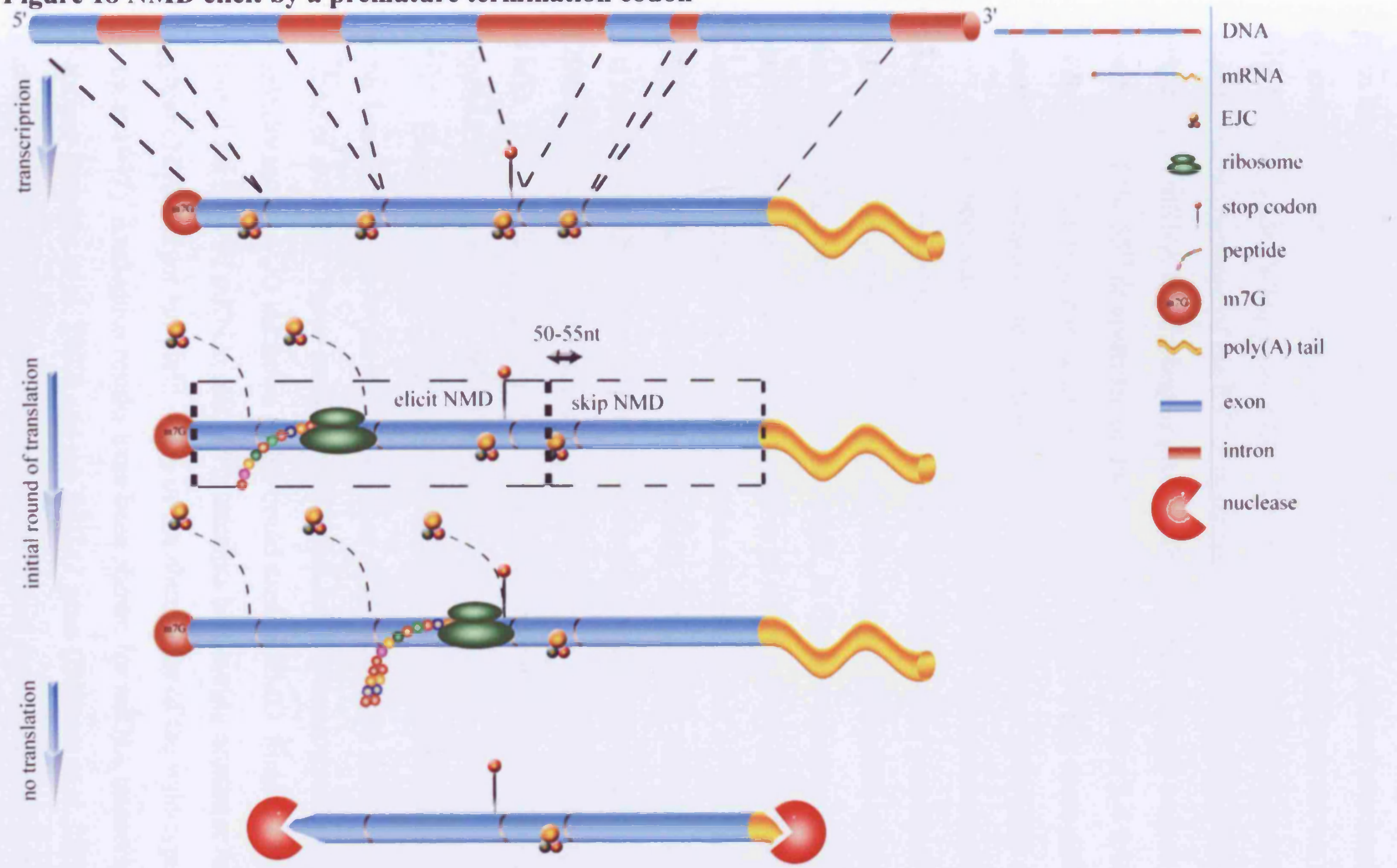


Figure 18 NMD elicit by a premature termination codon



5.1.5.3. Exceptions to the ~55nt boundary rule for triggering NMD

Even although it is commonly accepted that a PTC located more than 50-55nt upstream (5') of the last EEJ would generally elicit NMD, there are a few exceptions to this rule. Several studies have indicated that contrary to the ~55nt boundary rule, there could be a polar effect with respect to the position of the PTC. Hence, PTCs distal (in a 5' direction) from the last downstream intron trigger robust NMD, whereas proximally located PTCs generate a modest NMD response. Such a polar effect has been shown in the T-cell receptor beta (*TCRB*) gene (Wang et al. 2002). Wang et al. (2002) have shown that a distant nonsense codon (192nt upstream of the 55th nt upstream of the last intron) results in a ~50 fold reduction in mRNA level whereas proximally located nonsense codons (91nt and 142nt upstream of the 55th nt upstream of the last intron) led to only a 2-4 fold reduction in mRNA level. A possible explanation of this polar effect could be that more distal PTCs are associated with faster rates of deadenylation, than proximal ones (Cao and Parker 2003).

It has been suggested that introns might not be an absolute requirement for NMD to be triggered even although their presence might enhance the process of mRNA degradation. A transfected intronless *HEXA* minigene, containing a frameshift mutation or nonsense mutation in close proximity to the frameshift, in Chinese hamster ovary (CHO) cells, has been shown to yield half the normal wild-type mRNA level (Rajavel and Neufeld 2001). By contrast, the insertion of a spliceable intron downstream of a PTC in the naturally intronless *HSP70* gene results in the reduction of steady-state mRNA (Maquat and Li 2001). Therefore, it is unclear whether or not introns are essential for the triggering of NMD. Danckwardt et al. (2002) have suggested that aberrantly spliced mRNA transcripts are not always subject to NMD. These authors suggested that some genes contain *cis*-acting sequences that are required for triggering the NMD pathway.

Human β-globin (*HBB*) mRNAs that harbour nonsense mutations in the 5' region of exon 1 accumulate to levels similar to those of the wild-type mRNA (Inacio et al. 2004; Romao et al. 2000). These studies suggest that mRNA transcripts bearing PTCs in close proximity to the AUG initiation codon could escape NMD. Similar findings have been reported for the *TPI* mRNA; mRNA transcripts harbouring nonsense mutations are an apparent NMD target but yield ~84% of the abundance of the wild-type mRNA (Zhang and Maquat 1997). Analogous results have been shown for mRNA transcripts from the *RB1* gene (Sanchez-Sanchez et al. 2007) and the *BRCA1* gene (Buisson et al. 2006). These studies show that when an AUG codon is in close proximity and downstream of the premature termination codon, translation re-initiation could occur. Therefore, these apparent targets escape NMD and their cellular mRNA abundance is similar to the wild-type product.

Variability in NMD sensitivity has been reported from several studies, indicating that a tissue-specific response might exist. *CFTR* transcripts from different epithelial cell lines that carry the same nonsense mutation (W128X) have been shown to display different efficiencies in triggering NMD (Linde et al. 2007).

Because exceptions to the ~55nt boundary rule exist, additional and unidentified determinants that modulate the NMD sensitivity of these transcripts, might exist. These factors have yet to be discovered and our understanding of the mechanism of the NMD made more complete.

5.1.6. Why study the role of NMD in cancer?

Cancer is commonly viewed as a genetic disorder since sequence changes in somatic DNA are considered causative of neoplasms. Of the three basic types of gene that drive tumour development when mutated, i.e. oncogenes, genomic stability genes and tumour suppressor genes (Vogelstein and Kinzler 2004), the tumour suppressor genes generally require a 'biallelic gene inactivation' (Knudson 1971, 1978). Thus, one mutant allele, inherited through the germline together with subsequent somatic inactivation of the other allele has been the basis of the 'two-hit' hypothesis originally proposed by Knudson (Knudson 1971, 1978). However, subsequent studies have shown that some tumour suppressor genes are 'haploinsufficient for tumour suppression' (Payne and Kemp 2005). That is to say, one allele of the gene is insufficient to restore the function of the two copies of the wild-type product. By way of an example, a reduction in *TP53* gene dosage could be "sufficient to promote tumorigenesis" (Venkatachalam et al. 1998) via reduced level of apoptosis (Clarke et al. 1994) and reduced maintenance of genome integrity (Bouffler et al. 1995). Studies have also shown that some mutant forms of tumour suppressor genes exhibit dominant-negative effect over the wild-type product (examples of genes and references listed in 5.1.2). In this case, one mutant product of the gene binds to the wild-type protein and forms a non-functional complex. Therefore, there is only one mutant allele, but the end result is equivalent to a biallelic gene inactivation.

Depending upon the outcome of the NMD pathway (i.e. the decision as to whether or not to trigger NMD), nonsense mutations could confer a growth advantage on the cells through haploinsufficiency since some of those aberrant alleles, if transcribed (i.e. NMD skip), would give rise to protein products with limited or no function at all. By contrast, aberrant mRNA transcripts (e.g. harbouring nonsense mutations) could escape degradation through NMD, thereby releasing the potential for dominant-negative forms.

In the light of progress being made in the field of nonsense mutation read-through therapy (reviewed in Linde and Kerem 2008), exploring the similarities and differences in the distribution of potential NMD outcome is particularly timely.

5.1.7. Aims of the analysis

Bearing in mind the importance of nonsense mutations, it is perhaps surprising that few attempts have been made to explore the possible role of the NMD machinery in the development of cancer and more specifically in the context of tumour suppressor genes and their associated nonsense mutational spectrum.

The first and most important question to be addressed in this analysis is the involvement of NMD in the process of mutagenesis in a total of 17 different human tumour suppressor genes. The main objective of the analysis was to explore any similarities or differences that somatic and germline nonsense mutations might exhibit with respect to a potential/predicted NMD outcome. To accomplish this objective, a number of tasks were performed.

- **Definition of predicted NMD outcome**

Nonsense mutations were designated as either NMD elicit or NMD skip according to the commonly accepted ≥55nt rule explained in 5.1.5.2

- **Calculation of 'potential nonsense mutations'**

Potential nonsense mutations correspond to those codons in a gene coding sequence that could potentially give rise to a stop codon via a single base-pair substitution. The determination of potential nonsense mutations was performed in order to be able to determine whether or not observed nonsense mutations are non-randomly distributed with respect to predicted NMD outcome.

- **Assessment of the probability of finding nonsense mutations, with respect to predicted NMD status, by chance alone**

Observed nonsense mutations for each gene were subdivided into four categories. These included: exclusively somatic (only found in the soma); exclusively germline (only found in the germline); shared (found in both the soma and the germline); a combination of somatic, germline and shared. Each of these categories was compared to potential nonsense mutations, with respect to predicted NMD outcome. In addition, to assess the overall distribution of nonsense mutations (i.e. nonsense mutations in all genes) with respect to predicted NMD outcome, within each category, the nonsense mutations were combined for all genes (for further explanation, see 5.2.2.3).

- **Exploration of the similarities and differences in the positions of somatic and germline nonsense mutations with respect to predicted NMD outcome**

For each gene, the proportions of nonsense mutations were compared between the soma and the germline with respect to their predicted NMD outcome (i.e. NMD skip and/or NMD elicit). In a similar way, the proportions were calculated for the combination of somatic nonsense mutations observed for all genes and the corresponding combination of germline nonsense mutations for all genes. This was performed in order to assess the overall differences and similarities between the somatic and germline nonsense mutations found in all genes, with respect to predicted NMD outcome. In addition, for each gene the shared nonsense mutations were compared separately to the somatic and germline nonsense mutations with respect to predicted NMD outcome. This analysis was also performed for the combination of shared nonsense mutations for all genes and the combination of somatic, and the combination of germline nonsense mutations, for all genes.

These tasks were performed in order to provide meaningful answers to the following questions:

Are there more observed nonsense mutations (*viz.* somatic, germline and shared; combination of somatic, germline and shared nonsense mutations for **each gene and all genes**) predicted to skip or elicit NMD, than by would be expected by chance alone?

How do observed nonsense mutations (*viz.* somatic, germline and shared) compare to each other with respect to predicted NMD outcome?

Are there any differences/similarities between CpG- and non-CpG located nonsense mutations between somatic, germline, shared and potential nonsense mutations for each tumour suppressor gene and all genes combined?

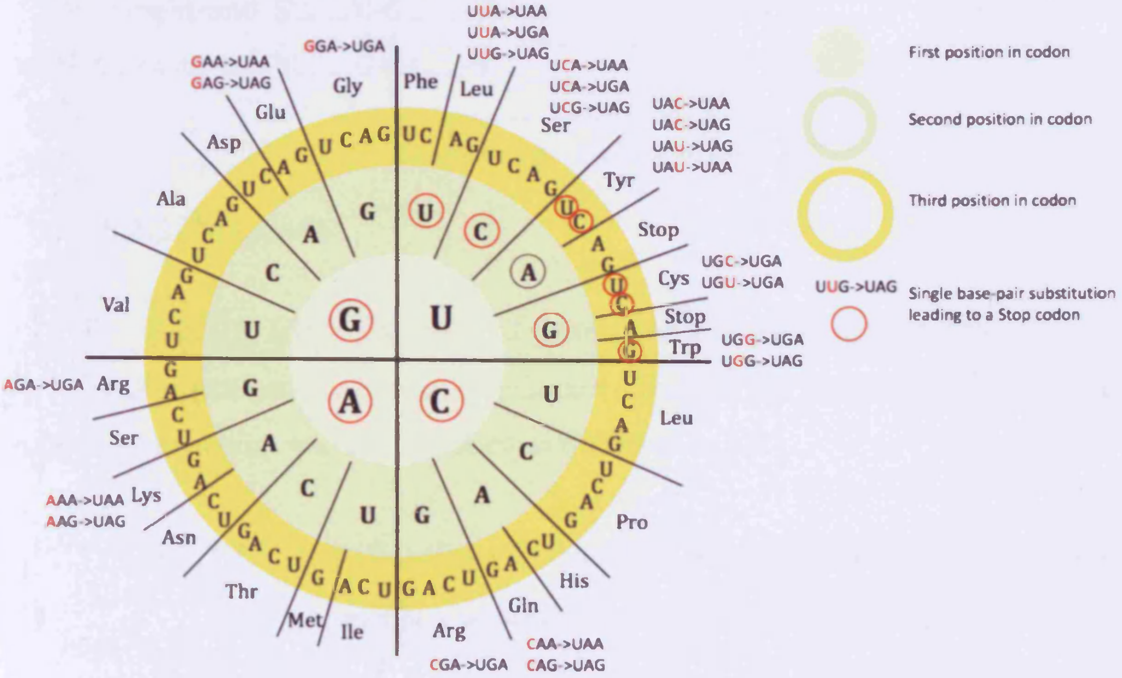
5.2. Materials and Methods

5.2.1. Materials

5.2.1.1. General definition of nonsense mutations

A nonsense mutation was defined as a single base-pair substitution that was responsible for the introduction of a stop codon into the coding region of a gene. Employing this definition, there are 23 possible single base-pair substitutions that could lead to the introduction of a premature stop codon. These substitutions are shown in Figure 19.

Figure 19 Possible single base-pair substitutions leading to a nonsense mutation (adapted from Frank-Kamenetski 1993)



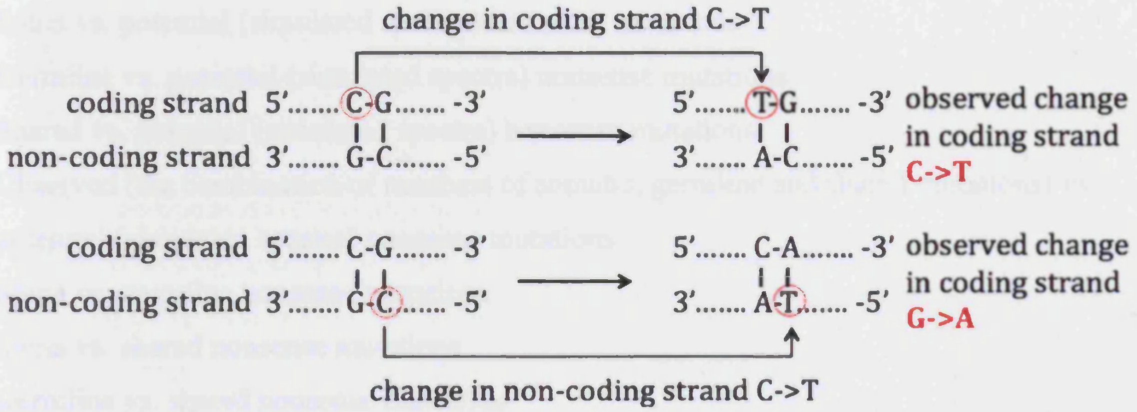
5.2.1.1.1. Labelling of somatic, germline and shared nonsense mutations

A detailed description of labelling of mutations is given in 3.3. A summary of the studied nonsense mutations is given in Table 18 and Table 19.

5.2.1.2. General definition of single base-pair substitutions generating nonsense mutations in CpG dinucleotides

CpG nonsense mutations were defined as C->T transitions found in the context of CpG-dinucleotides that lead to a stop codon. All observed nonsense mutations were logged according to the coding strand of DNA. Hence single base-pair substitutions in CpG dinucleotides on the non-coding strand would appear as G->A transitions (as shown in Figure 20).

Figure 20 Single base-pair substitutions in CpG dinucleotides


change in coding strand C->T
coding strand 5'.......C-G....... -3'
non-coding strand 3'....... G-C....... -5'
5'.......T-G....... -3'
3'....... A-C....... -5'
observed change in coding strand C->T
coding strand 5'....... C-G....... -3'
non-coding strand 3'....... G-C....... -5'
5'....... C-A....... -3'
3'....... A-T....... -5'
observed change in coding strand G->A
change in non-coding strand C->T

5.2.2. Methods

5.2.2.1. Identification of potential nonsense mutations

The identification of potential nonsense mutations and potential nonsense mutations in CpG-dinucleotides was accomplished as described in 3.5.

5.2.2.2. Identification of nonsense mutations that could potentially reduce mRNA abundance and discrimination from those that would not

In order to identify mRNA transcripts that would be predicted to be subject to degradation by the NMD apparatus, the ≥55nt rule (described in 5.1.5.2) was adopted. As a result, only those termination codons located ≥55nt upstream (*viz.* 5'; depicted in Figure 17 and Figure 18) of the most 3' exon-exon junction (measured after splicing) tend to give rise to a marked or complete reduction in mRNA abundance (Nagy and Maquat 1998).

I developed a computer program, which assigns a predicted NMD status (NMD skip or NMD elicit) for any given nonsense mutation. The predicted NMD status is based on the nucleotide position of the single base-pair substitution leading to a nonsense mutation relative to the nucleotide position of the most 3' exon-exon junction in a gene. Thus, for every nonsense mutation (*viz*. observed and all possible nonsense mutations) in all 17 tumour suppressor genes, a predicted NMD status was assigned.

5.2.2.3. Comparisons and calculation of statistical significance

In order to answer the questions posed in the aims of the analysis (Section 5.1.7), the following tests for each gene were performed:

Soma vs. potential (simulated spectra) nonsense mutations
Germline vs. potential (simulated spectra) nonsense mutations
Shared vs. potential (simulated spectra) nonsense mutations
Observed (the combination of numbers of somatic, germline and shared mutations) vs. potential (simulated spectra) nonsense mutations
Soma vs. germline nonsense mutations
Soma vs. shared nonsense mutations
Germline vs. shared nonsense mutations
Recurrent somatic vs. non-recurrent somatic nonsense mutations
Recurrent somatic shared vs. recurrent somatic non-shared nonsense mutations
Non-recurrent somatic shared vs. non-recurrent somatic non-shared nonsense mutations
Recurrent somatic shared vs. non-recurrent somatic non-shared missense mutations

In addition, the numbers of mutations in all genes were combined, only if nonsense mutations had the same label (i.e. somatic, germline, shared) to represent the combination of mutations in all genes. The aforementioned tests were also performed for the combination of mutations in all genes. Each of these comparisons was performed with respect to predicted NMD status (i.e. NMD elicit and NMD skip) and CpG-dinucleotide context. All mutations were categorized into two groups, i.e. 'within CpG dinucleotide' or not (i.e. not occurring in a CpG-dinucleotide); NMD elicit or not (i.e. NMD skip) and the tests were performed using a χ^2 statistic. For each of the tests, a χ^2 test statistic was calculated (see Chapter 3 (General methods) for description) to assess the statistical significance at the chosen alpha level ($\alpha = 0.05$). To allow for multiple hypothesis testing for the tests performed for each gene, 10,000 resampling permutations were performed (see Chapter 3 (General methods) for description) and corresponding p-value was termed permuted. To allow for multiple hypothesis testing, for the tests performed for CpG-dinucleotides and NMD-status analyses, a Bonferroni correction was applied. Therefore, each permuted p-value was multiplied by 2 (gene-wise $\alpha = 0.05/2$ or 0.025), to account for different tests within each gene (*viz.* NMD-status and location within CpG-dinucleotides). To allow for multiple hypothesis testing, for the tests performed for all genes, a Bonferroni correction was also applied. Therefore, each gene-wise p-value was multiplied by 17 (overall experiment-wise $\alpha = \frac{0.05}{2*17}$ or 0.0015).

I designed a computer program that automatically performs the χ^2 test statistic for each test along with the re-sampling permutations and Bonferroni corrections.

5.2.2.4. Calculation of power and effect size

Calculations of the power and associated effect sizes is described in 3.6.3. In order to keep the overall α at the 0.05 level, as multiple statistical tests were performed, the value of α used in the power calculations was set to 0.0001336898 (total number of tests performed 374; therefore, the Bonferroni-adjusted α, to account for multiple testing, was $\alpha = 0.05/374$, or 0.0001336898 for χ^2 test statistic). As Bonferroni correction is considered to be a conservative correction for multiple testing, the power calculated for the χ^2 statistical tests is a conservative estimate.

5.3. Results

The results are presented in comparison-wise fashion. Due to the overwhelming quantity of results that were generated during the work described, only summaries of statistically significant results are discussed and presented in the form of Tables and Figures (Table 23, Table 24 and Figure 21).

Nevertheless, the presented tables capture the results obtained for all comparisons performed, along with the directionality of the statistically significant results observed and power calculations. For further information, complete results for all comparisons performed are to be found in the Supplementary Tables.

In order to facilitate readability, whenever a comparison was statistically significant, it was substituted with the words "significant" or "significantly"; gene-wise statistically significant p-values were substituted with the symbol p_G and experiment-wise statistically significant p-values were substituted with the symbol p_E. In addition, whenever a comparison exhibited gene-wise and experiment-wise statistical significance, only the experiment-wise p-values were given and whenever a comparison exhibited only gene-wise (but not experiment-wise) statistical significance, only the gene-wise p-value was listed. Additionally, all gene-wise and experiment-wise statistically significant results are graphically summarized in Figure 21, along with the direction of the result and power calculations, but it was not referenced throughout the Results section, in order to reduce repetition.

5.3.1. Nonsense mutations and NMD status

5.3.1.1. Somatic vs. potential nonsense mutations

Only the *TP53* gene exhibited significantly more ($p_E<0.0034$) NMD-elicit somatic nonsense mutations as compared to potential mutations (Table 23). Moreover, all (100%) of the *TP53* somatic nonsense mutations were predicted to be potential targets of the NMD machinery (i.e. NMD elicit). By contrast, significantly fewer potential *TP53* nonsense mutations were predicted to elicit NMD (~56%). It has to be noted that ~74% of all possible *TP53* nonsense mutations were observed as somatic nonsense mutations. Therefore, while it appears that nonsense mutations were a very frequent event in the *TP53* gene, all of them were predicted to lead to the complete elimination of the mutant copy of the gene.

5.3.1.2. Germline vs. potential nonsense mutations

Only the *APC* gene exhibited a significantly higher (p_E<0.0034) proportion of germline NMD-elicit nonsense mutations, as compared to potential nonsense mutations (Table 23). A much higher proportion of germline nonsense mutations in the *APC* gene were predicted to be a target of the NMD machinery (42%), than was predicted for the potential nonsense mutations (21%). It is noteworthy that even although there were significantly more germline nonsense mutations predicted to be a subject to NMD (compared to the potential nonsense mutations), there were more germline nonsense mutations that were predicted to skip NMD than those that generally would not (42% predicted to elicit NMD versus 58% predicted to skip NMD). This indicates that the predicted outcome for a large proportion of the germline nonsense mutations would be NMD skip, but there was an excess of nonsense mutations that were predicted to elicit NMD compared to that for potential nonsense mutations.

5.3.1.3. Shared vs. potential nonsense mutations

Only the *BRCA1* gene exhibited a statistically significant deviation (p_G=0.017) in the distribution of predicted NMD status, but did not reach experiment-wise significance (p_E=0.30), when shared were compared to potential nonsense mutations (Table 23). The *BRCA1* gene had significantly more shared nonsense mutations predicted to skip NMD (33%), than potential nonsense mutations (2%). Further examination of the distribution of shared nonsense mutations in the *BRCA1* gene showed that there were only 6 shared nonsense mutations in total (4 were predicted to elicit NMD and 2 were predicted to skip NMD). In addition, there was ~85% statistical power to detect an experiment-wise significant result. Thus, it is likely that the shared nonsense mutations did not attain experiment-wise significance as a result of very small difference in the proportions of predicted NMD status between shared and potential nonsense mutations (effect size=0.19).

Even although only the *BRCA1* gene showed a trend of more shared mutations predicted to skip NMD, the combination of shared nonsense mutations for all genes, showed a similar trend in the same direction (p_G=0.036). More shared mutations were predicted to skip NMD (~26%) as compared to potential nonsense mutations (~16%).

Therefore, nonsense mutations found in both the soma and the germline are relatively more likely to skip NMD than potential nonsense mutations, at least for the *BRCA1* gene. In addition, the following genes showed a higher proportion of shared mutations predicted to skip NMD: *APC* (86%), *ATM* (33%) and *VHL* (45%) genes, as compared to potential nonsense mutations (79%, 2% and 29% for the *APC*, *ATM* and *VHL* genes respectively), but did not exhibit significant results.

5.3.1.4. Somatic vs. germline nonsense mutations

Only the *APC* gene showed a significant deviation in terms of the distribution of somatic and germline nonsense mutations, with respect to predicted NMD outcome (p_E<0.0034). An excess of germline nonsense mutations (42%) predicted to elicit NMD was noted as compared with the somatic nonsense mutations (6%). This finding is extremely interesting as there are reports in the literature that the first hit - the germline mutation in familial adenomatosis polyposis (FAP) - may influence and possibly even direct the type and position of the somatic hit (second hit; Latchford et al. 2007). For discussion of these results, see 5.4.1.1.

5.3.1.5. Germline vs. shared nonsense mutations

No individual gene exhibited a significant result, when germline were compared to shared nonsense mutations, with respect to predicted NMD outcome. Nevertheless, the combination of germline nonsense mutations for all genes exhibited significantly (p_G=0.0074) more mutations predicted to elicit NMD (~87%) as compared to the combination of shared mutations for all genes (~74%), even although the result did not attain experiment-wise significance (p_E=0.126). Therefore, a trend was evident in the distribution of germline and shared nonsense mutations, with respect to NMD status. Thus, for some genes, a higher proportion of germline nonsense mutations was subject to NMD (*APC*, *ATM*, *BRCA1* and *PTEN*) than was the case with shared nonsense mutations.

5.3.1.6. Combination of somatic, germline and shared nonsense mutations in individual genes

The *APC* and *TP53* were the only genes that exhibited a statistically significant difference in the distribution of predicted NMD status of observed nonsense mutations (the combination of somatic, germline and shared nonsense mutations as described in Section 5.1.7), when compared to the combination of potential nonsense mutations. A summary of the statistically significant results is presented in Table 23. Interestingly, for these genes, a significant excess of observed nonsense mutations predicted to elicit NMD was noted by comparison to potential nonsense mutations.

The *APC* gene showed many more observed nonsense mutations predicted to elicit NMD (31%) than the potential nonsense mutations (21%). As with the results reported in Section 5.3.1.2, even although there were significantly more observed nonsense mutations predicted to be subject to NMD (as compared to the potential nonsense mutations), there were still more observed nonsense mutations that were predicted to skip NMD than those that would not (31% predicted to elicit NMD versus 69% predicted to skip NMD). This indicates that the predicted outcome for a large proportion of the observed nonsense mutations would be NMD skip. However, there was an excess of nonsense mutations predicted to elicit NMD as compared to the potential nonsense mutations.

The *TP53* gene showed significantly more observed nonsense mutations predicted to elicit NMD (100%) as compared to potential nonsense mutations (56%). Indeed, all 97 *TP53* nonsense mutations (*viz*. somatic, germline and shared) were predicted to elicit NMD. As a result, all *TP53* nonsense mutations were very likely to lead to the complete elimination of the associated mutant mRNA.

5.3.1.7. Remainder of the comparisons performed

No statistically significant results were obtained for the remainder of comparisons performed for both nonsense mutations in individual genes and the combination of mutations in all genes. These included:

Somatic vs. shared nonsense mutations.

Recurrent vs. non-recurrent somatic nonsense mutations.

Recurrent and shared vs. recurrent non-shared nonsense mutations.

Non-recurrent and shared vs. non-recurrent non-shared somatic nonsense mutations.

Recurrent and shared vs. non-recurrent non-shared nonsense mutations.

Germline vs. shared nonsense mutations.

It should be noted that none of these comparisons showed enough statistical power (≥80%) to detect an experiment-wise significant result. Either a small effect size and/or a paucity of nonsense mutations must have contributed to these results.

5.3.2. Nonsense mutations in the context of CpG dinucleotides in the 17 tumour suppressor genes

5.3.2.1. Germline vs. potential nonsense mutations

A number of genes (*viz. APC, ATM, BRCA2, CDH1, NF1, TSC1* and *TSC2*) exhibited a statistically significant deviation in the distribution of germline CpG-located nonsense mutations as compared with the CpG-located potential nonsense mutations (p_G ranging from <0.0002 to 0.0062; Table 24). In addition, the results for the *ATM* and *NF1* genes attained a level of experiment-wise statistical significance (p_G<0.0034 for both genes). All the above genes exhibited an excess of CpG-located germline nonsense mutations as compared to potential nonsense mutations. The proportions of the CpG-located germline mutations ranged from 5% (*APC* and *BRCA2*) to 22% (*CDH1* and *ATM*). By contrast, the proportions of CpG-located potential nonsense mutations ranged from ~0% (*ATM, BRCA2, CHD1, NF1, TSC1* and *TSC2*) to ~1% (*APC*). Furthermore, the combination of germline nonsense mutations for all genes also exhibited an excess of CpG-located mutations (~7% within CpG dinucleotides), as compared to the combined CpG-located potential nonsense mutations for all genes (~0% within CpG-dinucleotides; p_E<0.0034).

Therefore, germline nonsense mutations for a number of genes are very likely to have resulted, albeit indirectly, from the heavy intra-genic methylation of CpG-dinucleotides.

5.3.2.2. Shared vs. potential nonsense mutations

A number of genes (*viz. APC, ATM, BRCA1, CDH1, NF1, NF2, PTEN, RB1* and *WT1*) exhibited statistically significant deviations in their distributions of shared CpG-located nonsense as compared to CpG-located potential nonsense mutations (p_E ranging from <0.0034 to 0.0068). All these genes exhibited an excess of CpG-located shared nonsense mutations as compared with potential nonsense mutations. The proportions of CpG-located shared mutations ranged from 17% (*BRCA1*) to 100% (*WT1*). By contrast, the proportions of CpG-located potential nonsense mutations ranged from ~0% (*ATM, BRCA1, CDH1, NF1, PTEN* and *WT1*) to ~1% (*APC, NF2* and *RB1*).

The comparison of shared nonsense mutations, combined for all genes, revealed a significantly higher proportion of CpG-located mutations as compared to potential nonsense mutations (p_E<0.0034), with ~37% of the shared nonsense mutations combined for all genes found in CpG-located dinucleotides versus ~0% for the CpG-located potential nonsense mutations (Table 24). This result indicates that the preponderance of shared nonsense mutations occur within CpG-located dinucleotides.

5.3.2.3. Somatic vs. shared nonsense mutations

The *APC*, *NF2*, *PTEN* and *TP53* genes exhibited preferential location of shared nonsense mutations within CpG-dinucleotides (p_G ranging from 0.0104 to 0.0384) as compared to somatic CpG-located nonsense mutations, although none of results attained experiment-wise significance. Nevertheless, in the case of these genes, none of the somatic nonsense mutations (0%) were located in CpG-dinucleotides, whereas the proportions of CpG-located shared mutations ranged from 23% (*PTEN*) to 44% (*TP53*). Therefore, it would seem that CpG-located nonsense mutations are predominantly shared mutations, rather than exclusively somatic nonsense mutations. In fact, almost all genes provided examples of somatic nonsense mutations that were found in non-CpG dinucleotides (~100%), the exception being the *PTCH* gene with just 1 CpG-located somatic nonsense mutation (~14% of *PTCH* gene mutations). Thus it was not surprising that the combination of somatic nonsense mutations for all genes exhibited a significantly smaller proportion of CpG-located nonsense mutations (p_E<0.0034; ~0%) as compared to the combination of shared nonsense mutations for all genes (~37% CpG-located).

5.3.2.4. Germline vs. shared nonsense mutations

The *APC*, *NF1*, *NF2* and *RB1* genes exhibited a preferential location of shared nonsense mutations within CpG-dinucleotides (p_G ranged from 0.0034 to 0.0384) as compared to germline CpG-located nonsense mutations, although none of the results attained experiment-wise significance. Nevertheless, all of these genes showed that a substantial proportion of the shared nonsense mutations (ranging from 29% to 90% for the *APC* and *NF1* genes respectively) were located in CpG-dinucleotides, whereas the proportions of CpG-located germline mutations ranged from 0% (*NF2*) to 9% (*NF1*).

Furthermore, the combination of shared nonsense mutations for all genes were also preferentially found within CpG-dinucleotides as compared to the combination of germline CpG-located nonsense mutations (p_E<0.0034; ~37% and ~7% for the shared and germline mutations respectively).

5.3.2.5. Recurrent somatic nonsense mutations

Only the *NF2* gene exhibited a significant result, when recurrent were compared to non-recurrent somatic mutations (p_G=0.0376), but this did not attain experiment-wise significance (p_E=0.639). The recurrent somatic mutations showed many more CpG-located mutations (~33%) as compared to non-recurrent somatic mutations (~0%).

The *TP53* gene showed a preferential location of recurrent and shared somatic mutations within CpG-dinucleotides (p_G=0.0174; 44% within CpG-dinucleotides) as compared to recurrent but not shared somatic mutations (0% within CpG-dinucleotides), although the result did not attain experiment-wise significance (p_E=0.296).

The *NF2* and *PTEN* genes showed a significantly higher proportion of recurrent and shared CpG-located somatic mutations (p_G 0.022/0.236 for the *NF2*/*PTEN* genes) as compared to CpG-located non-recurrent and non-shared somatic nonsense mutations.

Moreover, the combination of recurrent and shared mutations for all genes were found to be preferentially located within CpG-dinucleotides as compared to both the combination of recurrent, non-shared (p_E<0.0034) and the combination of non-recurrent, non-shared nonsense mutations for all genes (p_E<0.0034). Thus, amongst recurrent somatic nonsense mutations, shared mutations were more likely to be found within CpG-dinucleotides (~46%), than non-shared mutations (~0%). In addition, among non-recurrent somatic nonsense mutations, shared mutations were also more likely to be found within CpG-dinucleotides (~31%), than non-shared mutations (~1%).

5.3.2.6. Remainder of comparisons performed

No statistically significant results were obtained for the rest of the comparisons performed, for both comparisons in individual genes and the combination of nonsense mutations for all genes. These included:

Somatic vs. germline nonsense mutations.

Somatic vs. shared nonsense mutations.

Germline vs. shared nonsense mutations.

Recurrent somatic vs. non-recurrent shared nonsense mutations with respect to CpG-located and non-CpG located nonsense mutations.

Recurrent shared vs. non-recurrent shared nonsense mutations with respect to CpG-located and non-CpG located nonsense mutations.

Somatic CpG-located vs. shared CpG-located nonsense mutations with respect to recurrent and non-recurrent status of nonsense mutations.

5.4. Discussion

5.4.1. Nonsense mutations and nonsense-mediated mRNA decay (NMD)

Considering the importance of nonsense mutations associated with cancer, it is rather surprising that the distribution of nonsense mutations within tumour suppressor genes with respect to their likely NMD outcome has not so far been explored. This Chapter represents an attempt to shed light on the role that NMD might have played in helping to define the observed somatic nonsense mutational spectrum in 17 human tumour suppressor genes.

Depending upon whether or not NMD is elicited, any given nonsense mutation will have very different consequences for the expression of the gene involved. If elicited, NMD will automatically lead to a 'loss-of-function' due to the degradation of the affected mRNA species (haploinsufficiency). By contrast, a failure to elicit NMD ensures that the mRNA escapes NMD potentially resulting in the synthesis of an abnormal prematurely truncated protein that could, at least in principle, exert a 'gain-of-function' (dominant-negative) effect.

For the vast majority of reported nonsense mutations in human genes, there are no empirical data (e.g. from *in vitro* assays) that would serve to demonstrate unambiguously whether or not these lesions have elicited NMD. However, in the absence of such data, one can nevertheless employ the '~55nt boundary rule' (Maquat 2004; Nagy and Maquat 1998), by taking account of the relative position of a specific nonsense mutation with respect to the position of the last exon of a given gene to predict whether or not NMD is likely to have been elicited. In the absence of experimental (laboratory) verification, it is clear that the application of the '55 nucleotide rule' has the potential to be over-simplistic and hence inaccurate. However, it is clear that the retrospective experimental verification of NMD for all the mutations considered here would clearly not be feasible at this juncture.

The above *caveat* notwithstanding, nonsense mutations from 17 tumour suppressor genes were, on the sole basis of their relative genic locations, allocated to two distinct groups viz. *NMD-elicit* and *NMD-escape*. The relative proportions of the two groups were then compared for germline mutations, somatic mutations, mutations present in both the germline and the soma (i.e. shared mutations), the germline and somatic mutations combined, and a set of 'potential nonsense mutations' (codons that could be converted to termination codons by a single base-pair substitution). Significant differences were observed for two tumour suppressor genes (*APC* and *TP53*). These will now be considered separately.

5.4.1.1. *APC*

The *APC* gene encodes a large multidomain protein that plays an important role in regulating β-catenin and in mediating intracellular adhesion (Fearnhead et al. 2001). It is commonly accepted to conform to Knudson's 'two-hit' hypothesis and may therefore be regarded as a classical tumour suppressor. The germline and somatic *APC* mutational spectra have been known for some time to be different (Fearnhead et al. 2001; Lamlum et al. 1999). Over 60% of all somatic *APC* mutations occur within <10% of the gene coding sequence between codons 1281 and 1556 (of the 2843 amino acid-encoding open reading frame) in the so-called ***mutation cluster region*** (MCR; Beroud and Soussi 1996; Cheadle et al. 2002; Miyoshi et al. 1992). The *APC* somatic mutational spectrum is characterized by ~30% nonsense mutations and ~60% frameshifts. Within the MCR, there are two hotspots for nonsense mutations at codons 1309 and 1450 (Beroud and Soussi 1996). The majority of germline *APC* mutations are nonsense or frameshift; although fairly uniformly distributed between codons 200 and 1460 (Crabtree et al. 2003), hotspots occur at codons 1061 and 1309 (Leggett et al. 1997) that together account for about a third of inherited *APC* mutations (Beroud and Soussi 1996; Cetta et al. 2000). It is now thought that the type of germline *APC* mutation can play a role in determining the nature of the second (somatic) *APC* mutation. Thus, if the germline mutation occurs outside the region between codons 1194 and 1392, the second hit is likely to be a truncating mutation within the MCR (Lamlum et al. 1999). It is clear that if the germline *APC* mutation exerts an influence on the nature and/or location of the subsequent somatic *APC* mutation, the two mutational spectra are likely to be more distinct than if there were no such influence. In the context of this study, I was interested in assessing whether the differences between the germline and somatic *APC* mutational spectra might help to account for the observed frequency differences in predicted NMD.

Analysis of the *APC* gene showed that significantly more germline nonsense mutations were predicted to elicit NMD (42% NMD elicit) as compared to either the somatic (6% NMD elicit) or potential nonsense mutations (21% NMD elicit). Therefore, many more inherited *APC* nonsense mutations would be predicted to result in the degradation of nonsense mutation-bearing mRNA transcripts than might be expected by chance alone. Indeed, 56% of all germline nonsense mutations were predicted to skip NMD (thereby avoiding mRNA degradation and potentially leading to the synthesis of C-terminally truncated proteins). This result concurs with those of Mort et al. (2008) who found, in a large meta-analysis of some 5316 nonsense mutations in 380 different human genes causing inherited disease, that the proportion of disease-causing nonsense mutations predicted to elicit NMD was significantly higher than among potential nonsense mutations, implying that nonsense mutations that elicit NMD are more likely to come to clinical attention.

By contrast, the vast majority (94%) of somatic nonsense mutations are predicted to skip NMD (i.e. potential COOH-terminally truncated proteins). Indeed, only 6% of somatic nonsense mutations are predicted to elicit NMD as compared to a proportion of 21% for potential nonsense mutations. Thus, in sharp contradistinction to the situation pertaining with *APC* germline nonsense mutations, *APC* somatic nonsense mutations appear to exhibit an excess of NMD escape (potential gain-of-function) mutations [or put another way, a paucity of NMD elicit (potential loss-of-function) mutations]. This suggests that somatic nonsense mutations which lead to the synthesis of truncated forms of the *APC* protein (courtesy of their being encoded by mRNAs bearing premature termination codons but which have nevertheless escaped NMD) could have been selected for during the process of tumorigenesis by dint of their ability to confer a proliferative advantage upon the cells expressing them. One way in which this might operate would be if the truncated protein were to exert a dominant negative effect, thereby conferring oncogenic properties upon the *APC* gene. Since the oligomerization domain (residues 6-57) is located at the almost invariably included NH_2-terminal end of the APC protein (Fearnhead et al. 2001), some nonsense mutations could exert a dominant negative effect via dimerization of the truncated protein product with the wild-type protein, thereby reducing the amount of the wild-type product available to the cell. Indeed, it has been shown that truncating *APC* mutations can exert a dominant negative effect on the wild-type product (codon 1309; Dihlmann et al. 1999). Two missense variants (I1370K and E1317Q) have also been shown a dominant-negative effect (Fearnhead et al. 2001; Frayling et al. 1998). Furthermore, a N-terminal domain, between amino-acids 782-1018 immediately adjacent to the armadillo domain, has been shown to interact with the last 300 C-terminal amino-acids, but also with itself (Li et al. 2008). As a result, amino-terminally truncated proteins could interact with full-length *APC* protein (i.e. heterozygous state) and contribute to deregulation of early tumour cells, thus compromising normal migration of intestinal epithelial cells (Li et al. 2008).

On the other hand, if dominant-negative *APC* mutant forms were to play a major role in the process of tumorigenesis, it would be logical to suppose that null *APC* proteins would have to be present in a substantial proportion, as these would potentially display an effect on the cells similar to dominant-negative protein forms. Instead, homozygous null colon cancers are virtually non-existent (McCartney et al. 2006) and mice with one truncated (putative dominant-negative) and a wild-type copy, do not present polyps or tumours (Oshima et al. 1995). Mechanistically, *APC* plays its tumour suppressor role by binding (and hence down-regulating) soluble β-catenin, the key effector of the Wnt signalling pathway. *APC*'s β-catenin-binding sites are situated within three 15-amino acid repeats (residues 1020-1169) and a series of seven 20-amino acid repeats (residues 1265-2035). This means that although the majority of truncated mutant APC proteins contain the three 15-amino acid repeats, they lack either all or most of the 20-amino acid repeats (Fearnhead et al. 2001), implying that the loss of β-catenin-binding ability is important for the loss of tumour suppressor function. β-catenin is involved in the transcriptional activation of the *c-myc* (*MYC*; He et al. 1998) and *cyclin D1* (*CCDN1*; Shtutman et al. 1999; Tetsu and McCormick 1999) oncogenes, both of which regulate cell cycle progression. Thus, inactivation of *APC* is likely to promote cell proliferation by indirectly increasing the level of β-catenin (Sieber et al. 2000). Therefore, lack of *APC* null alleles could potentially be explained by "just-right" (Albuquerque et al. 2002) β-catenin levels, whereby levels above a certain threshold could potentially be lethal to cells (McCartney et al. 2006). Indeed, homozygous *Apc* mutant mice die during gastrulation (Moser et al. 1995), although one could argue that a wild-type *Apc* protein is necessary for embryonic development, but may not be lethal in cells of a fully developed organism. It has also been suggested that in fruit fly, the armadillo repeat 5 (ARM5) may have a special importance and contribute significantly to the overall structure of the arm-repeats (McCartney et al. 2006). This is potentially very interesting, as armadillo-repeat 5 (ARM5 ends at codon 628; ARM6 starts at codon 644; Xing et al. 2004) is N-terminally immediately adjacent to the border of a predicted NMD elicit/skip status of the mutant allele (codon 634). Whether this is a mere coincidence or whether it bears any functional relevance to humans with respect to potential involvement of the NMD machinery, remains to be addressed.

An alternative model to explain how truncated protein products might confer a cellular proliferative advantage would be if the NH_2-terminal end of the APC protein were to contain domains that promote cell division, whilst the C-terminal end of the protein contained domains that repress cell division. C-terminally truncated proteins (i.e. nonsense mutations at codons 750 and 1309, both at a considerable distance from the ~55nt boundary) have been shown to have a 'profound proliferation effect' (Tighe et al. 2004). Furthermore, full length *APC* protein inhibits DNA replication, by directly binding to DNA through a region that maps between codons 2140 and 2421 (Qian et al. 2008). Both the somatic nonsense mutations in the MCR region (codons 1281-1556) and most germline nonsense mutations (codons 200-1460) would potentially result in C-terminally truncated proteins that completely lack this region. Thus, C-terminally truncated proteins, as a result of the most common nonsense mutations, would potentially be insufficient to inhibit DNA replication, hence are very likely to promote cellular proliferation.

5.4.1.2. *TP53*

The *TP53* gene, also commonly referred to as the 'guardian of the genome' (Vousden 2000), is a multifunctional transcription factor that among many other functions regulates cell cycle progression, targets for apoptosis cells with an unacceptable amount of DNA damage, interacts with key proteins responsible for DNA transcription, replication and repair (Levine 1997; Vogelstein and Kinzler 2004). The key role of *TP53* is shown by the fact that ~50% of all human cancers harbour mutations in the *TP53* gene (Soussi and Beroud 2001). Interestingly, in contrast to other tumour suppressor genes, the majority (>80%) of lesions in *TP53* are missense mutations (The UMD *p53* mutation `database, http://p53.free.fr/ Soussi and Beroud 2001). Therefore, most of the inactivating mutations in the *TP53* gene are associated with a full-length gene product. *p53* contains several domains: activation domain 1 (residues 1-42), activation domain 2 (residues 43-63), proline-rich domain (64-91), DNA-binding domain (residues 100-300), domain responsible for nuclear localization and containing the export signal (residues 316-325), tetramerization domain (residues 326-356) and C-terminal basic domain (residues 364-393 Zhu et al. 2000). The majority of mutations found in *TP53* are localized in the DNA-binding domain, and these serve to modify the protein's contact with its target DNA sequence. Even although >80% of the lesions found are missense variants (Soussi and Beroud 2001), nonsense mutations and frameshift mutations still comprise ~8% and ~11% of lesions, respectively.

The above analysis of the *TP53* gene has shown that there were significantly more somatic nonsense mutations predicted to elicit NMD (100% NMD elicit) than would be expected on the basis of a comparison with the potential nonsense mutations (56% NMD elicit). Therefore, it would appear that, during tumorigenesis, somatic nonsense mutations in the *TP53* gene that would trigger NMD (leading to degradation of the *TP53* mRNA from the mutation-bearing allele) have been selected for. Indeed, a total of 87 observed somatic nonsense mutations were predicted to elicit NMD and none were predicted to escape NMD. This is a rather surprising finding, as there are numerous reports that mutant *TP53* forms could exert a dominant-negative effect over the wild-type or exhibit a gain-of-function. Even though these dominant-negative forms reported involve missense variants exclusively, it is perhaps surprising that there are no truncating nonsense mutations that could potentially give rise to dominant-negative-forms.

One possible explanation may be provided by animal studies, which show that, *TP53*$^{-/-}$ mice (i.e. mice with both alleles inactivated) do not develop carcinomas with particularly high frequency as compared to *TP53*$^{\text{missense variant/-}}$ or *TP53*$^{\text{missense variant/+}}$ mice (Lang et al. 2004; Olive et al. 2004). These authors suggested that inactivation of *TP53* is insufficient for tumour progression and that a gain-of-function is actually required. In addition, during tumour progression, *TP53* has been suggested to 'evolve' under strong positive selection (Glazko et al. 2004). Generally, in an evolutionary context, evidence of positive selection is taken as being indicative of the acquisition of new functions (Koonin et al. 2005). Thus far, no studies have ever reported a potential gain-of-function (i.e. dominant-negative effect over the wild-type) associated with any of the nonsense mutations found in *TP53*.

One plausible explanation of the spectrum of somatic nonsense mutations with respect to predicted NMD outcome is that gene-dosage may affect the selection of nonsense mutations. A reduction in gene-dosage (i.e. inactivation of one of the *TP53* alleles) has been shown to be sufficient for tumour development (Venkatachalam et al. 1998). Furthermore, mice with constitutive *TP53* deletions have been shown to be insufficient for apoptosis (Clarke et al. 1994) and maintenance of genome integrity (Bouffler et al. 1995).

Alternatively, most truncated *p53* proteins (made possible as a consequence of NMD skipping) would ultimately retain several functionally important domains, due to the fact that the last exon, that bears functional importance to the NMD pathway, is only 81bp in length. Thus, potentially only the tetramerization domain and the C-terminal basic domain would be functionally compromised in a truncated, mutant form of the *p53* protein. Therefore, it seems plausible that a potentially truncated protein could be partially functional, assuming the existence of a sufficiently stable product that could support transcription and translation. It would seem that such truncated proteins are not selected during tumorigenesis, as none of the somatic nonsense mutations present in the *TP53* gene were predicted to skip NMD. Indeed, a truncated *p53* protein bearing an unusual 7bp insertion in a choroid plexus tumour, has been shown to have completely lost transactivation and transrepression of target genes, such as *CDKN1A* and *CDKN2A*, but has nevertheless partially (~65% of wild-type protein) retained the ability to induce apoptosis in lymphocytes (Rutherford et al. 2002).

It should be noted that the mere prediction of potential NMD-elicit nonsense mutations would not necessarily mean that mutant protein would be completely removed from the cell. Although, the product of the mutant *p53* allele, comprising the 7bp insertion, reported by Rutherford et al. (2002), could not be detected as an expressed protein, the mechanism responsible for the partially retained ability to induce apoptosis remains unknown. In addition, a mutant *p53* protein comprising 770delT, was surprisingly reported to be as abundant as a wild-type copy (Anczukow et al. 2008), although its abundance was markedly increased when NMD was inhibited. Furthermore, a correlation has been reported, of increased expression levels of *CDK* inhibitor p21$^{\mathrm{CIP1/WAF1}}$ and nonsense mutations in the *TP53* gene (Mousses et al. 2001). The *CDK* inhibitor p21$^{\mathrm{CIP1/WAF1}}$ is part of the *p53*-dependent DNA damage response pathway; thus, it is likely that some transcriptional activation function was retained in these mutant *TP53* alleles that comprised nonsense mutations.

5.4.2. Nonsense mutations in the context of CpG dinucleotides

The enzymatic methylation of cytosine to 5-methylcytosine constitutes an epigenetic modification which is associated with a variety of different biological processes including gene regulation (i.e. gene expression), X-chromosome inactivation, genomic imprinting and development (Pfeifer 2000). DNA methylation occurs predominantly at CpG dinucleotides, which have been known for some time to be a mutational hotspot in both the germline and the soma (Cooper et al. 1995; Krawczak et al. 1998). Their hypermutability is estimated to be 6-7 times the base mutation rate (Cooper et al. 1995). Their hypermutability is largely dependent upon the methylation status of CpG dinucleotides, even although some reports have suggested a limited correlation (Millar et al. 1998). Therefore, patterns of methylation may influence the location and frequency of both somatic and germline nonsense mutations in the context of CpG dinucleotides. Furthermore, hypermethylation has been shown in sperm cells as compared to oocytes, but both are hypomethylated by comparison to somatic tissues (Allegrucci et al. 2005). However, CpG dinucleotides have been found to exhibit differences in their methylation status both within and between individuals (Millar et al. 1998).

The analysis of CpG-located nonsense mutations demonstrated that virtually all (98%) somatic CpG-located nonsense mutations (Table 21) were also found in the germline, as compared to ~33% of somatic nonsense mutations (irrespective of whether or not they occur in a CpG-dinucleotide; Table 20) and ~23% of the somatic non-CpG-located mutations (Table 22). As a result, CpG-located nonsense mutations are more likely to be shared between the soma and the germline, than non-CpG-located nonsense mutations. This was further confirmed, by the comparison with potential nonsense mutations, which indicated that shared nonsense mutations were more likely to be found within CpG-dinucleotides. This observation was made not only for the combination of nonsense mutations in all genes, but also in a number of individual genes (*viz. APC*, *ATM*, *BRCA1*, *CDH1*, *NF1*, *NF2*, *RB1* and *WT1*). Moreover, for some genes (*viz. BRCA1*, *CDKN2A*, *PTEN*, *STK11*, *TP53*, *VHL* and *WT1*), all possible CpG-located nonsense mutations were present in either the germline and/or the soma. Additionally, for the *CDKN2A*, *PTEN*, *TP53*, *VHL* and *WT1*, all possible CpG-located nonsense mutations were present in both the soma and the germline. This indicates that mutation hotspots are commonly shared between the germline and the soma and a sizeable proportion of these are located in CpG dinucleotides. In addition, no significant differences were observed between the germline and the soma with respect to CpG-located nonsense mutations for all genes, implying that the methylation status of these genes is likely to be very similar between the germline and the soma. However, insufficient statistical power meant that we could not rule out the possibility that a paucity of mutation could have contributed for the observed results.

Conversely, more than 81% of all possible nonsense mutations within CpG dinucleotides were present among the observed germline nonsense mutations (*viz.* germline and shared) as opposed to 36% in the soma (*viz.* somatic and shared). Therefore, one conclusion that can be drawn is that CpG sites in the germline are relatively more methylated than in the soma. This contrasts with at least one study which appears to show precisely the opposite, i.e. that the germline is hypomethylated by comparison to the soma (Allegrucci et al. 2005). An alternative explanation may however be that the smaller proportion of somatic CpG-located nonsense mutations could be due to the smaller number of somatic nonsense mutations examined as compared to the numbers of germline nonsense mutations.

Somatic nonsense mutations were found to be randomly distributed (as determined by reference to potential nonsense mutations) with respect to their occurrence in CpG and non-CpG-dinucleotides. Nevertheless, there was not enough statistical power to detect an experiment-wise significant result, due to the fact that there was only 1 CpG-located nonsense mutation (*PTCH*) and the great majority of somatic nonsense mutations (~99.6%) were non-CpG located. Furthermore, the great majority of somatic CpG-located nonsense mutations (98%) were also found in the germline, hence they were labelled shared. In addition, germline nonsense mutations were preferentially found within CpG-dinucleotides (*APC*, *BRCA2*, *CDH1*, *NF1*, *PTCH*, *TSC1* and *TSC2*).

We may infer that highly methylated CpG dinucleotides are present in both the germline and the soma, since if these CpG dinucleotides were not methylated, they would be no more mutable than any other dinucleotide. These results indicate that for some genes, specific CpG sites might be hypermethylated in the germline as compared to the soma, but for others the methylation status is potentially very similar. In addition, as reported previously, some variation of methylation levels in both the germline and the soma is evident (Millar et al. 1998; Trasler 1998).

The analysis of CpG-located nonsense mutations demonstrated that all CpG-located recurrent somatic nonsense mutations (i.e. those nonsense mutations that have been reported more than once) were also found in the germline. Conversely, all CpG-located shared nonsense mutations were found to be recurrent. Not only do somatic nonsense mutations recur in CpG-located dinucleotides, but all (100%) of these recurrent somatic CpG-located nonsense mutations have also been noted in the germline. Thus, CpG dinucleotides are generalized mutational hotspots due to methylation-mediated deamination, and their hypermutability is evident not only on account of their giving rise to recurrent somatic mutations but also because they give rise to germline mutations.

Taking the observed results altogether, one could pose a few additional questions for the future: why are virtually all somatic CpG-located nonsense mutations also found in the germline? Is endogenous mutagenesis the most important factor responsible for the high proportion of CpG-dinucleotides amongst nonsense mutations? Are exogenous factors also responsible for non-CpG located nonsense mutations?

The above findings could in principle be explicable at the level of the individual CpG dinucleotide, if data on the methylation status of all CpG sites were available. At present, data on the methylation status of only a few of the 17 human tumour suppressor genes are available and most of the published studies relate solely to CpG-dinucleotides located in promoter regions. Nevertheless, CpG-dinucleotides in exons 5-8 in the *TP53* gene have been shown to be completely methylated in all tissues and cells examined by Tornaletti and Pfeifer (1995), but also exon 4 (Magdinier et al. 2002). In addition, codon 248 in the *TP53* gene, a mutation hotspot in both the soma and the germline, has been shown to be methylated in all tissues and cells examined (Magewu and Jones 1994). Similarly, the promoter region and exon 1 of the *WT1* gene have also been shown to be extensively methylated (Laux et al. 1999). In addition, a number of genes have been shown to exhibit promoter hypermethylation, such as *BRCA1* (Baldwin et al. 2000; Esteller et al. 2000; Radpour et al. 2009), *CDKN2A* (Radpour et al. 2009; Wang et al. 2009), *PTEN* (Montiel-Duarte et al. 2008; Zhu et al. 2009). Since most of these studies only involve methylation patterns in relation to tumour tissues (i.e. somatic tissues), it is difficult to extrapolate intra-genic methylation patterns in the germline. Nevertheless, these studies provide support, in the case of a few genes, of a likely heavy intra-genic methylation in at least somatic tissues.

As mentioned earlier, CpG-located nonsense mutations in both the soma and the germline are very likely to have resulted from methylation-mediated deamination. Nevertheless, a number of studies have shown that epigenetic modifications could be modulated by exogenous factors. Specifically, these include tobacco smoke, metals, arsenic, ionizing UVA & UVB radiation, aflatoxin B1, alcohol, etc. (Fleming et al. 2008). In particular, human non-melanoma cancers have been shown to exhibit UV-specific mutational patterns. These patterns comprise C->T transitions at methyl CpG-associated dipyrimidine sites (Ikehata and Ono 2007). Therefore, it is likely that some of the CpG-located nonsense mutations could be due to exogenous factors, such as UV light. Some reports suggest that environmental factors, such as aflatoxin B1 increase mutation frequency in some codons, but not others (Chan et al. 2003); thus, structural sequence context may play an important role. Furthermore, G->T transversions at CpG-dinucleotides are very likely to be a mutational signature associated with exposure to tobacco smoke (Yoon et al. 2001). However, a closer look at the CpG-located nonsense mutations in the 17 tumour suppressor genes studied, revealed that no G->T transversions at CpG-dinucleotides were observed. Furthermore, only 3 such transversions would potentially lead to the introduction of a stop codon out of 23 possible (Figure 19). Thus, one could only speculate as to the importance of these G->T transversions, with respect to CpG-located nonsense mutations.

It has to be noted that the increased mutation frequency in CpG-dinucleotides as a result of carcinogens is largely dependent on the methylation status of the CpG-dinucleotides. As a result, methylation is required for the mutational-signature patterns associated with carcinogens (Pfeifer 2006).

Similar methylation patterns between the soma and the germline could potentially be explained by inherited epigenetic modifications, whereby a "hypothetically" heavy intra-genic germline methylation is transmitted to somatic cells. Such methylated CpG-dinucleotides, present in both the soma and the germline would be substrates for both spontaneous endogenous deamination and carcinogens. Indeed, CpG methylation has been shown to be heritable in Mendelian and non-Mendelian fashion (Fleming et al. 2008).

Mendelian inherited modifications are very likely to be result from incomplete epigenetic re-programming in the germline and have been shown in hereditary nonpolyposis colorectal cancer (Chan et al. 2006). Nevertheless, such "transgenerational" inheritance of epigenetic events could be also due to carcinogens (Anway et al. 2005; Chang et al. 2006; Fleming et al. 2008). Therefore, both endogenous mechanisms and carcinogens are quite likely to play an important role in determining methylation patterns in both the soma and the germline and hence ultimately, the mutability of methylated cytosines in both cell lineages.

Table 18 Distribution of germline and somatic nonsense mutations in the 17 human tumour suppressor genes

	Codons	Possible missense mutations	Possible CpG mutations						Somatic									Germline			
Gene	Number	F_P	F_{PCpG}	$\frac{F_{PCpG}}{F_P}$	F_S	F_{SCpG}	F_{SR}	F_{SH}	$\frac{F_S}{F_P}$	$\frac{F_{SCpG}}{F_S}$	$\frac{F_{SCpG}}{F_{PCpG}}$	$\frac{F_{SR}}{F_S}$	$\frac{F_{SH}}{F_S}$	F_G	F_{GCpG}	F_{GH}	$\frac{F_G}{F_P}$	$\frac{F_{GCpG}}{F_G}$	$\frac{F_{GCpG}}{F_{PCpG}}$	$\frac{F_{GH}}{F_G}$
APC	2844	1278	27	0.02	79	8	35	28	0.06	0.10	0.30	0.44	0.35	180	16	28	0.14	0.09	0.59	0.16
ATM	3057	1480	21	0.01	7	2	0	3	0.00	0.29	0.10	0.00	0.43	75	18	3	0.05	0.24	0.86	0.04
BRCA1	1864	803	4	0.00	9	1	0	6	0.01	0.11	0.25	0.00	0.67	121	4	6	0.15	0.03	1.00	0.05
BRCA2	3419	1594	5	0.00	1	0	0	0	0.00	0.00	0.00	0.00	0.00	76	4	0	0.05	0.05	0.80	0.00
CDH1	883	263	4	0.02	7	1	0	2	0.03	0.14	0.25	0.00	0.29	11	3	2	0.04	0.27	0.75	0.18
CDKN2A	157	27	2	0.07	18	2	3	5	0.67	0.11	1.00	0.17	0.28	7	2	5	0.26	0.29	1.00	0.71
NF1	2819	1089	19	0.02	14	9	0	10	0.01	0.64	0.47	0.00	0.71	115	18	10	0.11	0.16	0.95	0.09
NF2	596	251	7	0.03	42	6	18	18	0.17	0.14	0.86	0.43	0.43	43	6	18	0.17	0.14	0.86	0.42
PTCH	1448	480	5	0.01	9	1	0	2	0.02	0.11	0.20	0.00	0.22	27	2	2	0.06	0.07	0.40	0.07
PTEN	404	183	3	0.02	56	3	19	13	0.31	0.05	1.00	0.34	0.23	28	3	13	0.15	0.11	1.00	0.46
RB1	929	420	14	0.03	27	7	9	15	0.06	0.26	0.50	0.33	0.56	76	11	15	0.18	0.14	0.79	0.20
STK11	434	143	1	0.01	10	0	1	3	0.07	0.00	0.00	0.10	0.30	27	1	3	0.19	0.04	1.00	0.11
TP53	394	129	4	0.03	96	4	85	9	0.74	0.04	1.00	0.89	0.09	10	4	9	0.08	0.40	1.00	0.90
TSC1	1165	439	7	0.02	1	0	0	0	0.00	0.00	0.00	0.00	0.00	37	6	0	0.08	0.16	0.86	0.00
TSC2	1808	551	7	0.01	1	0	0	1	0.00	0.00	0.00	0.00	1.00	74	6	1	0.13	0.08	0.86	0.01
VHL	214	73	2	0.03	15	2	4	11	0.21	0.13	1.00	0.27	0.73	27	2	11	0.37	0.07	1.00	0.41
WT1	450	161	3	0.02	3	3	0	3	0.02	1.00	1.00	0.00	1.00	14	3	3	0.09	0.21	1.00	0.21
Total	22885	9364	135	0.01	395	49	174	129	0.04	0.12	0.36	0.44	0.33	948	109	129	0.10	0.11	0.81	0.14

F_i is the number of mutations, where $i \in \{P, PCpG, S, SCpG, SR, SH, G, GCpG, GH\}$

P-possible, $PCpG$-possible CpG-located , S-somatic , $SCpG$-somatic CpG-located, SR-somatic recurrent, SH-somatic shared, G-germline, $GCpG$-germline CpG-located, GH-germline shared

Values marked in red denote those genes that made a relatively large contribution to the corresponding mutational spectrum

Table 19 Shared recurrent nonsense mutations and shared nonsense mutations found in CpG-dinucleotides

Mutations	Shared	CpG-located		Recurrent		Recurrent and CpG-located			
Gene	F_{SH}	F_{CpG}	F_{CpG}/F_{SH}	F_{REC}	F_{REC}/F_{SH}	F_{REC_CpG}	F_{REC_CpG}/F_{SH}	F_{REC_CpG}/F_{CpG}	F_{REC_CpG}/F_{REC}
APC	28	8	0.29	17	0.61	6	0.21	0.75	0.35
ATM	3	2	0.67	0	0.00	0	0.00	0.00	N/A
BRCA1	6	1	0.17	0	0.00	0	0.00	0.00	N/A
BRCA2	0	0	N/A	0	N/A	0	N/A	N/A	N/A
CDH1	2	1	0.50	0	0.00	0	0.00	0.00	N/A
CDKN2A	5	2	0.40	2	0.40	1	0.20	0.50	0.50
NF1	10	9	0.90	0	0.00	0	0.00	0.00	N/A
NF2	18	6	0.33	10	0.56	6	0.33	1.00	0.60
PTCH	2	0	0.00	0	0.00	0	0.00	N/A	N/A
PTEN	13	3	0.23	8	0.62	3	0.23	1.00	0.38
RB1	15	7	0.47	7	0.47	5	0.33	0.71	0.71
STK11	3	0	0.00	1	0.33	0	0.00	N/A	0.00
TP53	9	4	0.44	9	1.00	4	0.44	1.00	0.44
TSC1	0	0	N/A	0	N/A	0	N/A	N/A	N/A
TSC2	1	0	0.00	0	0.00	0	0.00	N/A	N/A
VHL	11	2	0.18	3	0.27	1	0.09	0.50	0.33
WT1	3	3	1.00	0	0.00	0	0.00	0.00	N/A
Total	129	48	0.37	57	0.44	26	0.20	0.54	0.46

F_i is the number of mutations, where $i \in \{SH, CpG, REC, REC\_CpG\}$

SH-shared, CpG- CpG-located , REC-recurrent , REC_CpG-recurrent CpG-located

Values marked in red denote genes that made a relatively large contribution to the corresponding mutational spectrum

Table 20 Distribution of somatic, germline and shared nonsense mutations

Mutations	Nonsense	Somatic				Germline				Shared					
Gene	F_N	F_S	F_S / F_N	F_S / F_{ST}	F_S / F_{NT}	F_G	F_G / F_N	F_G / F_{GT}	F_G / F_{NT}	F_{SH}	F_{SH} / F_N	$\frac{F_{SH}}{F_{SH} + F_S}$	$\frac{F_{SH}}{F_{SH} + F_G}$	F_{SH} / F_{SHT}	F_{SH} / F_{NT}
APC	231	51	0.221	0.192	0.042	152	0.658	0.186	0.125	28	0.35	0.35	0.16	0.217	0.023
ATM	79	4	0.051	0.015	0.003	72	0.911	0.088	0.059	3	0.04	0.43	0.04	0.023	0.002
BRCA1	124	3	0.024	0.011	0.002	115	0.927	0.140	0.095	6	0.05	0.67	0.05	0.047	0.005
BRCA2	77	1	0.013	0.004	0.001	76	0.987	0.093	0.063	0	0.00	0.00	0.00	0.000	0.000
CDH1	16	5	0.313	0.019	0.004	9	0.563	0.011	0.007	2	0.13	0.29	0.18	0.016	0.002
CDKN2A	20	13	0.650	0.049	0.011	2	0.100	0.002	0.002	5	0.25	0.28	0.71	0.039	0.004
NF1	119	4	0.034	0.015	0.003	105	0.882	0.128	0.086	10	0.08	0.71	0.09	0.078	0.008
NF2	67	24	0.358	0.090	0.020	25	0.373	0.031	0.021	18	0.27	0.43	0.42	0.14	0.015
PTCH	34	7	0.206	0.026	0.006	25	0.735	0.031	0.021	2	0.06	0.22	0.07	0.016	0.002
PTEN	71	43	0.606	0.162	0.035	15	0.211	0.018	0.012	13	0.18	0.23	0.46	0.101	0.011
RB1	88	12	0.136	0.045	0.010	61	0.693	0.074	0.050	15	0.17	0.56	0.20	0.116	0.012
STK11	34	7	0.206	0.026	0.006	24	0.706	0.029	0.020	3	0.09	0.30	0.11	0.023	0.002
TP53	97	87	0.897	0.327	0.072	1	0.010	0.001	0.001	9	0.09	0.09	0.90	0.070	0.007
TSC1	38	1	0.026	0.004	0.001	37	0.974	0.045	0.030	0	0.00	0.00	0.00	0.000	0.000
TSC2	74	0	0.000	0.000	0.000	73	0.986	0.089	0.060	1	0.01	1.00	0.01	0.008	0.001
VHL	31	4	0.129	0.015	0.003	16	0.516	0.020	0.013	11	0.36	0.73	0.41	0.085	0.009
WT1	14	0	0.000	0.000	0.000	11	0.786	0.013	0.009	3	0.21	1.00	0.21	0.023	0.002
total	NT 1214	ST 266	0.219	1.000	0.219	GT 819	0.675	1.000	0.675	SHT 129	0.11	0.33	0.14	1.000	0.106

F_i is the number of mutations, where $i \in \{N,S,G,GT,NT,SH,SHT\}$

N-nonsense, S- somatic , G-germline , GT-germline total, NT-nonsense total, SH-shared, SHT-shared total

Values marked in red denote genes that made a relatively large contribution to the corresponding mutational spectrum

Table 21 Distribution of CpG-located somatic, germline and shared nonsense mutations

	Codons	Possible nonsense mutations	Possible nonsense CpG mutations		Somatic	Germline	Shared	Recurrent somatic non-shared	Recurrent somatic shared							
Gene	Number	F_P	F_{PCpG}	$\frac{F_{PCpG}}{F_P}$	F_S	F_G	F_{SH}	F_{RS}	F_{RSH}	$\frac{F_S}{F_{PCpG}}$	$\frac{F_G}{F_{PCpG}}$	$\frac{F_{SH}}{F_{PCpG}}$	$\frac{F_{SH}}{F_{SH}+F_S}$	$\frac{F_{SH}}{F_{SH}+F_G}$	$\frac{F_{RS}}{F_{SH}}$	$\frac{F_{RSH}}{F_{SH}}$
APC	2844	1278	27	0.02	0	8	8	0	6	0.00	0.30	0.30	1.00	0.50	N/A	0.75
ATM	3057	1480	21	0.01	0	16	2	0	0	0.00	0.76	0.10	1.00	0.11	N/A	0.00
BRCA1	1864	803	4	0.00	0	3	1	0	0	0.00	0.75	0.25	1.00	0.25	N/A	0.00
BRCA2	3419	1594	5	0.00	0	4	0	0	0	0.00	0.80	0.00	N/A	0.00	N/A	N/A
CDH1	883	263	4	0.02	0	2	1	0	0	0.00	0.50	0.25	1.00	0.33	N/A	0.00
CDKN2A	157	27	2	0.07	0	0	2	0	1	0.00	0.00	1.00	1.00	1.00	N/A	0.50
NF1	2819	1089	19	0.02	0	9	9	0	0	0.00	0.47	0.47	1.00	0.50	N/A	0.00
NF2	596	251	7	0.03	0	0	6	0	6	0.00	0.00	0.86	1.00	1.00	N/A	1.00
PTCH	1448	480	5	0.01	1	2	0	0	0	0.20	0.40	0.00	0.00	0.00	0.00	N/A
PTEN	404	183	3	0.02	0	0	3	0	3	0.00	0.00	1.00	1.00	1.00	N/A	1.00
RB1	929	420	14	0.03	0	4	7	0	5	0.00	0.29	0.50	1.00	0.64	N/A	0.71
STK11	434	143	1	0.01	0	1	0	0	0	0.00	1.00	0.00	N/A	0.00	N/A	N/A
TP53	394	129	4	0.03	0	0	4	0	4	0.00	0.00	1.00	1.00	1.00	N/A	1.00
TSC1	1165	439	7	0.02	0	6	0	0	0	0.00	0.86	0.00	N/A	0.00	N/A	N/A
TSC2	1808	551	7	0.01	0	6	0	0	0	0.00	0.86	0.00	N/A	0.00	N/A	N/A
VHL	214	73	2	0.03	0	0	2	0	1	0.00	0.00	1.00	1.00	1.00	N/A	0.50
WT1	450	161	3	0.02	0	0	3	0	0	0.00	0.00	1.00	1.00	1.00	N/A	0.00
total	22885	9364	135	0.01	1	61	48	0	26	0.01	0.45	0.36	0.98	0.44	0.00	0.54

F_i is the number of mutations, where $i \in \{P, PCpG, S, G, SH, RS, RSH\}$

P-possible, $PCpG$- possible CpG-located , S-somatic , G-germline, SH-shared, RS-recurrent somatic non-shared, RSH-recurrent somatic shared

Values marked in red denote genes that made a relatively large contribution to the corresponding mutational spectrum

Table 22 Distribution of non-CpG located somatic, germline and shared nonsense mutations

	Codons	Possible nonsense mutations	Possible nonsense CpG mutations		Somatic	Germline	Shared	Recurrent somatic non-shared	Recurrent somatic shared							
Gene	Number	F_P	F_{PCpG}	$\frac{F_{PCpG}}{F_P}$	F_S	F_G	F_{SH}	F_{RS}	F_{RSH}	$\frac{F_S}{F_P}$	$\frac{F_G}{F_P}$	$\frac{F_{SH}}{F_P}$	$\frac{F_{SH}}{F_{SH}+F_S}$	$\frac{F_{SH}}{F_{SH}+F_G}$	$\frac{F_{RS}}{F_{SH}}$	$\frac{F_{RSH}}{F_{SH}}$
APC	2844	1278	27	0.02	51	144	20	18	11	0.04	0.11	0.02	0.28	0.12	0.35	0.55
ATM	3057	1480	21	0.01	4	56	1	0	0	0.00	0.04	0.00	0.20	0.02	0.00	0.00
BRCA1	1864	803	4	0.00	3	112	5	0	0	0.00	0.14	0.01	0.63	0.04	0.00	0.00
BRCA2	3419	1594	5	0.00	1	72	0	0	0	0.00	0.05	0.00	0.00	0.00	0.00	N/A
CDH1	883	263	4	0.02	5	7	1	0	0	0.02	0.03	0.00	0.17	0.13	0.00	0.00
CDKN2A	157	27	2	0.07	13	2	3	1	1	0.48	0.07	0.11	0.19	0.60	0.08	0.33
NF1	2819	1089	19	0.02	4	96	1	0	0	0.00	0.09	0.00	0.20	0.01	0.00	0.00
NF2	596	251	7	0.03	24	25	12	8	4	0.10	0.10	0.05	0.33	0.32	0.33	0.33
PTCH	1448	480	5	0.01	6	23	2	0	0	0.01	0.05	0.00	0.25	0.08	0.00	0.00
PTEN	404	183	3	0.02	43	15	10	11	5	0.23	0.08	0.05	0.19	0.40	0.26	0.50
RB1	929	420	14	0.03	12	57	8	2	2	0.03	0.14	0.02	0.40	0.12	0.17	0.25
STK11	434	143	1	0.01	7	23	3	0	1	0.05	0.16	0.02	0.30	0.12	0.00	0.33
TP53	394	129	4	0.03	87	1	5	76	5	0.67	0.01	0.04	0.05	0.83	0.87	1.00
TSC1	1165	439	7	0.02	1	31	0	0	0	0.00	0.07	0.00	0.00	0.00	0.00	N/A
TSC2	1808	551	7	0.01	0	67	1	0	0	0.00	0.12	0.00	1.00	0.01	N/A	0.00
VHL	214	73	2	0.03	4	16	9	1	2	0.05	0.22	0.12	0.69	0.36	0.25	0.22
WT1	450	161	3	0.02	0	11	0	0	0	0.00	0.07	0.00	N/A	0.00	N/A	N/A
total	22885	9364	135	0.01	265	758	81	117	31	0.03	0.08	0.01	0.23	0.10	0.44	0.38

F_i is the number of mutations, where $i \in \{P, PCpG, S, G, SH, RS, RSH\}$

P-possible, $PCpG$- possible CpG-located , S-somatic , G-germline, SH-shared, RS-recurrent somatic non-shared, RSH-recurrent somatic shared

Values marked in red denote genes that made a relatively large contribution to the corresponding mutational spectrum

Table 23 Summary of statistically significant results for nonsense mutations, with respect to NMD status

Group 1 vs. Group 2	Gene	Group 1				Group 2				original $\chi 2$[3]	original p-value[3]	Permuted p-value[4]	gene-wise p-value[5]	experiment-wise p-value[6]	W(effect size)[7]	Power[8]
		NMD elicit		NMD skip		NMD elicit		NMD skip								
		N[1]	F[2]	N[1]	F[2]	N[1]	F[2]	N[1]	F[2]							
shared vs. pot	ALL	96	0.74	33	0.26	6879	0.84	1271	0.16	9.54	2.01E-03	1.78E-02	3.56E-02	6.05E-01	0.034	23.26
germ vs. shared	ALL	710	0.87	109	0.13	96	0.74	33	0.26	13.18	2.83E-04	3.70E-03	7.40E-03	1.26E-01	0.118	42.50
germ vs. pot	*APC*	64	0.42	88	0.58	225	0.21	822	0.79	30.84	2.81E-08	<1.00E-04	<2.00E-04	<3.40E-03	0.160	95.85
obs vs. pot	*APC*	71	0.31	160	0.69	225	0.21	822	0.79	9.09	2.57E-03	1.96E-02	3.92E-02	6.66E-01	0.084	21.06
soma vs. germ	*APC*	3	0.06	48	0.94	64	0.42	88	0.58	22.66	1.93E-06	<1.00E-04	<2.00E-04	<3.40E-03	0.334	82.66
shared vs. pot	*BRCA1*	4	0.67	2	0.33	664	0.98	15	0.02	23.81	1.07E-06	8.70E-03	1.74E-02	2.96E-01	0.186	85.53
soma vs. pot	*TP53*	87	1.00	0	0.00	18	0.56	14	0.44	43.14	5.10E-11	<1.00E-04	<2.00E-04	<3.40E-03	0.602	99.70
obs vs. pot	*TP53*	97	1.00	0	0.00	18	0.56	14	0.44	47.60	5.22E-12	<1.00E-04	<2.00E-04	<3.40E-03	0.607	99.90

[1]- Number; [2]-Frequency; Gene-wise or experiment-wise statistically significant results are marked in red; Pot.- Potential, Obs.- Observed, Soma.- Somatic

Table 24 Summary of statistically significant results for nonsense mutations, with respect to CpG-dinucleotides

Group 1 vs. Group 2	Gene	Group 1				Group 2				original χ^2[3]	original p-value[3]	Permuted p-value[4]	gene-wise p-value[5]	experiment-wise p-value[6]	W(effect size)[7]	Power[8]
		In CpG		Not in CpG		In CpG		Not in CpG								
		N[1]	F[2]	N[1]	F[2]	N[1]	F[2]	N[1]	F[2]							
germ vs. pot	ALL	61	0.07	758	0.93	25	0.00	8125	1.00	399.66	6.54E-89	<1.00E-04	<2.00E-04	<3.40E-03	0.211	100.00
shared vs. pot	ALL	48	0.37	81	0.63	25	0.00	8125	1.00	1978.72	~0.00E+0	<1.00E-04	<2.00E-04	<3.40E-03	0.489	100.00
obs vs. pot	ALL	110	0.09	1104	0.91	25	0.00	8125	1.00	569.88	~0.00E+0	<1.00E-04	<2.00E-04	<3.40E-03	0.247	100.00
soma vs. germ	ALL	1	0.00	265	1.00	61	0.07	758	0.93	18.64	1.58E-05	5.00E-04	1.00E-03	1.70E-02	0.131	69.07
soma vs. shared	ALL	1	0.00	265	1.00	48	0.37	81	0.63	108.46	2.13E-25	<1.00E-04	<2.00E-04	<3.40E-03	0.524	100.00
germ vs. shared	ALL	61	0.07	758	0.93	48	0.37	81	0.63	97.01	6.92E-23	<1.00E-04	<2.00E-04	<3.40E-03	0.320	100.00
Rec shared vs. rec non-shared	ALL	26	0.46	31	0.54	0	0.00	117	1.00	62.74	2.35E-15	<1.00E-04	<2.00E-04	<3.40E-03	0.600	100.00
Non-rec shared vs. non-rec non-shared	ALL	22	0.31	50	0.69	1	0.01	148	0.99	46.50	9.18E-12	<1.00E-04	<2.00E-04	<3.40E-03	0.459	99.86
Rec shared vs. non-rec non-shared	ALL	26	0.46	31	0.54	1	0.01	148	0.99	73.12	1.22E-17	<1.00E-04	<2.00E-04	<3.40E-03	0.596	100.00
germ vs. pot	*APC*	8	0.05	144	0.95	11	0.01	1036	0.99	15.10	1.02E-04	5.80E-03	1.16E-02	1.97E-01	0.112	52.66
shared vs. pot	*APC*	8	0.29	20	0.71	11	0.01	1036	0.99	118.96	1.07E-27	<1.00E-04	<2.00E-04	<3.40E-03	0.333	100.00
obs vs. pot	*APC*	16	0.07	215	0.93	11	0.01	1036	0.99	31.59	1.90E-08	4.00E-04	8.00E-04	1.36E-02	0.157	96.42
soma vs. shared	*APC*	0	0.00	51	1.00	8	0.29	20	0.71	16.21	5.66E-05	5.20E-03	1.04E-02	1.77E-01	0.453	58.20
germ vs. shared	*APC*	8	0.05	144	0.95	8	0.29	20	0.71	15.86	6.82E-05	5.20E-03	1.04E-02	1.77E-01	0.297	56.48
Rec shared vs. non-rec non-shared	*APC*	6	0.35	11	0.65	0	0.00	33	1.00	13.24	2.75E-04	7.70E-03	1.54E-02	2.62E-01	0.514	42.80
germ vs. pot	*ATM*	16	0.22	56	0.78	3	0.00	1398	1.00	260.51	1.33E-58	<1.00E-04	<2.00E-04	<3.40E-03	0.421	100.00
shared vs. pot	*ATM*	2	0.67	1	0.33	3	0.00	1398	1.00	372.53	5.26E-83	<1.00E-04	<2.00E-04	<3.40E-03	0.515	100.00
obs vs. pot	*ATM*	18	0.23	61	0.77	3	0.00	1398	1.00	272.36	3.47E-61	<1.00E-04	<2.00E-04	<3.40E-03	0.429	100.00
shared vs. pot	*BRCA1*	1	0.17	5	0.83	0	0.00	679	1.00	113.33	1.83E-26	<1.00E-04	<2.00E-04	<3.40E-03	0.407	100.00
germ vs. pot	*BRCA2*	4	0.05	72	0.95	1	0.00	1516	1.00	62.48	2.69E-15	3.10E-03	6.20E-03	1.05E-01	0.198	100.00
obs vs. pot	*BRCA2*	4	0.05	73	0.95	1	0.00	1516	1.00	61.65	4.11E-15	3.10E-03	6.20E-03	1.05E-01	0.197	100.00
germ vs. pot	*CDH1*	2	0.22	7	0.78	1	0.00	246	1.00	35.69	2.31E-09	8.40E-03	1.68E-02	2.86E-01	0.373	98.44
shared vs. pot	*CDH1*	1	0.50	1	0.50	1	0.00	246	1.00	61.25	5.04E-15	1.00E-03	2.00E-03	3.40E-02	0.496	100.00
obs vs. pot	*CDH1*	3	0.19	13	0.81	1	0.00	246	1.00	33.76	6.22E-09	8.40E-03	1.68E-02	2.86E-01	0.358	97.68
germ vs. pot	*NF1*	9	0.09	96	0.91	1	0.00	969	1.00	73.73	8.98E-18	<1.00E-04	<2.00E-04	<3.40E-03	0.262	100.00
shared vs. pot	*NF1*	9	0.90	1	0.10	1	0.00	969	1.00	791.98	~0.00E+0	<1.00E-04	<2.00E-04	<3.40E-03	0.899	100.00
obs vs. pot	*NF1*	18	0.15	101	0.85	1	0.00	969	1.00	139.55	3.34E-32	<1.00E-04	<2.00E-04	<3.40E-03	0.358	100.00
germ vs. shared	*NF1*	9	0.09	96	0.91	9	0.90	1	0.10	45.86	1.27E-11	1.70E-03	3.40E-03	5.78E-02	0.631	99.84
shared vs. pot	*NF2*	6	0.33	12	0.67	1	0.01	183	0.99	52.70	3.89E-13	<1.00E-04	<2.00E-04	<3.40E-03	0.511	99.97
obs vs. pot	*NF2*	6	0.09	61	0.91	1	0.01	183	0.99	12.82	3.43E-04	1.04E-02	2.08E-02	3.54E-01	0.226	40.55
soma vs. shared	*NF2*	0	0.00	24	1.00	6	0.33	12	0.67	9.33	2.25E-03	1.92E-02	3.84E-02	6.53E-01	0.471	22.23
germ vs. shared	*NF2*	0	0.00	25	1.00	6	0.33	12	0.67	9.68	1.86E-03	1.92E-02	3.84E-02	6.53E-01	0.475	23.96
rec vs. non-rec	*NF2*	6	0.33	12	0.67	0	0.00	24	1.00	9.33	2.25E-03	1.88E-02	3.76E-02	6.39E-01	0.471	22.23
Rec shared vs. non-rec non-shared	*NF2*	6	0.60	4	0.40	0	0.00	16	1.00	12.48	4.11E-04	1.04E-02	2.08E-02	3.54E-01	0.693	38.71
shared vs. pot	*PTEN*	3	0.23	10	0.77	0	0.00	112	1.00	26.48	2.66E-07	2.00E-04	4.00E-04	6.80E-03	0.460	90.77
soma vs. shared	*PTEN*	0	0.00	43	1.00	3	0.23	10	0.77	10.48	1.20E-03	1.25E-02	2.50E-02	4.25E-01	0.433	28.04

Rec shared vs. non-rec non-shared	*PTEN*	3	0.38	5	0.62	0	0.00	32	1.00	12.97	3.16E-04	1.25E-02	**2.50E-02**	4.25E-01	0.569	41.38
shared vs. pot	*RB1*	7	0.47	8	0.53	3	0.01	329	0.99	107.39	3.66E-25	<1.00E-04	**<2.00E-04**	**<3.40E-03**	0.556	100.00
obs vs. pot	*RB1*	11	0.12	77	0.88	3	0.01	329	0.99	29.03	7.12E-08	3.00E-04	**6.00E-04**	**1.02E-02**	0.263	94.16
germ vs. shared	*RB1*	4	0.07	57	0.93	7	0.47	8	0.53	15.65	7.64E-05	7.10E-03	**1.42E-02**	2.41E-01	0.454	55.41
soma vs. shared	*TP53*	0	0.00	87	1.00	4	0.44	5	0.56	40.35	2.13E-10	6.80E-03	**1.36E-02**	2.31E-01	0.648	99.43
Rec shared vs. rec non-shared	*TP53*	4	0.44	5	0.56	0	0.00	76	1.00	35.45	2.62E-09	6.80E-03	**1.36E-02**	2.31E-01	0.646	98.36
germ vs. pot	*TSC1*	6	0.16	31	0.84	1	0.00	400	1.00	54.91	1.26E-13	1.51E-02	**3.02E-02**	5.13E-01	0.354	99.98
obs vs. pot	*TSC1*	6	0.16	32	0.84	1	0.00	400	1.00	53.42	2.69E-13	1.51E-02	**3.02E-02**	5.13E-01	0.349	99.98
germ vs. pot	*TSC2*	6	0.08	67	0.92	1	0.00	476	1.00	32.32	1.31E-08	1.15E-02	**2.30E-02**	3.91E-01	0.242	96.90
obs vs. pot	*TSC2*	6	0.08	68	0.92	1	0.00	476	1.00	31.86	1.65E-08	1.15E-02	**2.30E-02**	3.91E-01	0.240	96.60
shared vs. pot	*WT1*	3	1.00	0	0.00	0	0.00	147	1.00	150.00	1.73E-34	<1.00E-04	**<2.00E-04**	**<3.40E-03**	1.000	100.00
obs vs. pot	*WT1*	3	0.21	11	0.79	0	0.00	147	1.00	32.10	1.47E-08	1.10E-03	**2.20E-03**	**3.74E-02**	0.447	96.76

[1]- Number; [2]-Frequency; Gene-wise or experiment-wise statistically significant results are marked in red; Pot.- Potential, Obs.- Observed, Soma.- Somatic, Rec.- Recurrent, Non-rec.- Non-recurrent

Figure 21 Summary of statistically significant results for nonsense mutations, with respect to CpG-dinucleotides and NMD status

Parameter		*ALL*	*APC*	*ATM*	*BRCA1*	*BRCA2*	*CDH1*	*CDKN2A*	*NF1*	*NF2*	*PTCH*	*PTEN*	*RB1*	*STK11*	*TP53*	*TSC1*	*TSC2*	*VHL*	*WT1*
Soma vs. pot	NMD														▨↑				
	CpG																		
Germ vs. pot	NMD		▨↑																
	CpG	▨↑	↑	🟩↑		🟩↑	🟩↑		🟩↑							🟩↑	🟩↑		
Shared vs. pot	NMD	↓			🟩↓														
	CpG	▨↑	▨↑	▨↑	▨↑		▨↑		▨↑	▨↑		▨↑	▨↑						▨↑
Obs vs. pot	NMD		↑												▨↑				
	CpG	▨↑	▨↑	▨↑	🟩	🟩↑	🟩↑		▨↑	↑			▨↑			🟩↑	🟩↑		▨↑
Soma vs. germ	NMD		▨↓																
	CpG	🟨↓																	
Soma vs. shared	NMD																		
	CpG	▨↓	↓							↓		↓			🟩↓				
Germ vs. shared	NMD	↑	↑																
	CpG	▨↓	↓						🟩↓	↓			↓						
Rec vs. non-rec	NMD																		
	CpG									↑									
Rec shared vs. rec non-shared	NMD																		
	CpG	▨↑													🟩↑				
Non-rec shared vs. non-rec non-shared	NMD																		
	CpG	▨↑																	
Rec shared vs. non-rec non-shared	NMD																		
	CpG	▨↑	↑							↑		↑							

Legend: ↑ or ↓ denotes the direction of the gene- or experiment-wise statistically significant result. The direction is with respect to the first group in the comparison. A Grey shaded box represents a experiment-wise statistically significant result, a non-shaded arrow (i.e. ↑ or ↓) represents a gene-wise statistically significant result, a Green shaded box represents ≥80% power to detect a statistically significant result for the comparison and associated effect size, Yellow shaded box represents ≤80% power and experiment-wise statistically significant result. Soma- Somatic, Germ- Germline, Obs.- Observed (somatic, germline and shared), Pot.- Potential, Rec.- Recurrent, Non-rec.- Non-recurrent;

6. Micro-lesions

6.1. Introduction

6.1.1. Importance of micro-lesions and their functional consequences

Cancer is perceived as a genetic disorder because DNA sequence changes are considered causative of neoplasms. Mutations in three basic types of gene drive tumour development: oncogenes, genomic stability genes and tumour suppressor genes (Vogelstein and Kinzler 2004). Generally 'biallelic gene inactivation' is required for tumour suppressor gene inactivation (Knudson 1971). Thus, one mutant allele, inherited through the germline, and subsequent inactivation of the other allele in the soma has been the basis of the 'two-hit' hypothesis originally proposed by Knudson (Knudson 1971, 1978).

Micro-deletions, micro-insertions and micro-indels (inserted and/or deleted nucleotides ≤20bp) are an important part of the mutational spectrum associated with cancer predisposition and tumour development. When causing frameshifts within the coding sequence of genes (i.e. when the length in base-pairs of the deleted or inserted bases is not divisible by three), these lesions invariably have drastic consequences for the function of the protein. Abrupt termination of translation would be expected when the reading frame is changed, due to the triplet nature of the genetic code (Crick et al. 1961; Yanofsky 2007). Thus, 'hidden' (out-of-frame) stop codons terminate mRNA translation (Seligmann and Pollock 2004). Seligmann and Pollock (2004) have proposed an 'ambush hypothesis', suggesting that 'hidden' stop codons (stop codons in -1 and +1 shifted reading frames) are frequently selected for, depending on adjacent codons and the synonymous codon state. These authors have suggested that codons with increased potential to form 'hidden' stops have greater usage frequency. The latter finding is compatible with the fact that translation termination following a frameshift would be likely to be beneficial to the cell, by reducing the energy waste on non-functional proteins or by reducing a cytotoxic effect (Seligmann and Pollock 2004). Indeed, it has been estimated that translation would be terminated on average ~15 codons following a frameshift (Itzkovitz and Alon 2007). Additionally, mRNAs with potential premature termination codons would be subject to quality control mechanisms that could potentially reduce or even eliminate the production of faulty proteins through the rapid degradation of the affected mRNA (Gonzalez et al. 2001; Holbrook et al. 2004; Lejeune and Maquat 2005; Maquat 2002, 2004; Vasudevan et al. 2002; Wilkinson 2005). Thus, the functional consequence of most frameshifts would be a truncated protein: truncated protein with markedly decreased concentration or no protein altogether. These types of mutations are often termed ‘loss-of-function’ mutations (Haber and Harlow 1997). Alternatively, some mutations could lead to a ‘gain-of-function’ where they give rise to dominant negative forms over the wild-type. Certain mutant forms of the *APC* gene could exert a dominant negative effect over the wild-type product (Dihlmann et al. 1999). By way of example, some mutant forms of *TP53* gene lead to a ‘loss-of-function’ whereas others give rise to a potent dominant negative over the wild-type forms (Junk et al. 2008).

On the other hand, in-frame (i.e. the number of added or subtracted nucleotides is divisible by 3) micro-deletions, micro-insertions and micro-indels would be expected to have less severe consequence for the function of the protein. In support of this assertion, a comprehensive meta-analysis of micro-deletions and micro-insertions causing inherited human genetic disease (Ball et al. 2005) revealed that the in-frame lesions of size 3bp and 6bp exhibit markedly lower frequencies than expectation. Since such mutations appear to come to clinical attention less frequently, it may be inferred that such micro-lesions are less likely to cause human disease. This notwithstanding, adding or subtracting amino acids, termed ‘protein tinkering’, could play an important role in carcinogenesis (Gonzalez et al. 2007). In support, all recorded mutations associated with non-small lung cancer in the Epidermal Growth Factor Receptor Mutation Database are in-frame (*EGFR* Database: http://www.cityofhope.org/mdl/egfr/Pages/default.aspx; Gu et al. 2007).

6.1.2. Endogenous mutagenesis

6.1.2.1. Direct repeats

Several lines of evidence are supportive of the non-random occurrence of micro-lesions (*viz.* micro-deletions, micro-insertions and micro-indels) and the potential involvement of endogenous mechanisms of mutagenesis. Early work by Efstratiadis et al. (1980) identified short (2-8bp) direct repeats (Figure 22a) around the endpoints of deletions. Efstratiadis et al. (1980) hypothesised a ‘slipped mispairing’ mechanism to explain the role of direct repeats in the process of mutagenesis, as depicted in Figure 23. The presence of such sequences could facilitate ‘slipped mispairing’ and thus promote deletions. This model could explain deletions of one copy of the repeat and the intervening sequence between the repeats.

Figure 22 Examples of repetitive elements: direct repeat (a); inverted repeat (b); mirror repeat (c); G-quartet (d); runs of identical nucleotides (e)

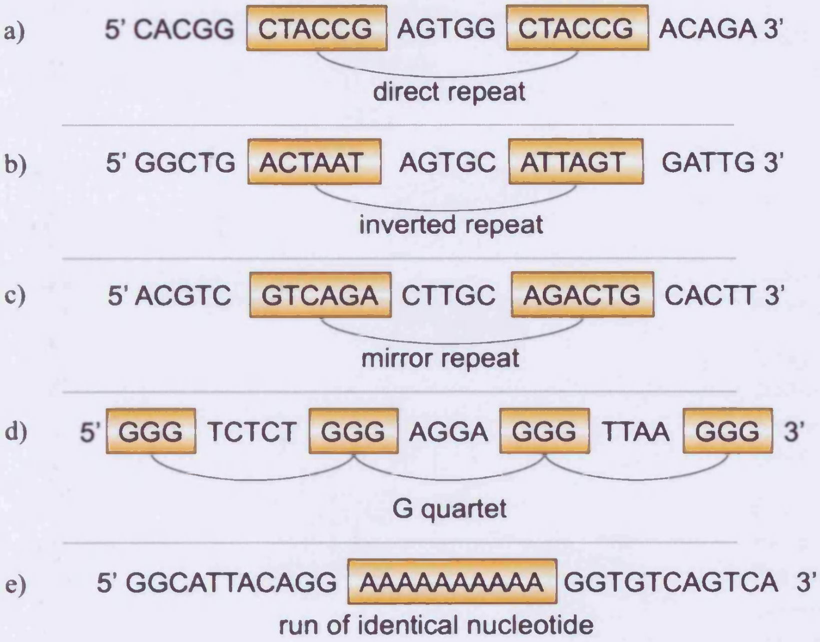
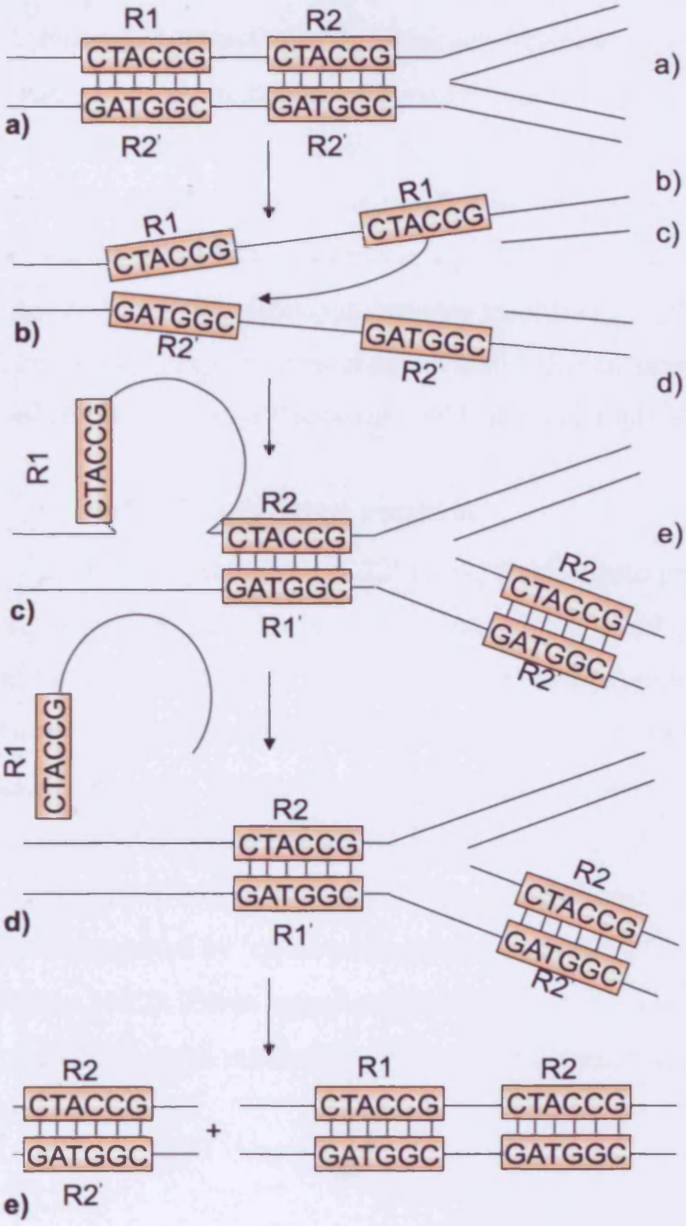

a)
5' CACGG CTACCG AGTGG CTACCG ACAGA 3'
direct repeat
b)
5' GGCTG ACTAAT AGTGC ATTAGT GATTG 3'
inverted repeat
c)
5' ACGTC GTCAGA CTTGC AGACTG CACTT 3'
mirror repeat
d)
5' GGG TCTCT GGG AGGA GGG TTAA GGG 3'
G quartet
e)
5' GGCATTACAGG AAAAAAAAAA GGTGTCAGTCA 3'
run of identical nucleotide

Figure 23 The 'slipped mispairing mechanism' for generation of micro-lesions during DNA replication (adapted from Cooper and Krawczak 1993; Efstratiadis et al. 1980)



a) Double stranded DNA containing a direct repeat (R1 and R2)

b) Single stranded DNA

c) R2 repeat mispairs with the complementary R1' repeat, producing single stranded loop

d) Single stranded loop containing repeat R1 excised and repaired by DNA repair enzymes

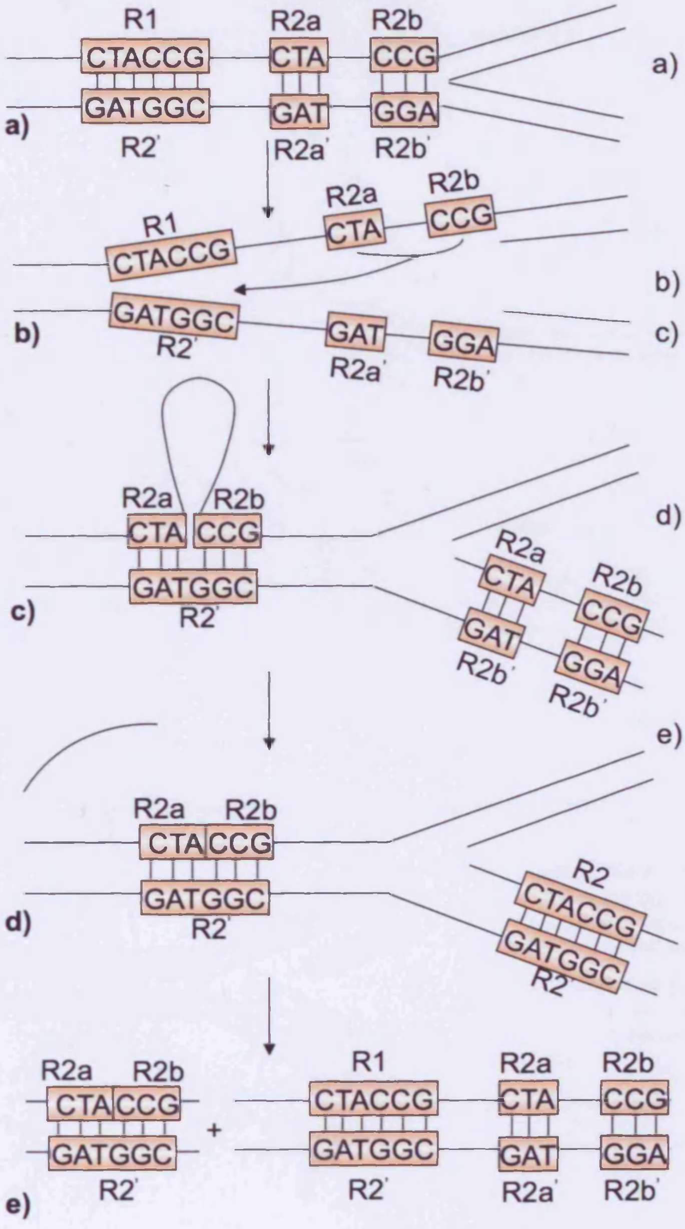
e) End of DNA replication. As a result only one of the newly derived double stranded DNA contains the full sequence of the repeat (R1 and R2). The second daughter double stranded DNA lacks sequence R1 and intervening sequence between R1 and R2

It has been noted that the proposed 'slipped mispairing' mechanism could not readily explain a number of deletions, where apparent involvement of direct repeats is noticeable. Thus, a modified version of the 'slipped mispairing' has been proposed (Cooper and Krawczak , depicted in Figure 24). This modified model suggests that the mispairing is only an intermediate and the deletion of one copy of the repeat as proposed in the original model does not occur. The intermediate is proposed only to last long enough to promote formation of a second copy repeat, followed by excision of the intervening sequence and subsequent end joining. Studies on triplet repeat expansion show that both insertions (expansions) and deletions (contractions) are extremely likely and are mediated by the repeats (Bowater and Wells 2001; Sinden et al. 2002; Wells et al. 2005). These studies have also suggested that expansions and contractions of repeats are dependent upon the replication origin. Expansions are generated when more stable slipped structures are formed on the newly synthesized Okazaki fragments (Hebert et al. 2004; Wells et al. 2005). Contractions on the other hand are generated when slipped structures are formed on the template strand for replication. The triplet repeats could be considered as simple short (3bp) direct repeats; thus, direct repeats may not only promote deletions, but also insertions (Wells et al. 2005), as shown in Figure 26. The slipped mispairing mechanism could also account for insertion events of one base frameshifts (Cooper and Krawczak 1993; Kunkel 1990; Ripley 1990).

6.1.2.2. Inverted repeats

Inverted repeats (Figure 22b) have the intrinsic property of forming secondary structures, such as hairpins in single-stranded DNA and cruciforms in double-stranded DNA (Bzymek and Lovett 2001; Wells 2007). An inverted repeat has self-complementarity within the same DNA strand, which allows the strand to fold back on itself and form a hairpin secondary structure (as shown in Figure 27a). Following the self-complementarity of the DNA code, palindromes could also be formed (two symmetrical hairpin structures on both sides of the DNA molecule, Figure 27b). A mechanism has been proposed that could explain deletions promoted by 'quasi-palindromic' sequences facilitated by inverted repeats (Figure 25; Ripley 1982). These 'quasi-palindromic' sequences represent imperfect inverted repeats, where there are extra nucleotides in one of the strands in the hairpin formation (Figure 25, region B).

Figure 24 Modified version of the 'slipped mispairing mechanism' for generation of deletions during DNA replication (adapted from Cooper and Krawczak 1993)



a) Double stranded DNA containing a direct repeat (R1) and homologous interrupted sequence (R2a and R2b)

b) Single stranded DNA

c) R2a and R2b sequences mispair with the complementary R2' repeat, producing single stranded loop

d) Single stranded loop, containing intervening sequence between R2a and R2b is excised and repaired by DNA repair enzymes

e) End of DNA replication. As a result only one of the newly derived double stranded DNA contains the original sequence (R1 and R2a + R2b). The second daughter double stranded DNA lacks the intervening sequence between R2a and R2b

Figure 25 Deletions and insertions mediated by quasi-palindromic sequences (modified after Cooper and Krawczak 1993; Ripley 1982)

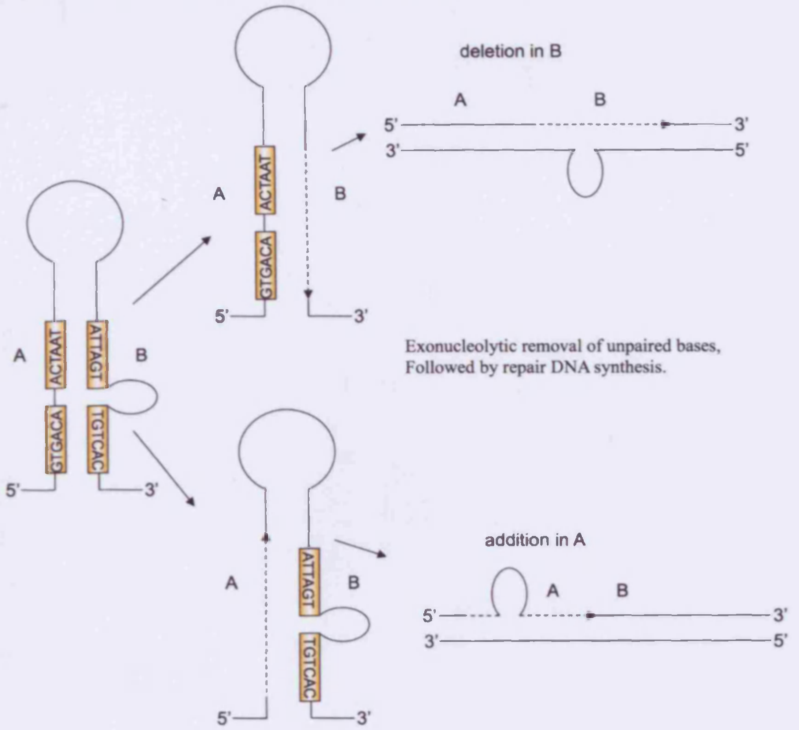
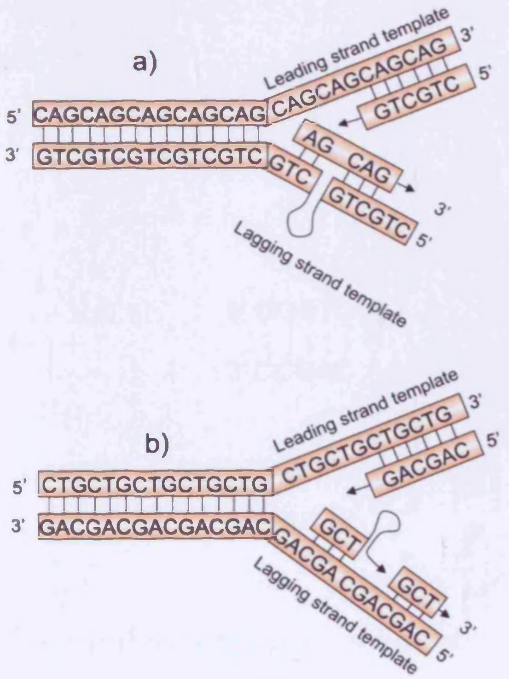


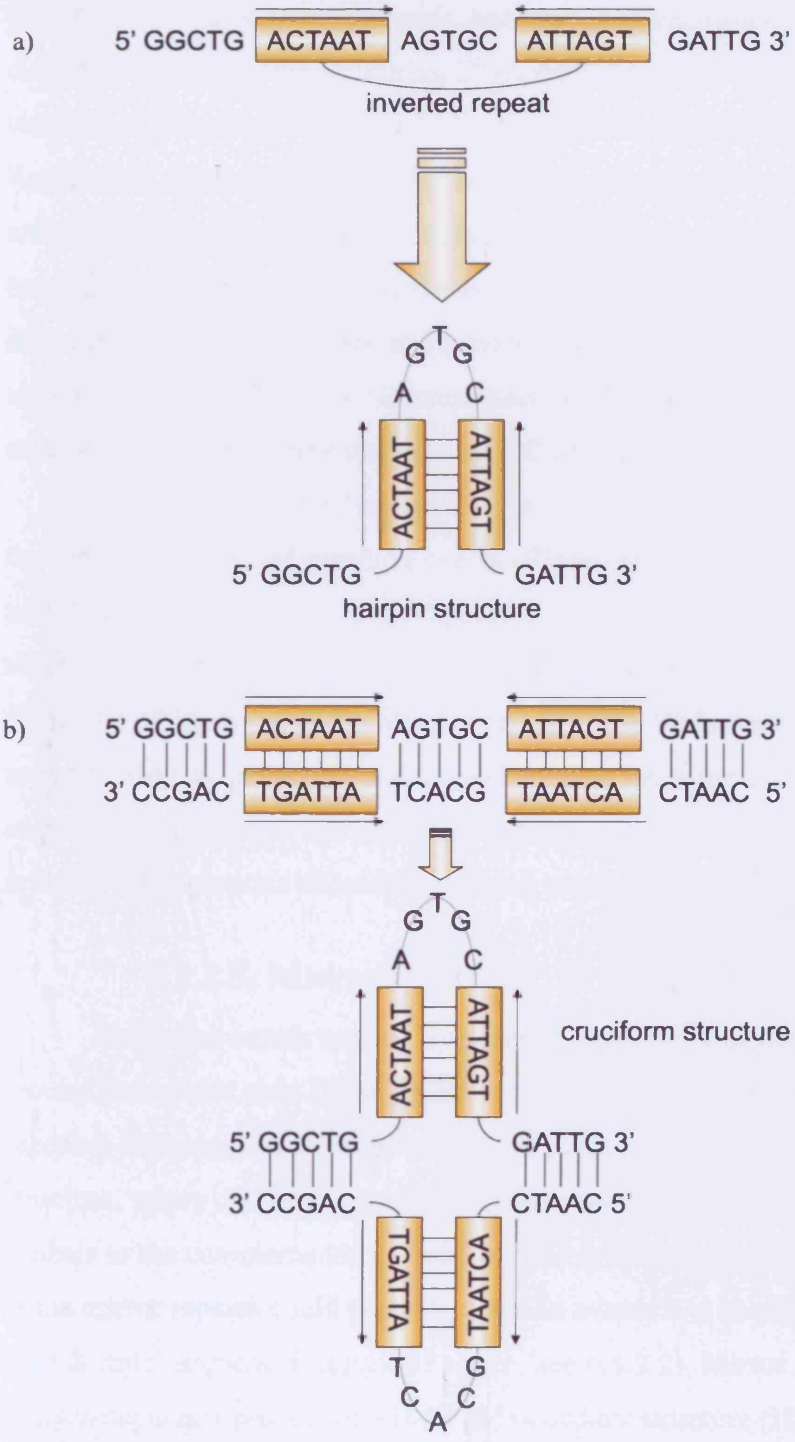
Figure 26 Triplet repeat expansion and contraction



a) Deletion event in the newly synthesized lagging strand, due to a stable hairpin loop structure in the template strand

b) Expansion event due to a hairpin loop structure in the newly synthesized lagging strand

Figure 27 Palindromes and hairpins and inverted repeats (adapted from Bzymek and Lovett 2001)



This model provides a mechanistic explanation for both deletion and insertion events. Exonucleolytic removal of unpaired bases in region B followed by DNA repair synthesis templated by region A (rather than region B on the complementary strand) would lead to deletion of the unpaired bases (Figure 24). Removal of region A and subsequent DNA repair synthesis is templated by region B (rather than region A on the complementary strand) would result in insertion of the mispaired nucleotides (Figure 25, region A). These events may not only be limited to enzymatic repair, but slipped mispairing could potentially also be involved during DNA replication. If, during DNA replication, the primer strand dissociates from the template, a hairpin structure could form and subsequent extension would lead to a mutation. Replication slipped mispairing on the lagging strand could also form after stalling at a hairpin and subsequent misalignment at a direct repeat nearby. Perfect hairpin DNA structures placed between direct repeats increase the deletion rate by up to fourfold when compared to tandem direct repeats alone (Bzymek and Lovett 2001). Furthermore, Bzymek et al. (2001) have shown that defects in the polymerase unit of DNA polymerase III increase the mutation rate mediated by inverted repeats by up to 100 times.

An alternative mechanism (i.e. "strand-switching") has been proposed which explains the same deletions and insertion events (Ripley 1982). The 'strand-switching' model shows that during DNA replication the displaced DNA strand (complementary to the template strand) could itself become a template. Thus, DNA synthesis continues, templated by the displaced DNA strand. Resolution or repair of the 'branched' DNA (Ripley 1982), could be accomplished by hairpin removal, and hence no mutation occurs. If quasi-palindromic sequences take part, or the 'strand-switching' occurs not exactly between the two inverted repeats, then deletions or insertions could result by incorrect use of template.

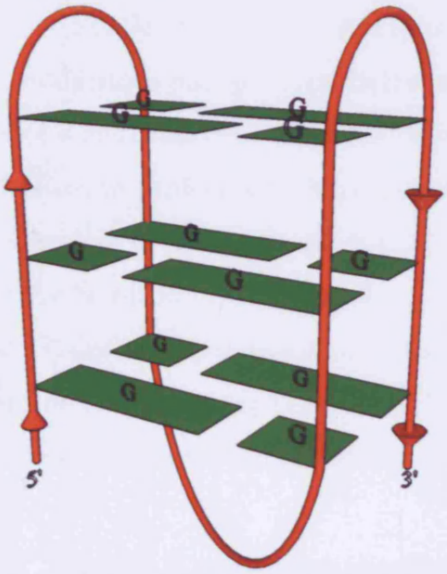
6.1.2.3. Mirror repeats

Sequence motifs termed 'symmetric elements' (mirror repeats) have also been noted around breakpoint ends (Krawczak and Cooper 1991). These authors proposed that these repeated elements could promote deletion events through an intermediate Möbius loop-like structure, where one DNA strand dissociates and twists through a half-turn and then re-anneals to the complementary strand in reverse orientation (Figure 29). Possible mismatches in the mirror repeats could promote deletion events in a similar way to the 'quasi-palindromic' sequences described above (see 6.1.2.2). Mirror repeats could adopt 'intramolecular triplexes' or a H-DNA secondary structure (Htun and Dahlberg 1988). These triplexes have been shown to form *in vitro* (Lyamichev et al. 1985) and *in vivo* (Kohwi et al. 1992; Kohwi and Panchenko 1993; Lee et al. 1989; Ussery and Sinden 1993). The triplexes affect replication fidelity, as DNA polymerase stalls at these secondary structures. In addition it has been shown that they stimulate homologous recombination by bringing direct repeats closer together (Kohwi and Panchenko 1993; Rooney and Moore 1995). It has also been shown that H-DNA could induce double-strand breaks in DNA, thereby increasing genomic instability (mutation frequencies) by up to 20-fold (Wang and Vasquez 2004).

6.1.2.4. C/G quartets

It has been shown that closely spaced runs of Gs could adopt unusual non B-DNA conformations (Bacolla et al. 2001). They are commonly termed G-quartets, 'tetraplexes' or 'tetrads' (Bacolla et al. 2001; Wells 2007). These tracts of G-quartets could be brought close together in a single stranded DNA and could form a four-stranded DNA secondary structure (Figure 28).

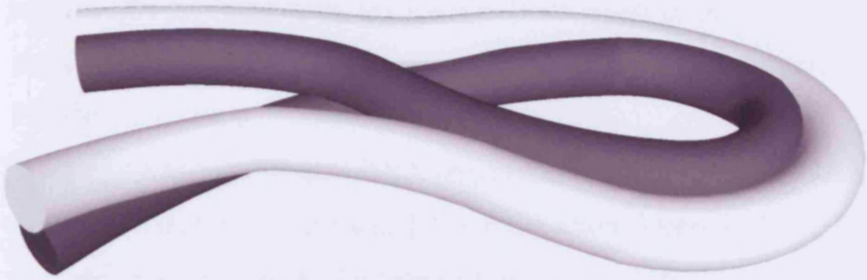
Figure 28 G-quartets (modified after Wells 2007)



It has been shown that the non-B DNA secondary structures (i.e. triplexes) are responsible for mutagenesis rather than the sequence *per se* (Wells 2007; Wojciechowska et al. 2006). They could also bring together direct repeats that could be present on both sides of the triplex structure (Shukla and Roy 2006). This could potentially lead to 'slipped mispairing' and might also induce homologous recombination between the direct repeats (Shukla and Roy 2006). In addition, these unusual structures could be recognized by nucleotide excision repair, causing double-strand breaks and replication fork collapse. Nucleotide excision repair enzymes recognize distortions in the DNA duplex and also chemical modification of single-stranded DNA (Luo et al. 2000). This would also result in recognition and removal of secondary DNA structures and subsequent repair of the gaps by the mismatch repair pathway

(Wells et al. 2005). It has been shown that repair of these gaps by the mismatch repair pathway could lead to deletions in triplet repeat sequences (Jaworski et al. 1995).

Figure 29 Möbius loop-like DNA structure (modified after Cooper and Krawczak 1993)



6.1.2.5. Runs of identical nucleotides

Studies have shown that runs of identical nucleotides are a major factor contributing to endogenous mutagenesis (Ball et al. 2005; Cooper and Krawczak 1993; Greenblatt et al. 1996; Kondrashov and Rogozin 2004). The 'slipped mispairing' mechanism has been proposed to explain why these monotonic sequences are mutagenic (see 'slipped mispairing' mechanism in direct repeats, 6.1.2.1). It has been noted that runs of identical nucleotides are mostly involved in small deletions and insertions (1-2bp). Monotonic runs of 2-5bp account for 83% of all 1bp deletions and insertions (Greenblatt et al. 1996). In another study, a considerable proportion (59/84; -1 frameshifts) were found to occur within monotonic sequences (Cooper and Krawczak 1993).

6.1.3. Exogenous mechanisms of mutagenesis

The human genome is constant subject to a variety of modifying agents (mutagens or toxins; Hagan and Sharrocks 2002). These include oxidation (e.g. reactive oxygen species), radiation (e.g. UV light , gamma and X rays), a plethora of chemicals (e.g. nitrosamines, aromatic amides, polycyclic hydrocarbons, etc.) and food toxins (i.e. aphlatoxin B1; Pineau et al. 2008), to name a few. These mutagens have an enormous impact on the integrity of the DNA. Reactive oxygen species, as a result of normal metabolic processes and numerous external sources (Bertram and Hass 2008), could directly attack DNA thereby generating a variety of mutagenic lesions; including oxidized bases, abasic sites, single-strand breaks (SSBs) and double-strand-breaks (DSBs; Breen and Murphy 1995; Sankaranarayanan and Wassom 2005). Ionizing radiation could also produce reactive oxygen species as well as directly induce SSBs, DSBs, DNA-DNA and DNA-protein links (Sankaranarayanan and

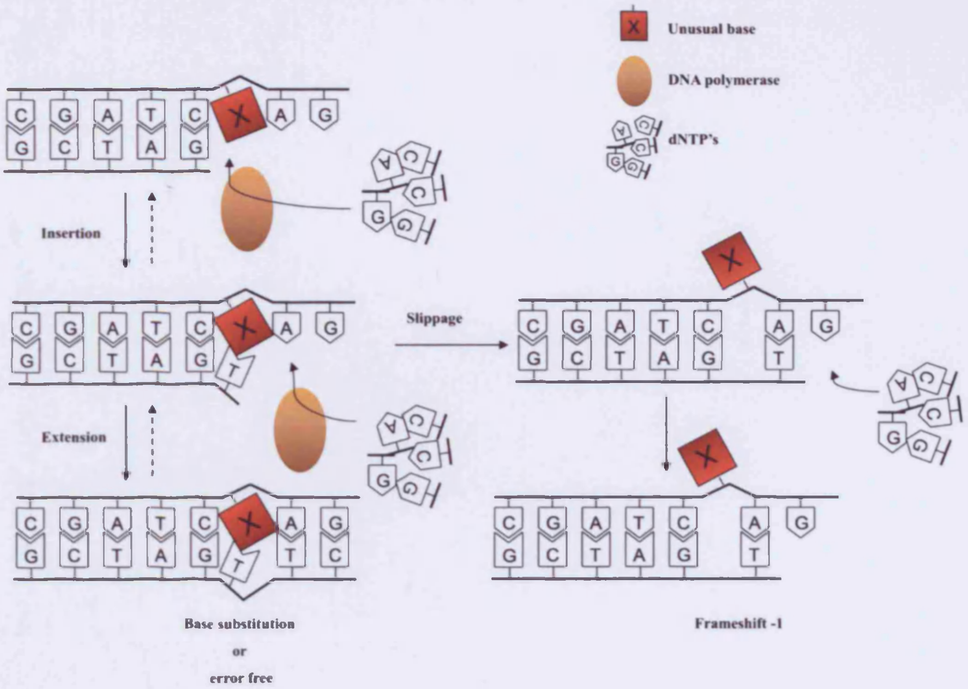
Wassom 2005). A variety of repair mechanisms are involved in the repair process of DNA lesions. These include homologous recombination (HR) and non-homologous recombination repair pathways to repair DSBs (Haber 2000; Jackson 2002; Takata et al. 1998). Non-homologous recombination repair, also known as non-homologous end joining (NHEJ), is the mechanism predominantly involved in DSB repair (Honma et al. 2003). NHEJ requires a short sequence (1-10bp) or no sequence homology at all (Critchlow and Jackson 1998; Pfeiffer et al. 2000; Tsukamoto and Ikeda 1998).

In addition, special DNA polymerases could bypass DNA lesions ('translesion synthesis') that block and stall DNA replication (Pages and Fuchs 2002). The repair mechanisms involved in the repair of mutagenic lesions are efficient, but also error-prone. Nucleotides are often lost, when broken DNA ends are modified in order to be joined by the NHEJ repair mechanism (Ferguson and Alt 2001; Lieber et al. 2003; Pfeiffer et al. 2000). 'Translesion synthesis' could potentially lead to frameshift mutations (Pages and Fuchs 2002). Figure 30 depicts the steps involved in 'translesion synthesis' and its potential to introduce frameshift mutations. The mechanism mainly comprises two steps: insertion and extension. When the DNA polymerase encounters an unusual base (e.g. abasic site, 8-oxo-guanine, B(α)P-N_2-dG, dG-C8-AAF, dG-C8-AF, dG-N_7AFB1, etc.), it will experience difficulty in finding a complementary deoxyrbonucleotide (dNTP) to incorporate. Thus, addition of an incorrect nucleotide is not uncommon. Moreover, if the neighbouring nucleotide on the opposite strand is complementary to the incorrectly added one, slippage may occur resulting in a -1 frameshift mutation (deletion of 1bp; Pages and Fuchs 2002).

6.1.4. Exogenous versus endogenous mechanisms of mutagenesis

The distinction between the action of environmental agents and an endogenous cause of DNA damage may not be so clear. In particular, some exogenous mutagens could induce slippage mutations in tetranucleotide repeats; thus, some sporadic mutations might reflect DNA damage caused by carcinogens (Slebos et al. 2002). Also the introduction of abasic sites in triplet-repeat tracts or the presence of mutagens (e.g. mitomycin C, cyclophosphamide and radiation) induces a higher rate with respect to triplet-repeat expansion (Lyons-Darden and Topal 1999; Pineiro et al. 2003; Zhang et al. 2002). In addition, similarities have been noted between radiation-induced mutations and spontaneous mutation slippage (Niwa 2006).

Figure 30 'Translesion synthesis' mechanism and its potential capacity to introduce frameshift mutations (after Pages and Fuchs 2002)



6.1.5. Somatic and germline mutations in tumour-suppressor genes

It is evident that endogenous mechanisms and various mutagens from endogenous or exogenous sources operate to influence the mutation spectra (*viz.* micro-deletions, micro-insertions and micro-indels). It has also been shown that mutational spectra resulting from the action of environmental mutagens can exhibit marked similarities with mutation spectra considered to be caused by endogenous mechanisms (Lyons-Darden and Topal 1999; Pineiro et al. 2003; Slebos et al. 2002; Zhang et al. 2002). Therefore, mutations could be caused by endogenous mutational mechanisms and exogenous mutagens with or without the interaction between them.

Mutations associated with the malignant transformation of normal cells could arise somatically or be inherited through the germline (Marshall et al. 1997). Germline mutations are generally meiotic in nature, whereas somatic mutations occur predominantly during mitosis. Despite the difference of origin, they often both exhibit similar repeat instability (Sturzeneker et al. 2000; Tijsterman et al. 2002) as well as a similar frequency of homozygosity (Assie et al. 2008). In response to ionizing radiation, both the germline and the soma show similar damage rates (Vilenchik and Knudson 2000). In addition, a review by Erickson (2003) suggests that mutations in genes other than cancer-associated genes might exhibit frequencies similar in the germline and the soma. Then again, it has been shown that the germline exhibits extreme minisatellite instability, whereas this mutational mechanism is rare in the soma (Buard et al. 2000). Also, there might be a direct relationship between the first (i.e. germline) hit and the second (somatic) hit. This relationship could be expressed in terms of the position of the germline hit influencing the position of the second, somatic, hit (Tijsterman et al. 2002).

It is quite surprising that relatively few studies have sought to compare the germline and somatic mutational spectra associated with micro-lesions (*viz.* micro-deletions, micro-insertions and micro-indels). When they have been performed, similarities but also differences in their relative frequency of occurrence and putative mechanisms have been found. In a meta-analysis, Marshall et al. (1997) found many more somatic 1bp deletions (22.3%) as compared to the germline (5.7%). This notwithstanding, no difference was found in any other group of mutations (i.e. insertions and deletions >2bp). Marshall et al. (1997) explored this question further by suggesting that the differences might be due to exposure to different environmental mutagens or differences in efficiency of DNA repair enzymes. Nevertheless, remarkable similarities shared between the soma and the germline have been shown in terms of micro-indels (Gonzalez et al. 2007), suggesting the predominant role of endogenous mechanisms operating to influence both the germline and somatic occurrence of micro-indels (i.e. strand switching and slippage caused by translesion DNA synthesis polymerases).

Clearly the mutational spectrum associated with the neoplastic transformation of normal cells is likely to be a consequence of both endogenous mechanisms and exogenous mutagens. Studying putative mechanisms underlying somatic and germline mutational spectra are thus extremely important since it could lead not only to earlier diagnosis but also to a better understanding of mechanisms underlying tumour progression. In addition, any similarities or differences in the germline and somatic mutational spectra could shed new light on the relative importance of exogenous and endogenous mutagenesis.

6.1.6. Aims of the analysis

The first and most important question addressed in this analysis is the involvement of repetitive elements in the process of mutagenesis in the 17 human tumour suppressor genes studied. The main objective of the analysis was to explore any similarities or differences that somatic and germline micro-lesions (*viz.* micro-deletions, micro-insertions and micro-indels) might exhibit with respect to their occurrence in repetitive elements. To accomplish this objective, a number of tasks were performed.

- **Analysis of the repetitivity of the studied tumour suppressor genes**

Repetitive elements, the most commonly implicated in the process of endogenous mutagenesis were sought in the extended cDNA sequences of the studied genes. These repetitive elements included repeats (i.e. direct, inverted and mirror), C/G quartets and runs of identical nucleotides (RINS).

- **Assessment of the probability of finding micro-lesions in the vicinity of repetitive elements by chance alone.**

Observed micro-lesions for each gene were subdivided into three categories. These included: exclusively somatic (only found in the soma); exclusively germline (only found in the germline); shared (found in both the soma and the germline). For each of these categories, a spectrum of micro-lesions was simulated. This simulated spectrum was used to assess the distribution of micro-lesions with respect to their occurrence in repetitive sequence elements in randomly selected mutations. In addition, to assess the overall distribution of micro-lesions (i.e. micro-lesions in all genes) with respect to their occurrence in repetitive elements, within each category, the micro-lesions were combined for all genes.

- **Explore the similarities and differences of somatic and germline micro-lesions with respect to their occurrence in repetitive elements.**

For each gene, the proportions of micro-lesions found in the vicinity of repetitive elements were compared between the soma and the germline. In addition, for each gene the shared micro-lesions were compared separately to the somatic and germline micro-lesions with respect to the positions of repetitive elements.

These analyses were designed to provide meaningful answers to the following questions:
Are there more observed mutations (*viz*. somatic, germline and shared; combination of somatic, germline and shared micro-lesions for **each gene and all genes**) found in the vicinity of repetitive elements, than would be expected by chance alone?
How do observed mutations (*viz*. somatic germline and shared) compare with each other with respect to their occurrence in the vicinity of repeats?

6.2. Materials and Methods

6.2.1. Materials

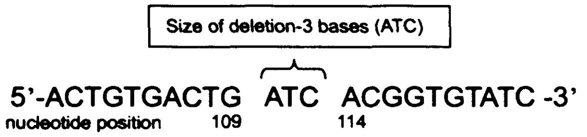
6.2.1.1. Labelling of micro-lesions

For detailed description of labelling of mutations, see 3.3. A summary of the studied micro-lesions is shown in Table 44.

6.2.1.2. Micro-deletions

Micro-deletions were defined by the positions of two breakpoints in the gene sequence and deleted bases between these two breakpoints (as shown in Figure 31). The distance *D* between the two breakpoints was set to ≤20 base-pairs (bp) and ranged from 1-20bp. For each gene, the micro-deletions were assigned a label (i.e. somatic, germline and shared).

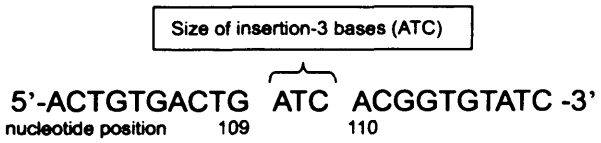
Figure 31 Example of a micro-deletion


Size of deletion-3 bases (ATC)
5'-ACTGTGACTG ATC ACGGTGTATC -3'
nucleotide position
109
114

6.2.1.3. Micro-insertions

Micro-insertions were defined by the position of a single breakpoint and inserted bases of size ≤20bp (ranging from 1-20), as shown in Figure 32. For the purposes of the analysis, the inserted bases were not taken into consideration, only the positions of the breakpoint.

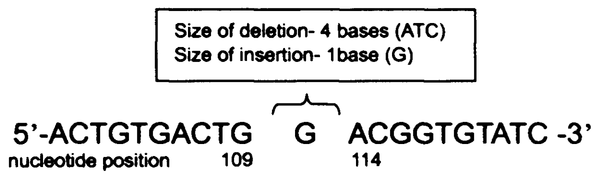
Figure 32 Example of a micro-insertion


Size of insertion-3 bases (ATC)
5'-ACTGTGACTG ATC ACGGTGTATC -3'
nucleotide position
109
110

6.2.1.4. Micro-indels

Micro-indels were defined by the positions of two breakpoints and inserted bases between the breakpoints. The lengths of both the deleted and the inserted bases were set to ≤20bp (ranging from 1-20), as shown in Figure 33. For the purposes of the analysis, the inserted bases were not taken into consideration, only the positions of the breakpoints.

Figure 33 Example of a micro-indel


Size of deletion- 4 bases (ATC)
Size of insertion- 1base (G)
5'-ACTGTGACTG G ACGGTGTATC -3'
nucleotide position 109
114

6.2.2. Methods

6.2.2.1. General definitions of repetitive elements

A DNA pattern or DNA substring is a sequence of nucleotides $\{A, C, G, T\}$ with an arbitrary length. Repetitive DNA sequence refers to a DNA substring or DNA pattern that is found multiple times (at least twice) throughout the sequence in question. Most frequent repetitive elements associated with mutagenesis are: repeats (*viz.* direct, inverted and mirror repeats), C/G quartets and runs of identical nucleotides (RINS).

6.2.2.1.1. Repeats

Repeats are defined by a pair of DNA substrings (5' and 3' parts) of length (*m*) and distance (*D*) between the parts of the repeats. Both parts of the repeat are found on the same strand of DNA. The length *m* of the 5' and 3' parts of the repeats was set to be ≥6bp, ≥7bp, and ≥8bp. The distance (*D*) between the 5' and 3' parts of the repeats was set to be ≤20bp (ranging from 0-20bp; 0bp when one part of the repeat abuts the other part). In addition to the aforementioned sizes, the number of mutations within regions that include the repeats themselves and ±5bp of flanking sequences away from the repeats were also considered. These included ≥6±5bp, ≥7±5bp, and ≥8±5bp regions.

6.2.2.1.1.1. Direct repeats

A direct repeat was defined as a DNA substring or a DNA pattern that is found twice on the same strand of DNA with an arbitrary distance between them. A direct repeat used in this analysis was defined as two copies of exactly the same DNA pattern/sequence (Figure 22a) that are found in the extended cDNA of the studied genes. When the repeats are entirely the same (e.g. CAGTTTA and CAGTTTA), they are said to be exact repeats. Only exact direct repeats were considered in this analysis. Multiple instances of exactly the same DNA patterns were considered as separate direct repeats (as shown in Figure 22a). Thus, each direct repeat consists of a pair of exactly the same DNA patterns (5' and 3' parts) of the same length (*m*) and distance (*D*) between the parts.

6.2.2.1.1.2. Inverted repeats

An inverted repeat was defined in a similar way to a direct repeat (see 6.2.2.1.1.1), with the exception that the second copy of the repeat (3' part) is the reverse complement of the first DNA pattern or substring (5' part; Lang 2007). Both parts of the inverted repeat are found on the same DNA strand (Figure 22b). Only exact inverted repeats were considered in this analysis (i.e. one copy of the inverted repeat is an exact reverse complement of the other; e.g. CAGTTA and TAACTG). Multiple instances were treated in an analogous way as explained in direct repeats (see 6.2.2.1.1.1). Inverted repeats that were also found to be direct repeats were excluded from the list of repeats. For example, an inverted repeat ATAT-ATAT is also a direct repeat. Thus, only the direct repeat was considered.

6.2.2.1.1.3. Mirror repeats

A mirror repeat was defined in an analogous way to a direct repeat (6.2.2.1.1.1), with the exception that the second copy of the repeat (3') is a mirror reflection of the first copy (5'). It is said that both parts of the mirror repeat have a centre of symmetry on a single strand of DNA (Figure 22c; Cooper and Krawczak 1991). Only exact mirror repeats were considered in the analysis (i.e. one copy of the mirror repeat is an exact mirror image of the other copy e.g. CAGTTA and ATTGAC). Mirror repeats that were also found to be direct repeats were excluded from the list of repeats. For example, a mirror repeat ATATA-ATATA is also a direct repeat. Thus, only the direct repeat was considered.

6.2.2.1.1.4. C/G quartets

C/G-quartets (Bacolla et al. 2004) were defined using the following equation:

Equation 5 C/G-quartets

$S_{(3-5)}N_{(1-5)}S_{(3-5)}N_{(1-5)}S_{(3-5)}N_{(1-5)}S_{(3-5)}$, where S could be either G or C and $N = \{A,C,G,T\}$.

An example of a G-quartet is given in Figure 22d.

6.2.2.1.1.5. Runs of identical nucleotides

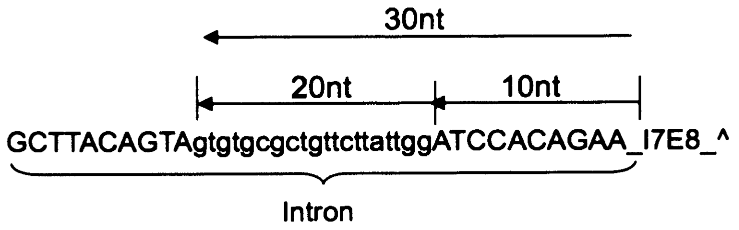
Runs of identical nucleotides (RINS), also known as 'homonucleotide runs' (Kondrashov and Rogozin 2004) or 'contiguous sequence' (Cooper and Krawczak 1993), were defined as non-interrupted sequence comprising the same nucleotide of length ≥4bp (e.g. AAAA, AAAAA, TTTT, etc.).

6.2.2.2. Simulation of potential micro-lesions

The number of breakpoints differs between the 3 types of lesions (i.e. micro-deletions and micro-indels having 2 breakpoints and micro-insertions having 1 breakpoint). It is logical that micro-deletions and indels would have a slightly better chance of coinciding with repetitive elements than micro-insertions on the basis that micro-deletions and indels have 2 breakpoints versus 1 breakpoint for the micro-insertions. Thus, the number of the different types of mutation (being micro-deletions, micro-insertions and indels) has to be known *a priori* in order to generate potential micro-lesions. By default, all positions in the cDNA sequence of the genes could serve as breakpoints of mutations (*viz*. micro-deletions, micro-insertions and micro-indels). The distance between the breakpoints in accordance with the already collected micro-lesions, could be 1-20bp for micro-deletions and micro-indels and 0bp for micro-insertions.

Some breakpoints could extend into the introns of the genes; thus, the leftmost or rightmost positions for a mutational breakpoint in the introns must be determined.
In order to calculate the leftmost or rightmost positions that a mutational breakpoint could extend into an intron, the most extreme case possible must be considered (leftmost position; Figure 34).

Figure 34 An example of the most extreme case of a micro-deletion with respect to a mutational breakpoint in the intron (micro-deletion in lower case)



The leftmost position of a mutational breakpoint that could occur in an intron is between nucleotide positions 30-31 (see Figure 36). A mutation in the vicinity of a repeat is defined in such a way that a breakpoint of the mutation has to overlap with the repeat (the repeat itself or between the repeats). Thus, the furthest position that a mutational breakpoint (for micro-deletions and micro-indels) could extend into an intron is:

10bp DNA reference + max 20bp deletion = 30bp (breakpoint between nucleotide positions 31-30), as shown in Figure 35 and Figure 36.

For micro-insertions:

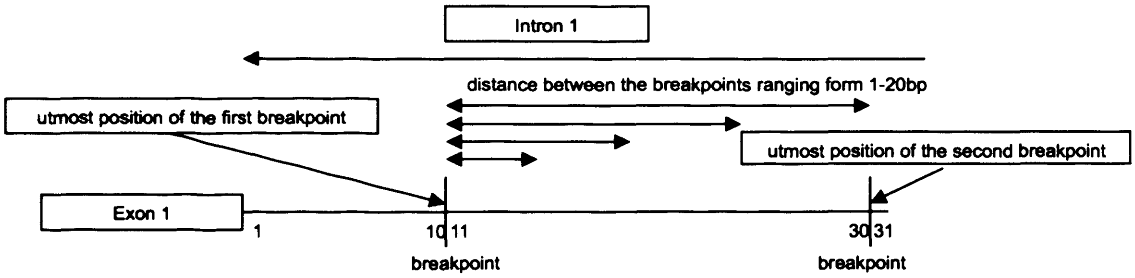
10bp (breakpoint between nucleotide positions 10-11)

The furthest (*viz.* leftmost and rightmost) positions where breakpoints of micro-deletions and micro-indels could extend into an intron could be split into two case scenarios:

An exon followed by an intron:

In this case, the first breakpoint could potentially occur anywhere in the exon and the first 10 nucleotides into the intron (breakpoint between nucleotides 10 and 11, Figure 35). The second breakpoint (towards the 3' end of the gene with respect to the first breakpoint) could potentially occur up to 30bp in the intron (breakpoint between nucleotides 30 and 31, Figure 35). The distance between these breakpoints is set to range from 1-20 nucleotides.

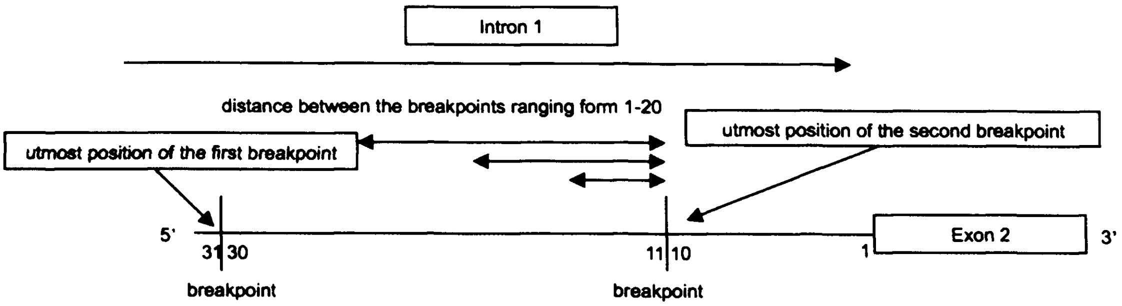
Figure 35 Potential breakpoints in micro-deletions and micro-indels (an exon followed by an intron)


Intron 1
distance between the breakpoints ranging form 1-20bp
utmost position of the first breakpoint
utmost position of the second breakpoint
Exon 1
1
10
11
30
31
breakpoint
breakpoint

An intron followed by an exon:

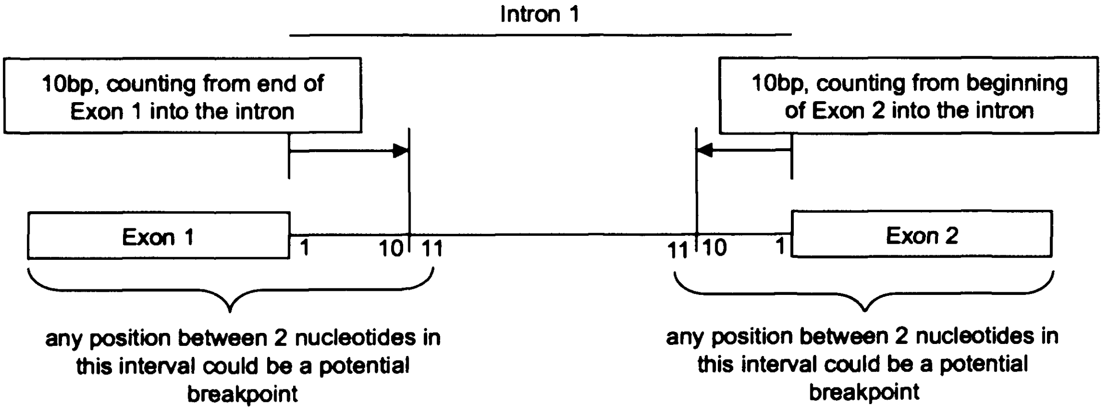
In this case, the first breakpoint could occur anywhere in the exon and the first 10nt of the intron (counting from the beginning of the exon into the intron). The leftmost position of the second breakpoint (towards the 5' end of the gene with respect to the first breakpoint) is the 30th nt (breakpoint between nucleotide positions 30-31) with a distance from the first breakpoint ranging from 1-20 nucleotides (as shown in Figure 36).

Figure 36 Potential breakpoints in micro-deletions and micro-indels (an intron followed by an exon)


Intron 1
distance between the breakpoints ranging form 1-20
utmost position of the first breakpoint
utmost position of the second breakpoint
5'
31 30
breakpoint
11 10
breakpoint
1
Exon 2
3'

For micro-insertions, the furthest (leftmost or rightmost) position into an intron where a breakpoint could potentially occur is between nucleotides 10-11, counting from the end of an exon into an intron and between nucleotides 10-11, counting from the beginning of an exon into the intron to account for the end of the intron (Figure 37).

Figure 37 Potential breakpoints in micro-insertions


Intron 1
10bp, counting from end of Exon 1 into the intron
10bp, counting from beginning of Exon 2 into the intron
Exon 1
1
10
11
11
10
1
Exon 2
any position between 2 nucleotides in this interval could be a potential breakpoint
any position between 2 nucleotides in this interval could be a potential breakpoint

Micro-insertions - 0 bases (by default, micro-insertions have only one breakpoint, hence the distance between the breakpoints is 0 bases).

6.2.2.3. Generation of extended cDNA sequences

As described in 6.2.2.2, the leftmost or rightmost breakpoint positions that a potential micro-lesion could extend into an intron is between nucleotides 30 and 31. Thus, the extended cDNA sequence needed for each intron is:

Equation 6 Extended cDNA sequence

$$REF_{(10bp)} + \max(del) + R_1 + \max(D) + R_2 = 65bp,$$

where REF is 10bp DNA reference, $\max(del)$ is the maximum length of the sequence deleted, R_1 is the size of the first part of the repeat (7bp, allowing for 1bp overlap), $\max(D)$

is the maximum distance in the repeat (20bp), R_2 is the size of the second part of the repeat (8bp).

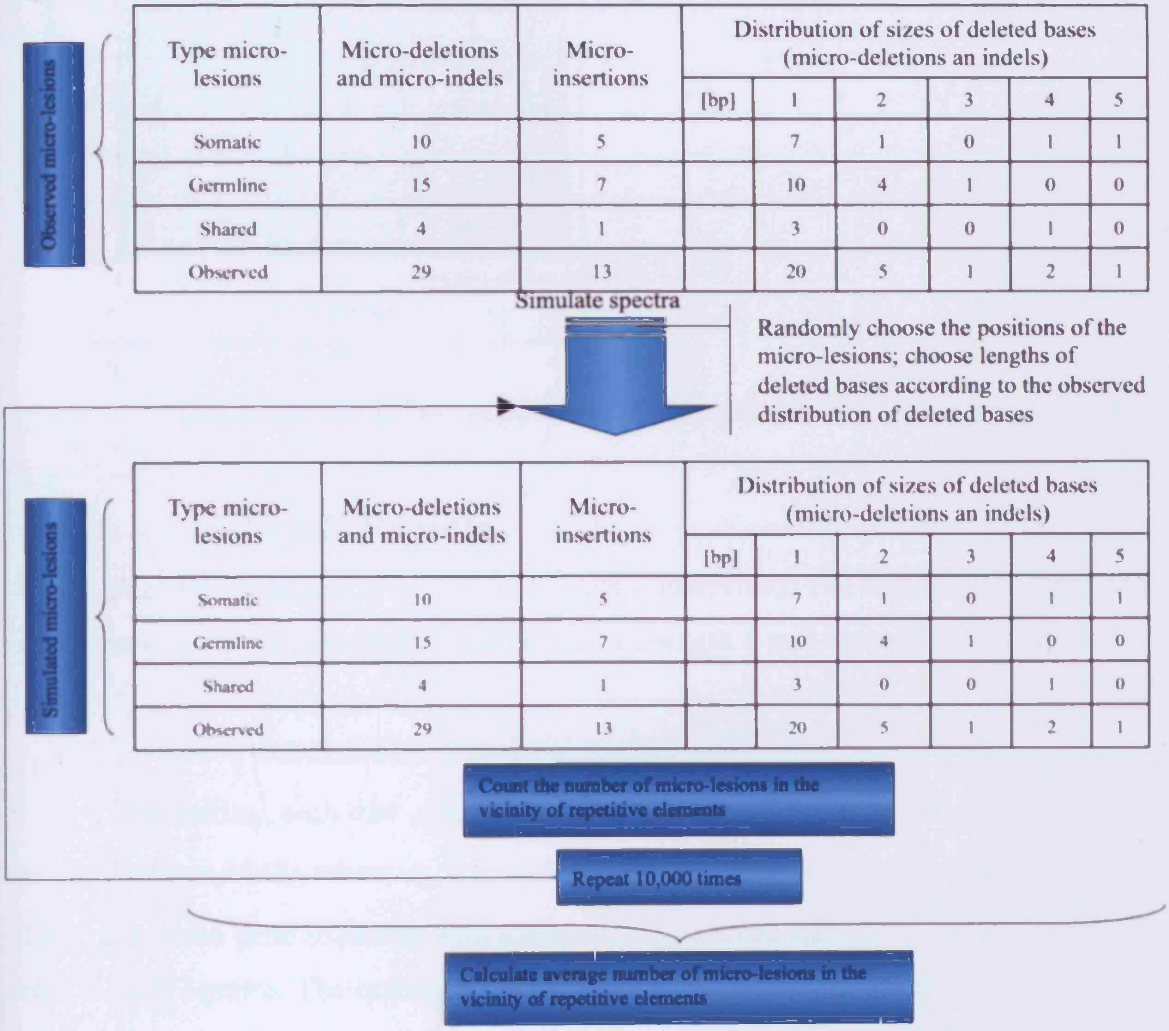
The actual size of the repeat could be bigger than the size found by the above formula, as one part of the repeat could abut the other. Thus, the actual combined size of both parts of the repeat could not extend more than 65bp + 20bp (no distance in the repeat). Thus, the extended cDNA sequence needed for each intron is 65+20=85bp. These 85bp of intronic sequence would also cover runs of identical nucleotides (defined as ≥4bp) and C/G quartets (maximum size of C/G quartets is 35bp, see Equation 5).

Generation of extended cDNA sequences is described in 3.4.

6.2.2.4. Generation of simulated spectra

In order to find out how many micro-lesions would be found in the vicinity of repetitive elements by chance alone, simulated spectra were used. I devised a computer program that automatically generates simulated spectra. For each gene, the number of micro-deletions, micro-insertions and indels were counted. Since the sizes of deleted bases in micro-deletions and indels are not uniformly distributed (as shown in Figure 39 - many more mutations are found with small lengths of deleted bases), the lengths of deleted bases could not be randomly chosen (i.e. simulate uniform distribution of lengths of deleted bases). Therefore for each gene, the distribution of lengths of deleted bases was assessed. The numbers of mutations within each size category of deleted bases (e.g. 1bp, 2bp, 3bp, up to the maximum size of the deleted bases) were counted and this represented the distribution of deleted bases. This distribution was used for choosing the size of the deleted bases for the generation of the simulated spectra. For each gene, on the basis of the number of mutations in each mutation category (i.e. somatic, germline and shared) and the distribution of sizes of deleted bases, mutational spectra were simulated. Thus, the positions of micro-lesions were chosen completely randomly and the distribution of sizes of deleted bases followed the distribution of the original micro-lesion spectra (for a flow chart, see Figure 38). This process of generating micro-lesion spectra was repeated 10,000 times as described above (i.e. 10,000 simulations). After each simulation, the number of mutations that occur either in repeats or in the vicinity of repetitive elements (as described above) was counted. After 10,000 simulations, the average number of mutations that occur either in the repeats or in the vicinity of the repetitive elements was calculated. These numbers were used to compare the observed number of mutations in the vicinity of repetitive elements with simulated spectra. In fact, the simulated spectra represent the number of mutations that would be found in the vicinity of repetitive elements if there were no association between micro-lesions and repetitive elements (i.e. that expected by chance alone).

Figure 38 Example of the process of generating simulated spectra



Type micro-lesions	Micro-deletions and micro-indels	Micro-insertions	Distribution of sizes of deleted bases (micro-deletions an indels)					
			[bp]	1	2	3	4	5
Somatic	10	5		7	1	0	1	1
Germline	15	7		10	4	1	0	0
Shared	4	1		3	0	0	1	0
Observed	29	13		20	5	1	2	1

Randomly choose the positions of the micro-lesions; choose lengths of deleted bases according to the observed distribution of deleted bases

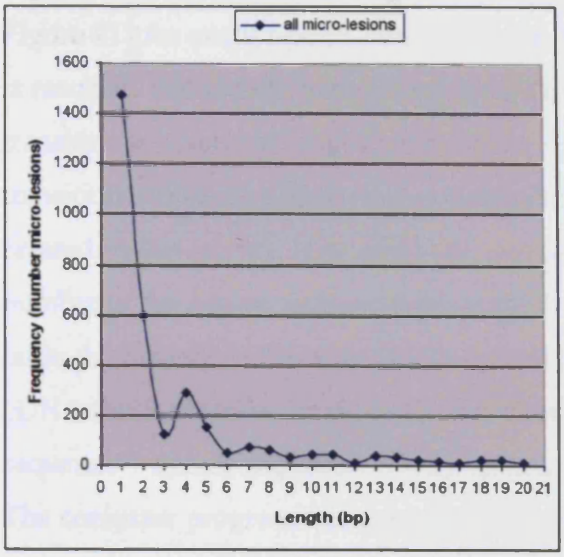
Type micro-lesions	Micro-deletions and micro-indels	Micro-insertions	Distribution of sizes of deleted bases (micro-deletions an indels)					
			[bp]	1	2	3	4	5
Somatic	10	5		7	1	0	1	1
Germline	15	7		10	4	1	0	0
Shared	4	1		3	0	0	1	0
Observed	29	13		20	5	1	2	1

Count the number of micro-lesions in the vicinity of repetitive elements

Repeat 10,000 times

Calculate average number of micro-lesions in the vicinity of repetitive elements

Figure 39 Distribution of lengths of micro-lesions (micro-deletions and indels)



6.2.2.5. Search for repetitive elements in the extended cDNA

6.2.2.5.1. Repeats

A novel algorithm to search for repeats was devised. The algorithm consists of three steps: pre-processing, search for repeats of fixed length *L* and extension of the repeats.

6.2.2.5.1.1. Pre-Processing

A substring, such that $a_ia_{i+1}..a_{i+l-1}$ of string $a_1a_2..a_L$ is termed an oligonucleotide or *l*-gram (Shannon 1948), where a_i is a nucleotide (i.e. A,C,G or T) within the oligoneucleotide. Thus, in a given gene sequence with a length *L* and a fixed size of the *l*-grams *l* ($l \le L$), there are $(L-l+1)$ *l*-grams. The extended cDNA sequence for each gene was used to generate all *l*-grams with a fixed size (*viz*. 6bp, 7bp and 8bp), as shown in Figure 40. For each gene, a tree-like structure also known as a trie-structure (Knuth 1973) was generated from the *l*-grams. A trie structure is a tree structure that is used to store strings (e.g. *l*-grams). It consists of numbered nodes and leaves (leaf is used to denote end-nodes or arcs). Trie has one node called a root node. Each arc is labelled by a symbol from *l*-gram. If several *l*-grams share a common prefix, then there is only one path leading from the root node to the node corresponding to the common prefixes; *l*-grams are represented and stored in end-nodes or leaves. An example of a trie-structure is given in Figure 41. The construction of the trie-structure was as follows:

Consecutively for each symbol of the *l*-gram starting from the root node (node 0 in Figure 41) for every *l*-gram, search the trie-structure until a mismatch is found or an end node is reached. The search was performed using a *goto* function that returns a success if a transition to a successive node is possible or a failure if no transition is possible. If no transition is possible, then a new arc labelled with the current symbol from the current node is created. Alternatively, if an end node is reached, the output function is updated with the number of the *l*-gram and positions of their occurrence. I devised a computer program that takes the extended cDNA sequence as well as the mapping file associated with the extended cDNA (both described in 6.2.2.3) and consecutively for all extended exons (85bp intronic sequence + exonic sequence + 85bp intronic sequence), generates the trie-based structure. The computer program was constructed in such a way that only three parameters are needed: extended cDNA, mapping file and size of the *l*-grams. Thus, for each gene, trie-based structures were generated (size of *l*-grams 6bp, 7bp and 8bp). A combination of an array and a hash was used to implement the trie-based structure. The array indices comprised the numbers of the nodes in the trie structure and each element in the array pointing to a hash comprising DNA symbols (A, C, G and T). The transition from a node to node (*goto* function) was implemented by assigning a pointer for each DNA symbol in the hash to an index in the array. The output function was implemented by using a hash. The keys in the hash represented the end nodes. Each number of end node (keys in the hash) pointed to an array with the positions of *l*-grams in the extended cDNA sequence

Figure 40 Example of generation of *l*-grams with a size 4bp and a sliding window of 1bp

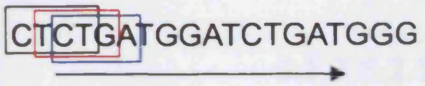
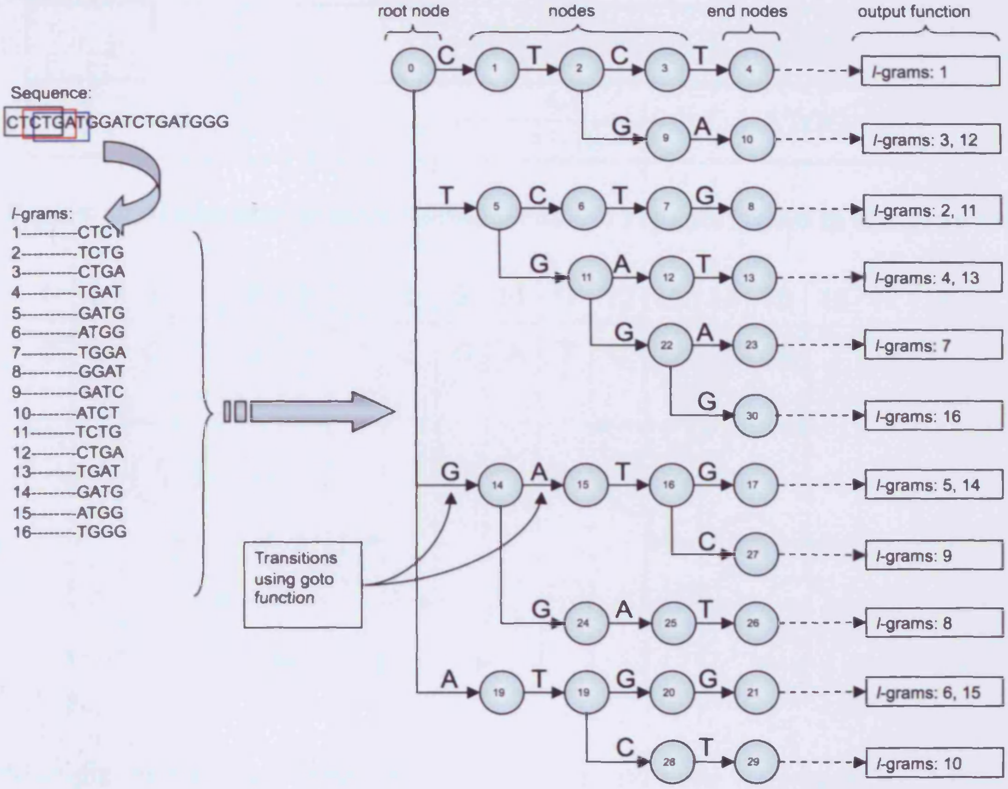


Figure 41 Example of a trie-based structure



6.2.2.5.2. Search for repeats with a fixed length L

I devised a computer program that searches the trie-based structure (as described in 6.2.2.5.1.1) for repeats (*viz*. direct, inverted and mirror repeats). For each gene, it takes the trie-based structure as an input and sequentially generates a list of the positions where repeats are found in the extended cDNA.

6.2.2.5.2.1.1. Direct repeats

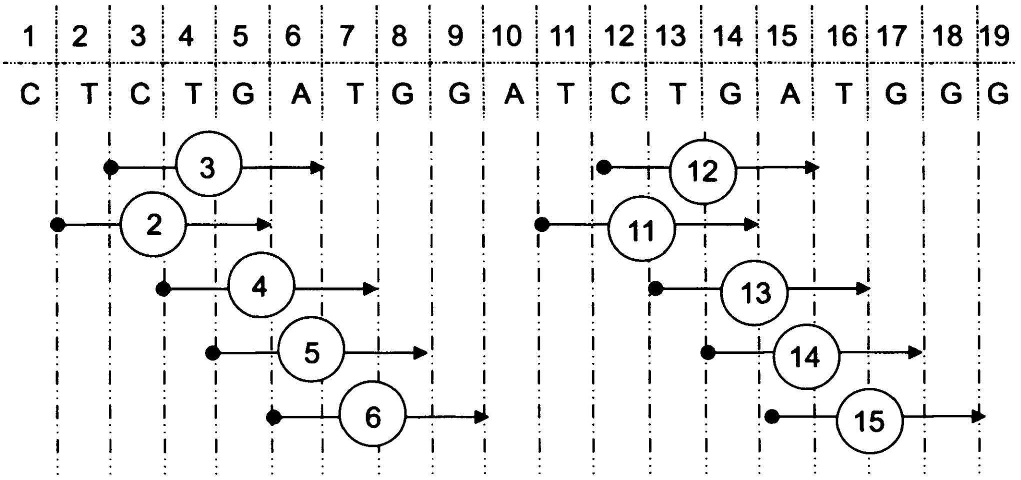
Finding direct repeats from the trie-based structures is straightforward. Direct repeats are found where the output function for a given end node produces positions of two or more *l*-grams (as seen in Table 25 and Figure 42).

Table 25 Direct repeats found at end nodes

Positions of 2 or more *l*-grams found at an end node	Gene positions	Sequence
3	3-6	CTGA
12	12-15	
2	2-5	TCTG
11	11-14	

4 13	4-7 13-16	TGAT
5 14	5-8 14-17	GATG
6 15	6-9 15-18	ATGG

Figure 42 Schematic representation of direct repeats found in the gene sequence



Multiple instances of *l*-grams found at a given end node were considered as separate direct repeats (as described in 6.2.2.1.1.1).

6.2.2.5.2.1.2. Inverted repeats

Finding inverted repeats was performed by searching the already generated trie-based structure (*l*-gram sizes: 6bp, 7bp and 8bp) with the reverse complement of the extended cDNA. Thus, *L-grams* were generated from the reverse complement (the reverse complement of the extended cDNA sequence was used as an input to the trie-based structure) with sizes: 6bp, 7bp and 8bp. The search procedure for every *l*-gram was as follows:

Start from the initial state (0) and *l*-gram *i*

Using the *goto* function, compare DNA symbols from the *l*-gram until a mismatch is found or an end node is reached.

If a mismatch is found, restart the search from the initial state (0) and *l*-gram *i*+1

If an end node is reached, output *i* and the positions found at the end node

Where inverted repeats were found, the actual genic position where the inverted repeat starts and ends were calculated using the following formulae:

Equation 7 Start position of inverted repeats

$P_{start} = L - K - 2$,

where P_{start} is the position where an inverted repeat starts, L is the gene size in bp and K is the number of the l-gram

Equation 8 End position of inverted repeats

$P_{end} = P_{start} + l - 1$,

where l is the fixed minimum size of the l-grams.

The first part of the inverted repeat (5') is the l-gram that is found at a given end node. Multiple instances of l-grams found at a given end node were considered as separate direct repeats (as described in 6.2.2.1.1.1).

6.2.2.5.2.1.3. Mirror repeats

The search for mirror repeats was performed in the same way as inverted repeats (described in 6.2.2.5.2.1.2), with the exception that the trie-based structure was searched using the reverse of the extended cDNA.

6.2.2.5.2.1.4. Extension of repeats

As shown in Figure 42, there were longer repeats than the fixed minimum size of l-grams exist. Thus, it was necessary to generate the longest possible repeats. I wrote a computer program that takes a list of repeats (*viz.* direct, inverted and mirror repeats) and generates the longest possible repeats. The program uses the following algorithm:

For l-grams found at end nodes, with a frequency ≥2, all possible pairs of l-grams were generated. These pairs of l-grams represent repeats, one l-gram the 5' part and the other l-gram in the pair the 3' part of the repeat. The so formed pairs were numerically sorted in ascending order by the position of the first part of the repeat. Consecutively, every pair of l-grams was compared with the rest of the pairs.

If P denotes the genic positions where l-grams start and end, then $P_S = start$ and $P_E = end$. If G is an l-gram pair, then G_{11} is the 5' part of the repeat and G_{12} is the 3' part of the repeat in the l-gram pair and the positions in an l-gram pair are: $P_S.G_{11}$, $P_E.G_{11}$, $P_S.G_{12}$ and $P_E.G_{12}$ (as shown in Figure 43). Let the distances D_1, D_2, D_3, D_4 be as follows:

$D_1 = P_S.G_{11} - P_S.G_{21}$,

$D_2 = P_E.G_{11} - P_E.G_{21}$,

$D_3 = P_S.G_{12} - P_S.G_{22}$ and

$D_4 = P_E.G_{12} - P_E.G_{22}$.

The following rules were used for extending the length of repeats:

Form a new repeat, if the following conditions are met:

$D_1 = D_3$ and $D_2 = D_4$ for direct repeats or

$D_1 = |D_3|$ and $D_2 = |D_4|$ for inverted and mirror repeats,

$[\min(P_S.G_{12}, P_S.G_{22}) - \max(P_E.G_{11}, P_E.G_{21})] > 0$, and

$P_S.G_{21} > P_S.G_{11}$ and $P_E.G_{21} < P_E.G_{11}$

Extension of repeats:

$P_S.G_{11} = \min(P_S.G_{11}, P_S.G_{21})$

$P_E.G_{11} = \max(P_E.G_{11}, P_E.G_{21})$

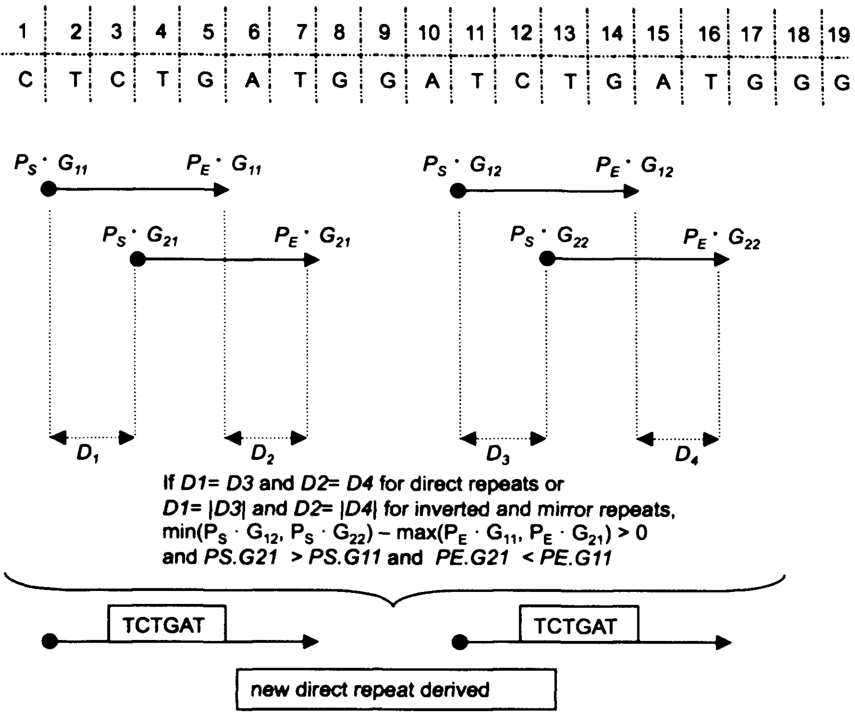
$P_S.G_{12} = \min(P_S.G_{12}, P_S.G_{22})$

$P_E.G_{12} = \max(P_E.G_{12}, P_E.G_{22})$

If a new repeat is formed:

Then, the newly formed extended repeat is added to the pool of *l*-gram pairs and the pair of *l*-grams that formed this repeat are deleted from the list of -*l*-gram pairs. Continue the comparisons with the newly formed extended repeat.

Figure 43 Extension of repeats



If *D1= D3* and *D2= D4* for direct repeats or
D1= |D3| and *D2= |D4|* for inverted and mirror repeats,
$\min(P_S \cdot G_{12}, P_S \cdot G_{22}) - \max(P_E \cdot G_{11}, P_E \cdot G_{21}) > 0$
and *PS.G21* > *PS.G11* and *PE.G21* < *PE.G11*

6.2.2.5.3. Search for C/G quartets

I designed a computer program that automatically searches the extended cDNA of each gene for C/G quartets. Starting from the beginning of the extended cDNA and sequentially for each extended exonic sequence (85bp intronic sequence around each exon and including the exonic sequence itself), the program finds stretches of Cs and Gs that are ≥3bp according to Equation 5. Then, if stretches of Cs and separately for Gs did not satisfy the criteria listed in Equation 5, they were removed from the list of C/G quartets.

6.2.2.5.4. Search for runs of identical nucleotides

In a similar fashion to that described in 6.2.2.5.3, a search for stretches of mononucleotides (*viz*. A, C, G and T) of length ≥4bp was performed. I designed a computer program that takes the extended cDNA as an input and generates the positions of all runs of identical nucleotides. In consecutive manner, the program reads the extended cDNA and records the positions and the lengths of stretches of mononucleotides.

6.2.2.6. Labelling micro-lesions with respect to their occurrence in repetitive elements

Every micro-lesion (*viz.* micro-deletions, micro-insertions and micro-indels) in each gene was labelled according to their occurrence in repetitive elements. Two labels were used: 1 if there is a repetitive element in the vicinity of the micro-lesion and 0 if there is no repetitive element in the vicinity of the micro-lesion. I designed a computer program such that when supplied with a list of micro-lesions with positions of the breakpoints and a list of repetitive elements with their positions, it automatically assigns appropriate labels (i.e. 1 or 0 as described above). Both lists are tab-delimited text files. The program works in an iterative way. Thus, for every micro-lesion, it scans the list of repeats and checks whether a mutational breakpoint occurs in the vicinity of a repetitive element or within repeats themselves (for description see 6.2.2.6.1 and 6.2.2.6.2).

6.2.2.6.1. Micro-lesions and repeats

Micro-lesions such that at least one breakpoint is found to coincide with repeats were labelled in the vicinity of repeats. To coincide with a repeat, a breakpoint had to overlap with any part of the repeat (i.e. 5' or 3' part) or lie in between the parts of the repeat (for repeat sizes of ≥6bp, ≥7bp and ≥8bp). In addition, for repeats ±5bp (i.e. ≥6±5bp, ≥7±5bp, ≥8±5bp), micro-lesions were considered to be in the vicinity of repeats if a mutational breakpoint is within 5bp of a repeat, but also including the rules for repeat sizes of ≥6bp, ≥7bp and ≥8bp.

6.2.2.6.2. Micro-lesions and runs of identical nucleotides

Micro-lesions were labelled in the vicinity of runs of identical nucleotides in the same way as described in 6.2.2.6.1. Instead of using the positions of repeats, the positions of runs of identical nucleotides were used. In addition, micro-lesions were considered to be in the vicinity of runs of identical nucleotides if a mutational breakpoint was within 5bp of a run, in addition to the rules for runs of identical nucleotides.

6.2.2.7. Comparisons and statistical significance

In order to answer the questions set out in the aims of the analysis (see Section 6.1.6), for each group of the studied repetitive elements (*viz.* repeats and runs of identical nucleotides) the following comparisons for each gene were performed:

Soma vs. simulated micro-lesions
Germline vs. simulated micro-lesions

Shared vs. simulated micro-lesions
Observed (*viz.* somatic, germline and shared mutations) vs. simulated micro-lesions
Soma vs. germline
Soma vs. shared
Germline vs. shared

For each of the tests, a χ^2 test statistic was calculated (see Chapter 3 (General methods) for description) to assess the statistical significance at the chosen significance level ($\alpha = 0.05$). To allow for multiple hypothesis testing for the tests performed for each gene, 10,000 resampling permutations were performed (see Chapter 3 (General methods) for description) and corresponding p-value was termed permuted. To allow for multiple hypothesis testing, for the tests performed for all repeat/runs of identical nucleotide sizes, a Bonferroni correction was applied. Therefore, each permuted p-value, associated with repeats was multiplied by 6 (repeat-wise $\alpha = 0.05/6$ or 0.0083), to account for different repeat sizes (*viz.* repeat sizes of ≥6bp, ≥7bp, ≥8bp, ≥6±5bp, ≥7±5bp and ≥8±5bp). In addition, each permuted p-value, associated with RINS, was multiplied by 2 (experiment-wise $\alpha = 0.05/2$ or 0.025), to account for the different RINS sizes (i.e. run sizes ≥4bp and ≥4±5bp). To allow for multiple hypothesis testing, for the tests performed for all genes, a Bonferroni correction was also applied. Therefore, each repeat-wise p-value was multiplied by 17 (overall experiment-wise $\alpha = \frac{0.05}{6*17}$ or 0.00049) and each run-wise p-value was multiplied by 2 (overall experiment-wise $\alpha = \frac{0.05}{2*17}$ or 0.0015).

I designed a computer program that automatically performs the χ^2 test statistic for each test along with the re-sampling permutations and Bonferroni corrections.

Furthermore, the numbers of mutations in all genes were combined, only if micro-lesions had the same label (i.e. somatic, germline, shared) to represent the combination of mutations in all genes. The aforementioned tests were also performed for the combination of mutations in all genes. Additional tests were performed for the combination of recurrent somatic mutations for all genes:

Somatic recurrent vs. somatic non-recurrent micro-lesions
Somatic recurrent and shared vs. somatic recurrent non-shared micro-lesions

Somatic non-recurrent and shared vs. somatic recurrent non-shared micro-lesions Somatic recurrent and shared vs. somatic non-recurrent non-shared micro-lesions

The combination of micro-lesions for all genes and the comparisons performed for those were considered as separate from the comparisons in individual genes. Therefore, to allow for multiple hypotheses testing, for the tests performed for all repeat/runs of identical nucleotide sizes, a Bonferroni correction was applied. Therefore, each permuted p-value associated with repeats/RINS was multiplied by 11 (repeat-wise $\alpha = 0.05/11$ or 0.0045), to account for different tests (see above, the test also included the comparisons for recurrent micro-lesions). To allow for multiple hypothesis testing, for the different sizes of repetitive elements performed, a Bonferroni correction was also applied. Therefore, each repeat-wise p-value was multiplied by 6 (overall experiment-wise $\alpha = \frac{0.05}{11*6}$ or 0.00076) and each run-wise p-value was multiplied by 2 (overall experiment-wise $\alpha = \frac{0.05}{11*2}$ or 0.0023).

6.3. Results

6.3.1. Repetitive elements in the studied genes

6.3.1.1. Repeats

Direct, mirror and inverted repeats and C/G-quartets were sought in the extended cDNA sequences of the 17 human tumour suppressor genes. The sizes of the repeats were as follows: ≥6bp, ≥7bp and ≥8bp; the distance between the repeats was set to ≤20bp (ranging from 0 to 20bp). A summary of the number and type of repeats found within the specified parameters in the studied genes is presented in Table 26 and Table 27. In total, 3591 repeats of size ≥6bp, 1179 repeats of size ≥7bp, and 499 repeats of size ≥8bp were found in all genes. These results clearly show that the smaller the minimum repeat size, the more repeats are found (Pearson's ρ=-0.95). At first glance, the number of repeats found varies between the genes (for repeats of size ≥6bp, the number ranged from 22 for the *CDKN2A* gene to 646 for the *ATM* gene).

However, there is very strong correlation between the size of the genes (bp) and the number of repeats (repeat size ≥6bp ρ=0.97, ≥7bp ρ=0.93 and ≥8bp ρ=0.81). These results show that despite the different absolute number of repeats in the different genes, the number of repeats relative to the gene size (bp) remains relatively similar in the different genes. In support of this finding, a recent study (Lawson and Zhang 2008) has revealed that the relative number of simple sequence repeats per sequence distance (i.e. megabase) is very similar between housekeeping and tissue-specific genes. It is evident that the repeats constitute a large proportion of the coding sequence of the genes. Repeats of size ≥6bp (excluding the distance between the two parts constituting a repeat) make up on average ~23% of the total length of the genes (results presented in Table 28 and Table 29). With increasing repeat size, the proportion of the gene length made up of repeats decreases, due to the smaller number of repeats. In other words, smaller repeats take up a larger proportion of the genes, due to the relatively larger number of repeats found. Thus, repeats with a size of ≥7bp constitute only ~9%, while repeats with a size of ≥8bp only encompass ~4% of the total length of the genes (Table 29).

By contrast, C and G quartets on average comprise only ~0.5% of the total length of the genes if the distance between the quartets is excluded and ~0.6% if the distance between the quartets is included. Despite the small proportion of the total gene length, there were 2 genes that had substantially more C/G quartets. These were the *STK11* gene with 46 C/G quartets (~3% of the extended cDNA sequence) and *TP53* with 90 C/G quartets respectively (~4% of the extended cDNA sequence). For the rest of the genes, the number of C/G quartets ranged from 0 to 18 (*TSC2* gene; ~0.77% of the extended cDNA sequence). Even though the C/G quartets encompass a relatively high proportion of the extended gene sequences in the *STK11* and *TP53*, none of the quartets is exonic (i.e. found in the exons of the genes). In fact, out of all 165 C/G quartets in all genes, only 7 were found in the exons of the genes (4.24%) and 6 out of those 7 exonic C/G quartets were found in the *TSC2* gene (~86%).

There was no apparent correlation between repeat length and the distance between the repeats (as shown in Figure 44). One apparent feature of all repeat sizes and repeat types is the most frequent distance within the repeats, which is 0bp. The two parts of these repeats abut each other.

6.3.1.2. Runs of identical nucleotides

In total, 2949 runs of mononucleotides of size ≥4bp in all genes were identified. These mononucleotides represent ~10% (ranging from 7.13% to 14.11%) of the total length of the genes (as shown in Table 30). Even though the number of mononucleotide runs varies between the different genes, there is a very strong correlation between the numbers of mononucleotide runs and the gene sizes in bp (Pearson's ρ=0.97). These results show that despite the different absolute number of mononucleotide runs in the different genes, the number of mononucleotide runs relative to the gene size (bp) remains constant. This finding is similar to the uniform distribution of repeats (number of repeats relative to the gene size, measured in bp) in these 17 human tumour suppressor genes (see 6.3.1.1). In support, it is considered that 'single-amino-acid tandem repeats' (repeated amino acids encoded by a single nucleotide- AAA AAA AAA) are abundant in mammalian proteins (Mularoni et al. 2007). In addition, Mularoni et al. (2007) have shown that proteins under strong selective constraints (i.e. highly conserved proteins) contain surprisingly high numbers of repeats. The most frequent mononucleotide runs are of size 4-7bp (Figure 45) and a negative correlation was observed between the frequency and size of the mononucleotide runs (Pearson's ρ=-0.47). Most frequent are runs of As and Ts. On average, for mononucleotides in all genes, runs of As account for 3.39% of the total gene length, whereas runs of Ts comprise 4.62%. On the other hand, runs of Cs and Gs are 3-4 times less frequently observed. Runs of Cs on average comprise 1.26% of the total gene length, whereas runs of Gs account for 1.04%.

6.3.2. Micro-lesions and repetitive elements

6.3.2.1. Micro-lesions and repeats

The positional co-localisation or co-occurrence of micro-lesions (*viz.* micro-deletions, micro-insertions and micro-indels) and repeats (*viz.* direct, inverted and mirror repeats and C/G quartets) with varying sizes (≥6bp, ≥7bp, ≥8bp, ≥6±5bp, ≥7±5bp, ≥8±5bp) and distance between comprising the repeats (ranging from 0 to 20bp) were analysed. The distribution of the micro-lesions with respect to their occurrence in repeats is detailed in Table 32. For all genes, on average ~35% of the micro-lesions were found in the vicinity of repeats of size ≥6bp. Interestingly, almost half (46.36%, Table 32) of the mutations were found to be in the vicinity of repeats when repeats of size ≥6±5bp were considered. In addition, there is an inverse relationship between the number of mutations and the minimum size of the repeats (Pearsons ρ=-0.97). Thus, increasing the minimum repeat size (e.g. ≥6bp vs. ≥7bp), results in a lower number of mutations found in the vicinity of repeats. This shows that more mutations are found in the vicinity of repeats when a smaller minimum repeat size is chosen.

Nevertheless, there is also a positive correlation between the number of repeats in a particular repeat size range (i.e. ≥6bp, ≥7bp and ≥8bp) and the number of mutations found in the vicinity of the repeats (Pearson's ρ=0.99). This implies that the proportions of mutations found in the vicinity of repeats is very similar to the total number of repeats, with respect to different repeat sizes. In addition, as shown in 6.3.1.1, the number of repeats decreases with increasing minimum repeat size. Thus, mutations are found less frequently in the vicinity of repeats when the minimum repeat sizes are increased. Nevertheless, the number of mutations found in the vicinity of repeats is proportional to the number of repeats. Interestingly, the proportion of mutations found in the vicinity of repeats in this analysis is relatively low as compared to reports by other authors. Ball et al. (2005) reported that 92% of the studied micro-deletions and micro-insertions co-localised with different types of repeats, namely direct, inverted, mirror and inversions of inverted repeats. It has to be said that Ball et al. (2005) studied >400 genes (micro-deletions) and >300 genes (micro-insertions). Similarly, in another study by Cooper and Krawczak (1993), all micro-deletions were found to be flanked or to lie within direct repeats. It should be noted that in contrast to this analysis, these studies have additionally searched repeats of sizes <6bp (i.e. 2, 3, 4 and 5bp), even although Cooper and Krawczak (1993) reported that direct repeats of size 2 and 3bp are underrepresented.

Thus, it could be speculated, as these repeats (i.e. ≤5bp) are underrepresented, that they might not mediate the occurrence of micro-lesions and their observation is due to chance occurrence. Alternatively, this particular dataset of 17 human tumour suppressor genes might not follow the same pattern observed in previous studies.

6.3.2.2. Micro-lesions and runs of identical nucleotides

The distribution of micro-lesions (*viz.* micro-deletions and micro-insertions) with respect to their occurrence in runs of mononucleotides was analysed. Micro-indels were excluded from this analysis as they are potentially mediated by different mechanisms than the micro-deletions and micro-insertions (Chuzhanova et al. 2003). The minimum size of the runs of mononucleotides was set to be ≥4bp. In addition, the numbers of mutations that occurred within regions comprising runs of mononucleotides themselves and ±5bp of flanking regions were also analysed. On average, ~11% (ranging from ~2.9% for the *VHL* gene to ~24.7% for the *BRCA2* gene) of the micro-lesions (*viz.* micro-deletions and micro-insertions) were found to be in the vicinity of runs of mononucleotides (Table 43). In addition, when runs of mononucleotides with ±5bp of flanking regions were also considered, the number of micro-lesions found in the vicinity of runs of identical nucleotides increased to ~24% (ranging from 6.8% for the *VHL* gene to ~40% for the *BRCA2* gene). Different types of micro-lesions (*viz.* somatic, germline and shared) for each genes were also analysed with respect to runs of identical nucleotides.

6.3.2.3. Somatic vs. simulated micro-lesions

For all repeat sizes, only 4 genes (*CDKN2A*, *PTEN*, *TP53* and *VHL*) exhibited statistically significant results, when somatic were compared to simulated micro-lesions. A summary of all significant results for somatic vs. simulated micro-lesions is presented in Table 34. There were many more somatic mutations (proportions of mutations found within or in the vicinity of repeats ranged from ~44% to ~67%) found in the vicinity of repeats (repeat sizes of ≥6bp, ≥7bp, ≥6±5bp and ≥7±5bp) for the *CDKN2A* gene, when compared to simulated (proportions of simulated mutations found within or in the vicinity of repeats ranged from ~23% to ~37%) micro-lesions (p_G ranged from 0.048 to 0.0012; p_E=0.02 for repeat size ≥6±5bp). The *PTEN* gene only showed significantly more somatic micro-lesions in the vicinity of repeats for repeat size ≥6±5bp (p_G=0.0096), with 54% and 36% for the somatic and simulated micro-lesions respectively. Interestingly, the *TP53* and *VHL* genes exhibited a significantly lower number of somatic mutations (~13% and ~3% found in the vicinity of repeats for the *TP53* and *VHL* respectively) in the vicinity of repeats (repeat size ≥7±5bp for both genes, p_G=0.0096 and 0.0114 for *TP53* and *VHL* respectively), as compared to simulated mutations (~20% and ~13% found in the vicinity of repeats for the *TP53* and *VHL* respectively).

Nevertheless, significantly more somatic mutations were found in the vicinity of repeats when somatic mutations were combined for all genes and compared to the simulated spectra (p_E=0.035 for repeat size of ≥6±5bp), with ~45% and 39% of the somatic and simulated micro-lesions found in the vicinity of repeats respectively. It is of note that only the repeat size of ≥6±5bp exhibited a significant result. In addition, none of the remainder of the comparisons had enough statistical power to detect an experiment-wise significant result.

Additionally, the somatic micro-lesions in the *APC*, *NF2*, *PTEN*, *TP53* and *VHL* genes comprised ~86% of all somatic mutations in all genes, with the somatic mutations in the *TP53* representing ~44% of the somatic mutations in all genes. Therefore, it is evident that there is a strong association of somatic micro-lesions and repeats in some genes (i.e. *APC* and *CDKN2A*), but for others (i.e. *TP53* and *VHL*) micro-lesions were less likely to be found in repeats than simulated mutations.

With respect to RINS, only the *APC* gene exhibited statistically significant result (RINS size ≥4±5bp; p_G=0.0158). Significantly more somatic micro-lesions (~40%) were noted in the vicinity of RINS, when compared to simulated mutations (~24%).

Therefore, these results suggest that in the case of somatic mutations the involvement of repeats is much more significant than runs of identical nucleotides.

6.3.2.4. Germline vs. simulated micro-lesions

Only the *BRCA2* gene exhibited a statistically significant result (repeat size ≥6±5bp, p_E=0.0204). Many more germline mutations (~56%) were found in the vicinity of repeats than simulated mutations (~42%). A summary of all significant results for germline vs. simulated micro-lesions is presented in Table 35. As with the somatic mutations, the germline micro-lesions combined for all genes exhibited many more mutations (~48%) in the vicinity of repeats (repeat size ≥6±5bp; p_E=0.00000005) than simulated mutations (~39%).

Furthermore, the germline micro-lesions in the *APC*, *BRCA1*, *BRCA2* and *NF1* genes comprised ~60% of all germline micro-lesions. As a result, it is likely that germline micro-lesions are strongly associated with repeats, not only in the *BRCA2* gene, but quite possibly to some degree in others, such as *APC*, *BRCA1* and *NF1*.

The *BRCA1* and *BRCA2* genes showed significantly more germline micro-lesions (~24% for RINS size in *BRCA1* ≥4bp; ~35% and ~44% for RINS sizes ≥4bp and ≥4±5bp in the *BRCA2*) found both within and in the vicinity of RINS (p_G=0.0206 RINS size ≥4±5bp *BRCA1* gene; p_G=0.013 and p_G=0.0072 for the *BRCA2* gene for RINS sizes ≥4bp and ≥4±5bp respectively), as compared to simulated micro-lesions (~14% and for RINS size in *BRCA1* ≥4bp; ~24% and ~32% for RINS sizes ≥4bp and ≥4±5bp n the *BRCA2*). Furthermore, the combination of germline micro-lesions for all genes showed significantly more micro-lesions (~12% and ~29% for RINS sizes ≥4bp and ≥4±5bp respectively) found in the vicinity of RINS (p_G=0.0474 and p_E=0.000031 for RINS sizes ≥4bp and ≥4±5bp respectively), as compared to simulated micro-lesions (~10% and ~24% for RINS sizes ≥4bp and ≥4±5bp respectively). Therefore, RINS are very likely to play a significant role in the positional occurrence of germline micro-lesions in at least two genes (i.e. *BRCA1* and *BRCA2*), but they are also likely to play some part in other genes as well.

6.3.2.5. Shared vs. simulated micro-lesions

With respect to repeats, no statistically significant difference was noted in any of the tests performed, although none of the comparisons had enough statistical power. It is very likely that lack of power was due to the paucity of shared mutations. Indeed, on average, shared mutations comprised only ~1.8% of the observed micro-lesions (*viz.* somatic, germline and shared).

No individual gene exhibited a significant difference between the proportions of shared and simulated micro-lesions, with respect to occurrence within or in the vicinity of RINS. It should be noted that there was not enough statistical power for any of the comparisons performed, with respect to RINS. Nevertheless, the combination of shared micro-lesions for all genes exhibited significantly (p_E=0.0223) more micro-lesions within RINS (~27%) as compared to simulated mutations (~7%). A summary of significant results for shared vs. simulated micro-lesions is presented in Table 36.

Thus, it may be concluded that the positional occurrence of shared micro-lesions is significantly influenced by runs of mononucleotides.

6.3.2.6. Somatic vs. germline micro-lesions

Only the *APC* gene exhibited a statistically significant result, when somatic were compared to germline micro-lesions. Significantly more somatic micro-lesions (p_E=0.0102 and p_G=0.0088 for RINS sizes ≥4bp and ≥4±5bp respectively) were noted within or in the vicinity of RINS (~19% and ~40% for RINS sizes ≥4bp and ≥4±5bp respectively) as compared to germline micro-lesions (~8% and ~26% for RINS sizes ≥4bp and ≥4±5bp respectively). A summary of all significant results for somatic vs. germline micro-lesions is presented in Table 38.

Interestingly, the combination of germline micro-lesions for all genes was significantly (p_E=0.00162) more likely to be found in the vicinity of RINS (~24%), than somatic micro-lesions (~30%).

Thus, it is evident that relatively more somatic mutations are found in the vicinity of RINS in comparison to germline micro-lesions for the *APC* gene. This notwithstanding, relatively more germline micro-lesions (i.e. the combination of germline micro-lesions for all genes) were found in the vicinity of RINS as compared to somatic micro-lesions. This result suggests that the difference in the distribution of mutations with respect to positions of RINS in the germline and in the soma may be widespread across a number of genes.

6.3.2.7. Somatic vs. shared micro-lesions

The *TP53* gene showed significantly more shared mutations (~50% and ~58% for RINS of sizes ≥4bp and ≥4±5bp respectively) found in the vicinity of RINS (RINS size ≥4bp p_E=0.0102 and RINS ≥4±5bp p_G=0.0416) as compared to somatic micro-lesions (~8% and ~24% for RINS of sizes ≥4bp and ≥4±5bp respectively). A summary of all significant results for somatic vs. shared micro-lesions is presented in Table 39.

Furthermore, the combination of shared micro-lesions showed significantly (p_E=0.0000572 and p_G=0.0471 for RINS of sizes ≥4bp and ≥4±5bp respectively) more mutations (~27% and ~39% for RINS of sizes ≥4bp and ≥4±5bp respectively) found in the vicinity of RINS than somatic micro-lesions (~10% and ~24% for RINS of sizes ≥4bp and ≥4±5bp respectively).

Thus, the shared mutations are much more likely to be found in the vicinity of RINS than the somatic micro-lesions in all genes and in particular the *TP53* gene.

6.3.2.8. Germline vs. shared micro-lesions

The *NF1* gene showed significantly more shared mutations (~67%) found in the vicinity of RINS of size ≥4bp (p_G=0.0238) as compared to somatic micro-lesions (~7%). Furthermore, the combination of shared micro-lesions showed significantly (p_E=0.0049) more mutations (~27%) found in the vicinity of RINS of size ≥4bp than somatic micro-lesions (~12%). A summary of all significant results for germline vs. shared micro-lesions is presented in Table 40.

Thus, the shared mutations are much more likely to be found in the vicinity of RINS than the germline micro-lesions in all genes and in particular the *NF1* gene.

6.3.2.9. Recurrent somatic micro-lesions

The combination of recurrent somatic micro-lesions for all genes exhibited a significantly higher number (~14%) of micro-lesions found in RINS of size ≥4bp (p_G=0.0393), as compared to non-recurrent somatic micro-lesions (~9%). Furthermore, the combination of somatic recurrent and shared micro-lesions for all genes showed many more mutations (~29%) found in RINS of size ≥4bp (p_G=0.0361) than somatic non-recurrent and non-shared micro-lesions (~9%). A summary of significant results for recurrent somatic micro-lesions is presented in Table 41.

Therefore, somatic micro-lesions recur in runs of identical nucleotides, but those are also shared between the germline and the soma.

6.3.2.10. Observed vs. simulated

The results for the combination of micro-lesions in the individual 17 genes were very much dependent on the number of somatic, germline and shared mutations. Thus, the mutational spectra in these genes could be separated into several groups: predominantly somatic micro-lesions (*CDKN2A*, *NF2*, *PTEN*, *TP53* and *VHL*); predominantly germline micro-lesions (*APC*, *ATM*, *BRCA1*, *BRCA2*, *NF1*, *PTCH*, *RB1*, *STK11*, *TSC1*, *TSC2* and *WT1*); similar proportions of somatic and germline micro-lesions (*CDH1*). The results for the combination of observed micro-lesions exhibited very similar patterns (e.g. direction of results and statistical significance) to the comparisons of the largest proportion of mutations in the individual genes. A summary of all significant results for observed vs. simulated micro-lesions is presented in Table 37.

6.4. Discussion

Numerous studies have reported the non-random occurrence of mutations. Indeed, sequence context has been shown to influence the specificity of insertions and deletions (Kunkel 1990; Ripley 1990). In addition, various studies have shown the relative importance of repetitive elements in the process of mediating endogenous mechanisms of mutagenesis (Ball et al. 2005; Chuzhanova et al. 2003; Cooper and Krawczak 1993; Greenblatt et al. 1996; Kondrashov and Rogozin 2004). The analysis in this chapter was designed to investigate the contribution of the local sequence environment (i.e. repetitive elements) in 17 human tumour suppressor genes to the associated micro-lesion spectra in the germline and the soma.

The extended cDNA sequences of the genes were searched for repetitive elements. This was accomplished by using a custom built novel computer algorithm. It allowed the identification and localisation of various types of repetitive element. These repetitive elements included direct repeats; inverted repeats; mirror repeats; C/G-quartets; and runs of identical nucleotides (RINS). The sizes of the repeats were ≥6bp, ≥7bp, ≥8bp, ≥6±5bp, ≥7±5bp, ≥8±5bp (maximum distance within the repeat ≤20bp) and the sizes of the runs of identical nucleotides were ≥4bp and ≥4±5bp. The repeats and runs of identical nucleotides were analysed separately so as to avoid overlap, but also to allow recognition of potential differences. The analysis showed that on average ~23% of the studied genes comprise repeats (repeat size ≥6bp). On the other hand, runs of identical nucleotides on average comprise ~10% of the total length (bp) of the studied tumour suppressor genes. These results are broadly consistent with those of a previous study which showed that ~9% of >22000 human genes studied comprise simple sequence repeats (di-, tetra-, penta-simple sequence repeats of <8 repeated units; Loire et al. 2009). Thus, the repetitive elements examined comprise a relatively large proportion of the studied genes. The analysis shows that the abundance of repetitive elements is related to gene size. Thus, the bigger the gene, the more repetitive elements it should have, but the relative length of the repetitive elements, with respect to individual gene sizes, is relatively similar between the different genes (~23% for repeats of size ≥6bp, ranging from ~17% for the *NF2* gene to ~26% for the *TP53* gene; ~10% for runs of identical nucleotides ≥4bp, ranging from ~5% for the *VHL* gene to ~14% for the *PTEN* gene). Micro-lesions were analysed with respect to the positions of the repetitive elements. An initial exploratory analysis revealed that a relatively large proportion of the micro-lesions could be accounted for by repetitive elements. It was discovered that on average for all genes,

~35% of micro-lesions are found in the vicinity of repeats (repeat size ≥6bp), whereas a much smaller proportion of the micro-lesions was accounted for by runs of identical nucleotides (~11%). These proportions increased considerably when micro-lesions were analysed ±5bp away from repeats and runs of identical nucleotides (~46 and ~24% for the repeats and runs of identical nucleotides respectively). In contrast to other studies, the proportions of micro-lesions found in the vicinity of repeats in this study, appear relatively small. Ball et al. (2005) reported that 92% of all studied micro-deletions and micro-insertions could be accounted for by various types of repeats. It has to be noted however that Ball et al. studied >300 genes. Therefore, it is likely that a smaller proportion of micro-lesions in the 17 human tumour suppressor genes studied are accounted for by repetitive elements. Similarly, in another study Cooper and Krawczak (1993) suggest that all studied micro-lesions were flanked or resided within repeats. Cooper and Krawczak (1993) however noted an under-representation of repeats of smaller sizes (i.e. 2 and 3bp) in association with the occurrence of micro-lesions. The findings presented here suggest that the lower the repeat size, the larger the number of repeats that will be found in the studied genes. Several studies have shown that the frequency of micro-lesions increases with the size of repetitive elements (Greenblatt et al. 1996; Kondrashov and Rogozin 2004; Vogler et al. 2006). Therefore, selecting a minimum repeat size of 6bp and a minimum mononucleotide run of 4bp could potentially explain the relatively smaller proportion of micro-lesions found within, or in the vicinity of, repetitive elements.

Taking into consideration the relatively large proportion of repetitive elements found in the studied 17 human tumour suppressor genes, it is inevitable that some of the micro-lesions would coincide with repetitive elements just by chance alone. Thus, finding repetitive elements in the vicinity of micro-lesions would not necessarily indicate that the micro-lesions were caused/mediated by repetitive elements. The approach applied in this analysis allows one to deduce whether the positional occurrence of micro-lesions is due simply to chance or whether micro-lesions indeed co-localise with repetitive elements (micro-lesions found in the vicinity of repetitive elements). In addition, the analysis was performed on somatic and germline micro-lesion spectra, thus allowing inference, but also a comparison and a contrast of the potential involvement of endogenous mechanisms of mutagenesis in both the soma and the germline.

The presented results showed that in the *CDKN2A* and the *PTEN* genes, many more somatic micro-lesions were found to be in the vicinity of repeats. Thus, it is very likely that endogenous mutagenesis, mediated by repeats, significantly influences the somatic micro-lesion spectrum in the *CDKN2A* and *PTEN* genes. Quite the opposite was found for the *TP53* and the *VHL* genes, with significantly fewer somatic mutations being found in the vicinity of repeats. This latter finding contrasts with various studies, which have shown that the majority of the micro-lesions in the *TP53* gene could be explained by the 'slipped mispairing' mechanism (Greenblatt et al. 1996; Tang et al. 2001). Therefore, it appears that mechanisms not involving repetitive elements are more likely to play an important part in shaping the somatic micro-lesion mutational spectrum in the *TP53* and *VHL* genes. These mechanisms could be exogenous in origin and hence could involve environmental carcinogens, such as reactive oxygen species, exposure to tobacco smoke, aflatoxin B1, environmental factors such as UV light and ionizing radiation (Barbour et al. 2006). It is noteworthy that the *TP53* extended cDNA gene sequence comprises ~4% C/G-quartets. All of these C/G-quartets were located within the intronic parts of the extended cDNA sequence. Furthermore, none of the micro-lesions were located in or within these C/G-quartets. Thus, it is likely that these repetitive elements could slightly increase the frequency of simulated micro-lesions and contribute to the overall results, at least for the *TP53* gene. In addition, tissue specificity has been reported to be a property of mutational spectra in the *TP53* gene (Glazko et al. 2004) and this specificity will be ignored when the micro-lesion spectrum is analysed as a whole.

Tissue specificity has been also shown in spontaneous micro-deletions and micro-insertions in the Big Blue transgenic mouse mutation detection system (*lacI* gene; Halangoda et al. 2001). The spectrum of these micro-lesions (i.e. pattern and size of distribution of micro-lesions) was found to be very similar to that of the *TP53* gene and the majority of these occur within mononucleotide runs. Thus, tumours with a different tissue origin may well have a different ratio of micro-lesions caused by environmental carcinogens and endogenous mechanisms. Furthermore, a recent study (Scaringe et al. 2008) has proposed a 'Tarzan' model of mutagenesis (see Figure 46). This model is reminiscent of the translesional synthesis mechanism, in that a large DNA adduct blocks the advancing replication. It suggests that the helicase unwinds nucleotides on the nascent strand, thereby allowing the translesion polymerase to synthesize additional nucleotides on the nascent strand. These additional nucleotides may be copied from either the nascent strand or the template strand and could serve to by-pass the DNA adduct. As a result some nucleotides will be incorrectly missed and others incorrectly added, giving rise to a micro-indel. Even though this mechanism only explains micro-indels, similar or other endogenous mechanisms [e.g. variety of error-prone polymerases, relaxed version of 'slipped mispairing' model (Kondrashov and

Rogozin 2004) or carcinogens may influence the mutational spectra in these two particular genes (i.e. *TP53* and *VHL*)].

The germline micro-lesions showed preferential co-localisation both within (and in the vicinity of) repetitive elements (repeats and RINS), as compared to simulated mutations. This was the case for the *BRCA2* gene and for the combination of germline micro-lesions for all genes. It is clear that the occurrence of germline micro-lesions is very likely to be influenced by endogenous mechanisms of mutagenesis, mediated by repetitive elements.

Despite the fact that the somatic micro-lesions in the *TP53* and *VHL* genes were less likely to be found in repetitive elements, shared micro-lesions (the combination of shared micro-lesions for all genes) were generally found to occur preferentially within RINS. As a result, micro-lesions found in both the soma and the germline were associated with runs of mononucleotides. Furthermore, the shared mutations in the *TP53* and *VHL* genes comprised ~34% of the shared micro-lesions in all genes. Therefore, runs of mononucleotides were very likely to play an important role in somatic and germline mutagenesis, as shared mutations are found in both the soma and the germline. In addition, some similarities between the mechanisms that generate germline and somatic mutations were evident. This finding concurs with a study that shows very similar frequencies between micro-insertions and micro-deletions in the mouse soma and the human germline (Gonzalez et al. 2007). It should be noted that frequency analysis can only indicate similarities or differences in the underlying mechanisms of mutagenesis.

Our results clearly demonstrate the relative importance of repeats in the process of mutagenesis, but also reveal the similarities in the underlying mechanisms between the somatic and germline mutational spectra.

Shared micro-lesions were also found to be more likely to be co-localised within RINS than the somatic *TP53* micro-lesions. Furthermore, shared mutations were preferentially found within RINS as compared to both the somatic and the germline mutations combined for all genes. As a result, mutational mechanisms appear to be shared between the germline and the soma. Additionally, a significant part of these shared mutational mechanisms were mediated through runs of identical nucleotides. Hence, it would appear that endogenous mutagenesis is an important factor in influencing the positional occurrence of both the somatic and germline micro-lesions.

The results presented herein, also indicated that somatic recurrent micro-lesions are more likely to be found within RINS than non-recurrent mutations. One could speculate that recurrent somatic mutations are more likely to be involved in tumour development than non-recurrent ones, due to the multiple independent observations. This would suggest that those micro-lesions mediated by RINS are more likely to play a role in tumour development than those micro-lesions that are not mediated by RINS. Bacolla and Wells (2009) have argued that recurrent mutations whose genes are involved in tumour development are more likely to contain repetitive elements. These repetitive elements have been shown to be associated with specific functions of the genes. Thus, C/G-quartets are involved in transcriptional initiation (Du et al. 2008; Huppert and Balasubramanian 2007), and RINS and simple sequence repeats are predominantly found in genes responsible for regulation of transcription and various cellular activities (Alba and Guigo 2004; Faux et al. 2005; Karlin et al. 2002). Furthermore, homopolymeric runs (e.g. runs of glutamic acid, alanine and leucine) have been associated with proteins responsible for DNA-binding, as well as transmembrane receptors and transcription factors (Alba and Guigo 2004). A few of the studied 17 human tumour suppressor genes are involved in DNA-binding (*TP53* and *WT1*); protein-protein and protein-DNA interactions (*BRCA2*); transcriptional activation and regulation (*BRCA1*, *APC* and *TP53*); interaction with cell-surface proteins (*NF2*); transcription factors (*TP53* and *WT1*) and cell receptors (*PTCH*) (Futreal et al. 2004; Knudson 2002; Sherr 2004; Vogelstein and Kinzler 2004). As a result, repetitive elements are likely to be involved in important functions of the genes, but nevertheless are also likely to be responsible or involved in the process of mutagenesis.

It was also found that not only were somatic recurrent mutations more likely to be found in RINS as compared to non-recurrent mutations, but they were also more likely to be found in the germline. Thus, somatic micro-lesions do not only recur in repetitive elements, such as RINS, but somatic recurrent micro-lesions that are also found in the germline have an even higher proportion of micro-lesions found in repetitive elements (i.e. RINS) . Therefore, this indicates that RINS are likely to be a shared mutational hotspot for both the soma and the germline on the basis that they recurred in the soma, but were also noted in the germline. Even although mutational mechanisms involving RINS appear to be shared between the germline and the soma, some differences were also noted. When micro-lesions were analysed with respect to positions of runs of identical nucleotides, a number of interesting findings were discovered. The *APC* gene exhibited many more somatic mutations within (or in the vicinity of) runs of identical nucleotides than the germline micro-lesions. This observation could perhaps be due to impaired mismatch repair. Patients with hereditary non-polyposis colorectal cancer and impaired mismatch repair genes exhibit a substantial excess of frameshift mutations (predominantly 1bp deletions; Huang et al. 1996) with a significant proportion (49%) of those are found in polyA runs. In addition, both sporadic and inherited gastrointestinal cancer associated with somatic or germline mutations in mismatch repair genes (*hMSH2* and *hMLH1*) display increased somatic frameshifts in polyC and polyA runs (Ohmiya et al. 2001). Inactivation of mismatch repair genes (*MSH-2* and *MSH-6*) in *C. elegans* suggests that spontaneous mutagenesis is increased in both the germline and the soma (Tijsterman et al. 2002). On the other hand, a closer look into the somatic and germline micro-lesions revealed that the distribution of 1bp deletions did not differ between the soma and the germline (~50% for both the somatic and germline micro-lesions). Hence, it is unlikely that impairment of mismatch repair could explain the difference between germline and somatic micro-lesions with respect to co-localisation within RINS, although this cannot be ruled out. An alternative mechanism that could potentially explain the difference between the somatic and germline micro-lesions, with respect to RINS, is a potential age-related shift in the efficacy of DNA-repair mechanisms. Indeed, such an age-related shift has been reported in unilateral sporadic vestibular schwannoma (Evans et al. 2005). Thus, in older patients, a higher proportion of frameshifts has been observed as compared to younger patients. Evans et al. (2005) have argued that this is most likely due to reduced mutation repair efficiency than an increased rate of mutagenesis.

By contrast, the combined germline mutations for all genes were more likely to be found in the vicinity of RINS as compared to somatic micro-lesions. It would appear that the occurrence of germline micro-lesions is more likely to be influenced by RINS than somatic mutations. One potential explanation could be that the proportion of somatic micro-lesions consequent to exogenous mutagenesis is higher as compared to the germline. Indeed, chemical carcinogens have been shown to be able to induce frameshifts (Lambert et al. 1992; Ripley 1990). Alternatively, the spectrum of somatic micro-lesions could be the result of more complex endogenous mutational mechanisms. These include the formation of quasi-palindromic loops, palindromic dyads and imperfect repetitive elements (Greenblatt et al. 1996). On the other hand, the combination of germline micro-lesions for all genes were more likely to be found in the vicinity in both repeats and RINS, where the combination of somatic micro-lesions were preferentially found only in the vicinity of repeats. Thus, defects in mutation-repair mechanisms (i.e. mismatch repair, non-homologous end joining, base-excision repair, etc. (Evans et al. 2005) are likely to influence the difference between the soma and germline with respect to RINS. A potentially impaired mismatch repair gene in the soma, could perhaps slightly increase the frequency of micro-lesions associated with repeats, but not runs of mononucleotides. Indeed, the proportion of somatic micro-lesions found within (or in the vicinity of) repeats was in almost all cases (all repeat sizes except ≥6bp) higher than the corresponding micro-lesions found in the germline. It is notable that none of the comparisons of somatic and germline micro-lesions exhibited statistically significant results, although insufficient statistical power to detect a difference could have contributed to the observed results.

It is evident that repetitive elements (i.e. repeats and runs of identical nucleotides) play an important role in the process of mutagenesis in these 17 tumour suppressor genes, when the mutational micro-lesion spectra are analysed as a whole. When analysed separately, the germline and the soma exhibit great similarities but also differences. Thus, it is important to distinguish between the two types of micro-lesions when analysing mechanisms that influence the process of mutagenesis. This finding appears not to concur with studies that have shown the importance of runs of identical nucleotides in the process of mutagenesis. In fact, most studies do not distinguish between somatic and germline micro-lesions, reporting only the effect of the runs of mononucleotides on the combined spectrum of germline and somatic mutations. In support, our results show that the combination of germline mutations for all genes reveal many more mutations found in the vicinity of runs of identical nucleotides and in the vicinity of repeats, as is the case with the combination of observed mutations for all genes. Furthermore, the combination of somatic micro-lesions revealed many more mutations found in the vicinity of repeats which was also noted in the combination of observed mutations for all genes.

The fact that most of the significant results are predominantly observed with shorter repeat sizes (e.g. minimum repeat size of 7bp would not include repeats of size 6bp) indicates that mutations are predominantly associated with short repeats rather than long repeats. Then again, there are more repeats with shorter sizes than longer ones. Thus, the difference in the distribution of mutations as compared to chance alone would be relatively small and would not yield statistically significant results. It is also intriguing that a relatively stronger association was observed of mutations found in the vicinity of repeats (i.e. ±5bp) in comparison to mutations found within repeats (e.g. a micro-lesions overlaps with a repeat). This is potentially quite interesting, as the 'slipped mispairing mechanism' would not readily explain this observation. Thus, it is possible that a proportion of the mutations mediated by repeats might be due to a mechanism that is similar to, but nevertheless distinct from, 'slipped mispairing'. A literature search did not yield mechanisms that could explain how repeats in close proximity to mutations could mediate mutations, although one report noted that most frameshift mutations in the human ubiquitin-B (*UBB*) and amyloid precursor protein (*APP*) genes were in close proximity to short simple repeats, but these were attributed to molecular misreading at the mRNA level (van Den Hurk et al. 2001). Then again, distant direct repeats (distance between repeats >20bp) could be brought closer together by flanking inverted repeats; thus, the involvement of the 'slipped mispairing' mechanism could not be ruled out in the cases of mutations found in the vicinity of repeats (i.e. ±5bp).

Table 26 Number and type of repeats found in the studied tumour suppressor genes (C and G quartets excluded)

Gene	Gene size (bp)	Number of direct repeats			Number of inverted repeats			Number of mirror repeats			all repeats		
		6bp	7bp	8bp	6bp	7bp	8bp	6bp	7bp	8bp	6bp	7bp	8bp
APC	11082	115	35	15	89	31	9	95	31	12	299	97	36
ATM	19711	269	85	32	199	61	22	178	52	23	646	198	77
BRCA1	9332	107	38	20	45	13	3	78	33	13	230	84	36
BRCA2	14677	146	45	12	126	29	13	144	34	10	416	108	35
CDH1	5369	58	15	4	31	5	1	33	11	3	122	31	8
CDKN2A	981	11	4	2	3	0	0	8	3	1	22	7	3
NF1	18147	210	86	48	125	41	13	163	58	27	498	185	88
NF2	4508	31	9	3	23	7	0	25	6	3	79	22	6
PTCH	8254	89	23	9	37	15	5	58	22	10	184	60	24
PTEN	2742	33	17	11	16	4	0	33	13	8	82	34	19
RB1	7377	123	51	40	74	22	9	94	44	19	291	117	68
STK11	2832	41	10	4	17	4	0	22	9	1	80	23	5
TP53	2882	50	27	21	12	3	1	22	6	1	84	36	23
TSC1	7065	94	45	29	25	8	0	50	16	7	169	69	36
TSC2	12394	111	32	9	74	17	5	93	23	8	278	72	22
VHL	1152	18	8	2	4	0	0	8	1	1	30	9	3
WT1	3050	33	15	7	23	6	1	25	6	2	81	27	10
Total	131555	1539	545	268	923	266	82	1129	368	149	3591	1179	499

Table 27 Number and type of G and C quartets found in the studied tumour suppressor genes

Gene	Number C quartets			Number G quartets			Number total quartets			Number Total quartets	
	C(3)	C(4)	C(5)	G(3)	CG(4)	G(5)	C quartets	G quartets	C/G quartets	In introns	In exons
APC	0	0	0	0	0	0	0	0	0	0	0
ATM	0	0	0	0	0	0	0	0	0	0	0
BRCA1	0	0	0	0	0	0	0	0	0	0	0
BRCA2	0	0	0	0	0	0	0	0	0	0	0
CDH1	0	0	0	0	0	0	0	0	0	0	0
CDKN2A	0	0	0	0	0	0	0	0	0	0	0
NF1	4	0	0	0	0	0	4	0	4	4	0
NF2	0	0	0	0	0	0	0	0	0	0	0
PTCH	3	0	0	0	0	0	3	0	3	3	0
PTEN	0	0	0	0	0	0	0	0	0	0	0
RB1	0	0	0	0	0	0	0	0	0	0	0
STK11	0	0	0	46	0	0	0	46	46	46	0
TP53	80	8	0	2	0	0	88	2	90	90	0
TSC1	4	0	0	0	0	0	4	0	4	3	1
TSC2	6	0	0	12	0	0	6	12	18	12	6
VHL	0	0	0	0	0	0	0	0	0	0	0
WT1	0	0	0	0	0	0	0	0	0	0	0
Total	97	8	0	60	0	0	105	60	165	158	7

Table 28 Length of repeats (including the distance between the repeats)

Repeats		**≥6bp**		**≥7bp**		**≥8bp**		**≥6±5bp**		**≥7±5bp**		**≥8±5bp**		**C quartets**		**G quartets**		**C quartets ±5bp**		**G quartets ±5bp**	
Gene	**Gene size [bp]**	**[bp]**	**%**	**[bp]**	**%**	**[bp]**	**%**	**[bp]**	**%**	**[bp]**	**%**	**[bp]**	**%**	**[bp]**	**%**	**[bp]**	**%**	**[bp]**	**%**	**[bp]**	**%**
APC	11082	4113	37.11	1681	15.17	749	6.76	5408	48.80	2303	20.78	1032	9.31	0	0.00	0	0.00	0	0.00	0	0.00
ATM	19711	7359	37.33	2917	14.80	1189	6.03	9616	48.78	3993	20.26	1661	8.43	0	0.00	0	0.00	0	0.00	0	0.00
BRCA1	9332	2986	32.00	1153	12.36	478	5.12	3968	42.52	1558	16.70	648	6.94	0	0.00	0	0.00	0	0.00	0	0.00
BRCA2	14677	5970	40.68	2167	14.76	839	5.72	7691	52.40	2937	20.01	1140	7.77	0	0.00	0	0.00	0	0.00	0	0.00
CDH1	5369	1582	29.47	471	8.77	131	2.44	2125	39.58	661	12.31	191	3.56	0	0.00	0	0.00	0	0.00	0	0.00
CDKN2A	981	339	34.56	181	18.45	76	7.75	411	41.90	240	24.46	96	9.79	0	0.00	0	0.00	0	0.00	0	0.00
NF1	18147	6261	34.50	2593	14.29	1146	6.32	8282	45.64	3575	19.70	1556	8.57	36	0.20	0	0.00	56	0.31	0	0.00
NF2	4508	1255	27.84	399	8.85	109	2.42	1684	37.36	562	12.47	159	3.53	0	0.00	0	0.00	0	0.00	0	0.00
PTCH	8254	2311	28.00	810	9.81	292	3.54	3109	37.67	1130	13.69	406	4.92	20	0.24	0	0.00	30	0.36	0	0.00
PTEN	2742	939	34.25	450	16.41	271	9.88	1217	44.38	588	21.44	371	13.53	0	0.00	0	0.00	0	0.00	0	0.00
RB1	7377	3142	42.59	1347	18.26	554	7.51	4027	54.59	1818	24.64	735	9.96	0	0.00	0	0.00	0	0.00	0	0.00
STK11	2832	1104	38.98	497	17.55	137	4.84	1472	51.98	683	24.12	187	6.60	90	3.18	0	0.00	130	4.59	0	0.00
TP53	2882	1075	37.30	513	17.80	210	7.29	1369	47.50	650	22.55	260	9.02	73	2.53	40	1.39	103	3.57	60	2.08
TSC1	7065	2275	32.20	1009	14.28	353	5.00	3011	42.62	1376	19.48	483	6.84	22	0.31	0	0.00	32	0.45	0	0.00
TSC2	12394	4409	35.57	1605	12.95	627	5.06	5768	46.54	2157	17.40	837	6.75	95	0.77	0	0.00	145	1.17	0	0.00
VHL	1152	410	35.59	161	13.98	69	5.99	545	47.31	211	18.32	89	7.73	0	0.00	0	0.00	0	0.00	0	0.00
WT1	3050	1078	35.34	454	14.89	171	5.61	1405	46.07	614	20.13	221	7.25	0	0.00	0	0.00	0	0.00	0	0.00
Total	131555	46608	35.43	18408	13.99	7401	5.63	61108	46.45	25056	19.05	10072	7.66	336	0.26	40	0.03	496	0.38	60	0.05

Table 29 Length of repeats (excluding the distance between the repeats)

min repeat size		≥6bp		≥7bp		≥8bp		≥6±5bp		≥7±5bp		≥8±5bp		C quartets		C quartets ±5bp		G quartets		G quartets ±5bp	
Gene	Gene size [bp]	[bp]	%	[bp]	%	[bp]	%	[bp]	%	[bp]	%	[bp]	%	[bp]	%	[bp]	%	[bp]	%	[bp]	%
APC	11082	2675	24.14	1091	9.84	509	4.59	4274	38.57	1762	15.90	799	7.21	0	0.00	0	0.00	0	0.00	0	0.00
ATM	19711	4862	24.67	2027	10.28	887	4.50	7612	38.62	3185	16.16	1376	6.98	0	0.00	0	0.00	0	0.00	0	0.00
BRCA1	9332	1955	20.95	791	8.48	344	3.69	3119	33.42	1234	13.22	519	5.56	0	0.00	0	0.00	0	0.00	0	0.00
BRCA2	14677	3744	25.51	1348	9.18	534	3.64	5894	40.16	2169	14.78	840	5.72	0	0.00	0	0.00	0	0.00	0	0.00
CDH1	5369	1062	19.78	346	6.44	110	2.05	1702	31.70	553	10.30	170	3.17	0	0.00	0	0.00	0	0.00	0	0.00
CDKN2A	981	243	24.77	131	13.35	67	6.83	334	34.05	195	19.88	87	8.87	0	0.00	0	0.00	0	0.00	0	0.00
NF1	18147	4082	22.49	1744	9.61	794	4.38	6509	35.87	2798	15.42	1217	6.71	26	0.14	0	0.00	46	0.25	0	0.00
NF2	4508	786	17.44	274	6.08	86	1.91	1285	28.50	443	9.83	136	3.02	0	0.00	0	0.00	0	0.00	0	0.00
PTCH	8254	1490	18.05	561	6.80	217	2.63	2408	29.17	898	10.88	331	4.01	14	0.17	0	0.00	24	0.29	0	0.00
PTEN	2742	578	21.08	300	10.94	188	6.86	918	33.48	448	16.34	292	10.65	0	0.00	0	0.00	0	0.00	0	0.00
RB1	7377	2091	28.34	930	12.61	393	5.33	3246	44.00	1450	19.66	589	7.98	0	0.00	0	0.00	0	0.00	0	0.00
STK11	2832	732	25.85	313	11.05	84	2.97	1148	40.54	508	17.94	134	4.73	66	2.33	0	0.00	109	3.85	0	0.00
TP53	2882	751	26.06	371	12.87	171	5.93	1130	39.21	528	18.32	231	8.02	52	1.80	36	1.25	88	3.05	56	1.94
TSC1	7065	1398	19.79	636	9.00	241	3.41	2246	31.79	1034	14.64	371	5.25	14	0.20	0	0.00	24	0.34	0	0.00
TSC2	12394	2793	22.54	997	8.04	412	3.32	4426	35.71	1563	12.61	622	5.02	68	0.55	0	0.00	118	0.95	0	0.00
VHL	1152	282	24.48	121	10.50	55	4.77	452	39.24	187	16.23	90	7.81	0	0.00	0	0.00	0	0.00	0	0.00
WT1	3050	686	22.49	280	9.18	119	3.90	1090	35.74	444	14.56	165	5.41	0	0.00	0	0.00	0	0.00	0	0.00
Total	131555	30210	22.96	12261	9.32	5211	3.96	47793	36.33	19399	14.75	7969	6.06	240	0.18	36	0.03	409	0.31	56	0.04

Figure 44 Distribution of lengths of repeats (direct, inverted and mirror) and the distance between repeats

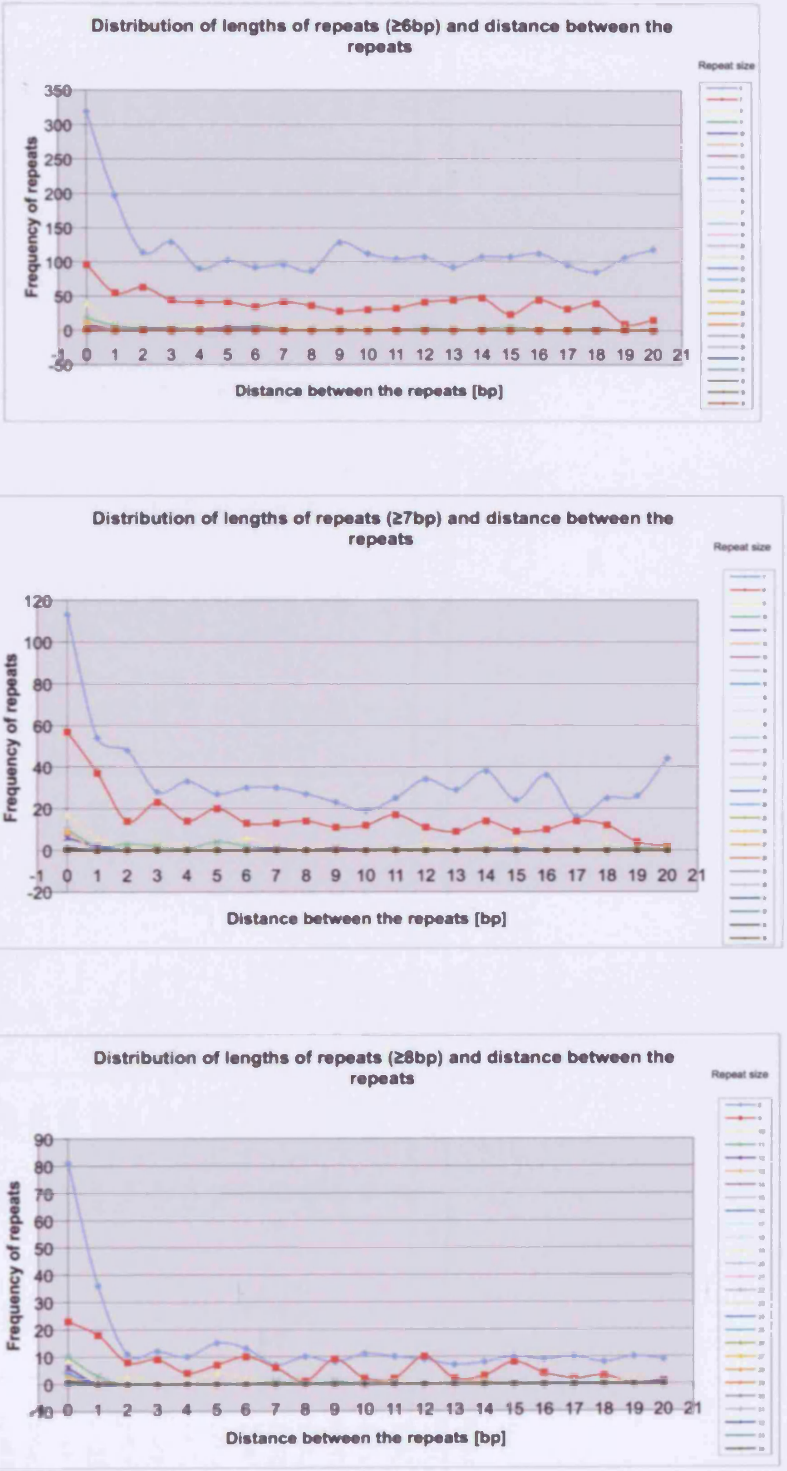


Table 30 Distribution of runs of identical nucleotides in the 17 tumour suppressor genes

		runs of A			runs of C			runs of G			runs of T			Total runs		
	Gene size		Length			Length			Length			Length			Length	
Gene	[bp]	N	[bp]	%	N	[bp]	%	N	[bp]	%	N	[bp]	%	N	[bp]	%
APC	11082	126	571	5.15	16	66	0.60	7	31	0.28	90	401	3.62	239	1069	9.65
ATM	19711	172	767	3.89	14	58	0.29	8	32	0.16	308	1499	7.60	502	2356	11.95
BRCA1	9332	80	356	3.81	22	92	0.99	9	39	0.42	82	375	4.02	193	862	9.24
BRCA2	14677	214	997	6.79	16	65	0.44	6	26	0.18	173	807	5.50	409	1895	12.91
CDH1	5369	20	88	1.64	33	144	2.68	18	73	1.36	41	178	3.32	112	483	9.00
CDKN2A	981	0	0	0.00	4	18	1.83	12	53	5.40	2	8	0.82	18	79	8.05
NF1	18147	151	690	3.80	21	99	0.55	21	86	0.47	239	1164	6.41	432	2039	11.24
NF2	4508	19	80	1.77	12	54	1.20	19	80	1.77	23	115	2.55	73	329	7.30
PTCH	8254	27	124	1.50	40	178	2.16	28	121	1.47	49	215	2.60	144	638	7.73
PTEN	2742	28	129	4.70	1	4	0.15	3	13	0.47	46	241	8.79	78	387	14.11
RB1	7377	81	391	5.30	11	49	0.66	8	34	0.46	108	557	7.55	208	1031	13.98
STK11	2832	4	17	0.60	23	102	3.60	42	182	6.43	2	8	0.28	71	309	10.91
TP53	2882	9	50	1.73	36	157	5.45	16	76	2.64	15	62	2.15	76	345	11.97
TSC1	7065	29	131	1.85	19	85	1.20	10	41	0.58	55	278	3.93	113	535	7.57
TSC2	12394	7	33	0.27	83	351	2.83	100	426	3.44	16	74	0.60	206	884	7.13
VHL	1152	2	10	0.87	4	16	1.39	2	8	0.69	6	27	2.34	14	61	5.30
WT1	3050	7	29	0.95	26	115	3.77	12	51	1.67	16	70	2.30	61	265	8.69
Total	131555	976	4463	3.39	381	1653	1.26	321	1372	1.04	1271	6079	4.62	2949	13567	10.31

N- Number of runs

Figure 45 Distribution of mononucleotides in all 17 tumour suppressor genes

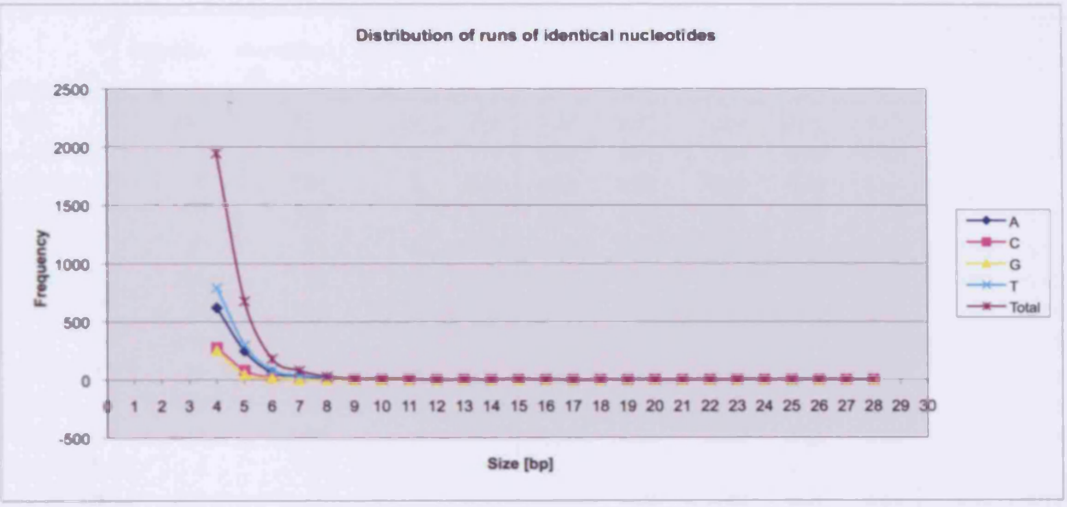


Table 31 Distribution of micro-lesions in the 17 tumour suppressor genes studied

	Number of micro-lesions										
Gene	Somatic F_S	Germline F_G	Shared F_{SH}	Total T	$\frac{F_S}{T}$	$\frac{F_G}{T}$	$\frac{F_{SH}}{T}$	$\frac{F_S}{T_S}$	$\frac{F_G}{T_G}$	$\frac{F_{SH}}{T_{SH}}$	$\frac{T}{T_T}$
APC	184	411	15	610	0.30	0.67	0.02	0.10	0.17	0.19	0.14
ATM	5	171	0	176	0.03	0.97	0.00	0.00	0.07	0.00	0.04
BRCA1	9	350	5	364	0.02	0.96	0.01	0.01	0.14	0.06	0.08
BRCA2	9	342	3	354	0.03	0.97	0.01	0.01	0.14	0.04	0.08
CDH1	15	21	0	36	0.42	0.58	0.00	0.01	0.01	0.00	0.01
CDKN2A	108	18	2	128	0.84	0.14	0.02	0.06	0.01	0.03	0.03
NF1	16	331	3	350	0.05	0.95	0.01	0.01	0.14	0.04	0.08
NF2	210	68	5	283	0.74	0.24	0.02	0.12	0.03	0.06	0.07
PTCH	21	82	0	103	0.20	0.80	0.00	0.01	0.03	0.00	0.02
PTEN	196	44	10	250	0.78	0.18	0.04	0.11	0.02	0.13	0.06
RB1	44	175	5	224	0.20	0.78	0.02	0.02	0.07	0.06	0.05
STK11	4	71	3	78	0.05	0.91	0.04	0.00	0.03	0.04	0.02
TP53	738	14	12	764	0.97	0.02	0.02	0.41	0.01	0.16	0.18
TSC1	1	82	0	83	0.01	0.99	0.00	0.00	0.03	0.00	0.02
TSC2	6	159	0	165	0.04	0.96	0.00	0.00	0.07	0.00	0.04
VHL	210	91	14	315	0.67	0.29	0.04	0.12	0.04	0.18	0.07
WT1	7	13	0	20	0.35	0.65	0.00	0.00	0.01	0.00	0.00
Total	1783 T_S	2443 T_G	77 T_{SH}	4303 T_T							

F_i is the number of micro-lesions, where $i \in \{S,G,SH\}$

T_i is the total number of micro-lesions, where $i \in \{S,G,SH,T\}$, and

S-somatic, G-germline , SH-shared , T-total (somatic, germline and shared)

Values marked in red denote genes that made a relatively large contribution to the corresponding mutational spectrum

Table 32 Number of mutations (micro-deletions, micro-insertions and micro-indels) found in the vicinity of repeats (direct, inverted and mirror repeats and C/G quartets)

		Mutations found in the vicinity of repeats											
	Size of repeats	**≥6 bp**		**≥7 bp**		**≥8 bp**		**≥6±5 bp**		**≥7±5 bp**		**≥8±5 bp**	
Gene	**N**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
APC	610	245	40.16	109	17.87	41	6.72	311	50.98	146	23.93	55	9.02
ATM	176	67	38.07	24	13.64	10	5.68	81	46.02	33	18.75	16	9.09
BRCA1	364	110	30.22	39	10.71	13	3.57	160	43.96	58	15.93	22	6.04
BRCA2	354	153	43.22	39	11.02	17	4.80	199	56.21	65	18.36	24	6.78
CDH1	36	10	27.78	6	16.67	2	5.56	12	33.33	8	22.22	2	5.56
CDKN2A	128	73	57.03	53	41.41	17	13.28	84	65.63	63	49.22	20	15.63
NF1	350	103	29.43	44	12.57	16	4.57	134	38.29	54	15.43	22	6.29
NF2	283	77	27.21	22	7.77	3	1.06	109	38.52	38	13.43	7	2.47
PTCH	103	40	38.83	10	9.71	4	3.88	53	51.46	14	13.59	4	3.88
PTEN	250	107	42.80	43	17.20	17	6.80	133	53.20	55	22.00	29	11.60
RB1	224	82	36.61	36	16.07	20	8.93	107	47.77	52	23.21	30	13.39
STK11	78	26	33.33	16	20.51	4	5.13	38	48.72	21	26.92	5	6.41
TP53	764	236	30.89	76	9.95	34	4.45	316	41.36	96	12.57	43	5.63
TSC1	83	30	36.14	19	22.89	14	16.87	44	53.01	29	34.94	16	19.28
TSC2	165	72	43.64	16	9.70	4	2.42	91	55.15	25	15.15	7	4.24
VHL	315	81	25.71	9	2.86	1	0.32	116	36.83	11	3.49	1	0.32
WT1	20	8	40.00	1	5.00	0	0.00	8	40.00	1	5.00	0	0.00
Total	4303	1520	35.31	562	13.05	217	5.04	1996	46.36	769	17.86	303	7.04

N- Number of micro-lesions

Table 33 Summary of statistically significant results, directionality and power calculations

Comparisons	Repeats or RINS	*ALL*	*APC*	*ATM*	*BRCA1*	*BRCA2*	*CDH1*	*CDKN2A*	*NF1*	*NF2*	*PTCH*	*PTEN*	*RB1*	*STK11*	*TP53*	*TSC1*	*TSC2*	*VHL*	*WT1*
Somatic vs. simulated	6bp							↑											
	7bp							↑											
	8bp																		
	6±5bp	🟨↑						▨↑				↑							
	7±5bp							🟩↑							↓			↓	
	8±5bp																		
	RINS±0bp																		
	RINS±5bp		↑																
Germline vs. simulated	6bp	🟨↑																	
	7bp																		
	8bp																		
	6±5bp	▨↑				🟨↑													
	7±5bp																		
	8±5bp																		
	RINS±0bp	↑				↑													
	RINS±5bp	▨↑			↑	↑													
Shared vs. simulated	6bp																		
	7bp																		
	8bp																		
	6±5bp																		
	7±5bp																		
	8±5bp																		
	RINS±0bp	🟨↑																	
	RINS±5bp																		
Observed vs. simulated	6bp	🟨↑						↑											
	7bp																	↓	
	8bp																		
	6±5bp	▨↑	↑			🟨↑		▨↑				↑							
	7±5bp							↑							↓			▨↓	
	8±5bp																	↓	
	RINS±0bp	🟨↑				↑													
	RINS±5bp	▨↑			↑	↑													
Soma vs. germ	6bp																		
	7bp																		
	8bp																		
	6±5bp																		
	7±5bp																		
	8±5bp																		
	RINS±0bp		▨↑																
	RINS±5bp	▨↓	↑																
Somatic vs. shared	6bp																		
	7bp																		
	8bp																		
	6±5bp																		
	7±5bp																		
	8±5bp																		
	RINS±0bp	▨↓													▨↓				
	RINS±5bp	↓													↓				
Germline vs. shared	6bp																		
	7bp																		
	8bp																		
	6±5bp																		
	7±5bp																		
	8±5bp																		
	RINS±0bp	🟨↓							↓										
	RINS±5bp																		

Legend: ↑ or ↓ denotes the direction of the gene- or experiment-wise statistically significant result. The direction is with respect to the first group in the comparison. A grey shaded box represents an experiment-wise statistically significant result, a non-shaded arrow (i.e. ↑ or ↓) represents a gene-wise statistically significant result, a green shaded box represents ≥80% power to detect a statistically significant result for the comparison and associated effect size, Yellow shaded box represents ≤80% power and experiment-wise statistically significant result. Soma- Somatic, Germ- Germline, Obs.- Observed (somatic, germline and shared), Pot.- Potential, Rec.- Recurrent, Non-rec.- Non-recurrent;

Table 34 Summary of statistically significant results for somatic vs. simulated micro-lesions, with respect to repetitive elements

		Repeats						RINS	
Gene	Parameter	≥6bp	≥7bp	≥8bp	≥6±5bp	≥7±5bp	≥8±5bp	≥4bp	≥4±5bp
ALL	Somatic [%]	33.60	12.39	4.66	45.09	16.55	6.39	9.73	24.13
	Simulated [%]	31.74	13.40	5.89	39.32	17.39	7.91	8.88	22.54
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	5.26E-03	1.00E+00	8.72E-01	1.00E+00	1.00E+00
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	3.15E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00
	Effect size	0.02	0.02	0.03	0.06	0.01	0.03	0.01	0.02
	Power [%]	1.43	0.68	4.28	54.96	0.35	5.34	1.46	2.65
CDKN2A	Somatic [%]	58.33	43.52	15.74	66.67	51.85	17.59	9.00	27.00
	Simulated [%]	37.04	23.15	11.11	36.11	23.15	13.89	10.00	28.00
	Gene-wise p-value	2.94E-02	4.80E-02	1.00E+00	1.20E-03	4.20E-03	1.00E+00	1.00E+00	1.00E+00
	Experiment-wise p-value	5.00E-01	8.16E-01	1.00E+00	2.04E-02	7.14E-02	1.00E+00	1.00E+00	1.00E+00
	Effect size	0.21	0.22	0.07	0.31	0.30	0.05	0.02	0.01
	Power [%]	36.21	37.80	0.64	84.29	80.81	0.31	0.20	0.17
PTEN	Somatic [%]	42.35	16.33	6.12	53.57	20.92	9.69	15.10	28.13
	Simulated [%]	30.61	11.73	7.65	35.71	13.78	10.71	11.98	21.88
	Gene-wise p-value	3.69E-01	1.00E+00	1.00E+00	9.60E-03	1.00E+00	1.00E+00	1.00E+00	9.30E-01
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.63E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00
	Effect size	0.12	0.07	0.03	0.18	0.09	0.02	0.05	0.07
	Power [%]	14.17	1.47	0.20	52.79	5.28	0.09	1.12	3.87
TP53	Somatic [%]	30.76	10.03	4.47	41.46	12.74	5.69	7.86	23.85
	Simulated [%]	32.11	14.63	6.37	41.33	19.92	8.13	9.49	26.29
	Gene-wise p-value	1.00E+00	1.91E-01	1.00E+00	1.00E+00	9.60E-03	1.00E+00	1.00E+00	1.00E+00
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.63E-01	1.00E+00	1.00E+00	1.00E+00
	Effect size	0.01	0.07	0.04	0.00	0.10	0.05	0.03	0.03
	Power [%]	0.17	21.35	3.03	0.05	59.73	5.06	1.92	1.79
VHL	Somatic [%]	23.81	2.86	0.00	35.71	3.33	0.00	3.35	6.22
	Simulated [%]	34.76	10.00	3.81	40.95	13.33	5.24	3.35	9.09
	Gene-wise p-value	3.43E-01	1.09E-01	2.04E-01	1.00E+00	1.14E-02	7.32E-02	1.00E+00	1.00E+00
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.94E-01	1.00E+00	1.00E+00	1.00E+00
	Effect size	0.12	0.15	0.14	0.05	0.18	0.16	0.00	0.05
	Power [%]	15.39	30.79	26.42	0.86	58.76	45.02	0.15	1.89

Table 35 Summary of statistically significant results for germline vs. simulated micro-lesions, with respect to repetitive elements

		Repeats						RINS	
Gene	Parameter	≥6bp	≥7bp	≥8bp	≥6±5bp	≥7±5bp	≥8±5bp	≥4bp	≥4±5bp
ALL	Germline [%]	36.55	13.63	5.40	47.52	19.03	7.70	12.25	29.71
	Simulated [%]	31.03	12.93	5.24	38.80	16.62	7.16	9.65	23.47
	Gene-wise p-value	4.89E-04	1.00E+00	1.00E+00	8.40E-09	3.02E-01	1.00E+00	4.74E-02	1.52E-05
	Experiment-wise p-value	2.94E-03	1.00E+00	1.00E+00	5.04E-08	1.00E+00	1.00E+00	9.47E-02	3.05E-05
	Effect size	0.06	0.01	0.00	0.09	0.03	0.01	0.04	0.07
	Power [%]	76.28	0.40	0.11	99.73	12.25	0.39	42.15	96.21
BRCA1	Germline [%]	29.43	10.57	3.43	43.71	16.00	6.00	12.43	34.62
	Simulated [%]	30.00	11.43	4.00	38.57	14.00	6.00	8.88	23.67
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.67E-01	2.06E-02
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.50E-01
	Effect size	0.01	0.01	0.02	0.05	0.03	0.00	0.06	0.12
	Power [%]	0.06	0.10	0.11	1.77	0.30	0.05	4.61	48.06

BRCA2	Germline [%]	42.98	11.40	4.97	55.85	18.42	7.02	24.10	43.98
	Simulated [%]	33.04	14.04	5.26	41.81	18.42	6.73	13.86	31.93
	Gene-wise p-value	1.69E-01	1.00E+00	1.00E+00	1.20E-03	1.00E+00	1.00E+00	1.30E-02	7.20E-03
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	2.04E-02	1.00E+00	1.00E+00	2.21E-01	1.22E-01
	Effect size	0.10	0.04	0.01	0.14	0.00	0.01	0.13	0.12
	Power [%]	20.96	0.71	0.06	57.36	0.05	0.06	57.32	50.74

Table 36 Summary of statistically significant results for shared vs. simulated micro-lesions, with respect to repetitive elements

		Repeats						RINS	
Gene	Parameter	≥6bp	≥7bp	≥8bp	≥6±5bp	≥7±5bp	≥8±5bp	≥4bp	≥4±5bp
ALL	Shared [%]	37.66	11.69	3.90	42.86	14.29	3.90	26.67	38.67
	Simulated [%]	32.47	11.69	5.19	40.26	16.88	6.49	6.67	21.33
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.12E-02	2.25E-01
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.23E-02	4.51E-01
	Effect size	0.05	0.00	0.03	0.03	0.04	0.06	0.27	0.19
	Power [%]	0.36	0.08	0.15	0.13	0.18	0.41	59.26	23.09

Table 37 Summary of statistically significant results for observed vs. simulated micro-lesions, with respect to repetitive elements

		Repeats						RINS	
Gene	Parameter	≥6bp	≥7bp	≥8bp	≥6±5bp	≥7±5bp	≥8±5bp	≥4bp	≥4±5bp
ALL	Observed [%]	35.35	13.08	5.07	46.43	17.92	7.09	11.45	27.52
	Simulated [%]	31.33	13.08	5.62	39.07	16.96	7.44	9.29	23.07
	Gene-wise p-value	8.38E-04	1.00E+00	1.00E+00	5.44E-11	1.00E+00	1.00E+00	1.36E-02	3.11E-05
	Experiment-wise p-value	5.03E-03	1.00E+00	1.00E+00	3.26E-10	1.00E+00	1.00E+00	2.71E-02	6.22E-05
	Effect size	0.04	0.00	0.01	0.07	0.01	0.01	0.04	0.05
	Power [%]	72.17	0.08	1.33	99.98	1.38	0.31	57.08	94.85
APC	Observed [%]	40.16	17.87	6.72	50.98	23.93	9.02	11.76	30.59
	Simulated [%]	33.11	14.92	6.72	40.00	19.84	9.02	9.92	24.54
	Gene-wise p-value	2.75E-01	1.00E+00	1.00E+00	3.00E-03	1.00E+00	1.00E+00	1.00E+00	1.55E-01
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	5.10E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00
	Effect size	0.07	0.04	0.00	0.11	0.05	0.00	0.03	0.07
	Power [%]	17.59	1.81	0.05	64.28	3.96	0.05	1.56	19.91
BRCA1	Observed [%]	30.22	10.71	3.57	43.96	15.93	6.04	12.22	34.38
	Simulated [%]	30.22	11.26	4.12	38.46	13.74	6.04	9.09	23.58
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	9.33E-01	1.82E-02
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.09E-01
	Effect size	0.00	0.01	0.01	0.06	0.03	0.00	0.05	0.12
	Power [%]	0.05	0.07	0.10	2.38	0.40	0.05	3.31	49.06
BRCA2	Observed [%]	43.22	11.02	4.80	56.21	18.36	6.78	24.71	44.48
	Simulated [%]	33.05	13.84	5.37	41.81	18.36	6.78	13.66	31.98
	Gene-wise p-value	1.08E-01	1.00E+00	1.00E+00	0.00E+00	1.00E+00	1.00E+00	8.00E-03	5.20E-03
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	0.00E+00	1.00E+00	1.00E+00	1.36E-01	8.84E-02
	Effect size	0.10	0.04	0.01	0.14	0.00	0.00	0.14	0.13
	Power [%]	24.18	0.95	0.09	63.61	0.05	0.05	69.10	57.66
CDKN2A	Observed [%]	57.03	41.41	13.28	65.63	49.22	15.63	7.63	27.12
	Simulated [%]	37.50	23.44	11.72	35.94	23.44	14.06	9.32	27.97
	Gene-wise p-value	2.94E-02	8.88E-02	1.00E+00	0.00E+00	4.20E-03	1.00E+00	1.00E+00	1.00E+00
	Experiment-wise p-value	5.00E-01	1.00E+00	1.00E+00	0.00E+00	7.14E-02	1.00E+00	1.00E+00	1.00E+00

	Effect size	0.20	0.19	0.02	0.30	0.27	0.02	0.03	0.01
	Power [%]	36.08	33.91	0.10	89.70	78.88	0.09	0.35	0.16
PTEN	Observed [%]	42.80	17.20	6.80	53.20	22.00	11.60	14.40	28.40
	Simulated [%]	30.80	11.60	7.60	35.60	14.00	10.80	11.93	21.81
	Gene-wise p-value	1.28E-01	1.00E+00	1.00E+00	**5.40E-03**	4.92E-01	1.00E+00	1.00E+00	6.32E-01
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	9.18E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00
	Effect size	0.12	0.08	0.02	0.18	0.10	0.01	0.04	0.08
	Power [%]	24.07	4.43	0.09	68.24	12.34	0.08	0.88	6.60
TP53	Observed [%]	30.89	9.95	4.45	41.36	12.57	5.63	8.54	24.57
	Simulated [%]	32.20	14.66	6.41	41.36	19.90	8.12	9.46	26.28
	Gene-wise p-value	1.00E+00	1.57E-01	1.00E+00	1.00E+00	**8.40E-03**	1.00E+00	1.00E+00	1.00E+00
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.43E-01	1.00E+00	1.00E+00	1.00E+00
	Effect size	0.01	0.07	0.04	0.00	0.10	0.05	0.02	0.02
	Power [%]	0.17	24.75	3.65	0.05	65.51	5.88	0.54	0.79
VHL	Observed [%]	25.71	2.86	0.32	36.83	3.49	0.32	2.91	6.80
	Simulated [%]	34.29	10.16	3.81	40.63	13.33	5.08	3.24	9.06
	Gene-wise p-value	4.68E-01	**1.92E-02**	1.52E-01	1.00E+00	**1.20E-03**	**3.54E-02**	1.00E+00	1.00E+00
	Experiment-wise p-value	1.00E+00	3.26E-01	1.00E+00	1.00E+00	**2.04E-02**	6.02E-01	1.00E+00	1.00E+00
	Effect size	0.09	0.15	0.12	0.04	0.18	0.15	0.01	0.04
	Power [%]	12.74	59.05	34.34	0.61	83.23	58.01	0.19	1.63

Table 38 Summary of statistically significant results for somatic vs. germline micro-lesions, with respect to repetitive elements

		Repeats						RINS	
Gene	Parameter	≥6bp	≥7bp	≥8bp	≥6±5bp	≥7±5bp	≥8±5bp	≥4bp	≥4±5bp
ALL	Somatic [%]	33.60	12.39	4.66	45.09	16.55	6.39	9.73	24.13
	Germline [%]	36.55	13.63	5.40	47.52	19.03	7.70	12.25	29.71
	Gene-wise p-value	5.16E-01	1.00E+00	1.00E+00	1.00E+00	4.12E-01	1.00E+00	1.25E-01	**8.12E-04**
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.50E-01	**1.62E-03**
	Effect size	0.03	0.02	0.02	0.02	0.03	0.02	0.04	0.06
	Power [%]	8.37	1.42	1.15	3.57	9.90	4.05	30.14	81.90
APC	Somatic [%]	36.41	17.93	8.15	53.80	27.72	13.04	19.89	39.78
	Germline [%]	41.61	17.76	5.84	49.88	22.38	7.06	7.52	26.07
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.97E-01	**6.00E-04**	**8.80E-03**
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**1.02E-02**	1.50E-01
	Effect size	0.05	0.00	0.04	0.04	0.06	0.10	0.18	0.14
	Power [%]	1.10	0.05	0.75	0.47	1.88	13.21	87.83	55.89

Table 39 Summary of statistically significant results for somatic vs. shared micro-lesions, with respect to repetitive elements

		Repeats						RINS	
Gene	Parameter	≥6bp	≥7bp	≥8bp	≥6±5bp	≥7±5bp	≥8±5bp	≥4bp	≥4±5bp
ALL	Somatic [%]	33.60	12.39	4.66	45.09	16.55	6.39	9.73	24.13
	Shared [%]	37.66	11.69	3.90	42.86	14.29	3.90	26.67	38.67
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**2.86E-05**	**4.71E-02**
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**5.72E-05**	9.42E-02
	Effect size	0.02	0.00	0.01	0.01	0.01	0.02	0.11	0.07
	Power [%]	0.43	0.09	0.12	0.15	0.23	0.65	95.03	42.28
TP53	Somatic [%]	30.76	10.03	4.47	41.46	12.74	5.69	7.86	23.85
	Shared [%]	33.33	8.33	0.00	33.33	8.33	0.00	50.00	58.33

	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**6.00E-04**	**4.16E-02**
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**1.02E-02**	7.07E-01
	Effect size	0.01	0.01	0.03	0.02	0.02	0.03	0.19	0.10
	Power [%]	0.06	0.06	0.31	0.18	0.13	0.42	97.74	33.68

Table 40 Summary of statistically significant results for germline vs. shared micro-lesions, with respect to repetitive elements

		Repeats						RINS	
Gene	Parameter	≥6bp	≥7bp	≥8bp	≥6±5bp	≥7±5bp	≥8±5bp	≥4bp	≥4±5bp
ALL	Germline [%]	36.55	13.63	5.40	47.52	19.03	7.70	12.25	29.71
	Shared [%]	37.66	11.69	3.90	42.86	14.29	3.90	26.67	38.67
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**2.45E-03**	1.00E+00
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**4.90E-03**	1.00E+00
	Effect size	0.00	0.01	0.01	0.02	0.02	0.02	0.08	0.03
	Power [%]	0.09	0.21	0.27	0.52	1.02	1.67	73.88	8.31
NF1	Germline [%]	30.21	13.29	4.83	38.67	16.31	6.65	7.12	21.98
	Shared [%]	0.00	0.00	0.00	33.33	0.00	0.00	66.67	66.67
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**2.38E-02**	4.78E-01
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.05E-01	1.00E+00
	Effect size	0.06	0.04	0.02	0.01	0.04	0.03	0.21	0.10
	Power [%]	0.94	0.25	0.10	0.06	0.33	0.13	75.10	9.14

Table 41 Summary of statistically significant results for recurrent somatic micro-lesions, with respect to repetitive elements

		Repeats						RINS	
Gene	Parameter	≥6bp	≥7bp	≥8bp	≥6±5bp	≥7±5bp	≥8±5bp	≥4bp	≥4±5bp
ALL	Somatic recurrent [%]	35.11	12.54	3.45	45.45	15.99	5.96	14.11	26.96
	Somatic non-recurrent [%]	33.27	12.36	4.92	45.01	16.67	6.49	8.76	23.50
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**3.93E-02**	1.00E+00
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.86E-02	1.00E+00
	Effect size	0.01	0.01	0.03	0.00	0.01	0.01	0.07	0.03
	Power [%]	0.46	0.31	1.26	0.09	0.12	0.21	44.49	4.03
ALL	Somatic recurrent non-shared [%]	36.08	13.06	3.44	47.08	16.84	6.19	12.71	26.12
	Somatic recurrent shared [%]	25.00	7.14	3.57	28.57	7.14	3.57	28.57	35.71
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	6.63E-01	1.00E+00	1.00E+00	4.42E-01	1.00E+00
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	8.84E-01	1.00E+00
	Effect size	0.07	0.07	0.00	0.11	0.07	0.03	0.13	0.06
	Power [%]	1.93	1.41	0.08	6.82	2.11	0.37	22.67	2.51
ALL	Somatic recurrent non-shared [%]	36.08	13.06	3.44	47.08	16.84	6.19	12.71	26.12
	Somatic non-recurrent shared [%]	44.90	14.29	4.08	51.02	18.37	4.08	25.53	40.43
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.27E-01	4.72E-01
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.53E-01	9.43E-01
	Effect size	0.06	0.06	0.01	0.03	0.01	0.03	0.13	0.11
	Power [%]	1.97	1.43	0.10	0.22	0.11	0.39	23.08	15.22
ALL	Somatic non-recurrent non-shared [%]	33.11	12.27	4.89	44.71	16.49	6.43	9.14	23.74
	Somatic recurrent shared [%]	25.00	7.14	3.57	28.57	7.14	3.57	28.57	35.71
	Gene-wise p-value	1.00E+00	1.00E+00	1.00E+00	9.74E-01	1.00E+00	1.00E+00	**3.61E-02**	1.00E+00
	Experiment-wise p-value	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.21E-02	1.00E+00
	Effect size	0.02	0.02	0.01	0.04	0.03	0.02	0.09	0.04
	Power [%]	0.97	0.69	0.13	4.79	2.06	0.44	66.30	5.69

Table 42 Repeats and recurrent somatic mutations

		Observed mutations																							
		≥6bp				≥7bp				≥8bp				≥6±5bp				≥7±5bp				≥8±5bp			
		In repeats		not in repeats		In repeats		not in repeats		In repeats		not in repeats		In repeats		not in repeats		In repeats		not in repeats		In repeats		not in repeats	
mutations	Total	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
somatic	1783	599	33.60	1184	66.40	221	12.39	1562	87.61	83	4.66	1700	95.34	804	45.09	979	54.91	295	16.55	1488	83.45	114	6.39	1669	93.61
germline	2443	893	36.55	1550	63.45	333	13.63	2110	86.37	132	5.40	2311	94.60	1161	47.52	1282	52.48	465	19.03	1978	80.97	188	7.70	2255	92.30
shared	77	29	37.66	48	62.34	9	11.69	68	88.31	3	3.90	74	96.10	33	42.86	44	57.14	11	14.29	66	85.71	3	3.90	74	96.10
observed	4305	1521	35.35	2782	64.65	563	13.08	3740	86.92	218	5.07	4085	94.93	1998	46.43	2305	53.57	771	17.92	3532	82.08	305	7.09	3998	92.91
somatic recurrent	347	119	34.3	228	65.71	42	12.1	305	87.9	12	3.46	335	96.54	153	44.1	194	55.91	53	15.3	294	84.73	20	5.76	327	94.24
somatic non recurrent	1436	480	33.4	956	66.57	179	12.5	1257	87.53	71	4.94	1365	95.06	651	45.3	785	54.67	242	16.9	1194	83.15	94	6.55	1342	93.45

N- Number of micro-lesions

Table 43 Distribution of micro-deletions and micro-insertions and runs of mononucleotides

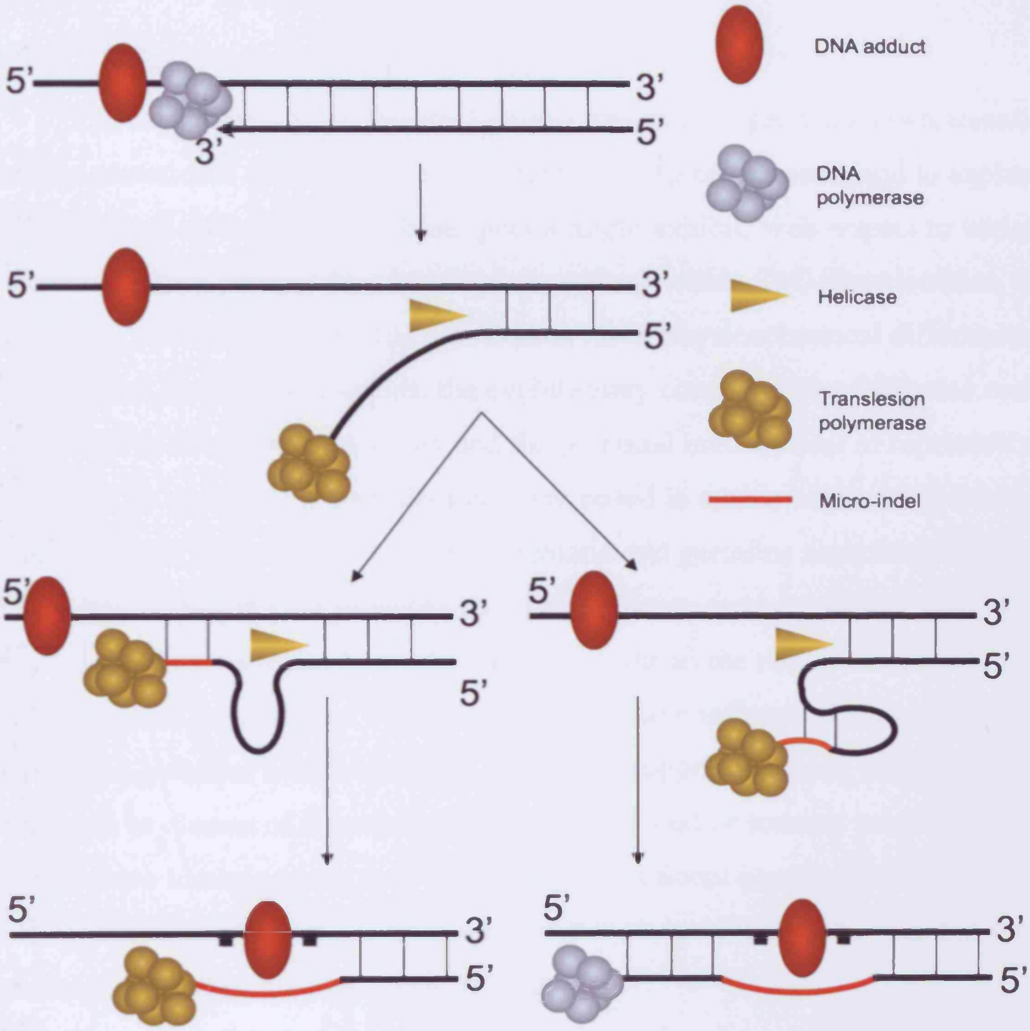
			Somatic mutations				Germline mutations				Shared mutations				Observed mutations			
	Mutations	Runs	Mutations in runs		Mutations not in runs		Mutations in runs		Mutations not in runs		Mutations in runs		Mutations not in runs		Mutations in runs		Mutations not in runs	
Gene	N	N	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
APC	595	677	36	19.89	145	80.11	30	7.52	369	92.48	4	26.67	11	73.33	70	11.76	525	88.24
ATM	162	1352	1	20.00	4	80.00	31	19.75	126	80.25	0	0.00	0	0.00	32	19.75	130	80.25
BRCA1	352	584	0	0.00	9	100.00	42	12.43	296	87.57	1	20.00	4	80.00	43	12.22	309	87.78
BRCA2	344	1054	3	33.33	6	66.67	80	24.10	252	75.90	2	66.67	1	33.33	85	24.71	259	75.29
CDH1	35	328	1	6.67	14	93.33	7	35.00	13	65.00	0	0.00	0	0.00	8	22.86	27	77.14
CDKN2A	118	61	9	9.00	91	91.00	0	0.00	16	100.00	0	0.00	2	100.00	9	7.63	109	92.37
NF1	342	1195	2	12.50	14	87.50	23	7.12	300	92.88	2	66.67	1	33.33	27	7.89	315	92.11
NF2	275	248	13	6.37	191	93.63	2	3.03	64	96.97	0	0.00	5	100.00	15	5.45	260	94.55
PTCH	94	492	4	20.00	16	80.00	13	17.57	61	82.43	0	0.00	0	0.00	17	18.09	77	81.91
PTEN	243	197	29	15.10	163	84.90	5	12.20	36	87.80	1	10.00	9	90.00	35	14.40	208	85.60
RB1	211	545	6	14.29	36	85.71	22	13.33	143	86.67	1	25.00	3	75.00	29	13.74	182	86.26
STK11	75	185	1	25.00	3	75.00	9	13.04	60	86.96	2	100.00	0	0.00	12	16.00	63	84.00
TP53	761	184	58	7.86	680	92.14	1	9.09	10	90.91	6	50.00	6	50.00	65	8.54	696	91.46
TSC1	79	393	0	0.00	1	100.00	8	10.26	70	89.74	0	0.00	0	0.00	8	10.13	71	89.87
TSC2	161	727	0	0.00	5	100.00	12	7.69	144	92.31	0	0.00	0	0.00	12	7.45	149	92.55
VHL	309	66	7	3.35	202	96.65	1	1.16	85	98.84	1	7.14	13	92.86	9	2.91	300	97.09
WT1	19	176	1	14.29	6	85.71	1	8.33	11	91.67	0	0.00	0	0.00	2	10.53	17	89.47
Total	4175	8464	171	9.73	1586	90.27	287	12.25	2056	87.75	20	26.67	55	73.33	478	11.45	3697	88.55

N- Number of micro-lesions

Table 44 Micro-deletions, micro-insertions and micro-indels

Gene	Type lesion	N Micro-deletions	Proportion of micro-deletions of total	N Micro-insertions	Proportion of micro-insertions of total	N Micro-indels	Proportion of micro-indels of total	Total N
APC	soma	137	74.46	44	23.91	3	1.63	184
	germ	284	69.10	115	27.98	12	2.92	411
	shared	15	100.00	0	0.00	0	0.00	15
ATM	soma	4	80.00	1	20.00	0	0.00	5
	germ	122	71.35	35	20.47	14	8.19	171
	shared	0	0.00	0	0.00	0	0.00	0
BRCA1	soma	6	66.67	3	33.33	0	0.00	9
	germ	255	72.86	83	23.71	12	3.43	350
	shared	3	60.00	2	40.00	0	0.00	5
BRCA2	soma	7	77.78	2	22.22	0	0.00	9
	germ	244	71.35	88	25.73	10	2.92	342
	shared	1	33.33	2	66.67	0	0.00	3
CDH1	soma	13	86.67	2	13.33	0	0.00	15
	germ	12	57.14	8	38.10	1	4.76	21
	shared	0	0.00	0	0.00	0	0.00	0
CDKN2A	soma	76	70.37	24	22.22	8	7.41	108
	germ	10	55.56	6	33.33	2	11.11	18
	shared	1	50.00	1	50.00	0	0.00	2
NF1	soma	13	81.25	3	18.75	0	0.00	16
	germ	218	65.86	105	31.72	8	2.42	331
	shared	3	100.00	0	0.00	0	0.00	3
NF2	soma	176	83.81	28	13.33	6	2.86	210
	germ	50	73.53	16	23.53	2	2.94	68
	shared	5	100.00	0	0.00	0	0.00	5
PTCH	soma	14	66.67	6	28.57	1	4.76	21
	germ	42	51.22	32	39.02	8	9.76	82
	shared	0	0.00	0	0.00	0	0.00	0
PTEN	soma	145	73.98	47	23.98	4	2.04	196
	germ	23	52.27	18	40.91	3	6.82	44
	shared	6	60.00	4	40.00	0	0.00	10
RB1	soma	30	68.18	12	27.27	2	4.55	44
	germ	112	64.00	53	30.29	10	5.71	175
	shared	4	80.00	0	0.00	1	20.00	5
STK11	soma	3	75.00	1	25.00	0	0.00	4
	germ	45	63.38	24	33.80	2	2.82	71
	shared	2	66.67	0	0.00	1	33.33	3
TP53	soma	504	68.29	234	31.71	0	0.00	738
	germ	8	57.14	3	21.43	3	21.43	14
	shared	8	66.67	4	33.33	0	0.00	12
TSC1	soma	1	100.00	0	0.00	0	0.00	1
	germ	53	64.63	25	30.49	4	4.88	82
	shared	0	0.00	0	0.00	0	0.00	0
TSC2	soma	3	50.00	2	33.33	1	16.67	6
	germ	110	69.18	46	28.93	3	1.89	159
	shared	0	0.00	0	0.00	0	0.00	0
VHL	soma	171	81.43	38	18.10	1	0.48	210
	germ	55	60.44	31	34.07	5	5.49	91
	shared	8	57.14	6	42.86	0	0.00	14
WT1	soma	4	57.14	3	42.86	0	0.00	7
	germ	8	61.54	4	30.77	1	7.69	13
	shared	0	0.00	0	0.00	0	0.00	0
Total	soma	1307	73.30	450	25.24	26	1.46	1783
	germ	1651	67.58	692	28.33	100	4.09	2443
	shared	56	72.73	19	24.68	2	2.60	77
Grand total		3014	70.04	1161	26.98	128	2.97	4303

Figure 46 Mechanism of generation of micro-indels (modified after Scaringe et al. 2008)



7. General discussion

7.1. Objectives

The main objective of this PhD project was to compare the known somatic and germline mutational spectra of 17 human tumour suppressor genes, and to explore the similarities and differences that these spectra might exhibit, with respect to various parameters. These parameters comprised: mutations within CpG dinucleotides, disease and non-disease associated nucleotide substitution rates, physicochemical differences between wild-type and mutant amino-acids, the evolutionary conservation of affected codons, the role of nonsense-mediated mRNA decay and the potential involvement of repetitive sequence elements. The questions to be addressed were posed in such a way as to provide a glimpse of the mechanisms that might influence the somatic and germline mutational spectra and their relative involvement in the process of tumour initiation and/or development. Thus, this work represents a formal attempt to try to shed some light on the relative importance of exogenous and endogenous mechanisms of mutagenesis that have influenced the occurrence of both somatic and germline mutations in the 17 tumour suppressor genes studied. Further, since some types or classes of mutations, either germline and/or somatic might be relatively more likely to drive tumorigenesis than others, some additional questions were also posed e.g. do such mutations exist and if so, are these mutations more likely to be found in the soma or the germline? Moreover, are they likely to recur?

To accomplish the main objective, the mutational spectra of 17 human tumour suppressor genes (i.e. *APC, ATM, BRCA1, BRCA2, CDH1, CDKN2A, NF1, NF2, PTCH, PTEN, RB1, STK11, TP53, TSC1, TSC2, VHL* and *WT1*) were sought, at the beginning of this PhD work (end of 2005). The selection of these 17 genes was based on their being known at the time to exhibit somatic mutations in one or more types of human cancer as well as germline mutations conferring an inherited predisposition to cancer. It should be noted that a considerable number of mutations in a given gene does not necessarily imply the significant involvement of that particular gene in tumour development. Nevertheless, for all these 17 genes, a sufficient body of biologically plausible evidence had been amassed to denote their involvement in tumorigenesis when mutated either in the germline or the soma. Because these genes are also some of the most commonly mutated genes in human cancers, the number of mutations available to study would be maximized. For example, ~50% of all studied cancers have been found to exhibit mutations in the *TP53* gene, indicative of its pivotal role of suppressing tumour development (Soussi and Beroud 2001).

The mutation spectra comprised single base-pair substitutions that introduced missense and nonsense mutations; and micro-lesions (*viz*. micro-deletions, micro-insertions and micro-indels; length of deleted and/or inserted nucleotides ≤20bp) within the coding regions of the genes. These sequence changes comprised both somatic and germline mutations. The germline mutations were obtained from the Human Gene Mutation Database (*HGMD*; http://www.hgmd.org; Stenson et al. 2003) whereas the somatic mutations were derived from literature searches, via PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez) and various somatic mutational databases (for detailed information see Chapter 2; General materials). In addition, somatic mutations that were reported more than once were recorded as being recurrent. Once the germline and somatic mutational spectra were directly compared for a particular gene, some mutations were found in both the soma and the germline. Therefore, in order to recognise the existence of such mutations, all mutations that were found in both the soma and the germline, in a particular gene, were removed and a single entry with the label 'shared' was created.

Overall, >4000 somatic and >4000 germline mutations were collected (summary of the collected mutations is presented in Table 45 and Table 46). Thus, the known mutations from the 17 human tumour suppressor genes under study, at the time of start of this PhD project (i.e. end of 2005), comprised >8000 mutations in total. With such a large number of mutations overall, it came as something of a surprise to find that some of the analyses undertaken exhibited relatively low statistical power to detect a significant difference (i.e. <80%). It should however be made clear that ~50% of all the collated somatic mutations (i.e. ~25% of all the mutations collected) were from the *TP53* gene. As a result, for some genes, a paucity of lesions was observed which limited the statistical power available in some cases. Therefore, for some of the comparisons performed, only a limited number of conclusions could be drawn. This notwithstanding, most of the analyses performed yielded some interesting findings. Some of these results confirmed the findings of previous research, whereas other results were relatively novel.

7.2. Summary of main conclusions

Some genes exhibited almost a complete absence of somatic mutations (i.e. *ATM*, *BRCA1*, *BRCA2*, *NF1*, *TSC1* and *TSC2*). In some of these cases evidence was presented that the paucity of somatic mutations could well be a consequence of mechanisms other than inactivation via somatic mutations, e.g. potential promoter hypermethylation for the *BRCA1* gene and dominant-negative effect of germline missense mutations in the case of the *ATM* gene. By contrast, only the *TP53* gene exhibited a paucity of germline mutations.

A small proportion of the combined mutations for all genes (~5%) were found to be shared between the germline and the soma, although the numbers slightly ranged between the different type of mutations. Thus, relatively greater proportions of the shared mutations were found amongst nonsense (~11%; ranged from 0% to 35%) and missense mutations (~7%; ranged from 0% to 12%), as compared to micro-deletions (~3%; ranged from 0% to 4%), micro-insertions (~2%; ranged from 0% to 8%) and micro-indels (~1%; ranged from 0% to 33%). Shared missense mutations were not merely coincidental. The shared missense mutations were found to be more likely to be drivers of tumorigenesis than either exclusively somatic or exclusively germline missense mutations. This was the case not only for the combined shared mutations across genes but also for individual genes, such as the *CDKN2A*, *TP53* and *VHL*.

A relatively large proportion of the combined somatic mutations for all genes (~33%; ranged from 0% to 53%) were found to have been independently observed more than once (i.e. they were found to recur), although the proportions ranged considerably between the different types of mutations. Thus, the largest proportions of recurrent mutations were found amongst the missense (~44%; ranged from 0% to 64%) and nonsense mutations (~43%; ranged from 0% to 89%), followed by micro-deletions (~20%; ranged from 0% to 32%), micro-insertions (~18%; ranged from 0% to 31%) and micro-indels (no mutations were found to recur). It is of note that the proportion of recurrent somatic mutations were mainly influenced by the numbers of recurrent mutations in the *TP53* gene, comprising ~78% of all recurrent mutations for all genes. The recurrence status of the somatic missense mutations was found to be heavily influenced by nucleotide context, but they were also likely to have been selected for their functional importance.

For some genes, it was found that intra-genic CpG-methylation was likely to have been frequently responsible for methylation-mediated deamination of 5-methylcytosine in CpG dinucleotides in both the soma and the germline (e.g. *TP53*). In addition, the CpG-dinucleotides comprised only ~1% of the extended cDNA sequence lengths combined for all genes, but missense and nonsense mutations found within CpG-dinucleotides comprised ~8% of all single base-pair substitutions for all genes. Even more, ~27 % of shared missense and nonsense mutations were found within CpG-dinucleotides, significantly more as compared to

~4% and ~8% for the somatic and germline mutations respectively (for more details see Chapter 5). Furthermore, CpG-located mutations were found to be more likely to be shared between the germline and the soma than non-CpG located mutations. Thus, for a number of genes, all possible CpG-located nonsense mutations were present in either the germline or the soma (i.e. *BRCA1*, *CDKN2A*, *PTEN*, *STK11*, *TP53*, *VHL* and *WT1*). This is likely to be consistent with the operation of this endogenous mutational mechanism in both the germline and the soma for the genes under study.

A significantly higher proportion (~34%) of the micro-lesions (both somatic and germline; repeats of size ≥6±5bp) combined for all genes were found within repetitive elements as compared to simulated micro-lesions (~32%; for more details see Chapter 6). Furthermore, shared micro-lesions were significantly more likely to be found in repetitive elements (i.e. runs of identical nucleotides) than both somatic and germline micro-lesions. Thus, it is likely that the mutational mechanisms responsible for these micro-lesions are shared between the germline and the soma and have probably been mediated by repetitive elements. Intriguingly, germline micro-lesions were found to be more likely to be influenced by repetitive elements than somatic micro-lesions. This could be a reflection of the probable higher proportion of mutations arising from the action of endogenous mutational mechanisms in the germline, as the germline is likely to be relatively protected from the action of exogenous mutagens compared to the soma.

7.3. Exogenous vs. endogenous mechanisms of mutagenesis

It is evident from numerous studies that the mutational spectrum in both the germline and the soma is influenced by the action of both endogenous mutational mechanisms and exogenous mutagens (e.g. carcinogens).

A similar mutational spectrum in the soma to that in the germline would argue strongly that the mechanisms that influence the spectra are very likely to be shared. Thus, the relatively large proportion of identical (i.e. in terms of the genic position and type of sequence change) mutations in both the soma and the germline (i.e. shared mutations) would suggest that those mutations are quite likely to have originated through the action of very similar mechanisms.

On average for all 17 genes studied, ~21% of shared missense mutations and ~37% of shared nonsense mutations were found within CpG dinucleotides (i.e. C->T on the coding and G->A on the non-coding DNA strand; for more details see Chapters 4 and 5). Therefore, a relatively large proportion of the single base-pair substitutions are quite likely attributable to endogenous deamination of 5-methylcytosine within CpG-dinucleotides (Pfeifer 2006) without having to invoke the action of exogenous mutagens or carcinogens. However, Pfeifer and Besaratinia (2009) have shown that methylated CpG-dinucleotides are a preferential target of various physical and chemical genotoxic agents (e.g. nitric oxide, benzo[α]pyrene, aflatoxin B1, etc.), at least in the context of the *TP53* gene. Thus, some of the mutations at CpG dinucleotides could well have resulted from the action of exogenous genotoxic agents or carcinogens as well as from the action of endogenous mechanisms. Their relative contribution is however as yet largely unknown.

It should be said that the *TP53* gene is likely to be a unique gene, in the sense that the majority of observed mutations are non-truncating mutations (i.e. missense) as opposed to other tumour-suppressor genes, where this is relatively uncommon. In addition, the analysis of micro-lesions suggested that observed micro-lesions combined for all genes were very likely to be associated with repetitive elements. Hence, it is likely that a relatively large proportion of the observed micro-lesions (both somatic and germline) are also likely to have resulted from endogenous mutagenesis, such as 'slipped-mispairing', 'quasi-palindromic' sequences, 'strand-switching' etc. (Bacolla and Wells 2009; Cooper and Krawczak 1993; Efstratiadis et al. 1980; Wells 2007; Wells et al. 2005). By contrast, the somatic micro-lesions in the *TP53* gene showed a trend in the opposite direction, i.e. relatively few somatic mutations were observed in repetitive elements. One could argue that relatively more somatic micro-lesions in the *TP53* gene are result of the actions of exogenous mutagenesis, in comparison to most of the tumour suppressor genes studied. Indeed, mutagens, such as reactive oxygen species, ionizing radiation, a variety of chemicals, food toxins, etc., have been shown to induce frameshifts, and both double and single-strand DNA breaks (Bertram and Hass 2008; Bertram et al. 2001; Breen and Murphy 1995; Sankaranarayanan and Wassom 2005). Nevertheless, shared micro-lesions in the *TP53* gene were preferentially found in runs of identical nucleotides, as compared to somatic micro-lesions. Thus, some *TP53* micro-lesions are likely to be the result of endogenous mechanisms and these are also likely to be found in both the soma and the germline.

It has to be said that a clear-cut distinction between carcinogens or toxins of endogenous origin (e.g. oxyradicals, nitric oxide, etc.) and exogenously derived carcinogens (e.g. cigarette and tobacco smoke, heavy metals, asbestos fibres, etc.) is probably untenable (Morley and Turner 1999; Valavanidis et al. 2009). Furthermore, the relative exposure of the germline and soma to environmental carcinogens is often largely unknown. One exception is

UV light in skin cancer, where the relative contribution can be ascertained since it is unlikely that UV light makes a significant contribution to the observed germline mutational spectrum because the germline is relatively well protected from exposure to UV light as compared to the soma. On the other hand, the somatic mutational spectrum of skin cells is very likely to be influenced by sunlight - indeed, it contains high frequency CC->TT tandem mutations (genome-wide CC->TT mutations; HPLC method; Mouret et al. 2008) - but it is relatively unlikely that UV light could directly induce mutagenic DNA damage (e.g. cyclobutane pyrimidine dimers and pyrimidine-pyrimidone photoproducts) in other tissues, mainly due to the absorption of the energy levels of UV light by the upper epidermal layers of the skin (Javeri et al. 2008). Nevertheless, the relative incidence of cyclobutane pyrimidine dimers, characteristic of UVC light, has not been found to correlate with the frequency of the most common mutations (Vreeswijk et al. 2009) and could be independent of the clinical phenotype (i.e. basal cell carcinoma; Heitzer et al. 2007). On the other hand, carcinogens are more likely to be found in various tissues, transported mainly via the blood stream, such as nitrosamines, heavy metals, alkaloids, etc. (Hoffmann et al. 2001; Taioli 2008). There has even been some evidence that air pollution could elevate germline mutation frequencies at expanded simple tandem repeats in mice, but whether these represent a good marker for other genomic regions or the cells of the human germline is largely unknown (Somers and Cooper 2009).

7.4. Structural characteristics of individual genes and observed mutational spectra

Private characteristics of individual genes are likely to influence the observed mutational spectrum. Thus, it is likely that certain features of a gene could influence the selection of the type and position of mutations. For example, the somatic and germline mutational spectra in the *APC* gene have been known for some time to be different (Fearnhead et al. 2001; Lamlum et al. 1999). It has also been shown that the type of the inherited (germline) mutation in the *APC* gene could play a role in determining the position and/or nature of the second (somatic) hit. Thus, if the germline mutation occurs outside the region between codons 1194 and 1392, the second hit is likely to be a truncating mutation within the Mutation Cluster Region (MCR; Lamlum et al. 1999). Furthermore, our analysis indicated that the majority of somatic and germline nonsense mutations in the *APC* gene are quite likely to have been selected for the loss of β-catenin-binding ability of the mutant protein product. Thus, specific sequence and structural characteristics of the *APC* gene (i.e. β-catenin-binding sites) are likely to play an important part in influencing the mutational spectrum of both the germline and somatic mutations.

The *TP53* gene is another gene in which gene-specific characteristics could well play an important role in shaping the mutational spectrum. Unlike any other tumour suppressor gene, the majority of lesions comprise missense mutations. These also cluster in the DNA-binding region of the gene (Soussi et al. 2005). It has been previously shown that mutations in the *TP53* gene are selected for the disruption of DNA-binding ability, i.e. the majority of mutations are located in the DNA-binding domain of the gene (Soussi and Beroud 2001; Soussi et al. 2005; Soussi and Wiman 2007). In addition, complete inactivation of the protein could be insufficient for tumorigenesis (Lang et al. 2004; Olive et al. 2004), hence the majority of lesions are non-truncating mutations (i.e. missense mutations). These observations have led to the idea that the *TP53* gene might be different from other tumour suppressor genes, not only because of its mutational spectrum, but also because it might not be a true tumour suppressor gene in the traditional sense. Tumour suppressor genes are generally regarded as sustaining a 'loss-of-function' mutations during tumour development, i.e. mutations are selected for their ability to partially or completely inactivate the wild-type protein product. Several studies have suggested a possible 'gain-of-function' or acquisition of oncogenic properties of the *TP53* (Kawamata et al. 2007; Strano et al. 2007), exemplifying this unusual tumour suppressor gene. It appears that *TP53* exhibits some duality with respect to oncogenic and tumour suppressor functions (Strano et al. 2007). Nevertheless, although nonsense mutations in the *TP53* gene are relatively uncommon (~8%), all were highly likely to have been selected for the complete inactivation of the protein. Therefore, not all mutations in the *TP53* gene are selected for a 'gain-of-function'.

It would seem that such private characteristics are confined to individual genes and are not commonly shared in the rest of the tumour suppressor genes studied. None of the other genes showed such extreme differences in their mutational spectrum. One could speculate that such private characteristics might be important in the context of other tumour suppressor genes, but for some of those, an insufficient number of mutations was observed. Thus, no definitive conclusions could be made, with respect to such private characteristics in other tumour suppressor genes.

7.5. Selection and mutability

As pointed out above, selection of the types and positions of mutations is likely to play a very important part in shaping the mutational spectrum. Nevertheless, in order for these mutations to be selected for their functional impact or consequences, the pool of mutations that selection exerts its influence upon, is likely to be critically dependent on mutability.

A study by Walter et al. (1998) have suggested that mutation frequencies in the germ cells (spermatogenic cells of all types) are relatively lower, as compared to somatic tissues. Thus, at least in mice, the spontaneous mutation frequency in male germ cells is likely to be lower than in somatic cells. Walter et al. (1998) have also suggested that the lower spontaneous frequency could be due to additional quality control mechanisms or checkpoints that could induce apoptosis within spermatogonic cells. Moreover, no increase in mutation frequencies in spermatogonia in mice, following ionizing irradiation, has been observed (Xu et al. 2008). Thus, differences in the mutation frequencies and hence corresponding mutational spectrum might differ between the soma and the germline, with respect to both spontaneous and induced mutations (e.g. environmental factors). Nevertheless, some similarities are also observed. Among others, an age-related shift in mutation frequency is observed in both somatic and germline cells in humans (Evans et al. 2005; Walter et al. 1998) and in mice (Walter et al. 1998). Evans et al. (2005) have argued that it could be mainly due to increasingly dysfunctional repair mechanisms with increase of age.

Throughout this PhD work, the 17 genes studied were regarded as, and assumed to be, classical tumour suppressor genes. As such, bi-allelic inactivation is required for tumour initiation and/or development, following Knudson's two-hit hypothesis (Knudson 1971, 1978). In addition, one functional allele of the tumour suppressor genes is likely to be sufficient for tumour suppression. Cancer is a somatic disorder, mainly occurring in post-reproductive age. On this basis, a major distinction between germline and somatic mutations may be made. Thus, germline mutations are very likely not selected against, due to the fact that selection has not yet acted upon the predisposing germline mutations, whereas somatic mutations are selected for their 'loss-of-function' effect on a cellular level (Stratton et al. 2009). Quite the opposite was found when somatic and germline missense mutations were analysed. Germline missense mutations were indeed more likely to have negative impact on the function of the proteins for the combined mutations in all genes (in comparison with somatic missense mutations) and specifically in the *ATM*, *BRCA1* and *VHL* genes. As a result, it seems that germline missense mutations are selected for their relatively more drastic consequences, than somatic missense mutations. Alternatively, germline missense mutations could have come to clinical attention, because these mutations could have given a predisposition or inherited risk, through various potential mechanisms. These mechanisms, include: haploinsufficiency, gene-dosage effect, increased somatic mutation frequency, etc. Further, the fact that one allele copy of the gene has already been inactivated, results in one less somatic hit required for tumour development. Furthermore, relatively fewer somatic mutations were observed in the *ATM* (~6% somatic mutations) and *BRCA1* (~3% somatic mutations) genes, as compared to other tumour suppressor genes.

Due to the relative paucity of second (somatic) hits in these genes, at least with the mutations analysed, one way to inactivate them is via germline bi-allelic inactivation. This is not likely in the case of tumour-suppressor genes. Furthermore, heterozygous *BRCA1* primary mammary epithelial cells have been shown to exhibit increased clonal growth and proliferation (Burga et al. 2009). Thus, one could speculate that a gene-dosage effect may exist for the *BRCA1* gene by which one functional allele is insufficient for tumour suppression. *ATM* knockout mice display a gene dose-dependent effect of the embryopathic effects of ionizing radiation (Bhuller and Wells 2006). Further, somatic LOH in the *ATM* gene is likely to be present in mammary carcinoma, but germline missense mutations occur with similar frequencies, whether somatic LOH is present or not (Feng et al. 2003). Thus, Feng et al. (2003) have suggested that LOH as a second (somatic) hit might not be a crucial step, at least in mammary carcinoma. Alternative mechanisms to bi-allelic inactivation also comprise dominant-negative effect over the wild-type product. Reports have suggested that some heritable missense mutations in the *ATM* gene could exhibit a dominant-negative effect over the wild-type protein product (Bakkenist and Kastan 2003; Gatti et al. 1999). Such germline mutations, displaying a dominant-negative effect could potentially explain the paucity of somatic mutations in the *ATM* gene.

Even although there is some evidence that the *ATM* and *BRCA1* genes may display duality with respect to their tumour suppressor functions, additional evidence supports their role as true tumour suppressor genes. Thus, *ATM*$^{-/-}$ mice are viable, although with impaired cell-cycle arrest, increased chromosome breaks and are radiosensitive (Gurley and Kemp 2001). In addition, *ATM* haploinsufficiency has been shown to exhibit little or no effect on the somatic or germline mutation rates (expanded simple tandem repeats; ESTR) in mice, although it is difficult to extrapolate intra-genic mutation rates from ESTRs (Somers and Cooper 2009). Since mice heterozygous for either *BRCA1* or *ATM* do not show increased susceptibility to mammary tumour development (Karabinis et al. 2001), the second somatic hit could be loss-of-heterozygosity. Further, somatic LOH or somatic bi-allelic deletions of the *BRCA1* gene have been shown in sebaceous gland carcinoma (Becker et al. 2008).

Further support for a tumour suppressor role for the *ATM* and *BRCA1* genes comes from the fact that the majority of germline mutations are truncating mutations (~74% and ~73% for the *ATM* and *BRCA1* genes respectively). Therefore, the majority of germline mutations are selected for their 'loss-of-function effect'. Thus, it is likely that a second hit could comprise LOH, gross gene rearrangement or suppression of gene expression through promoter hypermethylation.

Duality of tumour suppression and oncogenic functions are best exemplified by the mutational spectra in the *TP53* gene. As mentioned earlier, the majority of lesions are missense mutations. Thus, the selection of mutations is not only towards loss-of-function mutations (e.g. deletions, insertions, indels and nonsense mutations), but also gain-of-function mutations. This is supported by studies that have suggested that the loss-of-function mutations might not give sufficient enough growth advantage over the cells, and gain-of-function mutations are required (reviewed in Brosh and Rotter 2009). Thus, *TP53*-null mice have not shown an increased ability to form tumours (Brosh and Rotter 2009; Dittmer et al. 1993; Shaulsky et al. 1991; Wolf et al. 1984).

The fact that ~50% of all the somatic mutations studied herein were *TP53* mutations is likely to have influenced the overall result of ~55% of the combined somatic mutations in all genes being truncating lesions (i.e. nonsense mutations, micro-deletions, micro-insertions and micro-indels). Thus, a significantly higher proportion of germline mutations (~80%) are truncating mutations as compared to somatic mutations (~55%). It seems that most of the germline mutations are selected for their negative impact on the function on the protein and somatic mutations to some extent are not selected against. In addition, when *TP53* mutations are excluded, the proportion of somatic truncating mutations rises to ~68%, as compared to ~80% for the germline mutations, but still significantly more germline truncating mutations were observed. As a consequence, the relatively high proportion of *TP53* mutations that make up the somatic mutational spectrum have not skewed the overall results observed for the combined mutations for all genes. In addition, ~39% of all recurrent somatic micro-lesions combined for all genes are truncating mutations. When *TP53* recurrent mutations were excluded, ~77% of the recurrent somatic mutations were truncating. Thus, for most genes, but with the exception of *TP53*, most of the somatic mutations that recur are selected for their negative impact on the function of the protein.

It is well known that CpG dinucleotides are a mutational hotspot, mainly due to the spontaneous deamination of 5-methylcytosine (Pfeifer and Besaratinia 2009). In addition, estimates show that their hypermutability is 5 times the base mutation rate (Krawczak et al. 1998). Therefore, one would expect to observe a high proportion of CpG-located missense and nonsense mutations. Indeed, ~7% and ~9% of all missense and nonsense mutations respectively (combined for all genes) showed CpG-located mutations. It has to be said that not all CpG-located mutations are going to have the same impact on the function of the genes. It is likely that selection would also play an important part in the occurrence of CpG-located mutations. Both missense and nonsense germline mutations (combined for all genes) exhibited a relatively higher proportion of CpG-located mutations (~8% and ~7% for germline missense and nonsense mutations respectively) as compared to somatic mutations (~5% and <1% for the somatic missense and nonsense mutations). One explanation includes different methylation status of CpG-dinucleotides in the germline and the soma.

Alternatively, in case of similar methylation status, it is likely that selectional forces against CpG-located somatic mutations could have contributed to a lower proportion as compared to germline CpG-located mutations. This is further supported by the even lower proportion of recurrent somatic CpG-located missense mutations (~3%). Conversely, CpG-located recurrent nonsense mutations are selected for (~15%) as compared to exclusively somatic CpG-located nonsense mutations (<1%). In addition, mutability is likely to play equally important role in both the germline and the soma. This was supported by the observation that the highest proportion of CpG-located missense and nonsense mutations were found in shared mutations (~21% and ~37% for the missense and nonsense mutations respectively), as compared to exclusively somatic and exclusively germline mutations.

Surprisingly, very few mutations were shared between the soma and the germline with respect to micro-deletions, micro-insertions and micro-indels. One potential explanation includes the fact that micro-lesions could occur throughout the gene sequence with varying lengths (0-20bp) of affected nucleotides. Therefore, the number of possible mutations is much greater for micro-lesions than for missense and nonsense mutations; hence the probability of a micro-lesion to be found in both the soma and the germline is much lower than for shared missense and nonsense mutations.

Taking these results altogether, selection and mutability are the main forces that influence the mutational spectra in the germline and soma. As a result, some of the similarities, but also the differences between the germline and the somatic mutations are likely to be a result of differences or similarities in selection and/or mutability that operate on the mutational spectra in the 17 studied human tumour suppressor genes.

7.6. Contributions/Benefits of the PhD work with respect to cancer genetics

The presented herein PhD work is relevant to a number of mainstream hypotheses or ideas.

Some mutations are likely to confer relatively greater cellular growth and/or proliferative advantages, hence are more likely to contribute towards tumorigenesis, than others. As a result, mutations could be described as either being more likely to be 'drivers' of tumour development or more likely to be 'passengers' (Stratton et al. 2009). Thus, somatic cells could acquire 'passenger' mutations relatively early during clonal expansions that would then come to be present in a majority, or even all, cells in a tumour. Alternatively, cells that acquire 'passenger' mutations relatively late during clonal expansion, would display mosaicism when tumours are analysed. Most of these mutations would not confer growth and/or a proliferative advantage over the cells harbouring them. Hence, it is very likely that they are neither selected for, nor against. Consequently, 'passenger' mutations are very likely to be randomly distributed along the sequence of the cancer genes. On the other hand, 'driver' mutations confer growth and/or a proliferative advantage during clonal expansion and as such are very likely to be selected for. Identifying such mutations plays "a central role of cancer genome analysis" (Stratton et al. 2009). As a consequence, a great deal of effort by the whole cancer genetics community has been put into distinguishing such 'drivers' from 'passenger' mutations. Functional assays could help, by providing an insight into the functional consequences of a particular mutation, but are expensive (e.g. in terms of cost and labour) or do not exist for every mutation (Chan et al. 2007). As a consequence, a number of *in silico* algorithms have been designed, which are relatively inexpensive, to help to identify 'passenger' from 'drivers' mutations. These algorithms involve a number of parameters, such as physicochemical difference between wild-type and mutant amino acids, evolutionary conservation of affected codons, etc.

Our analysis, using some of these parameters suggests that some mutations are indeed more likely to have greater functional consequences than others. Thus, shared missense mutations displayed relatively greater functional impact than either somatic or germline missense mutations. In addition, germline missense mutations were more likely to confer more severe functional consequences than somatic mutations. One could rank these classes of mutations, with respect to their functional importance during tumour development. Shared missense mutations are more likely to be 'drivers' of tumorigenesis than germline mutations, which in turn are more likely to be of functional importance than somatic missense mutations. These results are further supported by studies showing that some or most of the somatic micro-lesions found in cancer genes are indeed 'passenger' mutations (Greenman et al. 2007).

On the other hand, nonsense mutations are generally considered to be detrimental to the gene function. Our analysis suggests that some additional factors play a role in determining the impact of nonsense mutations and not all mutations will have a similar functional impact. Nonsense mediated mRNA decay could be attributed to one of these factors. In most cases, mRNAs harbouring a stop codon would be degraded and their consequence would be similar to a null allele. Nevertheless, some nonsense mutations could escape degradation (e.g. nonsense mutations in the last exon or in close proximity to the AUG initiation codon), leading to a truncated protein product. Furthermore, these truncated products in some instances could be functionally intact (assuming stable and functional mRNA). Indeed, such mutations have been shown for the *RB1* and *BRCA1* genes (Buisson et al. 2006; Sanchez-Sanchez et al. 2007). In addition, it is likely that selection and functional characteristics of individual genes (e.g. the position of β-catenin binding sites in the *APC* gene) could also play an important role in distinguishing the functional importance of nonsense mutations. Similar reasoning could be applied to the functional consequences of micro-lesions (i.e. deletions, insertions and indels), where a stop codon, on average ~15 codons following a frameshift, is the likely consequence (Itzkovitz and Alon 2007).

This PhD work is particularly timely with respect to the whole-genome sequencing of cancer genomes (Greenman et al. 2007; Stratton et al. 2009). Thus, special attention has to be attributed to some classes or types of mutations (e.g. shared mutations), as these would be more likely to be relevant in tumour development and/or progression than others, at least within tumour suppressor genes.

7.7. Shortcomings

With hindsight, some shortcomings of the work have become evident.

The power analysis for all tests performed suggested that some of the comparisons had insufficient statistical power to detect an experiment-wise significant result. For some of these tests, a relatively small effect size was noted (i.e. <0.1). For others, the effect size was relatively large, but due to the paucity of mutations the comparisons did not reach statistical significance. It has to be said that ~50% of all somatic mutations were derived from the *TP53* gene and the *TP53* mutations comprised ~25% of all mutations for all genes. Therefore, with hindsight, the analyses performed in this PhD work should perhaps have been performed with a smaller number of genes, thereby reducing the number of tests performed. Such a scenario would have been more likely to yield greater statistical power; hence for some of the tests, conclusions could perhaps have been derived where it was not possible in this analysis. Nevertheless, the presented comparison of somatic and germline micro-lesions is the first of its kind and as such we could only make power calculations *post hoc* or retrospectively, rather than *a priori*. Thus, the power calculations and effect sizes could be used as a basis for future comparison of somatic and germline micro-lesions.

In Chapter 4, a number of parameters were used. These included Grantham physicochemical difference, disease and non-disease associated mutability rates and evolutionary conservation measures. The Grantham amino-acid difference, also known as the amino-acid physicochemical difference (Grantham 1974), is based on the difference of side-chain atomic composition, polarity and volume of two amino acids. Based only on a few parameters, the Grantham difference is likely to be an oversimplification. Indeed, other mathematically more complex scores have been created, such as MAPP (multivariate analysis of protein polymorphism; Stone and Sidow 2005) that combines amino-acid properties with multiple sequence alignments. Additionally, the AAindex database (http://www.genome.jp/dbget-bin/www_bfind?aaindex) has >500 indices that describe the individual amino acids (Kawashima et al. 2008). Thus, more accurate representation of amino-acid differences could be derived, although it would most likely involve mathematically and/or statistically more complicated calculations.

In a similar way to the Grantham difference, both the disease (Krawczak et al. 1998) and non-disease (Hess et al. 1994) associated nucleotide substitution rates used in the analysis of missense mutations are not likely to be completely accurate in representing the relative substitution rates. Some of the disease-associated single base-pair substitutions used by Krawczak et al. (1998) could be neutral or nearly neutral mutations with respect to functional/clinical importance. In addition, the non-disease associated substitution rates derived by Hess et al. (1994) were calculated using 311 aligned gene-pseudogene pairs of sequences. Sequence changes in pseudogene are commonly assumed to be biologically neutral. Nevertheless, some studies have suggested that some pseudogenes might be under some form of purifying selection; hence some of these pseudogenes could bear functional importance. Indeed, a pseudogene in mice (*Oct4* pseudogene; Lin et al. 2007) has been shown to be functional (i.e. expressed as mRNA) and displays a function in stem cell regulation. Therefore, some of the sequence changes in the gene-pseudogene pairs used in the calculation of the non-disease nucleotide substitution rates could potentially harbour changes that are non-neutral and bear functional importance or clinical relevance.

Only orthologous sequences from vertebrate species were used to derive evolutionary conservation estimates for every codon along the sequences of the 17 human tumour suppressor genes. Only sequences from vertebrate species were chosen to avoid amino-acid variation that could be correlated with functional differences of the associated products from relatively more phylogenetically distant species from humans (Miller and Kumar 2001). Thus, limiting ourselves to vertebrate species could have introduced chance variation into the values for the evolutionary variation measure. As a consequence, some codons would appear as evolutionarily conserved solely because of the limited number of sequences sampled. Calculations for the *BRCA1* gene suggest that when 7 species were used (mammals through fish; Abkevich et al. 2004; Tavtigian et al. 2008), up to 1 in 4 of the fully evolutionarily conserved positions may be invariant, simply due to chance.

The repetitive elements searched for in the extended cDNA sequences of the 17 human tumour suppressor genes studied, were defined *a priori* based on existing evidence that showed that these were the most common sequence elements found in and around the breakpoints of, and most likely to mediate the occurrence of, micro-lesions. During the micro-lesion analysis, no micro-lesions were noted in any of the discovered C/G-quartets. It is quite likely that most of these C/G quartets do not play a major role in mediating the occurrence of at least the studied micro-lesions that is micro-deletions, micro-insertions and micro-indels. Further, it is likely some potential micro-lesions could be positioned in these C/G-quartets and be counted as within, or in the vicinity of, repeats; hence could have contributed towards the observed results. Nevertheless, if we had *post factum* excluded those from the initial hypothesis and subsequently re-analysed the micro-lesions, it would have been tailoring the initial hypothesis according to the observed results. Therefore, those C/G-quartets were left in the group of repetitive elements studied, although it is likely that those could have slightly increased the proportion of potential micro-lesions found in or in the vicinity of repeats.

By analysing all repetitive elements and all observed micro-lesions in one homogenous sample, some important characteristics of specific sequence elements, such as mediating different type of micro-lesions, could have been missed. Thus, the different types of repetitive elements, that is direct, inverted and mirror repeats, could mediate and/or influence the occurrence of the different types of micro-lesions (i.e. micro-deletions, micro-insertions and micro-indels) with a different propensity. However, splitting the repetitive elements and micro-lesions into their counterparts, and analysing them separately, would have introduced 9 times as many statistical tests (3 types of repeats times 3 types of micro-lesions). This undoubtedly would have decreased the statistical power even further, although the effect sizes could well have been greater, thereby compensating for the increased number of statistical tests.

It has to be said that not all possible repetitive elements were analysed with respect to micro-lesions. It is possible that some other types of sequence elements are also likely to be involved in mediating the occurrence of micro-lesions. These sequence elements include di- and tri-nucleotide tandem repeats, various motifs (e.g. heptanucleotides CCCCCTG, TGRRKM, etc.), micro-indel hotspots (GTAAGT and its complement ACTTAC), etc. (Bacolla et al. 2004; Bacolla and Wells 2009; Ball et al. 2005; Cooper and Krawczak 1993; Wells 2007). As these sequence elements are not part of those most commonly found in or around breakpoints of micro-lesions, it is likely that their effect size is quite possibly smaller than the repetitive elements analysed here. Nevertheless, some of these elements could be involved in mediating the occurrence of micro-lesions.

With respect to the NMD analysis, the '55-nucleotide rule' (Nagy and Maquat 1998) is quite possibly oversimplified and is not representative of the biological fate of some of the mRNAs harbouring nonsense mutations. Mechanisms, such as translational re-initiation (Sanchez-Sanchez et al. 2007), polar effect (Wang et al. 2002), failure to trigger NMD (Inacio et al. 2004), are very likely to have been missed, by using the '55-nucleotide rule' (Nagy and Maquat 1998). It is of note that these mechanisms are very likely to be exceptions and have to be treated on an individual basis. Further, with advances in our understanding of NMD and similar mechanisms, a more complete picture is likely to emerge that will better describe the mechanisms that operate to influence the functionality of the protein products of genes harbouring premature stop codons.

7.8. Future work

Any study that sets out to compare mutational mechanisms underlying germline and somatic mutational spectra has to have a sufficient number of mutations. This PhD work was initiated at the end of 2005. Since then, numerous studies have reported additional mutations (both somatic and germline) that have been observed in tumour samples. As a consequence, any subsequent studies are likely to have increased statistical power (i.e. >80%) with which to derive relatively conclusive results. This applies especially to micro-lesions, that is micro-deletions, micro-insertions and micro-indels.

Recent, along with past, research suggests that soma-germline mutational interplay (Campbell 2009; Lamlum et al. 1999; Vortmeyer et al. 2002) could well be more widespread than anticipated. Such intricate interplay happens within the individuals, hence mutational spectra that is derived from different tumours (i.e. within separate individuals) and normal cells/tissues, is only a proxy to the actual mechanisms that operate within the cells of individual people. Therefore, one way to investigate the actual relationships is to have matched mutational data (i.e. somatic and germline mutational data) from tumours and normal cells/tissues from the same individual. A variation of such matched mutational data also includes a somatic bi-allelic inactivation. This would allow a direct comparison between the germline and the soma, with respect to relative impact on gene/protein structure and/or function.

A different aspect of the functional impact of somatic and germline mutations could involve secondary structure analysis of the affected protein products and mRNA. It is likely that different mutations would exhibit different effects on the secondary structure of proteins (Ng and Henikoff 2001; Wang and Moult 2001). Some amino-acids are crucial for the protein stability, such as 'buried residues' (Chen and Zhou 2005), whereas others, even large deletions, could be relatively tolerated within the protein (Khan and Vihinen 2007). Many of the proteins are fully functional in dimers or multimers, such as the *TP53* gene (Webber et al. 2009). Therefore, mutations in different positions would potentially have different functional consequences, with respect to protein folding and secondary structure.

Further to the secondary structure analysis, certain parts of the proteins (e.g. α-helices and β-sheets; Bhattacharjee and Biswas 2009) could be used to define clustering of mutations. Thus, mutations could be classified on their functional impact on the protein function, through disruption of secondary or tertiary structures.

Overall, the work presented herein has shown a glimpse of the similarities and differences between the germline and somatic mutational spectrum in the 17 human tumour suppressor genes studied. Some differences were found to be gene-specific, but some shared by all genes. In addition, the similarities that were found are likely to reflect mutational mechanisms that are shared between the germline and the soma.

Table 45 Summary of the mutations in the 17 human tumour suppressor genes studied

		Mutations																							
		Missense				Nonsense				Micro-deletions				Micro-insertions				Micro-indels				Total	Truncating		
		N_M	$\frac{N_M}{T}$	$\frac{N_M}{N_O}$	$\frac{N_M}{N_{GT}}$	N_N	$\frac{N_N}{T}$	$\frac{N_N}{N_O}$	$\frac{N_N}{N_{GT}}$	N_D	$\frac{N_D}{T}$	$\frac{N_D}{N_O}$	$\frac{N_D}{N_{GT}}$	N_{IS}	$\frac{N_{IS}}{T}$	$\frac{N_{IS}}{N_O}$	$\frac{N_{IS}}{N_{GT}}$	N_{ID}	$\frac{N_{ID}}{T}$	$\frac{N_{ID}}{N_O}$	$\frac{N_{ID}}{N_{GT}}$	N_O	$\frac{N_O}{N_{GT}}$	N_T	$\frac{N_T}{N_O}$
APC	Somatic	38	62.30	13.92	4.10	51	22.08	18.68	5.51	137	31.42	50.18	14.79	44	27.67	16.12	4.75	3	7.69	1.10	0.32	273	29.48	235	13.92
	Germline	22	36.07	3.61	2.38	152	65.80	24.96	16.41	284	65.14	46.63	30.67	115	72.33	18.88	12.42	36	92.31	5.91	3.89	609	65.77	587	3.61
	Shared	1	1.64	2.27	0.11	28	12.12	63.64	3.02	15	3.44	34.09	1.62	0	0.00	0.00	0.00	0	0.00	0.00	0.00	44	4.75	43	2.27
	Recurrent	4	10.26	5.26	0.43	35	44.30	46.05	3.78	33	21.71	43.42	3.56	4	9.09	5.26	0.43	0	0.00	0.00	0.00	76	27.84	72	5.26
	Total (T)	61	100.00	6.59	6.59	231	100.00	24.95	24.95	436	100.00	47.08	47.08	159	100.00	17.17	17.17	39	100.00	4.21	4.21	926 (N_{GT})		865	6.59
ATM	Somatic	11	12.79	55.00	3.23	4	5.06	20.00	1.17	4	3.17	20.00	1.17	1	2.78	5.00	0.29	0	0.00	0.00	0.00	20	5.87	9	55.00
	Germline	75	87.21	23.58	21.99	72	91.14	22.64	21.11	122	96.83	38.36	35.78	35	97.22	11.01	10.26	14	100.00	4.40	4.11	318	93.26	243	23.58
	Shared	0	0.00	0.00	0.00	3	3.80	100.00	0.88	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	3	0.88	3	0.00
	Recurrent	1	9.09	100.00	0.29	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	N/A	0.00	0.00	1	5.00	0	100.00
	Total (T)	86	100.00	25.22	25.22	79	100.00	23.17	23.17	126	100.00	36.95	36.95	36	100.00	10.56	10.56	14	100.00	4.11	4.11	341 (N_{GT})		255	25.22
BRCA1	Somatic	5	2.86	29.41	0.75	3	2.42	17.65	0.45	6	2.27	35.29	0.90	3	3.41	17.65	0.45	0	0.00	0.00	0.00	17	2.56	12	29.41
	Germline	169	96.57	26.66	25.49	115	92.74	18.14	17.35	255	96.59	40.22	38.46	83	94.32	13.09	12.52	12	100.00	1.89	1.81	634	95.63	465	26.66
	Shared	1	0.57	8.33	0.15	6	4.84	50.00	0.90	3	1.14	25.00	0.45	2	2.27	16.67	0.30	0	0.00	0.00	0.00	12	1.81	11	8.33
	Recurrent	0	0.00	0.00	0.00	0	0.00	0.00	0.00	2	22.22	100.00	0.30	0	0.00	0.00	0.00	0	#DIV/0!	0.00	0.00	2	6.90	2	0.00
	Total (T)	175	100.00	26.40	26.40	124	100.00	18.70	18.70	264	100.00	39.82	39.82	88	100.00	13.27	13.27	12	100.00	1.81	1.81	663 (N_{GT})		488	26.40
BRCA2	Somatic	20	18.87	66.67	3.72	1	1.30	3.33	0.19	7	2.78	23.33	1.30	2	2.17	6.67	0.37	0	0.00	0.00	0.00	30	5.59	10	66.67
	Germline	85	80.19	16.90	15.83	76	98.70	15.11	14.15	244	96.83	48.51	45.44	88	95.65	17.50	16.39	10	100.00	1.99	1.86	503	93.67	418	16.90
	Shared	1	0.94	25.00	0.19	0	0.00	0.00	0.00	1	0.40	25.00	0.19	2	2.17	50.00	0.37	0	0.00	0.00	0.00	4	0.74	3	25.00
	Recurrent	2	9.52	66.67	0.37	0	0.00	0.00	0.00	1	12.50	33.33	0.19	0	0.00	0.00	0.00	0	N/A	0.00	0.00	3	8.82	1	66.67
	Total (T)	106	100.00	19.74	19.74	77	100.00	14.34	14.34	252	100.00	46.93	46.93	92	100.00	17.13	17.13	10	100.00	1.86	1.86	537 (N_{GT})		431	19.74
CDH1	Somatic	14	42.42	41.18	16.47	5	31.25	14.71	5.88	13	52.00	38.24	15.29	2	20.00	5.88	2.35	0	0.00	0.00	0.00	34	40.00	20	41.18
	Germline	18	54.55	37.50	21.18	9	56.25	18.75	10.59	12	48.00	25.00	14.12	8	80.00	16.67	9.41	1	100.00	2.08	1.18	48	56.47	30	37.50
	Shared	1	3.03	33.33	1.18	2	12.50	66.67	2.35	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	3	3.53	2	33.33
	Recurrent	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	N/A	N/A	0.00	0	0.00	0	N/A
	Total (T)	33	100.00	38.82	38.82	16	100.00	18.82	18.82	25	100.00	29.41	29.41	10	100.00	11.76	11.76	1	100.00	1.18	1.18	85 (N_{GT})		52	38.82
CDKN2A	Somatic	170	73.28	58.42	44.74	13	65.00	4.47	3.42	76	87.36	26.12	20.00	24	77.42	8.25	6.32	8	80.00	2.75	2.11	291	76.58	121	58.42

	Germline	34	14.66	62.96	8.95	2	10.00	3.70	0.53	10	11.49	18.52	2.63	6	19.35	11.11	1.58	2	20.00	3.70	0.53	54	14.21	20	62.96
	Shared	28	12.07	80.00	7.37	5	25.00	14.29	1.32	1	1.15	2.86	0.26	1	3.23	2.86	0.26	0	0.00	0.00	0.00	35	9.21	7	80.00
	Recurrent	6	3.03	28.57	1.58	3	16.67	14.29	0.79	9	11.69	42.86	2.37	3	12.00	14.29	0.79	0	0.00	0.00	0.00	21	6.44	15	28.57
	Total (*T*)	232	100.00	61.05	61.05	20	100.00	5.26	5.26	87	100.00	22.89	22.89	31	100.00	8.16	8.16	10	100.00	2.63	2.63	380 (N_{GT})		148	61.05
	Somatic	2	2.35	9.09	0.36	4	3.36	18.18	0.72	13	5.56	59.09	2.35	3	2.78	13.64	0.54	0	0.00	0.00	0.00	22	3.97	20	9.09
	Germline	83	97.65	15.99	14.98	105	88.24	20.23	18.95	218	93.16	42.00	39.35	105	97.22	20.23	18.95	8	100.00	1.54	1.44	519	93.68	436	15.99
NF1	Shared	0	0.00	0.00	0.00	10	8.40	76.92	1.81	3	1.28	23.08	0.54	0	0.00	0.00	0.00	0	0.00	0.00	0.00	13	2.35	13	0.00
	Recurrent	0	0.00	0.00	0.00	0	0.00	0.00	0.00	1	6.25	100.00	0.18	0	0.00	0.00	0.00	0	N/A	0.00	0.00	1	2.86	1	0.00
	Total (*T*)	85	100.00	15.34	15.34	119	100.00	21.48	21.48	234	100.00	42.24	42.24	108	100.00	19.49	19.49	8	100.00	1.44	1.44	554 (N_{GT})		469	15.34
	Somatic	23	53.49	8.95	5.85	24	35.82	9.34	6.11	176	76.19	68.48	44.78	28	63.64	10.89	7.12	6	75.00	2.33	1.53	257	65.39	234	8.95
	Germline	20	46.51	17.70	5.09	25	37.31	22.12	6.36	50	21.65	44.25	12.72	16	36.36	14.16	4.07	2	25.00	1.77	0.51	113	28.75	93	17.70
NF2	Shared	0	0.00	0.00	0.00	18	26.87	78.26	4.58	5	2.16	21.74	1.27	0	0.00	0.00	0.00	0	0.00	0.00	0.00	23	5.85	23	0.00
	Recurrent	3	13.04	9.68	0.76	18	42.86	58.06	4.58	8	4.42	25.81	2.04	2	7.14	6.45	0.51	0	0.00	0.00	0.00	31	11.07	28	9.68
	Total (*T*)	43	100.00	10.94	10.94	67	100.00	17.05	17.05	231	100.00	58.78	58.78	44	100.00	11.20	11.20	8	100.00	2.04	2.04	393 (N_{GT})		350	10.94
	Somatic	13	35.14	31.71	7.47	7	20.59	17.07	4.02	14	25.00	34.15	8.05	6	15.79	14.63	3.45	1	11.11	2.44	0.57	41	23.56	28	31.71
	Germline	23	62.16	17.69	13.22	25	73.53	19.23	14.37	42	75.00	32.31	24.14	32	84.21	24.62	18.39	8	88.89	6.15	4.60	130	74.71	107	17.69
PTCH	Shared	1	2.70	33.33	0.57	2	5.88	66.67	1.15	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	3	1.72	2	33.33
	Recurrent	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	0	N/A
	Total (*T*)	37	100.00	21.26	21.26	34	100.00	19.54	19.54	56	100.00	32.18	32.18	38	100.00	21.84	21.84	9	100.00	5.17	5.17	174 (N_{GT})		137	21.26
	Somatic	201	81.71	45.68	35.45	43	60.56	9.77	7.58	145	83.33	32.95	25.57	47	68.12	10.68	8.29	4	57.14	0.91	0.71	440	77.60	239	45.68
	Germline	23	9.35	28.05	4.06	15	21.13	18.29	2.65	23	13.22	28.05	4.06	18	26.09	21.95	3.17	3	42.86	3.66	0.53	82	14.46	59	28.05
PTEN	Shared	22	8.94	48.89	3.88	13	18.31	28.89	2.29	6	3.45	13.33	1.06	4	5.80	8.89	0.71	0	0.00	0.00	0.00	45	7.94	23	48.89
	Recurrent	47	21.08	37.01	8.29	19	33.93	14.96	3.35	45	29.80	35.43	7.94	16	31.37	12.60	2.82	0	0.00	0.00	0.00	127	26.19	80	37.01
	Total (*T*)	246	100.00	43.39	43.39	71	100.00	12.52	12.52	174	100.00	30.69	30.69	69	100.00	12.17	12.17	7	100.00	1.23	1.23	567 (N_{GT})		321	43.39
	Somatic	22	37.29	28.21	5.93	12	13.64	15.38	3.23	30	20.55	38.46	8.09	12	18.46	15.38	3.23	2	15.38	2.56	0.54	78	21.02	56	28.21
	Germline	34	57.63	12.59	9.16	61	69.32	22.59	16.44	112	76.71	41.48	30.19	53	81.54	19.63	14.29	10	76.92	3.70	2.70	270	72.78	236	12.59
RB1	Shared	3	5.08	13.04	0.81	15	17.05	65.22	4.04	4	2.74	17.39	1.08	0	0.00	0.00	0.00	1	7.69	4.35	0.27	23	6.20	20	13.04
	Recurrent	1	4.00	8.33	0.27	9	33.33	75.00	2.43	2	5.88	16.67	0.54	0	0.00	0.00	0.00	0	0.00	0.00	0.00	12	11.88	11	8.33
	Total (*T*)	59	100.00	15.90	15.90	88	100.00	23.72	23.72	146	100.00	39.35	39.35	65	100.00	17.52	17.52	13	100.00	3.50	3.50	371 (N_{GT})		312	15.90
STK11	Somatic	17	36.17	60.71	10.69	7	20.59	25.00	4.40	3	6.00	10.71	1.89	1	4.00	3.57	0.63	0	0.00	0.00	0.00	28	17.61	11	60.71

	Germline	27	57.45	22.13	16.98	24	70.59	19.67	15.09	45	90.00	36.89	28.30	24	96.00	19.67	15.09	2	66.67	1.64	1.26	122	76.73	95	22.13
	Shared	3	6.38	33.33	1.89	3	8.82	33.33	1.89	2	4.00	22.22	1.26	0	0.00	0.00	0.00	1	33.33	11.11	0.63	9	5.66	6	33.33
	Recurrent	2	10.00	50.00	1.26	1	10.00	25.00	0.63	1	20.00	25.00	0.63	0	0.00	0.00	0.00	0	N/A	0.00	0.00	4	10.81	2	50.00
	Total (*T*)	47	100.00	29.56	29.56	34	100.00	21.38	21.38	50	100.00	31.45	31.45	25	100.00	15.72	15.72	3	100.00	1.89	1.89	159 (N_{GT})		112	29.56
TP53	Somatic	1138	92.37	57.97	54.37	87	89.69	4.43	4.16	504	96.92	25.67	24.08	234	97.10	11.92	11.18	0	0.00	0.00	0.00	1963	93.79	825	57.97
	Germline	6	0.49	28.57	0.29	1	1.03	4.76	0.05	8	1.54	38.10	0.38	3	1.24	14.29	0.14	3	100.00	14.29	0.14	21	1.00	15	28.57
	Shared	88	7.14	80.73	4.20	9	9.28	8.26	0.43	8	1.54	7.34	0.38	4	1.66	3.67	0.19	0	0.00	0.00	0.00	109	5.21	21	80.73
	Recurrent	781	63.70	71.98	37.31	85	88.54	7.83	4.06	162	31.64	14.93	7.74	57	23.95	5.25	2.72	0	N/A	0.00	0.00	1085	52.36	304	71.98
	Total (*T*)	1232	100.00	58.86	58.86	97	100.00	4.63	4.63	520	100.00	24.84	24.84	241	100.00	11.51	11.51	3	100.00	0.14	0.14	2093 (N_{GT})		861	58.86
TSC1	Somatic	2	22.22	50.00	1.54	1	2.63	25.00	0.77	1	1.85	25.00	0.77	0	0.00	0.00	0.00	0	0.00	0.00	0.00	4	3.08	2	50.00
	Germline	7	77.78	5.56	5.38	37	97.37	29.37	28.46	53	98.15	42.06	40.77	25	100.00	19.84	19.23	4	100.00	3.17	3.08	126	96.92	119	5.56
	Shared	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	0	N/A
	Recurrent	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	N/A	N/A	0.00	0	N/A	N/A	0.00	0	0.00	0	N/A
	Total (*T*)	9	100.00	6.92	6.92	38	100.00	29.23	29.23	54	100.00	41.54	41.54	25	100.00	19.23	19.23	4	100.00	3.08	3.08	130 (N_{GT})		121	6.92
TSC2	Somatic	0	0.00	0.00	0.00	0	0.00	0.00	0.00	3	2.65	50.00	0.91	2	4.17	33.33	0.61	1	25.00	16.67	0.30	6	1.83	6	0.00
	Germline	87	97.75	27.27	26.52	73	98.65	22.88	22.26	110	97.35	34.48	33.54	46	95.83	14.42	14.02	3	75.00	0.94	0.91	319	97.26	232	27.27
	Shared	2	2.25	66.67	0.61	1	1.35	33.33	0.30	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	3	0.91	1	66.67
	Recurrent	0	N/A	N/A	0.00	0	N/A	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	0.00	0	N/A
	Total (*T*)	89	100.00	27.13	27.13	74	100.00	22.56	22.56	113	100.00	34.45	34.45	48	100.00	14.63	14.63	4	100.00	1.22	1.22	328 (N_{GT})		239	27.13
VHL	Somatic	43	23.12	16.73	8.08	4	12.90	1.56	0.75	171	73.08	66.54	32.14	38	50.67	14.79	7.14	1	16.67	0.39	0.19	257	48.31	214	16.73
	Germline	98	52.69	47.80	18.42	16	51.61	7.80	3.01	55	23.50	26.83	10.34	31	41.33	15.12	5.83	5	83.33	2.44	0.94	205	38.53	107	47.80
	Shared	45	24.19	64.29	8.46	11	35.48	15.71	2.07	8	3.42	11.43	1.50	6	8.00	8.57	1.13	0	0.00	0.00	0.00	70	13.16	25	64.29
	Recurrent	5	0.12	20.00	0.94	4	1.00	16.00	0.75	14	0.08	56.00	2.63	2	0.05	8.00	0.38	0	0.00	0.00	0.00	25	9.73	20	20.00
	Total (*T*)	186	100.00	34.96	34.96	31	100.00	5.83	5.83	234	100.00	43.98	43.98	75	100.00	14.10	14.10	6	100.00	1.13	1.13	532 (N_{GT})		346	34.96
WT1	Somatic	1	2.50	12.50	1.35	0	0.00	0.00	0.00	4	33.33	50.00	5.41	3	42.86	37.50	4.05	0	0.00	0.00	0.00	8	10.81	7	12.50
	Germline	39	97.50	61.90	52.70	11	78.57	17.46	14.86	8	66.67	12.70	10.81	4	57.14	6.35	5.41	1	100.00	1.59	1.35	63	85.14	24	61.90
	Shared	0	0.00	0.00	0.00	3	21.43	100.00	4.05	0	0.00	0.00	0.00	0	0.00	0.00	0.00	0	0.00	0.00	0.00	3	4.05	3	0.00
	Recurrent	0	0.00	N/A	0.00	0	N/A	N/A	0.00	0	0.00	N/A	0.00	0	0.00	N/A	0.00	0	N/A	N/A	0.00	0	0.00	0	N/A
	Total (*T*)	40	100.00	54.05	54.05	14	100.00	18.92	18.92	12	100.00	16.22	16.22	7	100.00	9.46	9.46	1	100.00	1.35	1.35	74 (N_{GT})		34	54.05
ALL	Somatic	1720	62.18	45.55	20.69	273	22.36	7.23	3.28	1307	43.36	34.61	15.72	450	38.76	11.92	5.41	26	17.11	0.69	0.31	3776	45.42	2056	45.55

Germline	850	30.73	20.55	10.22	819	67.08	19.80	9.85	1651	54.78	39.92	19.86	692	59.60	16.73	8.32	124	81.58	3.00	1.49	4136	49.75	3286	20.55
Shared	196	7.09	48.76	2.36	129	10.57	32.09	1.55	56	1.86	13.93	0.67	19	1.64	4.73	0.23	2	1.32	0.50	0.02	402	4.84	206	48.76
Recurrent	852	44.47	61.38	10.25	174	43.28	12.54	2.09	278	20.40	20.03	3.34	84	17.91	6.05	1.01	0	0.00	0.00	0.00	1388	33.22	536	61.38
Total (T)	2766	100.00	33.27	33.27	1221	100.00	14.69	14.69	3014	100.00	36.25	36.25	1161	100.00	13.96	13.96	152	100.00	1.83	1.83	8314 (N_{GT})		5548	33.27

N_i is the number of micro-lesions, where $i \in \{M,N,D,IS,ID,O,GT\}$, and M-missense, N-nonsense , D-micro-deletions IS-micro-insertions, ID-micro-indels, O-sum of somatic, germline, shared or recurrent, GT-grand total

T-total (somatic, germline and shared)

All calculations for recurrent mutations are with respect to the numbers of somatic and shared mutations

Table 46 Summary of the CpG-located missense and nonsense mutations in the 17 human tumour suppressor genes studied

		Missense	Missense within CpG-dinucleotides		Nonsense	Nonsense within CpG-dinucleotides	
		N_M	N_{MCpG}	$\frac{N_{MCpG}}{N_M}$	N_N	N_{NCpG}	$\frac{N_{NCpG}}{N_N}$
APC	Somatic	38	1	2.63	51	0	0.00
	Germline	22	3	13.64	152	8	5.26
	Shared	1	0	0.00	28	8	28.57
	Recurrent	4	0	0.00	35	6	17.14
	Total (*T*)	61	4	6.56	231	16	6.93
ATM	Somatic	11	1	9.09	4	0	0.00
	Germline	75	6	8.00	72	16	22.22
	Shared	0	0	N/A	3	2	66.67
	Recurrent	1	1	100.00	0	0	N/A
	Total (*T*)	86	7	8.14	79	18	22.78
BRCA1	Somatic	5	0	0.00	3	0	0.00
	Germline	169	9	5.33	115	3	2.61
	Shared	1	0	0.00	6	1	16.67
	Recurrent	0	0	N/A	0	0	N/A
	Total (*T*)	175	9	5.14	124	4	3.23
BRCA2	Somatic	20	2	10.00	1	0	0.00
	Germline	85	9	10.59	76	4	5.26
	Shared	1	1	100.00	0	0	N/A
	Recurrent	2	0	0.00	0	0	N/A
	Total (*T*)	106	12	11.32	77	4	5.19
CDH1	Somatic	14	1	7.14	5	0	0.00
	Germline	18	4	22.22	9	2	22.22
	Shared	1	0	0.00	2	1	50.00
	Recurrent	0	0	N/A	0	0	N/A
	Total (*T*)	33	5	15.15	16	3	18.75
CDKN2A	Somatic	170	26	15.29	13	0	0.00
	Germline	34	2	5.88	2	0	0.00
	Shared	28	7	25.00	5	2	40.00
	Recurrent	6	0	0.00	3	1	33.33
	Total (*T*)	232	35	15.09	20	2	10.00
NF1	Somatic	2	0	0.00	4	0	0.00
	Germline	83	4	4.82	105	9	8.57
	Shared	0	0	N/A	10	9	90.00
	Recurrent	0	0	N/A	0	0	N/A
	Total (*T*)	85	4	4.71	119	18	15.13
NF2	Somatic	23	2	8.70	24	0	0.00
	Germline	20	2	10.00	25	0	0.00
	Shared	0	0	N/A	18	6	33.33
	Recurrent	3	1	33.33	18	6	33.33
	Total (*T*)	43	4	9.30	67	6	8.96
PTCH	Somatic	13	2	15.38	7	1	14.29
	Germline	23	2	8.70	25	2	8.00
	Shared	1	0	0.00	2	0	0.00
	Recurrent	0	0	N/A	0	0	N/A
	Total (*T*)	37	4	10.81	34	3	8.82

PTEN	Somatic	201	4	1.99	43	0	0.00
	Germline	23	1	4.35	15	0	0.00
	Shared	22	3	13.64	13	3	23.08
	Recurrent	47	3	6.38	19	3	15.79
	Total (T)	246	8	3.25	71	3	4.23
RB1	Somatic	22	3	13.64	12	0	0.00
	Germline	34	2	5.88	61	4	6.56
	Shared	3	1	33.33	15	7	46.67
	Recurrent	1	0	0.00	9	5	55.56
	Total (T)	59	6	10.17	88	11	12.50
STK11	Somatic	17	4	23.53	7	0	0.00
	Germline	27	2	7.41	24	1	4.17
	Shared	3	1	33.33	3	0	0.00
	Recurrent	2	1	50.00	1	0	0.00
	Total (T)	47	7	14.89	34	1	2.94
TP53	Somatic	1138	30	2.64	87	0	0.00
	Germline	6	0	0.00	1	0	0.00
	Shared	88	20	22.73	9	4	44.44
	Recurrent	781	23	2.94	85	4	4.71
	Total (T)	1232	50	4.06	97	4	4.12
TSC1	Somatic	2	0	0.00	1	0	0.00
	Germline	7	1	14.29	37	6	16.22
	Shared	0	0	N/A	0	0	N/A
	Recurrent	0	0	N/A	0	0	N/A
	Total (T)	9	1	11.11	38	6	15.79
TSC2	Somatic	0	0	N/A	0	0	N/A
	Germline	87	13	14.94	73	6	8.22
	Shared	2	2	100.00	1	0	0.00
	Recurrent	0	0	N/A	0	0	N/A
	Total (T)	89	15	16.85	74	6	8.11
VHL	Somatic	43	4	9.30	4	0	0.00
	Germline	98	2	2.04	16	0	0.00
	Shared	45	6	13.33	11	2	18.18
	Recurrent	5	0	0.00	4	1	25.00
	Total (T)	186	12	6.45	31	2	6.45
WT1	Somatic	1	0	0.00	0	0	N/A
	Germline	39	7	17.95	11	0	0.00
	Shared	0	0	N/A	3	3	100.00
	Recurrent	0	0	N/A	0	0	N/A
	Total (T)	40	7	17.50	14	3	21.43
ALL	Somatic	1720	80	4.65	273	1	0.37
	Germline	850	69	8.12	819	61	7.45
	Shared	196	41	20.92	129	48	37.21
	Recurrent	852	29	3.40	174	26	14.94
	Total (T)	2766	190	6.87	1221	110	9.01

${}_{i}CpG$ is the number of CpG-located mutations, where $i \in \{M,N\}$, and M-missense, N-nonsense

T-total (somatic, germline and shared)

All calculations for recurrent mutations are with respect to numbers of somatic and shared mutations

Bibliography

Abkevich, V. et al. 2004. Analysis of missense variation in human *BRCA1* in the context of interspecific sequence variation. *J Med Genet* 41(7), pp. 492-507.

Alba, M. M. and Guigo, R. 2004. Comparative analysis of amino acid repeats in rodents and humans. *Genome Res* 14(4), pp. 549-554.

Albuquerque, C. et al. 2002. The 'just-right' signaling model: *APC* somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade. *Hum Mol Genet* 11(13), pp. 1549-1560.

Allegrucci, C. et al. 2005. Epigenetics and the germline. *Reproduction* 129(2), pp. 137-149.

Alter, B. P. et al. 2007. Clinical and molecular features associated with biallelic mutations in *FANCD1/BRCA2*. *J Med Genet* 44(1), pp. 1-9.

Altman, D. G. 1991. *Practical statistics for medical research.* 1 ed. Chapman & Hall, p. 624.

Altschul, S. F. et al. 1990. Basic local alignment search tool. *J Mol Biol* 215(3), pp. 403-410.

Amrani, N. et al. 2004. A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature* 432(7013), pp. 112-118.

Anczukow, O. et al. 2008. Does the nonsense-mediated mRNA decay mechanism prevent the synthesis of truncated *BRCA1*, *CHK2*, and *p53* proteins? *Hum Mutat* 29(1), pp. 65-73.

Antoniou, A. C. et al. 2002. A comprehensive model for familial breast cancer incorporating *BRCA1*, *BRCA2* and other genes. *Br J Cancer* 86(1), pp. 76-83.

Anway, M. D. et al. 2005. Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* 308(5727), pp. 1466-1469.

Aretz, S. et al. 2004. Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the *APC* gene. *Hum Mutat* 24(5), pp. 370-380.

Assie, G. et al. 2008. Frequency of germline genomic homozygosity associated with cancer cases. *Jama* 299(12), pp. 1437-1445.

Azad, P. and Woodruff, R. C. 2006. Mutation and cloning efficiency. *Cloning Stem Cells* 8(4), pp. 237-239.

Bacolla, A. et al. 2001. *Pkd1* unusual DNA conformations are recognized by nucleotide excision repair. *J Biol Chem* 276(21), pp. 18597-18604.

Bacolla, A. et al. 2004. Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc Natl Acad Sci U S A* 101(39), pp. 14162-14167.

Bacolla, A. and Wells, R. D. 2009. Non-B DNA conformations as determinants of mutagenesis and human disease. *Mol Carcinog* 48(4), pp. 273-285.

Baker, K. E. and Parker, R. 2004. Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Curr Opin Cell Biol* 16(3), pp. 293-299.

Bakkenist, C. J. and Kastan, M. B. 2003. DNA damage activates *ATM* through intermolecular autophosphorylation and dimer dissociation. *Nature* 421(6922), pp. 499-506.

Baldwin, R. L. et al. 2000. *BRCA1* promoter region hypermethylation in ovarian carcinoma: a population-based study. *Cancer Res* 60(19), pp. 5329-5333.

Ball, E. V. et al. 2005. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat* 26(3), pp. 205-213.

Bamford, S. et al. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* 91(2), pp. 355-358.

Barbour, L. et al. 2006. Mutagenesis. *Methods Mol Biol* 313, pp. 121-127.

Barwell, J. et al. 2007. Lymphocyte radiosensitivity in *BRCA1* and *BRCA2* mutation carriers and implications for breast cancer susceptibility. *Int J Cancer* 121(7), pp. 1631-1636.

Becker, K. et al. 2008. Deletions of *BRCA1/2* and *p53* R248W gain-of-function mutation suggest impaired homologous recombination repair in fragile histidine triad-negative sebaceous gland carcinomas. *Br J Dermatol* 159(6), pp. 1282-1289.

Beerenwinkel, N. et al. 2007. Genetic progression and the waiting time to cancer. *PLoS Comput Biol* 3(11), p. e225.

Behm-Ansmant, I. et al. 2007. A conserved role for cytoplasmic poly(A)-binding protein 1 (*PABPC1*) in nonsense-mediated mRNA decay. *Embo J* 26(6), pp. 1591-1601.

Beroud, C. and Soussi, T. 1996. *APC* gene: database of germline and somatic mutations in human tumors and cell lines. *Nucleic Acids Res* 24(1), pp. 121-124.

Bertram, C. and Hass, R. 2008. Cellular responses to reactive oxygen species-induced DNA damage and aging. *Biol Chem* 389(3), pp. 211-220.

Bertram, G. et al. 2001. Endless possibilities: translation termination and stop codon recognition. *Microbiology* 147(Pt 2), pp. 255-269.

Besaratinia, A. et al. 2009. In vitro recapitulating of *TP53* mutagenesis in hepatocellular carcinoma associated with dietary aflatoxin B1 exposure. *Gastroenterology*.

Besaratinia, A. and Pfeifer, G. P. 2006. Investigating human cancer etiology by DNA lesion footprinting and mutagenicity analysis. *Carcinogenesis* 27(8), pp. 1526-1537.

Bhattacharjee, N. and Biswas, P. 2009. Structural patterns in alpha helices and beta sheets in globular proteins. *Protein Pept Lett* 16(8), pp. 953-960.

Bhuller, Y. and Wells, P. G. 2006. A developmental role for ataxia-telangiectasia mutated in protecting the embryo from spontaneous and phenytoin-enhanced embryopathies in culture. *Toxicol Sci* 93(1), pp. 156-163.

Blagosklonny, M. V. 2000. *p53* from complexity to simplicity: mutant *p53* stabilization, gain-of-function, and dominant-negative effect. *FASEB J* 14(13), pp. 1901-1907.

Bossi, L. and Ruth, J. R. 1980. The influence of codon context on genetic code translation. *Nature* 286(5769), pp. 123-127.

Bouffler, S. D. et al. 1995. Spontaneous and ionizing radiation-induced chromosomal abnormalities in *p53*-deficient mice. *Cancer Res* 55(17), pp. 3883-3889.

Bowater, R. P. and Wells, R. D. 2001. The intrinsically unstable life of DNA triplet repeats associated with human hereditary disorders. *Prog Nucleic Acid Res Mol Biol* 66, pp. 159-202.

Breen, A. P. and Murphy, J. A. 1995. Reactions of oxyl radicals with DNA. *Free Radic Biol Med* 18(6), pp. 1033-1077.

Brennan, P. et al. 2007. Uncommon *CHEK2* mis-sense variant and reduced risk of tobacco-related cancers: case control study. *Hum Mol Genet* 16(15), pp. 1794-1801.

Brosh, R. and Rotter, V. 2009. When mutants gain new powers: news from the mutant *p53* field. *Nat Rev Cancer*.

Buard, J. et al. 2000. Somatic versus germline mutation processes at minisatellite CEB1 (D2S90) in humans and transgenic mice. *Genomics* 65(2), pp. 95-103.

Buchholz, T. A. et al. 2002. Evidence of haplotype insufficiency in human cells containing a germline mutation in *BRCA1* or *BRCA2*. *Int J Cancer* 97(5), pp. 557-561.

Buhler, M. et al. 2006. EJC-independent degradation of nonsense immunoglobulin-mu mRNA depends on 3' UTR length. *Nat Struct Mol Biol* 13(5), pp. 462-464.

Buisson, M. et al. 2006. The 185delAG mutation (c.68_69delAG) in the *BRCA1* gene triggers translation reinitiation at a downstream AUG codon. *Hum Mutat* 27(10), pp. 1024-1029.

Burga, L. N. et al. 2009. Altered proliferation and differentiation properties of primary mammary epithelial cells from *BRCA1* mutation carriers. *Cancer Res* 69(4), pp. 1273-1278.

Bzymek, M. and Lovett, S. T. 2001. Evidence for two mechanisms of palindrome-stimulated deletion in *Escherichia coli:* single-strand annealing and replication slipped mispairing. *Genetics* 158(2), pp. 527-540.

Campbell, P. J. 2009. Somatic and germline genetics at the *JAK2* locus. *Nat Genet* 41(4), pp. 385-386.

Cao, D. and Parker, R. 2003. Computational modeling and experimental analysis of nonsense-mediated decay in yeast. *Cell* 113(4), pp. 533-545.

Carrington, M. et al. 1997. Novel alleles of the chemokine-receptor gene *CCR5*. *Am J Hum Genet* 61(6), pp. 1261-1267.

Carter, M. S. et al. 1995. A regulatory mechanism that detects premature nonsense codons in T-cell receptor transcripts in vivo is reversed by protein synthesis inhibitors in vitro. *J Biol Chem* 270(48), pp. 28995-29003.

Carvalho, M. et al. 2009. Analysis of a set of missense, frameshift, and in-frame deletion variants of *BRCA1*. *Mutat Res* 660(1-2), pp. 1-11.

Cetta, F. et al. 2000. Germline mutations of the *APC* gene in patients with familial adenomatous polyposis-associated thyroid carcinoma: results from a European cooperative study. *J Clin Endocrinol Metab* 85(1), pp. 286-292.

Chan, K. T. et al. 2003. In vitro aflatoxin B1-induced *p53* mutations. *Cancer Lett* 199(1), pp. 1-7.

Chan, P. A. et al. 2007. Interpreting missense variants: comparing computational methods in human disease genes *CDKN2A*, *MLH1*, *MSH2*, *MECP2*, and tyrosinase (*TYR*). *Hum Mutat* 28(7), pp. 683-693.

Chan, T. L. et al. 2006. Heritable germline epimutation of *MSH2* in a family with hereditary nonpolyposis colorectal cancer. *Nat Genet* 38(10), pp. 1178-1183.

Chan, W. M. and Poon, R. Y. 2007. The *p53* Isoform *Deltap53* lacks intrinsic transcriptional activity and reveals the critical role of nuclear import in dominant-negative activity. *Cancer Res* 67(5), pp. 1959-1969.

Chang, H. S. et al. 2006. Transgenerational epigenetic imprinting of the male germline by endocrine disruptor exposure during gonadal sex determination. *Endocrinology* 147(12), pp. 5524-5541.

Chang, Y.-F. et al. 2007. The Nonsense-Mediated Decay RNA Surveillance Pathway. *Annual Review of Biochemistry* 76(1).

Charames, G. S. and Bapat, B. 2003. Genomic instability and cancer. *Curr Mol Med* 3(7), pp. 589-596.

Cheadle, J. P. et al. 2002. Different combinations of biallelic *APC* mutation confer different growth advantages in colorectal tumours. *Cancer Res* 62(2), pp. 363-366.

Chen, H. and Zhou, H. X. 2005. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res* 33(10), pp. 3193-3199.

Chenevix-Trench, G. et al. 2002. Dominant negative *ATM* mutations in breast cancer families. *J Natl Cancer Inst* 94(3), pp. 205-215.

Cheung, L. W. et al. 2007. CpG/CpNpG motifs in the coding region are preferred sites for mutagenesis in the breast cancer susceptibility genes. *FEBS Lett* 581(24), pp. 4668-4674.

Chuzhanova, N. A. et al. 2003. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat* 21(1), pp. 28-44.

Clarke, A. R. et al. 1994. *p53* dependence of early apoptotic and proliferative responses within the mouse intestinal epithelium following gamma-irradiation. *Oncogene* 9(6), pp. 1767-1773.

Clarke, B. 1970. Selective constraints on amino-acid substitutions during the evolution of proteins. *Nature* 228(5267), pp. 159-160.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Second Edition ed. New York: Lawrence Erlbaum Associates, p. 567.

Collins, D. W. and Jukes, T. H. 1994. Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* 20(3), pp. 386-396.

Conti, E. and Izaurralde, E. 2005. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr Opin Cell Biol* 17(3), pp. 316-325.

Cooper, D. N. et al. 1995. *The nature and mechanisms of human gene mutation. In: The metabolic and molecular bases of inherited disease*. McGraw-Hill, New York.

Cooper, D. N. and Krawczak, M. 1991. Mechanisms of insertional mutagenesis in human genes causing genetic disease. *Hum Genet* 87(4), pp. 409-415.

Cooper, D. N. and Krawczak, M. 1993. Human gene mutation. p. 402.

Cooper, D. N. and Youssoufian, H. 1988. The CpG dinucleotide and human genetic disease. *Hum Genet* 78(2), pp. 151-155.

Corso, G. et al. 2007. Characterization of the P373L *E-cadherin* germline missense mutation and implication for clinical management. *Eur J Surg Oncol* 33(9), pp. 1061-1067.

Coulondre, C. et al. 1978. Molecular basis of base substitution hotspots in *Escherichia coli. Nature* 274, pp. 775-780.

Crabtree, M. et al. 2003. Refining the relation between 'first hits' and 'second hits' at the *APC* locus: the 'loose fit' model and evidence for differences in somatic mutation spectra among patients. *Oncogene* 22(27), pp. 4257-4265.

Crane, R. et al. 2004. Requirements for the destruction of human Aurora-A. *J Cell Sci* 117(Pt 25), pp. 5975-5983.

Crick, F. H. et al. 1961. General nature of the genetic code for proteins. *Nature* 192, pp. 1227-1232.

Critchlow, S. E. and Jackson, S. P. 1998. DNA end-joining: from yeast to man. *Trends Biochem Sci* 23(10), pp. 394-398.

Culbertson, M. R. 1999. RNA surveillance. Unforeseen consequences for gene expression, inherited genetic disorders and cancer. *Trends Genet* 15(2), pp. 74-80.

Cybulski, C. et al. 2008. Constitutional *CHEK2* mutations are associated with a decreased risk of lung and laryngeal cancers. *Carcinogenesis* 29(4), pp. 762-765.

Danckwardt, S. et al. 2002. Abnormally spliced I mRNAs: a single point mutation generates transcripts sensitive and insensitive to nonsense-mediated mRNA decay. *Blood* 99(5), pp. 1811-1816.

Dayhoff, M. O. et al. 1978. A model of evolutionary change in proteins. In: Dayhoff, M.O. ed. *Atlas of protein sequence and structure*. Vol. 5. Silver Spring, pp. 345-352.

Deans, A. J. et al. 2004. *Brca1* inactivation induces *p27(Kip1)*-dependent cell cycle arrest and delayed development in the mouse mammary gland. *Oncogene* 23(36), pp. 6136-6145.

Dihlmann, S. et al. 1999. Dominant negative effect of the *APC*1309 mutation: a possible explanation for genotype-phenotype correlations in familial adenomatous polyposis. *Cancer Res* 59(8), pp. 1857-1860.

Dittmer, D. et al. 1993. Gain of function mutations in *p53*. *Nat Genet* 4(1), pp. 42-46.

Dong, Z. and Fasullo, M. 2003. Multiple recombination pathways for sister chromatid exchange in *Saccharomyces cerevisiae*: role of *RAD1* and the *RAD52* epistasis group genes. *Nucleic Acids Res* 31(10), pp. 2576-2585.

Dostie, J. and Dreyfuss, G. 2002. Translation is required to remove Y14 from mRNAs in the cytoplasm. *Curr Biol* 12(13), pp. 1060-1067.

Drake, J. W. et al. 1998. Rates of spontaneous mutation. *Genetics* 148(4), pp. 1667-1686.

Drobetsky, E. A. et al. 1994. The mutational specificity of simulated sunlight at the aprt locus in rodent cells. *Carcinogenesis* 15(8), pp. 1577-1583.

Du, Z. et al. 2008. Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. *Genome Res* 18(2), pp. 233-241.

Dworkin, A. M. et al. 2009. Methylation not a frequent "second hit" in tumors with germline *BRCA* mutations. *Fam Cancer.*

Easton, D. F. et al. 1993. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* 52(4), pp. 678-701.

Efeyan, A. and Serrano, M. 2007. *p53*: guardian of the genome and policeman of the oncogenes. *Cell Cycle* 6(9), pp. 1006-1010.

Efstratiadis, A. et al. 1980. The structure and evolution of the human *beta-globin* gene family. *Cell* 21(3), pp. 653-668.

Elespuru, R. K. and Sankaranarayanan, K. 2007. New approaches to assessing the effects of mutagenic agents on the integrity of the human genome. *Mutat Res* 616(1-2), pp. 83-89.

Ellington, A. and Cherry, J. M. 2001. Characteristics of amino acids. *Curr Protoc Mol Biol* Appendix 1, p. Appendix 1C.

Epstein, C. J. 1967. Non-randomness of amino-acid changes in the evolution of homologous proteins. *Nature* 215(5099), pp. 355-359.

Erickson, R. P. 2003. Somatic gene mutation and human disease other than cancer. *Mutat Res* 543(2), pp. 125-136.

Esteller, M. et al. 2000. Promoter hypermethylation and *BRCA1* inactivation in sporadic breast and ovarian tumors. *J Natl Cancer Inst* 92(7), pp. 564-569.

Evans, D. G. et al. 2005. Age related shift in the mutation spectra of germline and somatic *NF2* mutations: hypothetical role of DNA repair mechanisms. *J Med Genet* 42(8), pp. 630-632.

Evers, B. and Jonkers, J. 2006. Mouse models of *BRCA1* and *BRCA2* deficiency: past lessons, current understanding and future prospects. *Oncogene* 25(43), pp. 5885-5897.

Farrell, W. E. and Clayton, R. N. 2003. Epigenetic change in pituitary tumorigenesis. *Endocr Relat Cancer* 10(2), pp. 323-330.

Farrugia, D. J. et al. 2008. Functional assays for classification of *BRCA2* variants of uncertain significance. *Cancer Res* 68(9), pp. 3523-3531.

Fasken, M. B. and Corbett, A. H. 2005. Process or perish: quality control in mRNA biogenesis. *Nat Struct Mol Biol* 12(6), pp. 482-488.

Faux, N. G. et al. 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. *Genome Res* 15(4), pp. 537-551.

Fearnhead, N. S. et al. 2001. The ABC of *APC*. *Hum Mol Genet* 10(7), pp. 721-733.

Feng, J. et al. 2003. Absence of somatic *ATM* missense mutations in 58 mammary carcinomas. *Cancer Genet Cytogenet* 145(2), pp. 179-182.

Ferguson, D. O. and Alt, F. W. 2001. DNA double strand break repair and chromosomal translocation: lessons from animal models. *Oncogene* 20(40), pp. 5572-5579.

Field, A. 2005. *Discovering Statistics using SPSS*. London: SAGE Publications Ltd, p. 816.

Fisher, R. H. 1990. *Statistical Methods, Experimental Design, and Scientific Inference*. J. H. Bennett ed. Oxford: OUP Oxford, p. 870.

Fleming, J. L. et al. 2008. The role of parental and grandparental epigenetic alterations in familial cancer risk. *Cancer Res* 68(22), pp. 9116-9121.

Fodde, R. and Smits, R. 2002. Cancer biology. A matter of dosage. *Science* 298(5594), pp. 761-763.

Forbes, S. et al. 2006. Cosmic 2005. *Br J Cancer* 94(2), pp. 318-322.

Forbes, S. A. et al. 2008. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* Chapter 10, p. Unit 10 11.

Ford, D. et al. 1998. Genetic heterogeneity and penetrance analysis of the *BRCA1* and *BRCA2* genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* 62(3), pp. 676-689.

Frank, T. S. et al. 2002. Clinical characteristics of individuals with germline mutations in *BRCA1* and *BRCA2*: analysis of 10,000 individuals. *J Clin Oncol* 20(6), pp. 1480-1490.

Frank-Kamenetski, M. D. 1993. *Unravelling DNA*. Wiley-VCH, p. 176.

Frayling, I. M. et al. 1998. The *APC* variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. *Proc Natl Acad Sci U S A* 95(18), pp. 10722-10727.

Friedberg, E. C. 2003. DNA damage and repair. *Nature* 421(6921), pp. 436-440.

Friedman, L. S. et al. 1994. Confirmation of *BRCA1* by analysis of germline mutations linked to breast and ovarian cancer in ten families. *Nat Genet* 8(4), pp. 399-404.

Frischmeyer, P. A. and Dietz, H. C. 1999. Nonsense-mediated mRNA decay in health and disease. *Hum Mol Genet* 8(10), pp. 1893-1900.

Futreal, P. A. et al. 2004. A census of human cancer genes. *Nat Rev Cancer* 4(3), pp. 177-183.

Gaffney, D. J. and Keightley, P. D. 2008. Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. *BMC Evol Biol* 8, p. 265.

Gatfield, D. et al. 2003. Nonsense-mediated mRNA decay in *Drosophila*: at the intersection of the yeast and mammalian pathways. *Embo J* 22(15), pp. 3960-3970.

Gatti, R. A. et al. 1999. Cancer risk in *ATM* heterozygotes: a model of phenotypic and mechanistic differences between missense and truncating mutations. *Mol Genet Metab* 68(4), pp. 419-423.

Glasker, S. et al. 2006. Second hit deletion size in von Hippel-Lindau disease. *Ann Neurol* 59(1), pp. 105-110.

Glazko, G. V. et al. 2006. Mutational hotspots in the *TP53* gene and, possibly, other tumor suppressors evolve by positive selection. *Biol Direct* 1, p. 4.

Glazko, G. V. et al. 2004. Mutation hotspots in the *p53* gene in tumors of different origin: correlation with evolutionary conservation and signs of positive selection. *Biochim Biophys Acta* 1679(2), pp. 95-106.

Glover, T. W. et al. 1991. Molecular and cytogenetic analysis of tumors in von Recklinghausen neurofibromatosis. *Genes Chromosomes Cancer* 3(1), pp. 62-70.

Gnarra, J. R. et al. 1997. Defective placental vasculogenesis causes embryonic lethality in *VHL*-deficient mice. *Proc Natl Acad Sci U S A* 94(17), pp. 9102-9107.

Goldgar, D. E. et al. 2004. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to *BRCA1* and *BRCA2*. *Am J Hum Genet* 75(4), pp. 535-544.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11(5), pp. 725-736.

Gonzalez, C. I. et al. 2001. Nonsense-mediated mRNA decay in *Saccharomyces cerevisiae*. *Gene* 274(1-2), pp. 15-25.

Gonzalez, K. D. et al. 2007. Somatic microindels: analysis in mouse soma and comparison with the human germline. *Hum Mutat* 28(1), pp. 69-80.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185(4154), pp. 862-864.

Greenblatt, M. S. et al. 1994. Mutations in the *p53* tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res* 54(18), pp. 4855-4878.

Greenblatt, M. S. et al. 1996. Deletions and insertions in the *p53* tumor suppressor gene in human cancers: confirmation of the DNA polymerase slippage/misalignment model. *Cancer Res* 56(9), pp. 2130-2136.

Greenman, C. et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* 446(7132), pp. 153-158.

Grippo, P. et al. 1968. Methylation of DNA in developing sea urchin embryos. *J Mol Biol* 36(2), pp. 195-208.

Groves, C. et al. 2002. Mutation cluster region, association between germline and somatic mutations and genotype-phenotype correlation in upper gastrointestinal familial adenomatous polyposis. *Am J Pathol* 160(6), pp. 2055-2061.

Gu, D. et al. 2007. Database of somatic mutations in *EGFR* with analyses revealing indel hotspots but no smoking-associated signature. *Hum Mutat* 28(8), pp. 760-770.

Gurley, K. E. and Kemp, C. J. 2001. Synthetic lethality between mutation in *Atm* and DNA-PK(cs) during murine embryogenesis. *Curr Biol* 11(3), pp. 191-194.

Haber, D. and Harlow, E. 1997. Tumour-suppressor genes: evolving definitions in the genomic age. *Nat Genet* 16(4), pp. 320-322.

Haber, J. E. 2000. Partners and pathwaysrepairing a double-strand break. *Trends Genet* 16(6), pp. 259-264.

Hagan, I. and Sharrocks, A. D. 2002. Understanding cancer: from the gene to the organism. Conference on genes and cancer. *EMBO Rep* 3(5), pp. 415-419.

Halangoda, A. et al. 2001. Spontaneous microdeletions and microinsertions in a transgenic mouse mutation detection system: analysis of age, tissue, and sequence specificity. *Environ Mol Mutagen* 37(4), pp. 311-323.

Han, S. Y. et al. 2000. Functional evaluation of *PTEN* missense mutations using in vitro phosphoinositide phosphatase assay. *Cancer Res* 60(12), pp. 3147-3151.

Han, Y. et al. 2007. The zinc finger domain of Wilms' tumor 1 suppressor gene (*WT1*) behaves as a dominant negative, leading to abrogation of *WT1* oncogenic potential in breast cancer cells. *Breast Cancer Res* 9(4), p. R43.

Hassan, N. M. et al. 2008. Presence of dominant negative mutation of *TP53* is a risk of early recurrence in oral cancer. *Cancer Lett* 270(1), pp. 108-119.

He, T. C. et al. 1998. Identification of *c-MYC* as a target of the *APC* pathway. *Science* 281(5382), pp. 1509-1512.

Hebert, M. L. et al. 2004. DNA double-strand breaks induce deletion of CTG.CAG repeats in an orientation-dependent manner in *Escherichia coli*. *J Mol Biol* 336(3), pp. 655-672.

Heitzer, E. et al. 2007. UV fingerprints predominate in the *PTCH* mutation spectra of basal cell carcinomas independent of clinical phenotype. *J Invest Dermatol* 127(12), pp. 2872-2881.

Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89(22), pp. 10915-10919.

Hess, S. T. et al. 1994. Wide variations in neighbor-dependent substitution rates. *J Mol Biol* 236(4), pp. 1022-1033.

Hilleren, P. and Parker, R. 1999. Mechanisms of mRNA surveillance in eukaryotes. *Annu Rev Genet* 33, pp. 229-260.

Hilton, J. L. et al. 2002. Inactivation of *BRCA1* and *BRCA2* in ovarian cancer. *J Natl Cancer Inst* 94(18), pp. 1396-1406.

Hoffmann, D. et al. 2001. The less harmful cigarette: a controversial issue. a tribute to Ernst L. Wynder. *Chem Res Toxicol* 14(7), pp. 767-790.

Hohenstein, P. and Fodde, R. 2003. Of mice and (wo)men: genotype-phenotype correlations in *BRCA1*. *Hum Mol Genet* 12 Spec No 2, pp. R271-277.

Holbrook, J. A. et al. 2004. Nonsense-mediated decay approaches the clinic. *Nat Genet* 36(8), pp. 801-808.

Honma, M. et al. 2003. Deletion, rearrangement, and gene conversion; genetic consequences of chromosomal double-strand breaks in human cells. *Environ Mol Mutagen* 42(4), pp. 288-298.

Howlett, N. G. et al. 2002. Biallelic inactivation of *BRCA2* in Fanconi anemia. *Science* 297(5581), pp. 606-609.

Htun, H. and Dahlberg, J. E. 1988. Single strands, triple strands, and kinks in H-DNA. *Science* 241(4874), pp. 1791-1796.

Huang, J. et al. 1996. *APC* mutations in colorectal tumors with mismatch repair deficiency. *Proc Natl Acad Sci U S A* 93(17), pp. 9049-9054.

Huppert, J. L. and Balasubramanian, S. 2007. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res* 35(2), pp. 406-413.

Hussain, S. P. and Harris, C. C. 1999. *p53* mutation spectrum and load: the generation of hypotheses linking the exposure of endogenous or exogenous carcinogens to human cancer. *Mutat Res* 428(1-2), pp. 23-32.

Ikehata, H. and Ono, T. 2007. Significance of CpG methylation for solar UV-induced mutagenesis and carcinogenesis in skin. *Photochem Photobiol* 83(1), pp. 196-204.

Inacio, A. et al. 2004. Nonsense mutations in close proximity to the initiation codon fail to trigger full nonsense-mediated mRNA decay. *J Biol Chem* 279(31), pp. 32170-32180.

Irwin, B. et al. 1995. Codon pair utilization biases influence translational elongation step times. *J Biol Chem* 270(39), pp. 22801-22806.

Itzkovitz, S. and Alon, U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res* 17(4), pp. 405-412.

Jackson, S. P. 2002. Sensing and repairing DNA double-strand breaks. *Carcinogenesis* 23(5), pp. 687-696.

Javeri, A. et al. 2008. Human 8-oxoguanine-DNA glycosylase 1 protein and gene are expressed more abundantly in the superficial than basal layer of human epidermis. *DNA Repair (Amst)* 7(9), pp. 1542-1550.

Jaworski, A. et al. 1995. Mismatch repair in *Escherichia coli* enhances instability of (CTG)n triplet repeats from human hereditary diseases. *Proc Natl Acad Sci U S A* 92(24), pp. 11019-11023.

Joerger, A. C. and Fersht, A. R. 2008. Structural biology of the tumor suppressor *p53*. *Annu Rev Biochem* 77, pp. 557-582.

Johnson, K. C. et al. 2002. Cellular transformation by a FERM domain mutant of the *Nf2* tumor suppressor gene. *Oncogene* 21(39), pp. 5990-5997.

Jones, A. V. et al. 2009. *JAK2* haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat Genet* 41(4), pp. 446-449.

Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. In: H.N., M. ed. *Mammalian protein metabolism.* New York: Academic Press, pp. 21-132.

Junk, D. J. et al. 2008. Different mutant/wild-type *p53* combinations cause a spectrum of increased invasive potential in nonmalignant immortalized human mammary epithelial cells. *Neoplasia* 10(5), pp. 450-461.

Kang, J. H. et al. 2001. Methylation in the *p53* promoter is a supplementary route to breast carcinogenesis: correlation between CpG methylation in the *p53* promoter and the mutation of the *p53* gene in the progression from ductal carcinoma in situ to invasive ductal carcinoma. *Lab Invest* 81(4), pp. 573-579.

Karabinis, M. E. et al. 2001. Heterozygosity for a mutation in *Brca1* or *Atm* does not increase susceptibility to ENU-induced mammary tumors in *Apc*(Min)/+ mice. *Carcinogenesis* 22(2), pp. 343-346.

Karlin, S. et al. 2002. Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci U S A* 99(1), pp. 333-338.

Kawamata, H. et al. 2007. Oncogenic mutation of the *p53* gene derived from head and neck cancer prevents cells from undergoing apoptosis after DNA damage. *Int J Oncol* 30(5), pp. 1089-1097.

Kawashima, S. et al. 2008. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 36(Database issue), pp. D202-205.

Khan, S. and Vihinen, M. 2007. Spectrum of disease-causing mutations in protein secondary structures. *BMC Struct Biol* 7, p. 56.

Khromova, N. V. et al. 2008. *p53* hot-spot mutants increase tumor vascularization via ROS-mediated activation of the HIF1/VEGF-A pathway. *Cancer Lett.*

Kilpivaara, O. et al. 2009. A germline *JAK2* SNP is associated with predisposition to the development of *JAK2*(V617F)-positive myeloproliferative neoplasms. *Nat Genet* 41(4), pp. 455-459.

Kim, J. et al. 2003. *BRCA1* associates with human papillomavirus type 18 E2 and stimulates E2-dependent transcription. *Biochem Biophys Res Commun* 305(4), pp. 1008-1016.

Kimura, M. 1991. Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proc Natl Acad Sci U S A* 88(14), pp. 5969-5973.

Kimura, M. and Ota, T. 1974. On some principles governing molecular evolution. *Proc Natl Acad Sci U S A* 71(7), pp. 2848-2852.

Kinzler KW, V. B. 2002. *The Genetic Basis of Human Cancer*. 2 ed. New York: McGraw-Hill.

Knudson, A. G. 2002. Cancer genetics. *Am J Med Genet* 111(1), pp. 96-102.

Knudson, A. G., Jr. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* 68(4), pp. 820-823.

Knudson, A. G., Jr. 1978. Retinoblastoma: a prototypic hereditary neoplasm. *Semin Oncol* 5(1), pp. 57-60.

Knuth, D. E. 1973. *The Art of Computer Programming*. Reading, MA: Addison-Wesley.

Kohwi, Y. et al. 1992. Intramolecular dG.dG.dC triplex detected in *Escherichia coli* cells. *J Mol Biol* 223(4), pp. 817-822.

Kohwi, Y. and Panchenko, Y. 1993. Transcription-dependent recombination induced by triple-helix formation. *Genes Dev* 7(9), pp. 1766-1778.

Kondrashov, A. S. and Rogozin, I. B. 2004. Context of deletions and insertions in human coding sequences. *Hum Mutat* 23(2), pp. 177-185.

Koonin, E. V. et al. 2005. *p53* gain-of-function: tumor biology and bioinformatics come together. *Cell Cycle* 4(5), pp. 686-688.

Krawczak, M. et al. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am J Hum Genet* 63(2), pp. 474-488.

Krawczak, M. and Cooper, D. N. 1991. Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. *Hum Genet* 86(5), pp. 425-441.

Kryukov, G. V. et al. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 80(4), pp. 727-739.

Kumar, S. et al. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5(2), pp. 150-163.

Kunkel, T. A. 1990. Misalignment-mediated DNA synthesis errors. *Biochemistry* 29(35), pp. 8003-8011.

Kuzmiak, H. A. and Maquat, L. E. 2006. Applying nonsense-mediated mRNA decay research to the clinic: progress and challenges. *Trends Mol Med* 12(7), pp. 306-316.

LaFave, M. C. and Sekelsky, J. 2009. Mitotic recombination: why? when? how? where? *PLoS Genet* 5(3), p. e1000411.

Lambert, I. B. et al. 1992. Carcinogen-induced frameshift mutagenesis in repetitive sequences. *Proc Natl Acad Sci U S A* 89(4), pp. 1310-1314.

Lamlum, H. et al. 1999. The type of somatic mutation at *APC* in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's 'two-hit' hypothesis. *Nat Med* 5(9), pp. 1071-1075.

Lane, D. P. 1992. Cancer. *p53*, guardian of the genome. *Nature* 358(6381), pp. 15-16.

Lang, D. M. 2007. Imperfect DNA mirror repeats in the *gag* gene of HIV-1 (*HXB2*) identify key functional domains and coincide with protein structural elements in each of the mature proteins. *Virol J* 4, p. 113.

Lang, G. A. et al. 2004. Gain of function of a *p53* hot spot mutation in a mouse model of Li-Fraumeni syndrome. *Cell* 119(6), pp. 861-872.

Latchford, A. et al. 2007. *APC* mutations in FAP-associated desmoid tumours are non-random but not 'just right'. *Hum Mol Genet* 16(1), pp. 78-82.

Laux, D. E. et al. 1999. Hypermethylation of the Wilms' tumor suppressor gene CpG island in human breast carcinomas. *Breast Cancer Res Treat* 56(1), pp. 35-43.

Lavin, M. F. et al. 2004. Functional consequences of sequence alterations in the *ATM* gene. *DNA Repair (Amst)* 3(8-9), pp. 1197-1205.

Lawson, M. J. and Zhang, L. 2008. Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5'-UTR region. *Gene* 407(1-2), pp. 54-62.

Le Hir, H. et al. 2000. The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *Embo J* 19(24), pp. 6860-6869.

Lee, J. S. et al. 1989. Triplex DNA in plasmids and chromosomes. *Gene* 82(2), pp. 191-199.

Leggett, B. A. et al. 1997. Severe upper gastrointestinal polyposis associated with sparse colonic polyposis in a familial adenomatous polyposis family with an *APC* mutation at codon 1520. *Gut* 41(4), pp. 518-521.

Lejeune, F. et al. 2002. The exon junction complex is detected on CBP80-bound but not eIF4E-bound mRNA in mammalian cells: dynamics of mRNP remodeling. *Embo J* 21(13), pp. 3536-3545.

Lejeune, F. and Maquat, L. E. 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr Opin Cell Biol* 17(3), pp. 309-315.

Levine, A. J. 1997. *p53*, the cellular gatekeeper for growth and division. *Cell* 88(3), pp. 323-331.

Levine, A. J. et al. 2004. *P53* is a tumor suppressor gene. *Cell* 116(2 Suppl), pp. S67-69, 61 p following S69.

Levine, R. L. et al. 2007. Role of *JAK2* in the pathogenesis and therapy of myeloproliferative disorders. *Nat Rev Cancer* 7(9), pp. 673-683.

Lewis, B. P. et al. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* 100(1), pp. 189-192.

Li, F. Q. et al. 2000. Selection of a dominant negative retinoblastoma protein (*RB*) inhibiting satellite myoblast differentiation implies an indirect interaction between *MyoD* and *RB*. *Mol Cell Biol* 20(14), pp. 5129-5139.

Li, L. et al. 2007. Hypoxia-inducible factor linked to differential kidney cancer risk seen with type 2A and type 2B *VHL* mutations. *Mol Cell Biol* 27(15), pp. 5381-5392.

Li, S. and Wilkinson, M. F. 1998. Nonsense surveillance in lymphocytes? *Immunity* 8(2), pp. 135-141.

Li, W. H. et al. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol* 2(2), pp. 150-174.

Li, Z. et al. 2008. Novel self-association of the *APC* molecule affects *APC* clusters and cell migration. *J Cell Sci* 121(Pt 11), pp. 1916-1925.

Liberzon, E. et al. 2004. Germ-line *ATM* gene alterations are associated with susceptibility to sporadic T-cell acute lymphoblastic leukemia in children. *Genes Chromosomes Cancer* 39(2), pp. 161-166.

Lieber, M. R. et al. 2003. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol* 4(9), pp. 712-720.

Lin, H. et al. 2007. Stem cell regulatory function mediated by expression of a novel mouse *Oct4* pseudogene. *Biochem Biophys Res Commun* 355(1), pp. 111-116.

Lindahl, T. 1974. An N-glycosidase from *Escherichia coli* that releases free uracil from DNA containing deaminated cytosine residues. *Proc Natl Acad Sci U S A* 71(9), pp. 3649-3653.

Linde, L. et al. 2007. Nonsense-mediated mRNA decay affects nonsense transcript levels and governs response of cystic fibrosis patients to gentamicin. *J Clin Invest* 117(3), pp. 683-692.

Linde, L. and Kerem, B. 2008. Introducing sense into nonsense in treatments of human genetic diseases. *Trends Genet* 24(11), pp. 552-563.

Little, M. et al. 1995. DNA binding capacity of the *WT1* protein is abolished by Denys-Drash syndrome *WT1* point mutations. *Hum Mol Genet* 4(3), pp. 351-358.

Loeb, L. A. and Harris, C. C. 2008. Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Res* 68(17), pp. 6863-6872.

Loire, E. et al. 2009. Hypermutability of genes in Homo sapiens due to the hosting of long mono-SSR. *Mol Biol Evol* 26(1), pp. 111-121.

Longman, D. et al. 2007. Mechanistic insights and identification of two novel factors in the *C. elegans* NMD pathway. *Genes Dev* 21(9), pp. 1075-1085.

Luo, C. et al. 2000. Recognition and incision of site-specifically modified C8 guanine adducts formed by 2-aminofluorene, N-acetyl-2-aminofluorene and 1-nitropyrene by UvrABC nuclease. *Nucleic Acids Res* 28(19), pp. 3719-3724.

Luo, L. et al. 1998. Ataxia-telangiectasia and T-cell leukemias: no evidence for somatic *ATM* mutation in sporadic T-ALL or for hypermethylation of the *ATM*-NPAT/E14 bidirectional promoter in T-PLL. *Cancer Res* 58(11), pp. 2293-2297.

Lyamichev, V. I. et al. 1985. A pH-dependent structural transition in the homopurine-homopyrimidine tract in superhelical DNA. *J Biomol Struct Dyn* 3(2), pp. 327-338.

Lyons-Darden, T. and Topal, M. D. 1999. Abasic sites induce triplet-repeat expansion during DNA replication in vitro. *J Biol Chem* 274(37), pp. 25975-25978.

Magdinier, F. et al. 2002. Epigenetic marks at *BRCA1* and *p53* coding sequences in early human embryogenesis. *Mol Hum Reprod* 8(7), pp. 630-635.

Magewu, A. N. and Jones, P. A. 1994. Ubiquitous and tenacious methylation of the CpG site in codon 248 of the *p53* gene may explain its frequent appearance as a mutational hot spot in human cancer. *Mol Cell Biol* 14(6), pp. 4225-4232.

Maglott, D. et al. 2005. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 33(Database issue), pp. D54-58.

Maher, E. R. and Kaelin, W. G., Jr. 1997. von Hippel-Lindau disease. *Medicine (Baltimore)* 76(6), pp. 381-391.

Mak, B. C. et al. 2005. Aberrant beta-catenin signaling in tuberous sclerosis. *Am J Pathol* 167(1), pp. 107-116.

Makalowski, W. and Boguski, M. S. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci U S A* 95(16), pp. 9407-9412.

Malkin, D. et al. 1990. Germ line *p53* mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 250(4985), pp. 1233-1238.

Maquat, L. E. 2002. Nonsense-mediated mRNA decay. *Curr Biol* 12(6), pp. R196-197.

Maquat, L. E. 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* 5(2), pp. 89-99.

Maquat, L. E. and Li, X. 2001. Mammalian heat shock *p70* and histone H4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay. *Rna* 7(3), pp. 445-456.

Marshall, B. et al. 1997. Germline versus somatic mutations of the *APC* gene: evidence for mechanistic differences. *Hum Mutat* 9(3), pp. 286-288.

Masutani, C. et al. 2000. Xeroderma pigmentosum variant: from a human genetic disorder to a novel DNA polymerase. *Cold Spring Harb Symp Quant Biol* 65, pp. 71-80.

McCartney, B. M. et al. 2006. Testing hypotheses for the functions of *APC* family proteins using null and truncation alleles in Drosophila. *Development* 133(12), pp. 2407-2418.

Medghalchi, S. M. et al. 2001. Rent1, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. *Hum Mol Genet* 10(2), pp. 99-105.

Mendell, J. T. et al. 2004. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat Genet* 36(10), pp. 1073-1078.

Menon, K. P. and Neufeld, E. F. 1994. Evidence for degradation of mRNA encoding alpha-L-iduronidase in Hurler fibroblasts with premature termination alleles. *Cell Mol Biol (Noisy-le-grand)* 40(7), pp. 999-1005.

Meric-Bernstam, F. 2007. Heterogenic loss of *BRCA* in breast cancer: the "two-hit" hypothesis takes a hit. *Ann Surg Oncol* 14(9), pp. 2428-2429.

Millar, D. S. et al. 1998. Variation of site-specific methylation patterns in the factor VIII (*F8C*) gene in human sperm DNA. *Hum Genet* 103(2), pp. 228-233.

Miller, L. H. et al. 1976. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N Engl J Med* 295(6), pp. 302-304.

Miller, M. P. and Kumar, S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet* 10(21), pp. 2319-2328.

Mitrovich, Q. M. and Anderson, P. 2000. Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in *C. elegans*. *Genes Dev* 14(17), pp. 2173-2184.

Mitui, M. et al. 2009. Functional and computational assessment of missense variants in the ataxia-telangiectasia mutated (*ATM*) gene: mutations with increased cancer risk. *Hum Mutat* 30(1), pp. 12-21.

Miyata, T. et al. 1979. Two types of amino acid substitutions in protein evolution. *J Mol Evol* 12(3), pp. 219-236.

Miyoshi, Y. et al. 1992. Somatic mutations of the *APC* gene in colorectal tumors: mutation cluster region in the *APC* gene. *Hum Mol Genet* 1(4), pp. 229-233.

Montiel-Duarte, C. et al. 2008. Resistance to Imatinib Mesylate-induced apoptosis in acute lymphoblastic leukemia is associated with *PTEN* down-regulation due to promoter hypermethylation. *Leuk Res* 32(5), pp. 709-716.

Morley, A. A. and Turner, D. R. 1999. The contribution of exogenous and endogenous mutagens to in vivo mutations. *Mutat Res* 428(1-2), pp. 11-15.

Mort, M. et al. 2008. A meta-analysis of nonsense mutations causing human genetic disease. *Hum Mutat* 29(8), pp. 1037-1047.

Moser, A. R. et al. 1995. Homozygosity for the Min allele of *Apc* results in disruption of mouse development prior to gastrulation. *Dev Dyn* 203(4), pp. 422-433.

Mote, P. A. et al. 2004. Germ-line mutations in *BRCA1* or *BRCA2* in the normal breast are associated with altered expression of estrogen-responsive proteins and the predominance of progesterone receptor A. *Genes Chromosomes Cancer* 39(3), pp. 236-248.

Mount, S. M. 2000. Genomic sequence, splicing, and gene annotation. *Am J Hum Genet* 67(4), pp. 788-792.

Mouret, S. et al. 2008. Differential repair of UVB-induced cyclobutane pyrimidine dimers in cultured human skin cells and whole human skin. *DNA Repair (Amst)* 7(5), pp. 704-712.

Mousses, S. et al. 2001. *p53* missense but not truncation mutations are associated with low levels of *p21(CIP1/WAF1)* mRNA expression in primary human sarcomas. *Br J Cancer* 84(12), pp. 1635-1639.

Muhlemann, O. et al. 2008. Recognition and elimination of nonsense mRNA. *Biochim Biophys Acta* 1779(9), pp. 538-549.

Muhlrad, D. and Parker, R. 1999. Aberrant mRNAs with extended 3' UTRs are substrates for rapid degradation by mRNA surveillance. *Rna* 5(10), pp. 1299-1307.

Mularoni, L. et al. 2007. Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats. *Genomics* 89(3), pp. 316-325.

Murphy, J. A. et al. 2004. The *CDKN2A* database: Integrating allelic variants with evolution, structure, function, and disease association. *Hum Mutat* 24(4), pp. 296-304.

Nagarajan, R. P. and Costello, J. F. 2009. Epigenetic mechanisms in glioblastoma multiforme. *Semin Cancer Biol* 19(3), pp. 188-197.

Nagy, E. and Maquat, L. E. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci* 23(6), pp. 198-199.

Neel, J. V. 1983. Frequency of spontaneous and induced "point" mutations in higher eukaryotes. *J Hered* 74(1), pp. 2-15.

Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3(5), pp. 418-426.

Ng, P. C. and Henikoff, S. 2001. Predicting deleterious amino acid substitutions. *Genome Res* 11(5), pp. 863-874.

Niwa, O. 2006. Indirect mechanisms of genomic instability and the biological significance of mutations at tandem repeat loci. *Mutat Res* 598(1-2), pp. 61-72.

Notaro, R. et al. 2000. Human mutations in glucose 6-phosphate dehydrogenase reflect evolutionary history. *FASEB J* 14(3), pp. 485-494.

Oguchi, K. et al. 2003. Missense mutation and defective function of *ATM* in a childhood acute leukemia patient with *MLL* gene rearrangement. *Blood* 101(9), pp. 3622-3627.

Ohmiya, N. et al. 2001. Germline and somatic mutations in *hMSH6* and *hMSH3* in gastrointestinal cancers of the microsatellite mutator phenotype. *Gene* 272(1-2), pp. 301-313.

Olcaydu, D. et al. 2009. A common *JAK2* haplotype confers susceptibility to myeloproliferative neoplasms. *Nat Genet* 41(4), pp. 450-454.

Olive, K. P. et al. 2004. Mutant *p53* gain of function in two mouse models of Li-Fraumeni syndrome. *Cell* 119(6), pp. 847-860.

Ollila, S. et al. 2008. Mechanisms of pathogenicity in human *MSH2* missense mutants. *Hum Mutat* 29(11), pp. 1355-1363.

Olshen, A. B. and Jain, A. N. 2002. Deriving quantitative conclusions from microarray expression data. *Bioinformatics* 18(7), pp. 961-970.

Olson, M. V. and Varki, A. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat Rev Genet* 4(1), pp. 20-28.

Oshima, M. et al. 1995. Evidence against dominant negative mechanisms of intestinal polyp formation by *Apc* gene mutations. *Cancer Res* 55(13), pp. 2719-2722.

Pages, V. and Fuchs, R. P. 2002. How DNA lesions are turned into mutations within cells? *Oncogene* 21(58), pp. 8957-8966.

Pamilo, P. and Bianchi, N. O. 1993. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol Biol Evol* 10(2), pp. 271-281.

Payne, S. R. and Kemp, C. J. 2005. Tumor suppressor genetics. *Carcinogenesis* 26(12), pp. 2031-2045.

Pfeifer, G. P. 2000. *p53* mutational spectra and the role of methylated CpG sequences. *Mutat Res* 450(1-2), pp. 155-166.

Pfeifer, G. P. 2006. Mutagenesis at methylated CpG sequences. *Curr Top Microbiol Immunol* 301, pp. 259-281.

Pfeifer, G. P. and Besaratinia, A. 2009. Mutational spectra of human cancer. *Hum Genet* 125(5-6), pp. 493-506.

Pfeifer, G. P. et al. 2005. Mutations induced by ultraviolet light. *Mutat Res* 571(1-2), pp. 19-31.

Pfeiffer, P. et al. 2000. Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis* 15(4), pp. 289-302.

Pfleger, C. M. and Kirschner, M. W. 2000. The KEN box: an *APC* recognition signal distinct from the D box targeted by *Cdh1*. *Genes Dev* 14(6), pp. 655-665.

Pineau, P. et al. 2008. Chromosome instability in human hepatocellular carcinoma depends on *p53* status and aflatoxin exposure. *Mutat Res.*

Pineiro, E. et al. 2003. Mutagenic stress modulates the dynamics of CTG repeat instability associated with myotonic dystrophy type 1. *Nucleic Acids Res* 31(23), pp. 6733-6740.

Pogo, A. O. and Chaudhuri, A. 2000. The Duffy protein: a malarial and chemokine receptor. *Semin Hematol* 37(2), pp. 122-129.

Pollock, P. M. et al. 1996. Compilation of somatic mutations of the *CDKN2* gene in human cancers: non-random distribution of base substitutions. *Genes Chromosomes Cancer* 15(2), pp. 77-88.

Powell, S. M. et al. 1992. *APC* mutations occur early during colorectal tumorigenesis. *Nature* 359(6392), pp. 235-237.

Pugacheva, E. N. et al. 2002. Novel gain of function activity of *p53* mutants: activation of the dUTPase gene expression leading to resistance to 5-fluorouracil. *Oncogene* 21(30), pp. 4595-4600.

Qian, J. et al. 2008. The *APC* tumor suppressor inhibits DNA replication by directly binding to DNA via its carboxyl terminus. *Gastroenterology* 135(1), pp. 152-162.

Qian, L. et al. 1993. T cell receptor-beta mRNA splicing: regulation of unusual splicing intermediates. *Mol Cell Biol* 13(3), pp. 1686-1696.

Radpour, R. et al. 2009. Methylation profiles of 22 candidate genes in breast cancer using high-throughput MALDI-TOF mass array. *Oncogene*.

Rajavel, K. S. and Neufeld, E. F. 2001. Nonsense-mediated decay of human *HEXA* mRNA. *Mol Cell Biol* 21(16), pp. 5512-5519.

Ramus, S. J. et al. 2007. Contribution of *BRCA1* and *BRCA2* mutations to inherited ovarian cancer. *Hum Mutat* 28(12), pp. 1207-1215.

Rasmussen, S. A. et al. 2000. Chromosome 17 loss-of-heterozygosity studies in benign and malignant tumors in neurofibromatosis type 1. *Genes Chromosomes Cancer* 28(4), pp. 425-431.

Ripley, L. S. 1982. Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. *Proc Natl Acad Sci U S A* 79(13), pp. 4128-4132.

Ripley, L. S. 1990. Frameshift mutation: determinants of specificity. *Annu Rev Genet* 24, pp. 189-213.

Robinson, M. J. and Cobb, M. H. 1997. Mitogen-activated protein kinase pathways. *Curr Opin Cell Biol* 9(2), pp. 180-186.

Romao, L. et al. 2000. Nonsense mutations in the human beta-globin gene lead to unexpected levels of cytoplasmic mRNA accumulation. *Blood* 96(8), pp. 2895-2901.

Rooney, S. M. and Moore, P. D. 1995. Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells. *Proc Natl Acad Sci U S A* 92(6), pp. 2141-2144.

Rosner, M. et al. 2003. Cell size regulation by the human *TSC* tumor suppressor proteins depends on *PI3K* and *FKBP38*. *Oncogene* 22(31), pp. 4786-4798.

Ruas, M. et al. 1999. Functional evaluation of tumour-specific variants of *p16INK4a/CDKN2A*: correlation with protein structure information. *Oncogene* 18(39), pp. 5423-5434.

Ruiz-Echevarria, M. J. and Peltz, S. W. 2000. The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. *Cell* 101(7), pp. 741-751.

Rutherford, J. et al. 2002. Investigations on a clinically and functionally unusual and novel germline *p53* mutation. *Br J Cancer* 86(10), pp. 1592-1596.

Rutter, J. L. et al. 2003. *CDKN2A* point mutations D153spl(c.457G>T) and IVS2+1G>T result in aberrant splice products affecting both *p16INK4a* and *p14ARF*. *Oncogene* 22(28), pp. 4444-4448.

Sage, E. et al. 1996. Mutagenic specificity of solar UV light in nucleotide excision repair-deficient rodent cells. *Proc Natl Acad Sci U S A* 93(1), pp. 176-180.

Sanchez-Sanchez, F. et al. 2007. Attenuation of disease phenotype through alternative translation initiation in low-penetrance retinoblastoma. *Hum Mutat* 28(2), pp. 159-167.

Sankaranarayanan, K. and Wassom, J. S. 2005. Ionizing radiation and genetic risks XIV. Potential research directions in the post-genome era based on knowledge of repair of radiation-induced DNA double-strand breaks in mammalian somatic cells and the origin of deletions associated with human genomic disorders. *Mutat Res* 578(1-2), pp. 333-370.

Scaringe, W. A. et al. 2008. Somatic microindels in human cancer: The insertions are highly error-prone and derive from nearby but not adjacent sense and antisense templates. *Hum Mol Genet*.

Schubbert, S. et al. 2007. Hyperactive Ras in developmental disorders and cancer. *Nat Rev Cancer* 7(4), pp. 295-308.

Schulz, W. A. and Hoffmann, M. J. 2009. Epigenetic mechanisms in the biology of prostate cancer. *Semin Cancer Biol* 19(3), pp. 172-180.

Scott, S. P. et al. 2002. Missense mutations but not allelic variants alter the function of *ATM* by dominant interference in patients with breast cancer. *Proc Natl Acad Sci U S A* 99(2), pp. 925-930.

Seligmann, H. and Pollock, D. D. 2004. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA Cell Biol* 23(10), pp. 701-705.

Shannon, C. E. 1948. A mathematic theory of communication. *The Bell System Techical Journal* 27, pp. pt I, 379-423, pt II,623-656.

Shaulsky, G. et al. 1991. Alterations in tumor development in vivo mediated by expression of wild type or mutant *p53* proteins. *Cancer Res* 51(19), pp. 5232-5237.

Sherr, C. J. 2004. Principles of tumor suppression. *Cell* 116(2), pp. 235-246.

Shtutman, M. et al. 1999. The *cyclin D1* gene is a target of the beta-catenin/LEF-1 pathway. *Proc Natl Acad Sci U S A* 96(10), pp. 5522-5527.

Shukla, A. K. and Roy, K. B. 2006. Rec A-independent homologous recombination induced by a putative fold-back tetraplex DNA. *Biol Chem* 387(3), pp. 251-256.

Sieber, O. M. et al. 2000. The adenomatous polyposis coli (*APC*) tumour suppressor--genetics, function and disease. *Mol Med Today* 6(12), pp. 462-469.

Simpson, L. and Parsons, R. 2001. *PTEN*: life as a tumor suppressor. *Exp Cell Res* 264(1), pp. 29-41.

Sinden, R. R. et al. 2002. Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *J Biosci* 27(1 Suppl 1), pp. 53-65.

Slebos, R. J. et al. 2002. Mutations in tetranucleotide repeats following DNA damage depend on repeat sequence and carcinogenic agent. *Cancer Res* 62(21), pp. 6052-6060.

Sokal, R. R. and Rohlf, F. J. 1995. *Biometry: the principles and Practice of Statistics in Biological Research*. 3 ed. New York: Freeman.

Somers, C. M. and Cooper, D. N. 2009. Air pollution and mutations in the germline: are humans at risk? *Hum Genet* 125(2), pp. 119-130.

Soussi, T. and Beroud, C. 2001. Assessing *TP53* status in human tumours to evaluate clinical outcome. *Nat Rev Cancer* 1(3), pp. 233-240.

Soussi, T. and Beroud, C. 2003. Significance of *TP53* mutations in human cancer: a critical analysis of mutations at CpG dinucleotides. *Hum Mutat* 21(3), pp. 192-200.

Soussi, T. et al. 2005. Reassessment of the *TP53* mutation database in human disease by data mining with a library of TP53 missense mutations. *Hum Mutat* 25(1), pp. 6-17.

Soussi, T. and Wiman, K. G. 2007. Shaping genetic alterations in human cancer: the *p53* mutation paradigm. *Cancer Cell* 12(4), pp. 303-312.

Stead, J. D. and Jeffreys, A. J. 2000. Allele diversity and germline mutation at the insulin minisatellite. *Hum Mol Genet* 9(5), pp. 713-723.

Stedman, H. H. et al. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature* 428(6981), pp. 415-418.

Steelman, L. S. et al. 2008. Suppression of *PTEN* function increases breast cancer chemotherapeutic drug resistance while conferring sensitivity to mTOR inhibitors. *Oncogene* 27(29), pp. 4086-4095.

Stefansson, O. A. et al. 2009. Genomic profiling of breast tumours in relation to *BRCA* abnormalities and phenotypes. *Breast Cancer Res* 11(4), p. R47.

Stenson, P. D. et al. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 21(6), pp. 577-581.

Stone, E. A. and Sidow, A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* 15(7), pp. 978-986.

Stormo, G. D. et al. 1986. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res* 14(16), pp. 6661-6679.

Strachan, T. and Read, A. 2004. *Human molecular genetics 3*. 3rd ed. Garland Publishing, p. 674.

Strano, S. et al. 2007. Mutant *p53* proteins: between loss and gain of function. *Head Neck* 29(5), pp. 488-496.

Stratton, M. R. et al. 2009. The cancer genome. *Nature* 458(7239), pp. 719-724.

Sturzeneker, R. et al. 2000. Microsatellite instability in tumors as a model to study the process of microsatellite mutations. *Hum Mol Genet* 9(3), pp. 347-352.

Szabo, C. I. et al. 1996. Human, canine and murine *BRCA1* genes: sequence comparison among species. *Hum Mol Genet* 5(9), pp. 1289-1298.

Szabo, C. I. et al. 2004. Understanding germ-line mutations in *BRCA1*. *Cancer Biol Ther* 3(6), pp. 515-520.

Taioli, E. 2008. Gene-environment interaction in tobacco-related cancers. *Carcinogenesis* 29(8), pp. 1467-1474.

Takano, A. et al. 2004. A missense variation in human casein kinase I epsilon gene that induces functional alteration and shows an inverse association with circadian rhythm sleep disorders. *Neuropsychopharmacology* 29(10), pp. 1901-1909.

Takata, M. et al. 1998. Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *Embo J* 17(18), pp. 5497-5508.

Tamasauskas, D. et al. 2001. A homologous naturally occurring mutation in *Duffy* and *CCR5* leading to reduced receptor expression. *Blood* 97(11), pp. 3651-3654.

Tang, R. et al. 2001. Mutations of *p53* gene in human colorectal cancer: distinct frameshifts among populations. *Int J Cancer* 91(6), pp. 863-868.

Tange, T. O. et al. 2004. The ever-increasing complexities of the exon junction complex. *Curr Opin Cell Biol* 16(3), pp. 279-284.

Tats, A. et al. 2008. Preferred and avoided codon pairs in three domains of life. *BMC Genomics* 9, p. 463.

Tavtigian, S. V. et al. 2006. Comprehensive statistical study of 452 *BRCA1* missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* 43(4), pp. 295-305.

Tavtigian, S. V. et al. 2008. In silico analysis of missense substitutions using sequence-alignment based methods. *Hum Mutat* 29(11), pp. 1327-1336.

Tetsu, O. and McCormick, F. 1999. Beta-catenin regulates expression of *cyclin D1* in colon carcinoma cells. *Nature* 398(6726), pp. 422-426.

Thomas, R. K. et al. 2007. High-throughput oncogene mutation profiling in human cancer. *Nat Genet* 39(3), pp. 347-351.

Thompson, J. D. et al. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* Chapter 2, p. Unit 2 3.

Tighe, A. et al. 2004. Truncating *APC* mutations have dominant effects on proliferation, spindle checkpoint control, survival and chromosome stability. *J Cell Sci* 117(Pt 26), pp. 6339-6353.

Tijsterman, M. et al. 2002. Frequent germline mutations and somatic repeat instability in DNA mismatch-repair-deficient *Caenorhabditis elegans. Genetics* 161(2), pp. 651-660.

Toledo, F. and Wahl, G. M. 2006. Regulating the *p53* pathway: in vitro hypotheses, in vivo veritas. *Nat Rev Cancer* 6(12), pp. 909-923.

Tornaletti, S. and Pfeifer, G. P. 1995. Complete and tissue-independent methylation of CpG sites in the *p53* gene: implications for mutations in human cancers. *Oncogene* 10(8), pp. 1493-1499.

Tost, J. 2009. DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. *Methods Mol Biol* 507, pp. 3-20.

Trasler, J. M. 1998. Origin and roles of genomic methylation patterns in male germ cells. *Semin Cell Dev Biol* 9(4), pp. 467-474.

Tsukamoto, Y. and Ikeda, H. 1998. Double-strand break repair mediated by DNA end-joining. *Genes Cells* 3(3), pp. 135-144.

Uchikawa, H. et al. 2006. Brain- and heart-specific *Patched-1* containing exon 12b is a dominant negative isoform and is expressed in medulloblastomas. *Biochem Biophys Res Commun* 349(1), pp. 277-283.

Upadhyaya, M. et al. 2008. Germline and somatic *NF1* gene mutation spectrum in *NF1*-associated malignant peripheral nerve sheath tumors (MPNSTs). *Hum Mutat* 29(1), pp. 74-82.

Upadhyaya, M. et al. 1997. Mutational and functional analysis of the neurofibromatosis type 1 (*NF1*) gene. *Hum Genet* 99(1), pp. 88-92.

Ussery, D. W. and Sinden, R. R. 1993. Environmental influences on the in vivo level of intramolecular triplex DNA in *Escherichia coli*. *Biochemistry* 32(24), pp. 6206-6213.

Valavanidis, A. et al. 2009. 8-hydroxy-2' -deoxyguanosine (8-OHdG): A critical biomarker of oxidative stress and carcinogenesis. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 27(2), pp. 120-139.

van Den Hurk, W. H. et al. 2001. Novel frameshift mutations near short simple repeats. *J Biol Chem* 276(15), pp. 11496-11498.

Varley, J. M. 2003. Germline *TP53* mutations and Li-Fraumeni syndrome. *Hum Mutat* 21(3), pp. 313-320.

Vasudevan, S. et al. 2002. Non-stop decay--a new mRNA surveillance pathway. *Bioessays* 24(9), pp. 785-788.

Venkatachalam, S. et al. 1998. Retention of wild-type *p53* in tumors from *p53* heterozygous mice: reduction of *p53* dosage can promote cancer formation. *Embo J* 17(16), pp. 4657-4667.

Vilenchik, M. M. and Knudson, A. G., Jr. 2000. Inverse radiation dose-rate effects on somatic and germ-line mutations and DNA damage rates. *Proc Natl Acad Sci U S A* 97(10), pp. 5381-5386.

Visnes, T. et al. 2008. Review. Uracil in DNA and its processing by different DNA glycosylases. *Philos Trans R Soc Lond B Biol Sci.*

Vitkup, D. et al. 2003. The amino-acid mutational spectrum of human genetic disease. *Genome Biol* 4(11), p. R72.

Vogelstein, B. and Kinzler, K. W. 2004. Cancer genes and the pathways they control. *Nat Med* 10(8), pp. 789-799.

Vogelstein, B. et al. 2000. Surfing the *p53* network. *Nature* 408(6810), pp. 307-310.

Vogler, A. J. et al. 2006. Effect of repeat copy number on variable-number tandem repeat mutations in *Escherichia coli* O157:H7. *J Bacteriol* 188(12), pp. 4253-4263.

Vorechovsky, I. et al. 1996. The *ATM* gene and susceptibility to breast cancer: analysis of 38 breast tumors reveals no evidence for mutation. *Cancer Res* 56(12), pp. 2726-2732.

Vortmeyer, A. O. et al. 2002. Somatic point mutation of the wild-type allele detected in tumors of patients with *VHL* germline deletion. *Oncogene* 21(8), pp. 1167-1170.

Vousden, K. H. 2000. *p53*: death star. *Cell* 103(5), pp. 691-694.

Vreeswijk, M. P. et al. 2009. Site-specific analysis of UV-induced cyclobutane pyrimidine dimers in nucleotide excision repair-proficient and -deficient hamster cells: Lack of correlation with mutational spectra. *Mutat Res* 663(1-2), pp. 7-14.

Wait, S. D. et al. 2004. Somatic mutations in VHL germline deletion kindred correlate with mild phenotype. *Ann Neurol* 55(2), pp. 236-240.

Walker, D. R. et al. 1999. Evolutionary conservation and somatic mutation hotspot maps of *p53*: correlation with *p53* protein structural and functional features. *Oncogene* 18(1), pp. 211-218.

Walker, D. R. and Koonin, E. V. 1997. SEALS: a system for easy analysis of lots of sequences. *Proc Int Conf Intell Syst Mol Biol* 5, pp. 333-339.

Walter, C. A. et al. 1998. Mutation frequency declines during spermatogenesis in young mice but increases in old mice. *Proc Natl Acad Sci U S A* 95(17), pp. 10015-10019.

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

Wan, P. T. et al. 2004. Mechanism of activation of the *RAF-ERK* signaling pathway by oncogenic mutations of B-RAF. *Cell* 116(6), pp. 855-867.

Wang, G. and Vasquez, K. M. 2004. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc Natl Acad Sci U S A* 101(37), pp. 13448-13453.

Wang, J. et al. 2002. Boundary-independent polar nonsense-mediated decay. *EMBO Rep* 3(3), pp. 274-279.

Wang, J. S. et al. 2009. DNA Promoter Hypermethylation of *p16* and *APC* Predicts Neoplastic Progression in Barrett's Esophagus. *Am J Gastroenterol.*

Wang, R. Y. et al. 1982. Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. *Biochim Biophys Acta* 697(3), pp. 371-377.

Wang, X. et al. 2006. Gene losses during human origins. *PLoS Biol* 4(3), p. e52.

Wang, Z. and Moult, J. 2001. SNPs, protein structure, and disease. *Hum Mutat* 17(4), pp. 263-270.

Webber, T. M. et al. 2009. Conformational detection of *p53*'s oligomeric state by FlAsH Fluorescence. *Biochem Biophys Res Commun* 384(1), pp. 66-70.

Welch, E. M. and Jacobson, A. 1999. An internal open reading frame triggers nonsense-mediated decay of the yeast SPT10 mRNA. *Embo J* 18(21), pp. 6134-6145.

Wells, R. D. 2007. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem Sci* 32(6), pp. 271-278.

Wells, R. D. et al. 2005. Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res* 33(12), pp. 3785-3798.

Wheelan, S. J. et al. 2001. Spidey: a tool for mRNA-to-genomic alignments. *Genome Res* 11(11), pp. 1952-1957.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics* 1, pp. 80-83.

Wilkinson, M. F. 2005. A new function for nonsense-mediated mRNA-decay factors. *Trends Genet* 21(3), pp. 143-148.

Willis, A. et al. 2004. Mutant *p53* exerts a dominant negative effect by preventing wild-type *p53* from binding to the promoter of its target genes. *Oncogene* 23(13), pp. 2330-2338.

Wojciechowska, M. et al. 2006. Non-B DNA conformations formed by long repeating tracts of *myotonic dystrophy type 1*, *myotonic dystrophy type 2*, and *Friedreich's ataxia* genes, not the sequences per se, promote mutagenesis in flanking regions. *J Biol Chem* 281(34), pp. 24531-24543.

Wolf, D. et al. 1984. Reconstitution of *p53* expression in a nonproducer Ab-MuLV-transformed cell line by transfection of a functional *p53* gene. *Cell* 38(1), pp. 119-126.

Xing, Y. et al. 2004. Crystal structure of a beta-catenin/*APC* complex reveals a critical role for *APC* phosphorylation in *APC* function. *Mol Cell* 15(4), pp. 523-533.

Xu, G. et al. 2008. Recovery of a low mutant frequency after ionizing radiation-induced mutagenesis during spermatogenesis. *Mutat Res* 654(2), pp. 150-157.

Yamada, H. et al. 2009. Identification and characterization of a novel germline *p53* mutation in a patient with glioblastoma and colon cancer. *Int J Cancer* 125(4), pp. 973-976.

Yamashita, A. et al. 2005. The role of SMG-1 in nonsense-mediated mRNA decay. *Biochim Biophys Acta* 1754(1-2), pp. 305-315.

Yampolsky, L. Y. et al. 2005. Distribution of the strength of selection against amino acid replacements in human proteins. *Hum Mol Genet* 14(21), pp. 3191-3201.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13(5), pp. 555-556.

Yang, Z. and Bielawski, J. P. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15(12), pp. 496-503.

Yanofsky, C. 2007. Establishing the triplet nature of the genetic code. *Cell* 128(5), pp. 815-818.

Yin, Y. and Shen, W. H. 2008. *PTEN*: a new guardian of the genome. *Oncogene* 27(41), pp. 5443-5453.

Yoon, J. H. et al. 2001. Methylated CpG dinucleotides are the preferential targets for G-to-T transversion mutations induced by benzo[a]pyrene diol epoxide in mammalian cells: similarities with the *p53* mutation spectrum in smoking-associated lung cancers. *Cancer Res* 61(19), pp. 7110-7117.

Zhang, J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J Mol Evol* 50(1), pp. 56-68.

Zhang, J. and Maquat, L. E. 1997. Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells. *Embo J* 16(4), pp. 826-833.

Zhang, Y. et al. 2002. Detection of radiation and cyclophosphamide-induced mutations in individual mouse sperm at a human expanded trinucleotide repeat locus transgene. *Mutat Res* 516(1-2), pp. 121-138.

Zhang, Y. and Xiong, Y. 1999. Mutations in human *ARF* exon 2 disrupt its nucleolar localization and impair its ability to block nuclear export of *MDM2* and *p53*. *Mol Cell* 3(5), pp. 579-591.

Zhu, J. et al. 2000. Definition of the *p53* functional domains necessary for inducing apoptosis. *J Biol Chem* 275(51), pp. 39927-39934.

Zhu, Y. et al. 2009. Involvement of *PTEN* promoter methylation in cerebral cavernous malformations. *Stroke* 40(3), pp. 820-826.