Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Mutational signatures: the patterns of somatic mutations hidden in cancer genomes☆

Ludmil B Alexandrov and Michael R Stratton

All cancers originate from a single cell that starts to behave abnormally due to the acquired somatic mutations in its genome. Until recently, the knowledge of the mutational processes that cause these somatic mutations has been very limited. Recent advances in sequencing technologies and the development of novel mathematical approaches have allowed deciphering the patterns of somatic mutations caused by different mutational processes. Here, we summarize our current understanding of mutational patterns and mutational signatures in light of both the somatic cell paradigm of cancer research and the recent developments in the field of cancer genomics.

Addresses
Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, United Kingdom

Corresponding author: Alexandrov, Ludmil B (la2@sanger.ac.uk)

Current Opinion in Genetics & Development 2014, **24**:52–60

This review comes from a themed issue on **Cancer genomics**

Edited by **David J Adams** and **Ultan McDermott**

0959-437/$ – see front matter,

http://dx.doi.org/10.1016/j.gde.2013.11.014

Introduction

Long before the discovery of the double helix [1], it was well established that ultraviolet light (UV) can cause tumours of the skin [2]. While the mechanism was unclear at this time, it was hypothesized that successive doses of UV radiation result in accelerating the relative rate of cell proliferation [3]. The paradigm shifting discovery that the genetic material is contained within a deoxyribonucleic acid led to many studies in the late 1950s and throughout the 1960s examining how organisms protect their DNA from endogenous and exogenous mutations, and a focus was given to ultraviolet induced mutations (reviewed in Ref. [4•]). It was established that exposure to UV light can lead to the formation of dimers of any two adjacent pyrimidine bases on the same DNA strand with a preference for thymine–thymine dimers [4•]. It was further shown that UV irradiation damage predominantly results in cytosine to thymine or cytosine–cytosine to thymine–thymine changes, preferentially occurring at these pyrimidine dimers (i.e. C > T or CC > TT DNA mutations at dipyrimidine sites) [5,6]. This was the first detailed characterization of the pattern of DNA changes occurring due to the activity of an exogenous mutagen and, as such, the very first description of a signature of a mutational process.

While these early studies established the mutational signature of UV light, it was unclear whether UV induced mutations are present and involved in the neoplastic expansion of human cancers. The development of the DNA sequencing technique with chain-terminating inhibitors by Sanger *et al.* [7] allowed rapid examination of the genetic material contained in cancer cells. In the early 1990s, two studies sequenced exons of the gene *TP53* [8•,9•] from several patients and provided experimental evidence that aflatoxin and UV light leave distinct patterns (consistent with the ones observed in experimental systems) of DNA mutations respectively in hepatocellular and squamous-cell carcinomas. These studies confirmed that the mutational signatures of carcinogens are left as 'evidence' in the genomes of cancer cells [10] thus spawning research which first examined the mutations across *TP53* and later across multiple genes and even whole cancer genomes in order to provide a better understanding of the mutational processes involved in human carcinogenesis.

Mutational patterns of *TP53*

Multiple independent studies used Sanger sequencing of some (or all) exons of a cancer gene to provide clues to the aetiology of both endogenous and exogenous factors of human carcinogenesis. *TP53* was usually selected for this analysis due to its high prevalence of somatic mutations in almost all tumour classes [11••]. Commonly, each of these studies involved multiple samples of a cancer type that were examined for somatic mutations in *TP53* (studies reviewed in Refs. [11••,12,13]). The *TP53* somatic mutations were aggregated, their spectrum was reported as specific for the given cancer type, and this spectrum was then compared to mutations generated experimentally in *in vitro* or *in vivo* systems [11••,13]. It should be noted that the mutational spectra of other genes, albeit rarely, were also used for such analysis [14].

These early studies revealed a significant heterogeneity of the *TP53* spectra across different cancer types, which allowed associating some patterns of mutation to known carcinogens. Here, we provide a brief summary of some of the more important findings while details could be found in Refs. [11••,12,13]. The *TP53* spectrum of skin carcinomas exhibited C > T and CC > TT mutations at dipyrimidines (all substitutions and dinucleotide substitutions are referred to by the pyrimidine(s) of the mutated Watson-Crick base pair). This was consistent with the *in vitro* described mutational signature of UV light. The *TP53* mutational spectrum derived from lung cancers in tobacco smokers was overwhelmed by C > A substitutions, which coincided with the class of mutation produced experimentally as a result of bulky adduct formation by tobacco carcinogens on guanine [15]. In other tobacco associated cancers, such as oesophageal and head and neck tumours, C > A mutations (while still ubiquitous) were less common while there was a significant increase of T > C mutations. Interestingly, in both smokers and non-smokers, C > T and C > G mutations at non-CpG sites were elevated when compared to all other cancer types, with bladder tumours harbouring the most C > G mutations [11••]. Additionally, it was demonstrated that C > A transversions were common in hepatocellular cancers and these mutations were believed to be associated with aflatoxin, a known carcinogen commonly found in food from southern Africa and Asia [16]. Lastly, all cancer types harboured at least some C > T mutations at CpG dinucleotides (mutated base underlined), a process attributed to the normal cellular event of deamination of 5-methylcytosine [11••].

The analyses of *TP53* spectra were the first attempts to bridge the gap between molecular cancer genetics and epidemiology [17]. The large number of studies examining *TP53* spectra required a computational resource to facilitate and retrieve the already identified somatic mutations. At first these data were managed by the researchers that were generating it but in 1994 the International Agency for Research on Cancer (IARC) started to maintain a database while providing a free access to it [17]. The first release of the IARC *TP53* database contained ~3 000 somatic mutations [18] while the most recent version (R16) released in November of 2012, which can be found at http://p53.iarc.fr/, contains almost 30 000 somatic mutations in *TP53*.

Though extremely informative, the data gathered from single gene studies have significant limitations. In these studies, the spectrum of a cancer type is reported by aggregating mutations from multiple samples. This may be adequate when a single mutational process generates the majority of mutations in the particular cancer (e.g. UV light is the predominant mutational process in melanoma [19••]). However, usually multiple mutational processes are operative in a single cancer sample, and combining their mutations generates a mixed composition of the patterns of somatic mutations. In most cases, reporting this jumbled spectrum is uninformative for the diversity of mutational processes operative in a single cancer type or in a single cancer sample [20••]. Moreover, the examined *TP53* exons are both under selection and also have a specific nucleotide sequence. This affects the opportunity for observing a somatic mutation and as such the reported spectrum can be a reflection of the processes of selection and/or the nucleotide architecture of the *TP53* gene in addition to the processes of mutation [21,22].

Two studies tried to overcome some of the single gene limitations by leveraging a targeted capillary sequencing approach of large number of genes. A survey of the 518 protein kinase genes in 25 human breast cancer samples revealed 92 somatic mutations (90 substitutions and 2 indels) in which C > T transitions and C > G transversions preceded by thymine (i.e. C > T and C > G at TpC, mutated base is underlined) occurred with a higher than expected frequency [23]. This survey was later expanded to 210 cancer samples and it revealed more than 1 000 somatic mutations with significant variations in their patterns across the examined twelve cancer types [24]. Only a small fraction of the mutations reported in these screens are likely to be affected by selection [25], thus indicating that the observed mutational patterns reflect the operative mutational processes in the analyzed samples and not the processes of negative or positive selection.

Mutational patterns identified in next generation sequencing data

The development of second-generation sequencing technologies allowed examination of cancer exomes (i.e. the combined protein coding exons) and even whole cancer genomes. Sequencing cancer exomes has been generally preferred as the majority of known cancer-causing driver somatic substitutions, indels, and copy number changes (although generally not rearrangements) [21] are located in protein coding genes. As the nucleotide sequence of protein coding genes is ~1% of the whole genome, analysis of exomes is considered an advantageous and cost effective methodology for discovering the genes involved in neoplastic development. As a result, many studies have focused predominantly on the generation and analysis of exome sequences [26].

Early next generation sequencing studies started revealing patterns of somatic substitutions in different cancer types. In 2010, two back-to-back studies in *Nature* reported the patterns of somatic mutations in a malignant melanoma [27•] and small cell lung carcinoma [28•]. As expected, a strong signature of tobacco carcinogens was found in the genome of the lung cancer, while the mutational signature of ultraviolet light overwhelmed