

MINIREVIEW

# Environmental exposures and mutational patterns of cancer genomes

Gerd P Pfeifer\*

## Abstract

The etiology of most human cancers is unknown. Genetic inheritance and environmental factors are thought to have major roles, and for some types of cancer, exposure to carcinogens is a proven mechanism leading to tumorigenesis. Sequencing of entire cancer genomes has not only begun to provide clues regarding functionally relevant mutations, but has also paved the way towards understanding the initial exposures leading to DNA damage, repair and eventually to mutation of specific sequences within a cancer genome. Two recent studies of melanoma and small cell lung cancer exemplify what type of information can be gained from cancer genome sequencing.

## Origins of human cancer

The origins of human cancers have environmental and hereditary components. Germline mutations of tumor suppressor genes found in cancer predisposition syndromes are prominent examples of inheritance and include well known tumor suppressor genes, such as the retinoblastoma gene (*RB*), *TP53*, the breast cancer genes *BRCA1* and *BRCA2*, the adenomatous polyposis coli gene *APC*, the mismatch repair genes *MLH1* and *MSH2* and a few others. Although mutations in these genes are very rare in the general population, they confer a high risk for developing the disease. Mutations in this group of genes account for only a small fraction of the excess cancer incidence in familial cancer. For some common cancers with significant aspects of heritability, such as prostate cancer, highly penetrant susceptibility genes are still unknown. For these reasons, attention has now shifted towards ascribing much of the observed familial cancer risk to polygenic models of predisposition in which variant alleles, each conferring a small added risk,

cooperate to produce a significant risk factor if several of the adverse alleles are inherited. Many of the high- and moderate-risk genetic mutations conferring enhanced cancer susceptibility in families occur in DNA repair genes, or DNA damage response genes in general, suggesting that some form of DNA damage or replication abnormality is often at the root of cancer initiation.

Recently, genome-wide association studies (GWASs) have provided large datasets for the identification of low-penetrance genes responsible for enhanced cancer susceptibility in the general population. Most of the major cancers have now been investigated by GWASs and close to 100 new cancer-susceptibility loci have been identified [1]. For some cancers with strong environmental components, such as lung cancer, only a few significant loci were found as a result of the overwhelming effect of cigarette smoking on cancer risk, but other cancer types (such as prostate cancer) have yielded many (over 20) such loci. Although GWASs are now capturing the excitement of the cancer genetics community and numerous high-profile studies with large sample sizes and ever-increasing genome coverage are being published, it should not be forgotten that the majority of the cancer risk is thought to be non-genetic (the risk is due to the environment) and this is true for the major human cancers, including prostate cancer, breast cancer and colorectal cancer, for which the heritability accounts for 42%, 27% and 35% of the phenotypic variance, respectively [2]. Thus, in most common cancers, environmental factors supersede the role of genetic inheritance.

Unfortunately, environmental components have convincingly been linked to human cancer in only a few select cases. Most notable are skin cancers associated with sunlight exposure and lung cancer associated with cigarette smoking. Over many decades, epidemiological and molecular studies have established and confirmed this link. Non-melanoma skin cancer is found on sun-exposed skin, and melanoma has been linked with intermittent or recreational sun exposure, in particular in early childhood [3]. Although the ultraviolet B (UVB, 290 to 320 nm) component of sunlight has generally been implicated in these cancers, a role of UVA (320 to 400 nm) cannot currently be excluded.

\*Correspondence: gpfeifer@coh.org  
Department of Cancer Biology, Beckman Research Institute, City of Hope, Duarte, CA 91010, USA

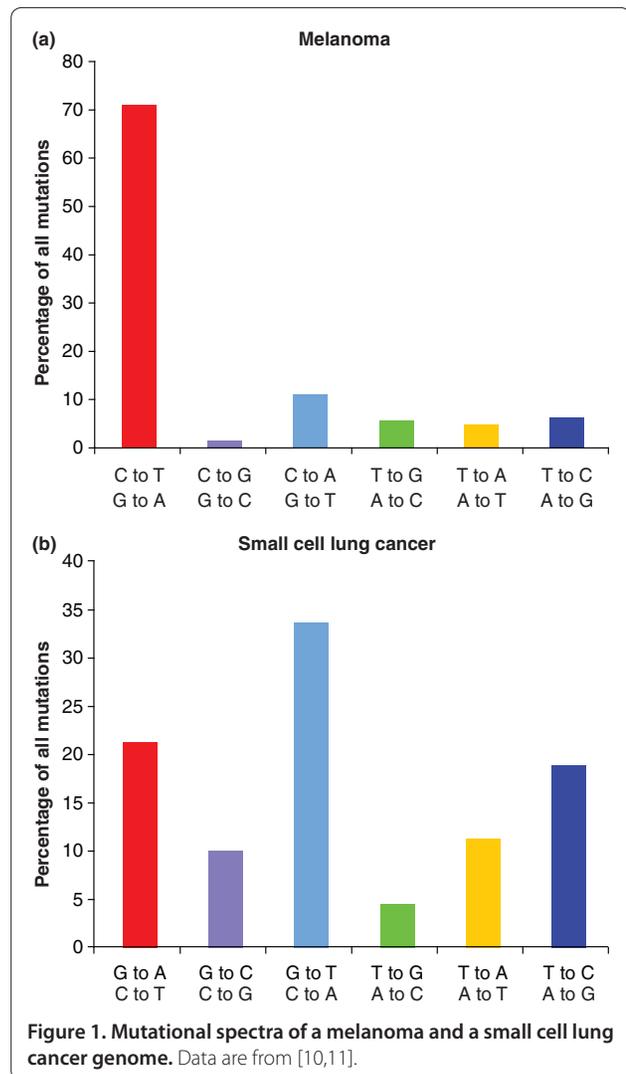
### TP53 mutations

A breakthrough in cancer etiology research has been the demonstration of exposure-specific mutational fingerprints in the *TP53* tumor suppressor gene and in a few other genes that are found mutated in human tumors at a substantial frequency [4,5]. These studies of *TP53* mutations have found UVB-specific mutations, C-to-T transitions at dipyrimidine sequences and CC-to-TT tandem mutations as hallmarks of sunlight exposure leading to non-melanoma skin cancers [6]. The CC-to-TT tandem mutations in *TP53* are almost never found in human tumors not related to sunlight. Similarly, G-to-T transversions, which are particularly enriched at methylated CpG (mCpG) dinucleotide sequences in *TP53*, are characteristic for smoking-associated lung cancers and are much less frequent in lung cancers of non-smokers, or in other cancers not related to smoking [7]. The mCpG-associated G-to-T transversions have been linked to one prominent class of cigarette smoke carcinogens, the polycyclic aromatic hydrocarbons, which have strong selectivity for forming DNA lesions at exactly these DNA sequences [8] and for inducing the same type of mutational events in *in vitro* systems. This mechanistically strengthens the link between smoking and lung cancer [7].

### Insights from whole-genome sequencing of human cancers

Moving beyond mutational studies of important cancer-relevant genes, such as *TP53*, it is now possible to conduct high-throughput sequencing of cancer genomes. Initial reports focusing on sequencing a large number of coding exons have been performed on several types of human cancer, including lung cancer [9]. This year, two articles in *Nature* have expanded our knowledge of environmental carcinogenesis by determining the sequence of the entire genomes of a small cell lung cancer (SCLC) and a melanoma cell line [10,11].

In the first study, the authors [10] sequenced the genome of a melanoma cell line using Illumina short sequence read technology. They identified over 30,000 base substitutions - relative to a lymphoblastoid cell line from the same patient - and various other events, including insertions, deletions, copy number changes and rearrangements. This study is the first comprehensive analysis of a solid tumor genome. Although definitive novel driver mutations in potential cancer-relevant genes were not identified from this single sample, the results gave important clues to the etiology and mechanistic history of how the mutations have arisen as a consequence of UV-induced DNA damage. By far the most common mutation was the C-to-T transition event, accounting for more than two-thirds of all mutations (Figure 1). A total of 92% of the C-to-T mutations occurred at the 3' base of a pyrimidine dinucleotide,



much higher than expected by chance. These mutations are characteristic of UVB-induced DNA damage [12]. The frequency of C-to-T and CC-to-TT mutations due to sunlight exposure is also known to be higher at CpG dinucleotides [12]. C-to-T substitutions (7.7%) and CC-to-TT double substitutions (10.0%) both showed elevated frequencies at CpG dinucleotides compared with that expected by chance (4.4%). Therefore, the mutation spectrum and sequence context indicate that most C-to-T somatic substitutions in the melanoma cell line can be attributed to ultraviolet-light-induced DNA damage.

The mutational landscape of this melanoma cell line is also shaped by DNA repair processes [10]. Nucleotide excision repair is the repair pathway responsible for removing UV-induced pyrimidine dimers. A specialized mechanism of transcription-coupled nucleotide excision repair removes pyrimidine dimers preferentially from active genes and specifically from the transcribed strand of active genes. This repair activity was reflected in the

distribution of C-to-T and CC-to-TT mutations in the melanoma genome, in which these types of mutations were more prominent on the non-transcribed DNA strand of active genes. Genes expressed at a high level showed a lower frequency of somatic mutations than genes expressed at a low level, on both the transcribed and non-transcribed strands. The authors [10] also reported lower mutation prevalence in exons than in introns, but this could be due to negative selection of coding sequence mutations.

The second study [11] focused on a SCLC genome. The authors [11] used the ABI SoliD sequencing platform to generate mate-pair shotgun sequences at more than 30x coverage of the tumor genome and a normal B lymphocyte reference genome from the same individual. This was the first whole-genome sequence of a human lung cancer specimen. Almost 23,000 somatic mutations were identified. The enormous statistical power of this dataset, not affected by selection, gave an elaborate picture of a mutational landscape sculpted by tobacco carcinogen exposure, its sequence preference and several types of DNA repair pathways. As with other similar studies, the fraction of non-synonymous substitutions within protein coding sequences of the cancer genome was not very different from that expected from random events. This means that many tumor genomes will need to be sequenced to identify true tumor-driving mutations. In the SCLC genome, obtained from a type of cancer almost always associated with tobacco smoking, G-to-T transversions were the most frequent changes observed (34% of all mutations; Figure 1). This frequency is remarkably similar to the pattern of substitutions observed in the *TP53* tumor suppressor gene in SCLC cases collected from the International Agency for Research on Cancer *TP53* mutation database [7] and suggests the involvement of tobacco carcinogens in mutation induction [13]. CpG dinucleotides were significantly enriched in the G-to-T mutation set compared with controls. This, again, is consistent with the *TP53* mutational spectra of smokers' lung cancer. G-to-C transversions were more enriched in unmethylated compartments of the genome and were often adjacent to A, that is, they occurred in the GpA sequence context. The origin of such specific G-to-C mutations is currently unknown but they have also been observed in other tumor types [9]. In keeping with what is known about G-to-T transversions in *TP53* and about transcription-coupled and strand-specific repair of bulky carcinogen DNA adducts, the authors [11] found that G-to-T transversions were strongly targeted to the non-transcribed DNA strand of active genes. Significantly lower mutation prevalence, on both transcribed and non-transcribed DNA strands, was observed in more highly expressed genes for G-to-T and also for other types of mutations, suggesting that, in addition to the

strand-specific repair pathway, a repair pathway exists that preferentially removes lesions from both strands of active genes [11].

Recently, Lee *et al.* [14] analyzed the genome of a lung adenocarcinoma using high-throughput sequencing by unchained combinatorial probe anchor ligation chemistry on self-assembling DNA nanoarrays. They found over 50,000 high-confidence single nucleotide variations in the tumor relative to normal lung. In this study as in the others [10,11], transversions at guanine (G to T) were the most common events (46% of all mutations), attesting to the role of tobacco carcinogens in shaping the mutational patterns in this tumor.

## Conclusions

The data presented in these reports [10,11] show the power of whole-genome sequencing to characterize at unprecedented levels of resolution and sequence coverage the many complex mutational signatures found in human cancers induced by environmental exposures. It is expected that additional whole-cancer-genome sequencing datasets will be forthcoming that will cover the same tumor type (to address inter-individual variation or different histological subtypes of cancer). Other cancers for which an environmental origin is known or suspected - for example, aflatoxin-associated liver cancer - will be extremely important to analyze. Furthermore, it is hoped that whole-genome mutational spectra for cancers of unknown etiology - for example, breast or pancreatic cancer - will bring forward new hypotheses regarding potential agents that have compatible mutational specificity and should further be investigated as causative agents of human cancer.

## Abbreviations

GWAS, genome-wide association study; mCpG, methylated CpG; SCLC, small cell lung cancer; UV, ultraviolet.

## Competing interests

The author declares that he has no competing interests.

Published: 16 August 2010

## References

1. Fletcher O, Houlston RS: **Architecture of inherited susceptibility to common cancer.** *Nat Rev Cancer* 2010, **10**:353-361.
2. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer - analyses of cohorts of twins from Sweden, Denmark, and Finland.** *N Engl J Med* 2000, **343**:78-85.
3. Leiter U, Garbe C: **Epidemiology of melanoma and nonmelanoma skin cancer - the role of sunlight.** *Adv Exp Med Biol* 2008, **624**:89-103.
4. Hainaut P, Hollstein M: **p53 and human cancer: the first ten thousand mutations.** *Adv Cancer Res* 2000, **77**:81-137.
5. Hussain SP, Harris CC: **Molecular epidemiology of human cancer: contribution of mutation spectra studies of tumor suppressor genes.** *Cancer Res* 1998, **58**:4023-4037.
6. Brash DE, Rudolph JA, Simon JA, Lin A, McKenna GJ, Baden HP, Halperin AJ, Ponten J: **A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma.** *Proc Natl Acad Sci USA* 1991, **88**:10124-10128.

7. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P: **Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers.** *Oncogene* 2002, **21**:7435-7451.
8. Denissenko MF, Pao A, Tang M-s, Pfeifer GP: **Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53.** *Science* 1996, **274**:430-432.
9. Pfeifer GP, Besaratinia A: **Mutational spectra of human cancer.** *Hum Genet* 2009, **125**:493-506.
10. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordonez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, *et al.*: **A comprehensive catalogue of somatic mutations from a human cancer genome.** *Nature* 2010, **463**:191-196.
11. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordonez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, *et al.*: **A small-cell lung cancer genome with complex signatures of tobacco exposure.** *Nature* 2010, **463**:184-190.
12. Pfeifer GP, You YH, Besaratinia A: **Mutations induced by ultraviolet light.** *Mutat Res* 2005, **571**:19-31.
13. Hecht SS: **Tobacco smoke carcinogens and lung cancer.** *J Natl Cancer Inst* 1999, **91**:1194-1210.
14. Lee W, Jiang Z, Liu J, Haverly PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, Ha C, Johnson S, Kennemer MI, Mohan S, Nazarenko I, Watanabe C, Sparks AB, Shames DS, Gentleman R, de Sauvage FJ, Stern H, Pandita A, Ballinger DG, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z: **The mutation spectrum revealed by paired genome sequences from a lung cancer patient.** *Nature* 2010, **465**:473-477.

doi:10.1186/gm175

Cite this article as: Pfeifer GP: Environmental exposures and mutational patterns of cancer genomes. *Genome Medicine* 2010, **2**:54.