Mutational spectra and signatures

David H. Phillips, King's College London

A variety of experimental systems have been used to study the endogenous and exogenous factors driving mutagenesis. Traditionally, experimental mutagenesis studies have been limited to the analysis of mutations in a single gene (*e.g. HPRT*, *lacZ*, *cII*, *TP53*), which were identified in tumours or by specifically selecting for the growth of mutated cells or clones from mutagen-treated cell populations. Because each tumour or cell clone harboured only one or two mutations in a particular gene, patterns of mutations were inferred through pooling data collected from many samples, sometimes from different experiments. Nevertheless, such studies have revealed patterns, or spectra, of mutations that are characteristic of particular environmental exposures; for example aflatoxin B_1 and liver cancer, aristolochic acid and urothelial cancer, UV and melanoma.

Massively-parallel next-generation sequencing (NGS) technology has resulted in an extraordinary increase in the speed and scale of sequencing, permitting the exploration of all protein-coding exons (exome sequencing) or whole genomes (whole-genome sequencing, WGS) in samples from patients or experimental model systems. This technology enables the detection of hundreds or even thousands of mutations in a single sample, increasing the power of each experiment considerably. Furthermore, the distribution of mutations throughout the genome can now be explored to gain further insights into mutagenic mechanisms. The complex biological insights buried within these large, multi-dimensional datasets can be dissected using mathematical separation approaches. In the COSMIC database, at least 30 distinct mutation signatures have been extracted from WGS data across 40 different cancer types, including some signatures associated with exposure to carcinogens, such as tobacco smoke in lung cancer and UV radiation in malignant melanoma. Many novel signatures have also been uncovered and the race is on to understand their aetiology.

Thus, whole genome sequencing of human tumours has revealed distinct patterns of mutation that hint at the causative origins of cancer. This can be tested by examining the mutational signatures induced in experimental systems by putative cancer-causing agents. Signatures have been generated in mutagen-exposed mouse embryo fibroblasts (MEFs) human induced pluripotent stem (hiPS) cells and other cell systems. The results reveal that each mutagen induces a characteristic mutation signature that, in some cases, matches a signature found in human tumours.

In an analysis of somatic mutations in cancers for which tobacco smoking confers an elevated risk, it was found that smoking is associated with increased mutation burdens of multiple different mutational signatures, which contribute to different extents in different tissues. One of these signatures, mainly found in tissues directly exposed to tobacco smoke, is attributable to misreplication of DNA damage caused by tobacco carcinogens, as it similar to that of benzo[a]pyrene, a tobacco carcinogen. Others likely reflect indirect activation of DNA editing by APOBEC cytidine deaminases and of an endogenous clock-like mutational process. The results are consistent with the proposition that smoking increases cancer risk by increasing the somatic mutation load although direct evidence for this mechanism is lacking in some cancer types. These proofs-of-principle suggest that mutational spectra will shed light on other causes of human cancer.