(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Biochimica et Biophysica Acta 1816 (2011) 199–208

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbacan

Review

Advances in carcinogenesis: A historical perspective from observational studies to tumor genome sequencing and TP53 mutation spectrum analysis

Thierry Soussi *

Karolinska Institute, Dept. of Oncology–Pathology Cancer Center Karolinska (CCK), SE-171 76 Stockholm, Sweden
Université Pierre et Marie Curie-Paris 6, 75005 Paris, France

ARTICLE INFO

Article history:
Received 20 June 2011
Received in revised form 12 July 2011
Accepted 13 July 2011
Available online xxxx

Keywords:
Mutation database
TP53 gene mutations
Cancer etiology
Novel generation sequencing
Molecular epidemiology

ABSTRACT

Tumor sequencing projects have been initiated over the last decade with the promising goal of identifying novel cancer genes and potential therapeutic targets. One of the unexpected findings of these projects was the discovery that cancer genomes contain thousands of passenger mutations that are irrelevant to tumor development and are coselected by a small number of driver mutations that constitute the true selection power in cancer progression. Although often discarded and considered to be irrelevant, the value of passenger mutations should not be underestimated, as they are the most important markers of the exposure to various carcinogens and are essential to assess the etiology of individual tumors.

Over the last century, the history of cancer epidemiology evolved in different stages and concepts from occupational observational studies beginning in the 18th century, *in vitro* and *in vivo* experimental analyses and cancer gene analyses, such as Ha-ras or TP53. Mutation spectra of passenger mutations from various types of cancers not only confirm the findings of molecular epidemiology analysis, but also reveal novel profiles that will extend this knowledge to single tumors in all types of cancer.

Contents

Abbreviations: 6,4-PPs, pyrimidine(6–4)pyrimidone photoproducts; AK, Actinic Keratosis; BaP, benzo(a)pyrene; BPDE, benzo(a)pyrene diol epoxide; BCC, Basal cell carcinoma; CPD, cyclobutane pyrimidine dimers; CRC, colorectal carcinoma; ER, estrogen receptor; ERG, ETS related Gene; EZH2, Histone-lysine N-methyltransferase; GBM, Glioblastoma; HCC, hepatocellular carcinoma; HER2, Human Epidermal growth factor Receptor 2; HGMD, Human Gene Mutation Database; ICGC, International Cancer Genome Consortium; NGS, novel generation sequencing; NSCLC, non small cell lung cancer; PR, Progesterone receptor; SCC, Squamous cell carcinoma; TMPRSS2, androgen-regulated trans-membrane protease, serine 2; TP53 RE, TP53 DNA response element

* Karolinska Institute, Dept. of Oncology–Pathology Cancer Center Karolinska (CCK), SE-171 76 Stockholm, Sweden. Tel.: +46 85 177 73 36.
E-mail address: thierry.soussi@ki.se.

0304-419X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbcan.2011.07.003

1. Introduction

During the first decade of the 21st century, there has been a revolution in DNA sequencing with the appearance of novel generation sequencing (NGS) and the release of hundreds of genome sequences from various species including the whole genome of single healthy individuals or the complex heterogeneous genome of tumors [1,2]. The power of the current NGS methodology can be compared to that of a magnifying glass allowing the scientist to look not only at the past with the sequencing of extinct species such as the Neanderthal genome but also to have a glimpse at the future with the recent sequencing of the genome of a normal twelve-week-old fetus using DNA released into maternal blood [3,4]. NGS methodology is also much more sensitive, allowing the detection of base changes in a heterogeneous DNA population with a sensitivity as high as 1 in 5000 copies. This is particularly useful in the field of cancer to identify rare clones that could be responsible for drug resistance or metastasis.

Conventional sequencing using Sanger's methodology has allowed the discovery of genetic alterations in cancer genes and NGS expands this knowledge by providing an accurate picture of each type of alteration including copy number variations, translocations or missense mutations [5]. A surprising finding is the much higher than expected prevalence of somatic single nucleotide substitutions, which constitute the most frequent genetic alteration detected in tumor genomes [6]. The majority of these mutations are passenger mutations (or hitchhiking mutations) that have no active role in cancer progression and are only coselected by the driver mutations, which constitute the true driving force for cell transformation [7]. Passenger mutations can be found in coding or noncoding regions of the genome and can be difficult to distinguish from driving mutations, but this distinction is essential in order to obtain an accurate picture of the cancer genome. Several statistical approaches have been developed to resolve this problem, such as comparing the observed to expected ratios of synonymous:non-synonymous variants. Alternatively, various bioinformatics methods can be used to indicate whether an amino acid substitution is likely to damage protein function on the basis of either conservation through species or whether or not the amino acid change is conservative.

TP53 mutations are found in approximately 50% of human cancers [8]. The TP53 protein is a transcription factor that binds a very loose DNA response element (TP53RE) found in several hundred genes that are differentially activated depending on the cell type, identity, and extent of damage, and various other parameters that have yet to be identified [9]. The unique feature of TP53 compared to other tumor suppressor genes is its mode of inactivation. While most tumor suppressor genes are inactivated by frameshift or nonsense mutations leading to absence of protein synthesis (or production of a truncated product), more than 80% of TP53 alterations are missense mutations that lead to the synthesis of a full-length protein that accumulates in the nucleus of the tumor cell. This selection to maintain mutant TP53 in tumor cells is believed to be required for both a dominant negative activity to inhibit wild-type TP53 expressed by the remaining allele, and for a gain of function that transforms mutant TP53 into a dominant oncogene [8].

One of the greatest contributions to the study of TP53 mutations has been provided by molecular epidemiology and its applications to human carcinogenesis [10,11]. These studies demonstrate a link between exposure to various types of carcinogens, specific mutational events in the TP53 gene and the development of specific cancers.

In this review, we will discuss how the data generated by the various cancer genome sequencing projects can be compared to the spectrum of TP53 mutations and how they will dramatically improve the accuracy of molecular epidemiology studies and extend our knowledge about the etiology of cancer.

2. Molecular epidemiology: from observational studies to cancer genome sequencing

Numerous observational studies on occupational cancers were conducted in the 18th and 19th centuries (Fig. 1 and Supplementary Table 1). The first half of the 20th century saw the development of numerous animal models that were used to demonstrate the carcinogenic potential of chemical substances. Yamagiwa was the first to show that an experimental cancer could be induced in the rabbit ear by repeated application of coal tar two to three times a week for several months [12]. These types of studies were repeated by hundreds of scientists, leading to the birth of experimental chemical carcinogenesis and to the establishment of lists of potential carcinogens [13,14]. The 1950s were marked by the discovery of the double helix and the demonstration that most carcinogens are able to bind to DNA [15]. The analysis of these carcinogen–DNA interactions at the molecular or atomic levels represents a major progress over the last 50 years. Information gleaned from these studies has helped to elucidate the mechanisms of mutation of many different carcinogens and, more specifically, their mutation spectra. These studies, originally carried out in *E. coli*, also paved the way for development of the Ames test, which is still routinely used today to screen suspicious compounds for mutagenic potential and which forms part of the regulatory testing requirements for new drugs [16]. This approach was then adapted to mammals, either in cell culture models or directly in animals, particularly rodents. However, animal studies must be interpreted with caution for various reasons. First, the metabolism (activation and detoxification) of carcinogens in rodents is very different from that in primates. Second, the methods of inducing experimental exposure in the laboratory often differ from conditions of natural, chronic exposure. Transgenic models in mice expressing a mutant oncogene or in mice nullizygous for a tumor suppressor gene also have their limitations. Mice expressing the mutant Ha-ras oncogene represent a striking example. Mutations in the Ha-ras gene are very frequent in murine skin tumors, and mice expressing a Ha-ras transgene have a high incidence of skin tumors and are highly sensitive to carcinogens. However, this relationship has not been confirmed by analysis of ras gene mutations in human skin cancers, as the mutation rate of this gene is relatively low. This example illustrates the limits and precautions involved in the interpretation of murine models.

Analysis of mutations in human cancer DNA first started with the K-ras gene and the demonstration that the spectrum of these mutations was tumor-specific [17]. Unfortunately, the number of missense mutations that activate the oncogenic activity of K-ras is restricted to a small number of codons (12, 13, 61 and, to a lesser extent, 146), therefore limiting the impact of these studies. However, analysis of the spectrum of TP53 mutations in various types of cancer has profoundly extended these studies by establishing, for the first time, a direct link between carcinogen exposure and several types of human cancer (Figs. 2 and 3). The most striking example is that of tandem mutations, specifically induced by ultraviolet radiation, which are only observed in skin cancers. The relationships between G->T transversion and lung cancer in smokers or mutation of codon 249 observed in aflatoxin B1-induced liver cancers are also very demonstrative (Fig. 3). These studies were possible because TP53 was the only gene that combined several specific features used to study the origin of carcinogenesis in a human population: 1) it is mutated in many types of cancers; 2) the mutation frequency is high; 3) the gene is predominantly modified by point mutations; 4) the gene is small enough to be relatively easy to analyze; 5) the number of codons that can be targeted by these mutations is very high (more than 200) and multiple types of substitution can be found at a single codon [10,11].

NGS has overcome the limitations of Sanger sequencing and allows high levels of data throughput that were not available only a few years ago. NGS therefore provides access to full cancer genome data allowing unbiased analysis of mutation spectra.

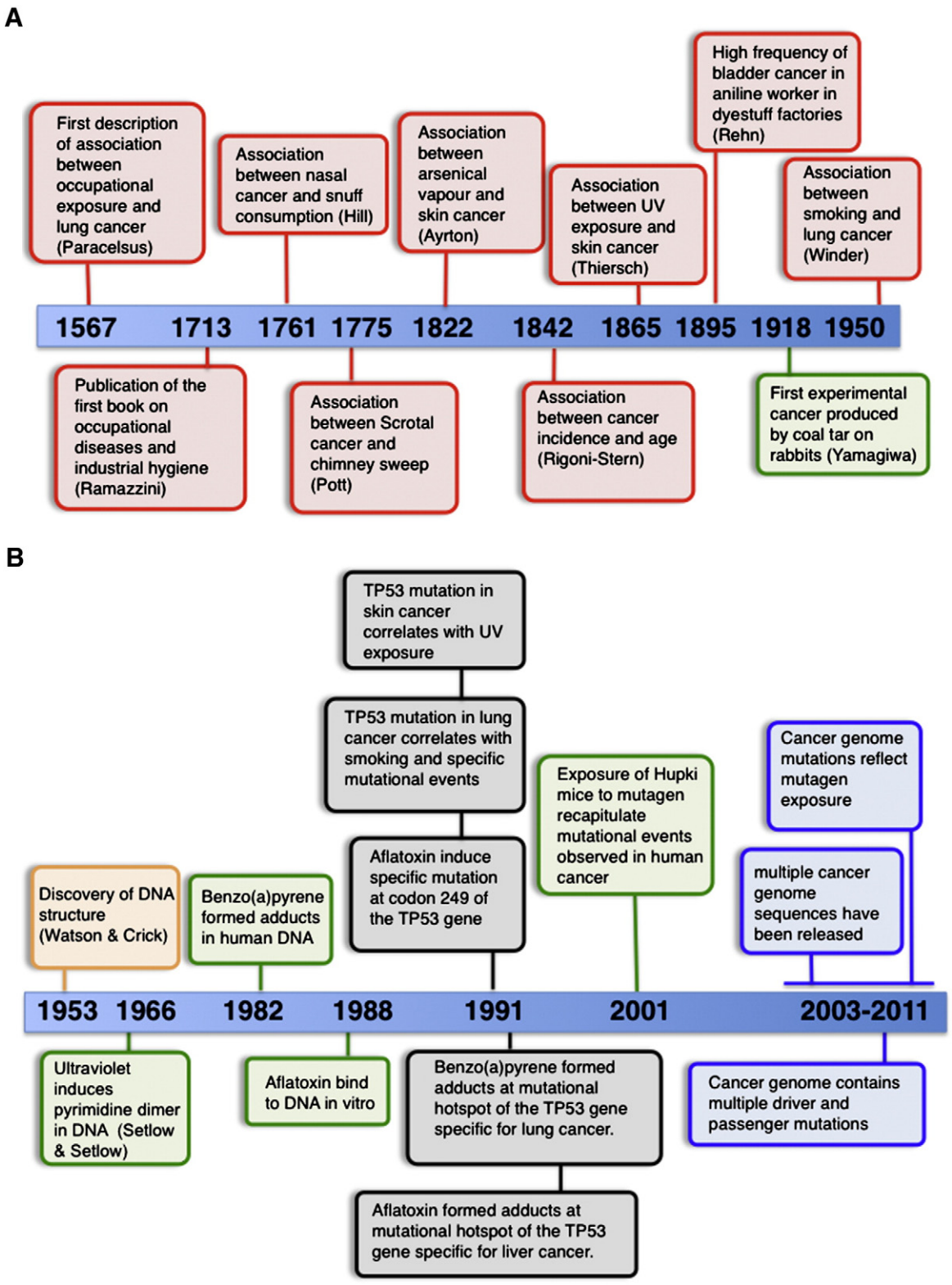

Fig. 1. Advances in carcinogenesis: Three centuries of analysis can be divided into occupational observational studies (Red), experimental studies (green), mutation profiling either in the TP53 gene (black) or in the whole genome (blue). References for these studies can be found in Supplementary Table 1.

3. Mutation spectra in brain and colorectal cancer

TP53 mutational events in glioblastoma and colorectal cancer are fairly similar with a high frequency of G:C->A:T transitions, predominantly localized at CpG dinucleotides (Fig. 4). A similar spectrum is also observed for TP53 mutations in other brain tumors such as astrocytoma (data not shown).

5-Methylcytosine at CpG dinucleotide is a DNA modification with important implications for the regulation of gene expression [18]. The coding sequence of the TP53 gene contains 42 CpG dinucleotides that

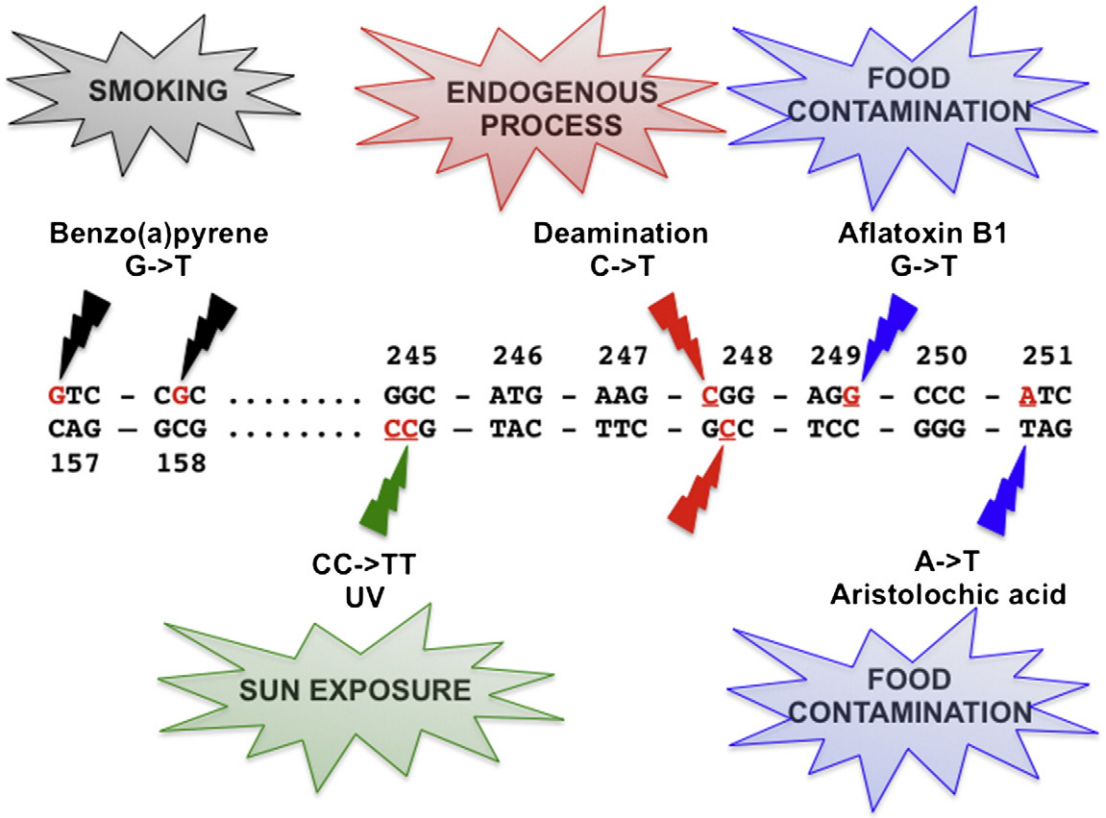

Fig. 2. Summary of carcinogens and mutational events that can alter the p53 gene. Only representative regions of the TP53 gene are depicted. *In vitro* studies have shown that BaP binds strongly to guanine residues 157 and 158 confirming the origin of the mutational hot spot observed in lung cancer from smoker patients. Similarly, Aflatoxin B1, a potent hepatocarcinogen found as a food contaminant, binds *in vitro* to codon 249 of the TP53 gene confirming the R249S hot spot found in liver cancer from exposed countries. Aristolochic acid forms a covalent adduct with adenine that leads to A->T transversion. This mutational event is specific for urothelial cancer from patients with Balkan Endemic Nephropathy, an environmental disease strongly associated with exposure to aristolochic acid. In skin cancer, the UV signature (tandem mutation at dipyrimidine sites and C->T transition) is highly specific. In several cancers associated with endogenous processes, the frequency of G:C->T:A transition at codons that include a CpG dinucleotide can be as high as 70%. Bases targeted by the various carcinogens are shown in red.

are scattered along the 10 exons. This frequency is very high compared to other eukaryotic genes. All these CpG are methylated in various cell types, but the function of TP53 gene methylation has never been addressed either *in vitro* or *in vivo* [19]. It is currently unknown whether or not these CpG are associated with the fine-tuning of p53 gene expression. Although they are methylated in normal cells, their status in human tumor is unknown. Most TP53 hot spots for mutations are localized at CpG sites with a mutation spectrum compatible with 5-methylcytosine deamination (Fig. 3). These hot spot codons, CGN at positions 175, 248 or 273, encode arginine residues important for TP53 structure and/or activity. It is interesting to note that arginine can also be encoded by AGG and AGA that have the same frequency of usage in human but are not targeted by methylation. It has not yet been determined whether or not there is a specific selection to keep CGN in the TP53.

Analysis of the human genome has revealed that CpG are present at only about 20% of their expected frequency, a paucity thought to be linked to an endogenous process due to the high mutation rate from the methylated cytosine [20]. It is generally assumed that the higher deamination rate of 5-methylcytosine leads to a T/G mismatch which is not efficiently repaired, resulting in a high rate of C->T transition. The link between the high level of G:C->A:T transition and this endogenous process is reinforced by the observation that the majority of mutations observed in hereditary genetic disease are G:C->A:T transitions, as shown by the analysis of the Human Gene Mutation Database (HGMD) that comprised 54.625 mutations in 2.113 genes (Fig. 5) [21]. An analysis of the mutation spectrum of snp in the human genome also shows a predominance of G:C->A:T transitions (Fig. 4).

Large-scale sequencing of CRC genomes has been restricted to exons. Two types of analysis have been performed *i.e.* sequencing that targets exons of specific gene families such as kinases (kinome) and phosphatases (phosphatome) or exomic sequencing that targets all potential exons of the human genome [6,22–24]. Exome sequencing performed in glioblastoma and colorectal cancer confirms that G:C->A:T transition at CpG sites is the predominant mutational event for missense mutations and fully supports the TP53 spectrum (Fig. 5). Although this observation suggests that deamination of 5-methylcytosine is a key event in brain and colorectal cancer, there are no data to indicate whether this event is more frequent in cancer cells or whether an endogenous or exogenous mutagen could lead to this mutation spectrum. Several non-exclusive mechanisms could be associated with this mutation spectrum. Firstly, it is possible that other genetic defects can alter the frequency of CpG deamination and/or their repair rate. Prostate cancer associated with a TMPRSS2-ERG translocation displays a much higher frequency of transition at CpG dinucleotides than in the absence of translocation (see below). A second possible mechanism involves mutation at the G residue, as a G->A transition will lead to a mutational event that cannot be distinguished from the C->T on the other strand.

4. Mutation spectra in lung cancer

Epidemiologic studies have provided solid evidence that the incidence of lung cancer, very rare at the beginning of the century, has increased in parallel with tobacco exposure, with a male/female lag-time due to the fact that cigarette use occurred later in women [25,26]. Cigarette smoke contains more than 3000 different

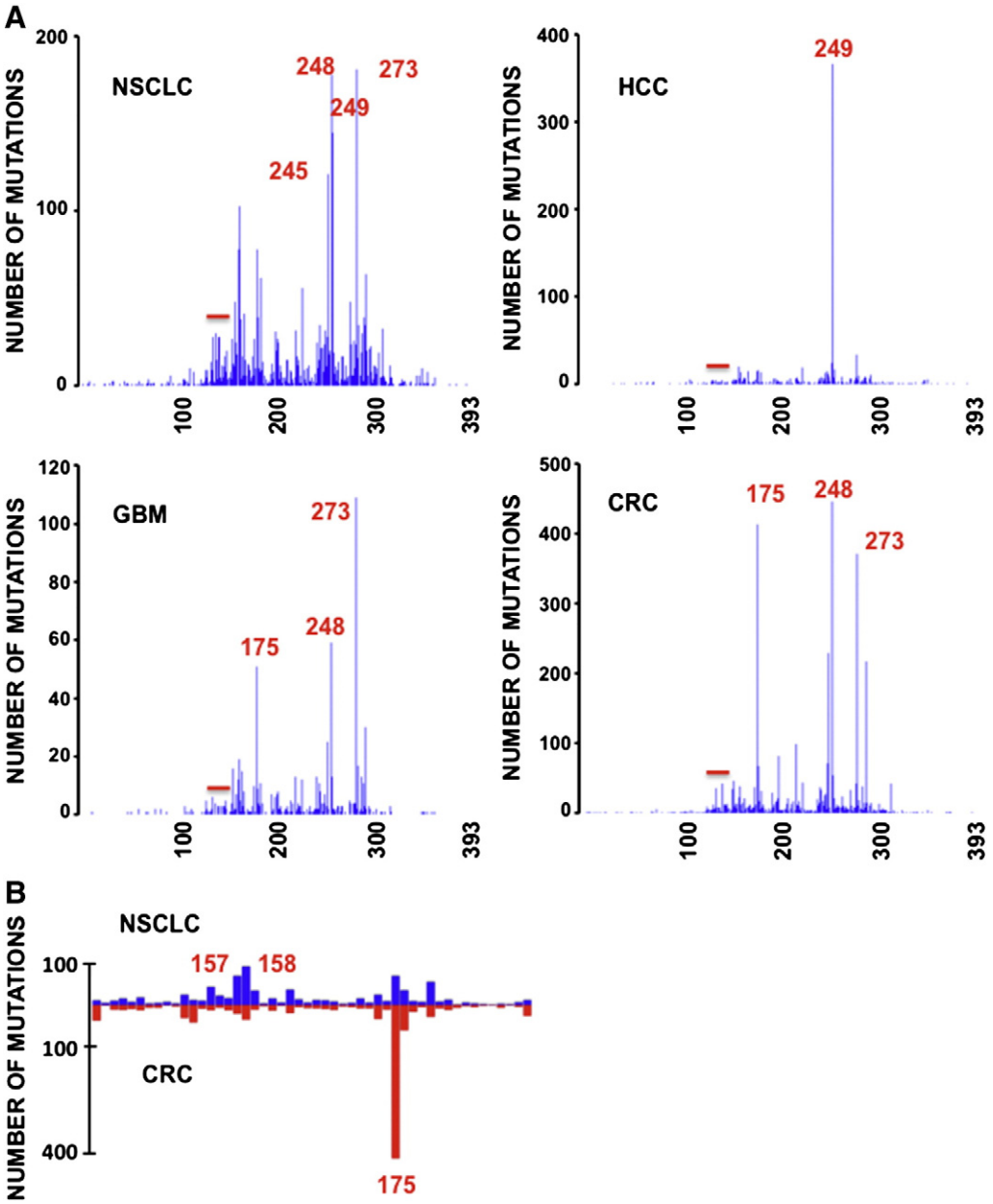

Fig. 3. Distribution of mutations in the TP53 protein (393 aa). A: Hot spots for somatic mutations can be explained both at the DNA level (the codon is highly susceptible to modification) and the protein level (the residue is essential for protein function). In the case of TP53, both explanations prevail. The hot spot at codon 175 (similar to the other hot-spot codons 245, 248, 273 and 282, shown in part A) contains a CpG dinucleotide and is a key residue to maintain the 3D structure of the TP53 (see text for more information). B: Mutations at codons 157 and 158 are specific for lung cancer and are not found in colorectal cancer. The low rate of mutations at codon 175 in lung cancer has not been explained. A similar number of mutations were used for this analysis. Hot spot regions that include codons 157 and 158 are indicated by a red line in part A. HCC: Hepatocellular carcinoma; NSCLC: Non-small cell lung cancer; GBM: glioblastoma; CRC: colorectal carcinoma. (Data from the UMDTP53 database: http://p53.free.fr).

substances, including certain substances with demonstrated carcinogenic activity in animals. In particular, a major component of cigarette smoke is benzo(a)pyrene (BaP), a mutagen also present in large quantities in soot and identified as a causal agent of scrotal cancer in chimney sweeps. There is experimental evidence that benzo(a) pyrene diol epoxide (BPDE), a mutagenic metabolite of BaP, forms guanine adducts which persist to generate G->T transversions.

We will not discuss in detail the impassioned and stormy debate over the relationship between smoking and lung cancer. Similarly, the relationship between TP53 mutation in lung cancer and exposure to BaP has also been the subject of heated controversies following reports that laboratories supported by the tobacco industry have tended to underestimate the impact of these analyses. This controversy originates from the inclusion of a dubious article in the TP53 database. Inclusion of these publications in studies that analyzed the relationship between mutational events in smokers and non-smokers led to results that were ambiguous and favorable to the tobacco industry [27]. More recent studies, using a specific verification algorithm and statistical analysis, have demonstrated that data from these articles are completely artefactual [28]. The TP53 mutation spectrum in lung cancer using curated databases shows a predominance of G:C->T:A transversions which are not present in other types of cancer and a very specific hot spot (codons 150–160 with a preference for codons 157 and 158) (Figs. 3 and 5A). There are several lines of evidence indicating that the transversions in the TP53 gene are due to carcinogens present in tobacco smoke [10]: i) only lung cancer (and, to a lesser extent, Head and Neck SCC and esophageal cancers which are also tobacco-related) has such a high rate of G:C>T:A transversions; ii) there is a linear correlation between this transversion rate and the number of cigarettes smoked; iii) non-

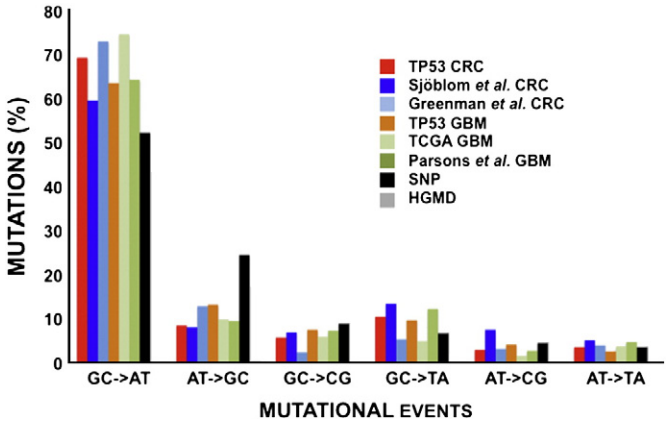

Fig. 4. Mutation spectrum in colorectal carcinoma and glioblastoma. Data were obtained from the UMD TP53 database (TP53), exome or whole genome sequence studies of human tumors. Statistical analysis shows that there are no significant differences between the spectra observed in colorectal carcinoma and glioblastoma. Full references and data are available in Supplementary Table 2.

smokers with lung cancer do not display this high transversion frequency (Fig. 5B); iv) treatment of human bronchial cells with BPDE induces the appearance of adducts in codons 157 and 158 [29], a guanine-rich region which is a hot spot for formation of these adducts.

Two recent lines of research have expanded these observations and confirmed the association between TP53 mutations in human cancer and carcinogen exposure. The first line of research concerns a novel experimental mouse model developed to assess the mutagenicity of various physical or chemical carcinogens. The Hupki mouse model system was constructed using gene-targeting technology and contains human wild-type TP53 gene from exons 4 to 9 in place of the homologous murine DNA sequences in both copies of the mouse TP53 gene [30]. The human region encodes the proline-rich domain and the

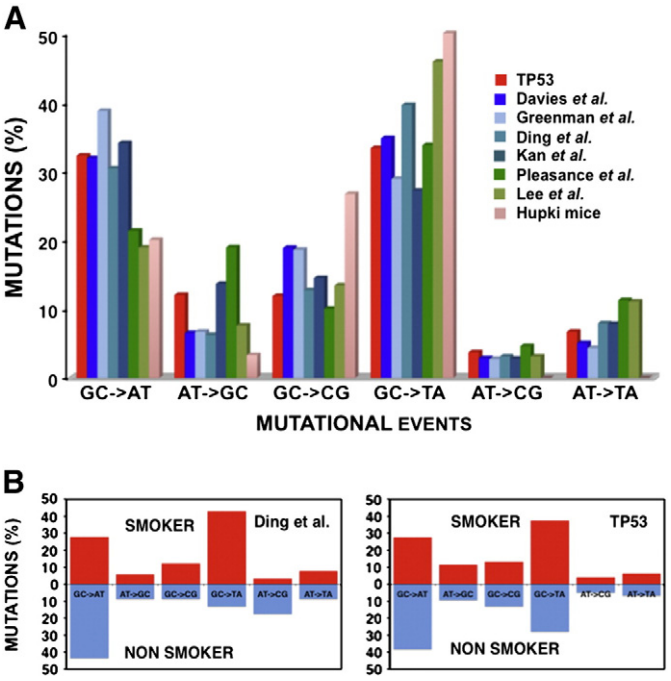

Fig. 5. A: Mutation spectrum in lung cancer. A: Data were obtained from the UMD TP53 database (TP53), Hupki cells treated with benzo(a)pyrene (Hupki), exome or whole genome sequence studies of human tumors. B: Mutation spectrum from smokers (red) and non smokers (blue) patients. Left: exomic study from Ding et al. (623 genes); Right: data from the UMD TP53 database. Statistical analysis shows that there are no significant differences between the spectra observed in lung cancer. Full references and data are available in Supplementary Table 2.

DNA-binding region. This chimeric gene remains under normal transcription regulation at the mouse locus. No dysfunction of TP53 activity including nuclear accumulation of TP53 protein after exposure to DNA-damaging agents, or transcriptional activation of known TP53 downstream targets, has been observed. This Hupki mouse develops normally, exhibits no apparent physiologic defects, remains fertile, and shows no susceptibility to spontaneous cancer [30]. Exposure of Hupki mouse embryonic fibroblasts to BaP leads to a high frequency of TP53 missense mutations and G->T transversions are the most frequent mutational events (Fig. 5A) [31]. Furthermore, these mutations are localized at codons 157 and 158 in exon 5, the exact same hot spot observed in lung cancer in heavy smokers [31].

The second line of research concerns the various large-scale sequencing analyses of lung cancer genomes, as both exomic and whole genome sequencing have been performed. Analysis of data from 4 independent exome studies performed in lung cancer shows that G:C->T:A transversions are the most frequent mutational event revealing that other genes can be used as probes for carcinogen exposure [6,32–34] (Fig. 5A). Whole cancer genome sequencing has also been performed on a lung cancer cell line (NCI-H209) and a smoker patient with NSCLC. Comparison with the sequence of the normal DNA from the same individuals provided a clear picture of more than 50,000 somatic mutational events specific to the tumor genomes. The majority of these events occurred in non-coding parts of the genome and can be considered to be passenger mutations that were simply coselected with the few driving mutations in cancer genes. Mutational profiling demonstrates a distribution that is undistinguishable from other distributions whether the analysis was restricted to a single gene such as TP53 or the multiple genes from exome sequencing (Fig. 5A). This high frequency of G:C->T:A transversion observed in lung cancer is never observed in cancer unrelated to smoking. The mutation spectrum observed in lung cancer is statistically different from the spectrum observed in CRC ($p<0.001$, Chi^2 test). This analysis indicates that lung cells are the target for thousands of mutagenic events induced by the various tobacco carcinogens, only a few of which target cancer genes, while the remainder are scattered in non-essential regions of the genome, but carry a specific mutagen signature. Whether or not some of these passenger mutations occur in normal cells before cellular transformation is currently unknown, but novel sequencing methodologies should rapidly resolve this question.

5. Mutation spectra in breast cancer

Breast cancer is a very complex disease with several levels of heterogeneity both in terms of etiology and the large numbers of histologic subtypes [35]. Way back in the 18th century (1713 to be exact), Bernardino Ramazzini observed that breast cancer was more frequent in nuns [36]. However, this correlation between nulliparity and cancer was only confirmed by physiologic and molecular evidence more than 250 years later, with the development of endocrinology and a better understanding of the role of hormones in breast cancer. Other factors such as smoking, diet, alcohol consumption or body size have also been associated with an increased, albeit low, risk of breast cancer. Genetic factors have also been shown to account for 27% of all breast cancers, including the highly penetrant mutations of BRCA1 and BRCA2 and several low penetrant genes that are either associated with carcinogen metabolism or DNA repair.

Breast cancer can be classified into more than 20 histologic subtypes with the highly prevalent invasive ductal carcinoma (65–80%), invasive lobular carcinoma (5–10%) and 17 distinct histologic special subtypes (up to 25%) [37]. Recent high-throughput microarray-based gene expression profiling has refined breast cancer subtypes and has identified molecular signatures associated with prognosis. On the other hand, progress in identification of the contribution of various risk

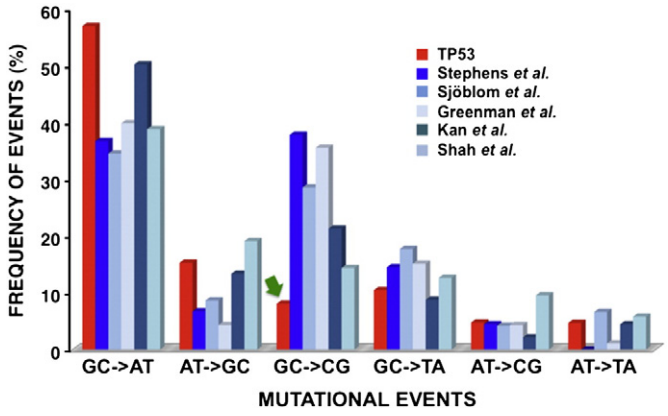

Fig. 6. Mutation spectrum in breast carcinoma. Data were obtained from the UMD TP53 database (TP53), exome or whole genome sequence studies of human tumors. Full references and data are available in Supplementary Table 2.

factors to the carcinogenesis of sporadic breast cancer has been very slow mainly due to the marked heterogeneity of the disease and its risk factors.

The spectrum of TP53 mutations in breast cancer displays a high frequency of G:C->A:T transitions and 50% of these transitions occur at CpG dinucleotides (Fig. 6). This lower frequency of CpG mutations compared to CRC and glioblastoma (75% and 70%, respectively) suggests a mechanism other than deamination of 5-methyl-cytosine. The spectrum of TP53 mutations in breast cancer is very heterogeneous among the various published studies, a feature that is not observed for other types of cancer. An analysis of the TP53 mutation spectrum in breast tumors from 15 geographically and ethnically diverse populations showed a significantly distinct pattern [38]. It is possible that the profile in low-risk populations (Japan) predominantly reflects a baseline, endogenous process, while mutagens present in high-risk populations might result in another spectrum indicating that a fraction of breast cancer mutations occur as a result of environmental exposure.

The mutation spectra derived from exome studies of various types of breast cancer or from the whole genome sequence of a breast cancer also display a high level of heterogeneity and differ from those of TP53 studies. Three independent exomic studies show a strikingly similar pattern with a high frequency of G:C->C:G transversions (30–40%) not observed in the TP53 gene (Fig. 6) [6,22,39]. Although a few mutagens that specifically induce this type of transversion have been identified, further studies will be necessary to pinpoint the mechanism resulting in this spectrum. Global analysis of TP53 mutations in all breast cancers from the database does not indicate a high frequency of GC->CG transversion, although three individual studies in white patients in the USA reported a high frequency of this transversion (19–33%) which is never observed in European or Japanese patients, emphasizing the importance of the geographical factors that contribute to breast cancer. A fourth exomic study conducted by Kan et al. on a series of 183 breast cancers showed a lower frequency of GC->CG transversion (20%). The authors analyzed three types of breast cancer, including 59 epidermal growth factor receptor 2 positive (HER2-positive), 65 Hormone receptor positive (ER/PR positive) and 59 triple-negative (ER/PR/HER2 negative) tumors [34]. Although, the global frequency of mutations in these three subtypes of cancer is similar, their mutation spectrum is different and the frequency of GC->CG transversions varied from 10% (ER/PR), to 28% (ER/PR/HER2 negative) and 30% (HER2 positive). It is therefore clear that pooling data from different types of breast cancer could mask a specific spectrum. A similar conclusion can be drawn when patients are derived from different geographical areas. In the three exomic analyses described above, the mutation spectrum was performed on global data without taking into account the various histologic subtypes.

To date, only two studies have described the mutation spectrum of a whole genome of a breast tumor: a lobular breast tumor (ER/PR positive) and a triple-negative basal-like breast tumor, but detailed information for analysis of mutational events is only available for this second tumor [40,41]. This spectrum, that includes both passenger and driver mutations, shows a high frequency of transitions (G:C->A:T and A:T->G:C), but also 15% of G:C->C:G transversions (Supplementary Fig. 5). The authors indicated that this spectrum was different from the genome data derived from a lobular carcinoma that displayed 30% of G:C->C:G transversions. It is therefore obvious that breast cancers present major differences in terms of mutation spectrum of reflecting the heterogeneity of their origin. Multiple non-confounding factors could explain this observation, such as exposure to different carcinogens, genetic variations in enzymes associated with xenobiotic metabolism, DNA repair or genetic defects in the BRCA1/BRCA2 genes. Only a detailed analysis of the mutation spectrum of multiple breast tumors corresponding to different histologic types and from different geographical areas will be able to reveal the molecular etiology of breast cancer.

6. Mutation spectra in skin cancer

Epidemiologic studies over several decades have identified UV light exposure as the most important risk factor for melanoma (and other skin cancers) but other factors such as family history of melanoma, skin type and pigmentation or occupations associated with the electronic and chemical industries are also important in the etiology of melanoma [42]. By direct excitation of the DNA molecule, UV generates DNA photoproducts, mostly cyclobutane pyrimidine dimers (CPDs) and pyrimidine(6–4)pyrimidone photoproducts (6,4-PPs) that generate typical mutations, namely C→T transitions, by misincorporation of adenine rather than cytosine during replication. Such mutations are commonly found in UV-induced skin cancers, but are only rarely observed in internal malignancies. These transitions, including the CC→TT tandem mutations, have been termed “UV-signature mutations” and pyrimidine dimers are generally accepted to be the major UV-related premutagenic lesions.

More than 70% of BCC and SCC harbor a TP53 gene mutation, but this frequency is much lower in melanoma, ranging from 10 to 25% [43–45]. The TP53 mutation spectrum in skin cancer is unique with a predominance of C>T transitions at pyrimidine dimers and a high frequency of tandem mutations, two events that have not been observed in internal tumors (Fig. 7). This mutation spectrum has also been observed in UV-induced tumors in the murine TP53 gene of normal mice or in the human TP53 genome of Hupki mice.

The full genome sequence of the Colo-328 melanoma cell lines is now available and the mutation spectrum of 76,085 mutations also displays a UV signature with a high frequency of tandem mutations and transitions at Py–Py sites (Fig. 7) [46]. As indicated for the NCI-H209 cell lines, the majority of mutational events observed in the Colo-829 genome can be considered to be passenger mutations, supporting their value in this type of analysis. The mutation spectrum of the 50,000 mutations described in the NCI-H209 cell line is totally different from the 76,085 mutations in the Colo-829 cell line, mostly comprising passenger mutations, indicating that they have been specifically shaped by the different mutagens associated with these two types of cancer.

7. Mutation spectra in hepatocellular carcinoma

The incidence of HCC varies widely according to geographic location and between ethnic groups. The predominant factors associated with HCC include chronic hepatitis B and C viral infection, chronic alcohol consumption, aflatoxin-B1-contaminated food and virtually all

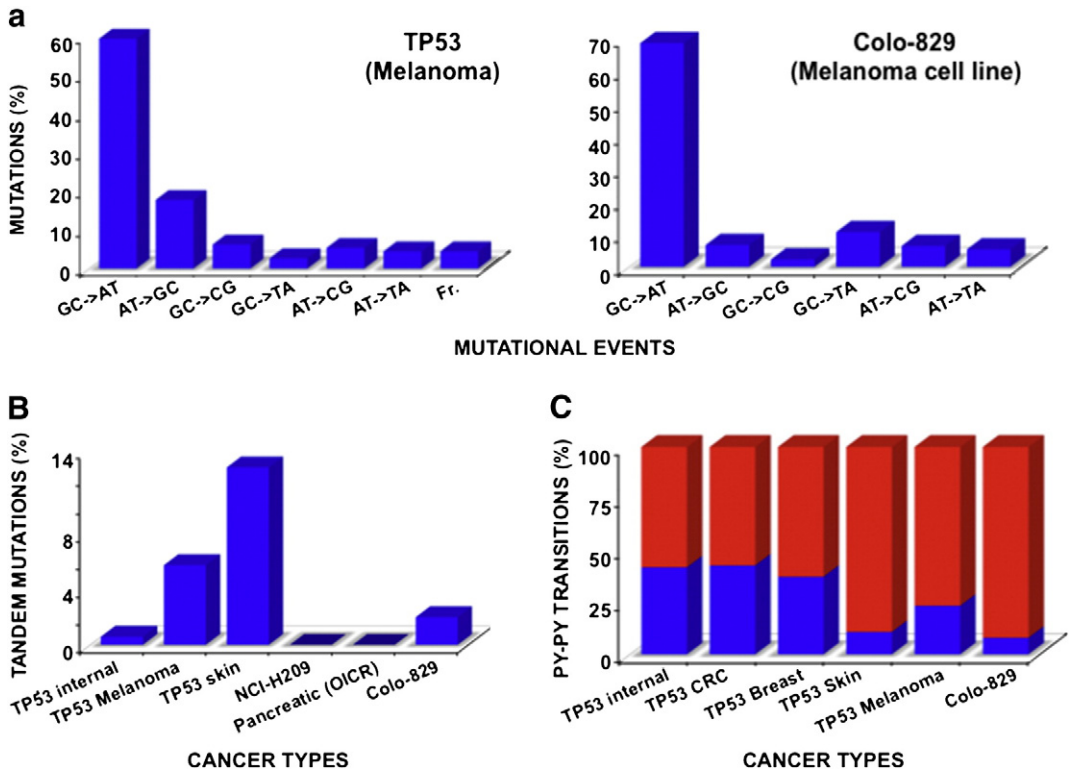

Fig. 7. Mutation spectrum in skin cancer. A: Mutation spectrum of the TP53 gene in melanoma (left) and the whole genome of the Colo-328 cell line (right). B: Frequency of tandem mutations in various cancers. TP53 internal: all TP53 mutations except skin cancer; TP53 melanoma: TP53 mutations in melanoma; TP53 skin: TP53 mutations in skin cancer (except melanoma); NCI-H209: mutation in the whole genome of the NCI-H209 cell line; pancreatic (OICR): mutation in the genome of a pancreatic tumor; Colo-328: mutations in the whole genome sequence of the Colo-328 cell line. C: Frequency of transition at pyrimidine–pyrimidine sequences (Py–Py sites) in various cancers. Red: Py–Py sites; Blue: non Py–Py sites. Full references and data are available in Supplementary Table 2. (Data from the UMDTP53 database: http://p53.free.fr).

cirrhosis-inducing conditions. Analysis of TP53 mutations in HCC has revealed a unique feature with the discovery of a very specific hot spot at codon 249 associated with the deleterious mutation R249S (Fig. 3). Worldwide epidemiologic studies have shown that the mutation in codon 249 is strictly specific to countries in which food is contaminated by aflatoxin B. In countries in which aflatoxin B-contaminated food is not consumed (including Europe and the USA), the rate of p53 mutations in hepatocellular carcinoma is low, and the mutations are scattered along the central part of p53, as in other types of cancer (Supplementary Fig. 4). *In vitro* and *in vivo* studies in human cells have demonstrated that aflatoxin B1 binds to codon 249. This observation, along with the fact that this R249S mutation is highly deleterious for p53 function, explains the existence of this mutational hot spot. It is not clear whether or not there is also a specific selection for this mutation in hepatocytes (Ref. [47] for a detailed review).

The spectrum of TP53 mutations in HCC in Europe or in HCC not related to aflatoxin B1 still shows a larger number of G:C->T:A transversions, suggesting that other, as yet unidentified mutagens could also be associated with the etiology of this cancer (Supplementary Fig. 4). Partial data from two on-going sequencing studies are available and also show a very heterogeneous mutation spectrum, but the number of available mutations is still very low (Supplementary Fig. 4). This heterogeneity most probably reflects the complex etiology of HCC and only complete sequencing of multiple tumors from different origins will allow comprehensive analysis of the mutation spectrum of HCC.

8. Mutation spectra in other types of cancer

Exomic or whole genome sequences from other types of cancer are also available. In prostate carcinoma, G:C->A:T transition has been found to be the major mutational event, whether the study is restricted to the TP53 gene, or is based on the whole exomic sequence or the whole genomic sequence (Supplementary Fig. 5). In the genomic sequencing study the whole genome of 7 prostate tumors has been decrypted and each one displayed a similar mutation spectrum except for localization of the G:C->A:T transition [48]. In the three tumors that displayed a TMPRSS2-ERG translocation, the frequency of transition at CpG dinucleotides was very high. The TMPRSS2-ERG fusion protein has been shown to induce a repressive epigenetic program via direct activation of the H3K27 methyltransferase EZH2. Whether or not this change in the methylation profile is specific to the gene in the androgen receptor pathway or whether it can be extended to the whole genome is not known, but variations in the methylation profile of the cancer genome could change its overall homeostasis and alter DNA repair of endogenous alterations.

The exomic sequence of 316 ovarian carcinoma has been established [49]. TP53 mutations were found in almost all tumors (96%) with a mutational profile identical to previous studies reported in the TP53 mutation database (Supplementary Fig. 6). Both were characterized with a high frequency of G:C->A:T transition and a high frequency (15%) of small insertions and deletions. This is the highest frequency of frameshift mutations found for any cancer in the database.

In hematologic malignancies, studies limited to TP53 or based on the whole genome show a very similar mutation spectrum [50]. All studies show a high frequency of transition at CpG dinucleotides (Supplementary Fig. 6). An as yet unresolved question concerns the origin of these transitions in various types of cancer. Are they due to a very common mutagenic process or only the fingerprints left by heterogeneous mechanisms?

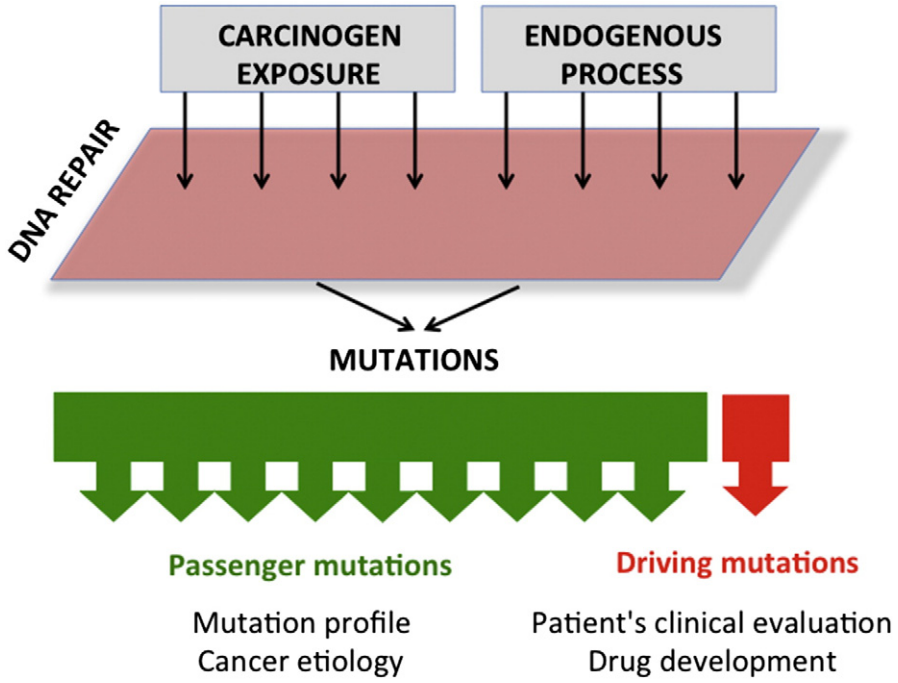

Fig. 8. Shaping mutations in human cancer. In the exposure step, exogenous or endogenous carcinogens lead to the generation of a wide range of DNA adducts. In a second step, most of the lesions are eliminated by DNA repair. In the fixation step, non-repaired lesions are transformed into stable mutations that are transmitted to daughter cells after cell division, resulting in three possible outcomes. First, if the mutation leads to a lethal phenotype, the cell dies and this counterselection prevents propagation of this specific alteration. A second outcome, which is the most common, is that the mutation does not result in any particular phenotype. This is the case for the majority of intergenic or intronic alterations, but also for certain intragenic mutations that target nonessential amino acid residues. Mutations targeting the third base of nucleotide codons also fall into this category, although it should be kept in mind that they can lead to defects in RNA stability or splicing. These "neutral mutations" are not involved in any selection process and are expanded at the rate of normal cell division of the original tissue. Nevertheless, if this particular cell is the target of a second alteration that leads to its clonal expansion, the first mutation will be coselected despite the absence of associated phenotype. Such mutations, called "passenger mutations," as opposed to the "driving mutations" that lead to clonal expansion, are very common, and their frequency is related to the efficiency of the mechanism controlling genetic stability. The third possible outcome for stable somatic mutations is a functional modification that results in a new phenotype that contributes to the neoplastic process. These true driving mutations are fortunately rare, but a selection process can be modulated by factors such as tissue specificity, the presence of other somatic genetic modifications, or the individual's genetic background. Both types of mutations are useful in tumor profiling, but accurate detection of driving mutations, the only mutations useful in clinical practice, is a difficult task.

9. Discussion and prospects

Various endogenous and exogenous mutagens can therefore account for the various mutations observed in human tumors. These mutagens shape the landscape of mutations and leave specific signatures (Fig. 8). Analysis of the various mutation spectra in the TP53 gene has been very fruitful and has revealed important information concerning the etiology of lung, liver, urothelial and skin cancer. With recent progress in sequencing methodology, it is now possible to extend this type of analysis to the whole genome, which not only confirms data derived from TP53 studies, but also demonstrates that all types of alterations, passenger and driver, can be used to track mutagen fingerprints. Therefore, the various biases that could interfere with TP53 studies, such as the low frequency of TP53 mutations in some cancers or the small size of the gene, can now be overcome and the use of passenger mutations for spectrum analysis offers the following advantages: i) they can be found in large numbers in a single tumor, allowing single tumor profiling. Current analyses indicate that 5000 to 50,000 passenger mutations can be found per tumor genome; ii) due to their origin, they are true neutral mutations preventing any bias in the mutation spectrum liable to occur during selection; iii) the lack of selective pressure also allows analysis of each type of cancer, avoiding the problem of specific gene inactivation in particular cancers; iv) the large size of the human genome makes this analysis independent of the target size or sequence allowing detailed analysis of sequence context that was not possible in small genes. This last point will be essential for detailed mutational spectra analysis. Experimental studies have demonstrated that the fingerprints leaved by all carcinogens were strongly influenced by the sequence context surrounding the targeted residue. Analysis of this context in the entire human genome will be of tremendous value to understand how and why DNA sequence will influence the formation of specific adduct. Unfortunately, this type of analysis could not be performed in the present analysis. The format of sequencing data available for the majority of exomic and full genomic sequencing is already partially interpreted and sequence context recovery was not possible. Sequence repositories with fully accessible data need to be established in order that scientist from every fields of investigation can performed every types of analysis. This problem of data format and data mining of whole genome sequences has been extensively discussed in a recent review by Chin et al. [51].

Passenger mutations are considered to be unwanted contaminants of tumor genomes, but the present analysis demonstrates their usefulness and that it would be a very serious mistake to disregard or delete these mutations from databases. Distinguishing driver mutations from passenger mutations is still a complex task, but pooling both types of information does not introduce any bias, as the number of driver mutations is extremely low and both types of mutations are somatic. The only problem at the present time concerns the possible contamination of cancer data by rare germline snp that have not been excluded but the recent improvement of the snp database, in terms of both quality and volume of data, should easily resolve this problem.

The International Cancer Genome Consortium (ICGC) has undertaken a large-scale cancer genome analysis of more than 25,000 tumors in 50 different types/subtypes of cancer [52]. With the forthcoming release of the third generation sequencer, the number of tumor genome sequences will increase tremendously and it is not unrealistic to predict that tumor genome sequencing could be

performed routinely in the next decade. Analysis of this information in each type of cancer and at the level of individual patients would constitute a major breakthrough to identify the etiology of human cancer. Like an anthropologist faced with a collection of old fossil bones, our last remaining daunting task will be the reconstitution of all of the pathways that led to these heterogeneous mutation spectra.

Acknowledgements

Our work is supported by Cancerföreningen i Stockholm, Cancerfonden and the Swedish Research Council (VR). I am grateful to Ted Liefeld for help in providing information on in Myeloma mutations.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.bbcan.2011.07.003.

References

[1] M. Meyerson, S. Gabriel, G. Getz, Advances in understanding cancer genomes through second-generation sequencing, Nat. Rev. Genet. 11 (2010) 685–696.
[2] U. McDermott, J.R. Downing, M.R. Stratton, Genomics and the continuum of cancer care, N. Engl. J. Med. 364 (2011) 340–350.
[3] R.E. Green, et al., A draft sequence of the Neandertal genome, Science 328 (2010) 710–722.
[4] Y.M. Lo, et al., Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus, Sci. Transl. Med. 2 (2010) 61ra91.
[5] P.A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, M.R. Stratton, A census of human cancer genes, Nat. Rev. Cancer 4 (2004) 177–183.
[6] C. Greenman, et al., Patterns of somatic mutation in human cancer genomes, Nature 446 (2007) 153–158.
[7] D.A. Haber, J. Settleman, Cancer: drivers and passengers, Nature 446 (2007) 145–146.
[8] T. Soussi, K.G. Wiman, Shaping genetic alterations in human cancer: the p53 mutation paradigm, Cancer Cell 12 (2007) 303–312.
[9] S.L. Harris, A.J. Levine, The p53 pathway: positive and negative feedback loops, Oncogene 24 (2005) 2899–2908.
[10] M.S. Greenblatt, W.P. Bennett, M. Hollstein, C.C. Harris, Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis, Cancer Res. 54 (1994) 4855–4878.
[11] T. Soussi, C. Beroud, Significance of TP53 mutations in human cancer: a critical analysis of mutations at CpG dinucleotides, Hum. Mutat. 21 (2003) 192–200.
[12] K. Yamagiwa, K. Ichikawa, Experimental study of the pathogenesis of carcinoma, J. Cancer Res. 3 (1918) 1–21.
[13] L.A. Loeb, C.C. Harris, Advances in chemical carcinogenesis: a historical review and prospective, Cancer Res. 68 (2008) 6863–6872.
[14] A. Luch, Nature and nurture — lessons from chemical carcinogenesis, Nat. Rev. Cancer 5 (2005) 113–125.
[15] P. Brookes, P.D. Lawley, Evidence for the binding of polynuclear aromatic hydrocarbons to the nucleic acids of mouse skin: relation between carcinogenic power of hydrocarbons and their binding to deoxyribonucleic acid, Nature 202 (1964) 781–784.
[16] B.N. Ames, W.E. Durston, E. Yamasaki, F.D. Lee, Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection, Proc. Natl. Acad. Sci. U. S. A. 70 (1973) 2281–2285.
[17] S. Rodenhuis, R.J. Slebos, A.J. Boot, S.G. Evers, W.J. Mooi, S.S. Wagenaar, P.C. van Bodegom, J.L. Bos, Incidence and possible clinical significance of K-ras oncogene activation in adenocarcinoma of the human lung, Cancer Res. 48 (1988) 5738–5741.
[18] M.R. Rountree, K.E. Bachman, J.G. Herman, S.B. Baylin, DNA methylation, chromatin inheritance, and cancer, Oncogene 20 (2001) 3156–3165.
[19] S. Tornaletti, G.P. Pfeifer, Complete and tissue-independent methylation of CpG sites in the p53 gene: implications for mutations in human cancers, Oncogene 10 (1995) 1493–1499.
[20] A.P. Bird, DNA methylation and the frequency of CpG in animal DNA, Nucleic Acids Res. 8 (1980) 1499–1504.
[21] P.D. Stenson, E.V. Ball, K. Howells, A.D. Phillips, M. Mort, D.N. Cooper, The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics, Hum. Genomics 4 (2009) 69–72.
[22] T. Sjoblom, et al., The consensus coding sequences of human breast and colorectal cancers, Science 314 (2006) 268–274.
[23] D.W. Parsons, et al., An integrated genomic analysis of human glioblastoma multiforme, Science 321 (2008) 1807–1812.
[24] R. McLendon, et al., Comprehensive genomic characterization defines human glioblastoma genes and core pathways, Nature 455 (2008) 1061–1068.
[25] E.L. Wynder, E.A. Graham, Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases, J. Am. Med. Assoc. 143 (1950) 329–336.
[26] R.N. Proctor, Tobacco and the global lung cancer epidemic, Nat. Rev. Cancer 1 (2001) 82–86.
[27] J.M. Boullin, p53 tumour suppressor gene and the tobacco industry, Lancet 365 (2005) 567.
[28] T. Soussi, B. Asselain, D. Hamroun, S. Kato, C. Ishioka, M. Claustres, C. Beroud, Meta-analysis of the p53 mutation database for mutant p53 biological activity reveals a methodologic bias in mutation detection, Clin. Cancer Res. 12 (2006) 62–69.
[29] F. Cui, M.V. Sirotin, V.B. Zhurkin, Impact of Alu repeats on the evolution of human p53 binding sites, Biol. Direct 6 (2011) 2.
[30] J.L. Luo, Q. Yang, W.M. Tong, M. Hergenhahn, Z.Q. Wang, M. Hollstein, Knock-in mice with a chimeric human/murine p53 gene develop normally and show wild-type p53 responses to DNA damaging agents: a new biomedical research tool, Oncogene 20 (2001) 320–328.
[31] Z. Liu, K.R. Muehlbauer, H.H. Schmeiser, M. Hergenhahn, D. Belharazem, M.C. Hollstein, p53 mutations in benzo(a)pyrene-exposed human p53 knock-in murine fibroblasts correlate with p53 mutations in human lung tumors, Cancer Res. 65 (2005) 2583–2587.
[32] H. Davies, et al., Somatic mutations of the protein kinase gene family in human lung cancer, Cancer Res. 65 (2005) 7591–7595.
[33] L. Ding, et al., Somatic mutations affect key pathways in lung adenocarcinoma, Nature 455 (2008) 1069–1075.
[34] Z. Kan, et al., Diverse somatic mutation patterns and pathway alterations in human cancers, Nature 466 (2010) 869–873.
[35] A. Prat, C.M. Perou, Deconstructing the molecular portraits of breast cancer, Mol. Oncol. 5 (2011) 5–23.
[36] H. Dolezalova, B. Vojtesek, J. Kovarik, Epitope analysis of the human p53 tumour suppressor protein, Folia Biol. (Praha) 43 (1997) 49–51.
[37] B. Weigelt, F.C. Geyer, J.S. Reis-Filho, Histological types of breast cancer: how special are they? Mol. Oncol. 4 (2010) 192–208.
[38] A. Hartmann, H. Blaszyk, J.S. Kovach, S.S. Sommer, The molecular epidemiology of p53 gene mutations in human breast cancer, Trends Genet. 13 (1997) 27–33.
[39] P.J. Stephens, et al., Complex landscapes of somatic rearrangement in human breast cancer genomes, Nature 462 (2009) 1005–1010.
[40] S.P. Shah, et al., Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution, Nature 461 (2009) 809–813.
[41] L. Ding, et al., Genome remodelling in a basal-like breast cancer metastasis and xenograft, Nature 464 (2010) 999–1005.
[42] D.E. Brash, Sunlight and the onset of skin cancer, Trends Genet. 13 (1997) 410–414.
[43] D.E. Brash, J.A. Rudolph, J.A. Simon, A. Lin, G.J. McKenna, H.P. Baden, A.J. Halperin, J. Ponten, A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma, Proc. Natl. Acad. Sci. U. S. A. 88 (1991) 10124–10128.
[44] J.P. Moles, C. Moyret, B. Guillot, P. Jeanteur, J.J. Guilhou, C. Theillet, N. Basset-Seguin, p53 gene mutations in human epithelial skin cancers, Oncogene 8 (1993) 583–588.
[45] N. Dumaz, C. Drougard, A. Sarasin, L. Daya-Grosjean, Specific UV-induced mutation spectrum in the p53 gene of skin tumors from DNA-repair-deficient xeroderma pigmentosum patients, Proc. Natl. Acad. Sci. U. S. A. 90 (1993) 10529–10533.
[46] E.D. Pleasance, et al., A comprehensive catalogue of somatic mutations from a human cancer genome, Nature 463 (2010) 191–196.
[47] F. Staib, S.P. Hussain, L.J. Hofseth, X.W. Wang, C.C. Harris, TP53 and liver carcinogenesis, Hum. Mutat. 21 (2003) 201–216.
[48] M.F. Berger, et al., The genomic complexity of primary human prostate cancer, Nature 470 (2011) 214–220.
[49] D. Bell, et al., Integrated genomic analyses of ovarian carcinoma, Nature 474 (2011) 609–615.
[50] M.A. Chapman, et al., Initial genome sequencing and analysis of multiple myeloma, Nature 471 (2011) 467–472.
[51] L. Chin, W.C. Hahn, G. Getz, M. Meyerson, Making sense of cancer genomic data, Genes Dev. 25 (2011) 534–555.
[52] T.J. Hudson, et al., International network of cancer genome projects, Nature 464 (2010) 993–998.