Review

CellPress

Dissecting genetic and environmental mutation signatures with model organisms

Romulo Segovia[1,2*], Annie S. Tam[1,2*], and Peter C. Stirling[1,2]

[1] Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, V5Z 1L3, Canada
[2] Department of Medical Genetics, University of British Columbia, Vancouver V6T 1Z3, Canada

Deep sequencing has impacted on cancer research by enabling routine sequencing of genomes and exomes to identify genetic changes associated with carcinogenesis. Researchers can now use the frequency, type, and context of all mutations in tumor genomes to extract mutation signatures that reflect the driving mutational processes. Identifying mutation signatures, however, may not immediately suggest a mechanism. Consequently, several recent studies have employed deep sequencing of model organisms exposed to discrete genetic or environmental perturbations. These studies exploit the simpler genomes and availability of powerful genetic tools in model organisms to analyze mutation signatures under controlled conditions, forging mechanistic links between mutational processes and signatures. We discuss the power of this approach and suggest that many such studies may be on the horizon.

Mutation signatures and mutational processes in cancer genomes

Mutational processes in cancer have long been an area of intense interest and are summarized clearly in a recent review [1]. Historically, the association between a suspected carcinogen and the types of mutations it may cause were determined at specific loci, such as the frequently mutated *TP53* (tumor protein p53) locus in human cancer cells [2–4]. The type of mutations identified, together with their context, is also called the mutation spectrum of a carcinogen and it provides information about the range of events that occur as a direct result of exposure. With advances in whole-genome sequencing (WGS) technologies and progressively improving bioinformatics pipelines, it is now possible to sequence entire cancer genomes and call variants with high sensitivity. The impact of next-generation sequencing in understanding cancer mutation signatures became clear in 2010 when WGS of a melanoma revealed a signature of mutations indicating exposure to UV light [5], and sequencing of a small cell lung cancer genome revealed mutations consistent with tobacco exposure [6]. These were landmark studies that revealed mutation signatures derived from a single cancer genome could be a powerful indicator of the mutational processes that shaped tumorigenesis. Patient and treatment history can certainly help to guide the interpretation of mutational spectra and correlate molecular mechanisms of carcinogens to tumor mutation patterns (e.g., arsenic-exposed lung cancer [7]). However, in many instances in the clinic, prior exposure to a carcinogen or the contributions of different genome-destabilizing factors are unknown. These unknown factors confound the interpretation of WGS results, and delineating causative factors of tumor genome instability is difficult without first understanding the independently-contributing processes that drive tumor evolution (Figure 1). Despite these difficulties, with the surge of cancer genome-sequencing projects, it is now possible to extract significant mutation signatures appearing across many cancer genomes (Box 1). Indeed, rigorous computational approaches exist to extract parallel mutation signatures from pan-cancer data [8,9] or for a specific cancer type with multiple samples [10,11]. This method has tremendous power to categorize causes of mutations in cancer and enable researchers to exploit that information to improve prevention or guide prognosis and treatment [1,12,13].

Mutational mechanisms implicated in specific tumor types through mutation signatures include age and genetic perturbations in genome maintenance proteins. In particular, defects in *BRCA1* (breast cancer 1, early onset) and *BRCA2* showed genomic signatures that were strongly associated with these mutations, indicating a signature unique to DNA repair by homologous recombination [8]. An interesting and potentially important new mechanism to emerge from mutation signature analysis is the role of dysregulated cytidine deaminase activity as a causative factor of clustered C>T and C>G mutations in the TpC context in breast cancer and other cancer genomes [8,11,14,15]. This process of clustered mutation generation has been named 'kataegis' and has been linked quite conclusively to the action of cytidine deaminases of the APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide) family [8,11,14–19]. It is clear the mutation signatures identified in cancers are a consequence of DNA damage and subsequent DNA maintenance responses. However, linking a specific DNA repair process

Keywords: mutation signature; cancer genomics; mutagens; deep sequencing; model organisms
*These authors contributed equally.

0168-9525/
 http://dx.doi.org/10.1016/j.tig.2015.04.001

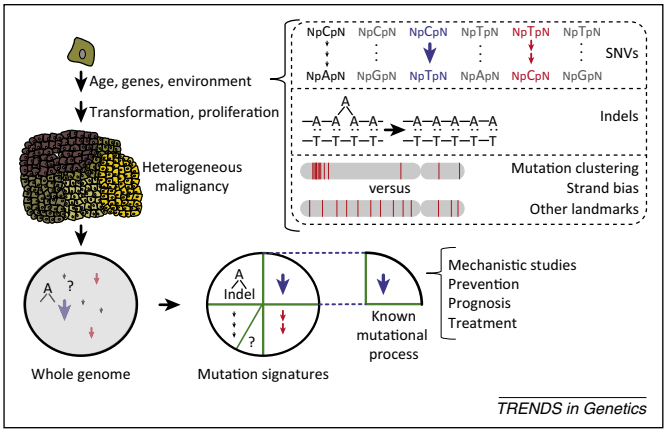

Figure 1. Dissecting the complexity of mutation signatures in tumors. A cell lineage (green) is exposed to various mutational processes before oncogenesis (i.e., age-associated, genetic, environmental). After transformation and during proliferation additional cancer-specific mutational processes occur. These mutational processes may have different signatures of single-nucleotide variants (black, blue, and red arrows), insertions and deletions (indels), or in the position of mutations relative to higher-order genomic features (e.g., the next closest mutation, transcriptional status, etc.). This leads to a heterogeneous tumor, from which whole-genome sequencing captures the composite of all mutational processes (grey shaded circle). Mutation signatures can be extracted computationally [1,8,9] and in some cases linked to known drivers of mutational processes (blue arrow). This information can guide further mechanistic studies; if the process is environmental, it can guide cancer prevention; if the process is correlated with clinical data, it can guide prognosis; or, if the process is ongoing, it could guide treatment options.

to a genomic signature is not directly feasible because there are parallel DNA repair pathways that act at varying times, and changing mutagenic exposures may alter the genomic landscape. Differentiating past and ongoing mutational processes in cancer genomes will be important in understanding the contributing factors that lead to carcinogenesis, and in matching the tumor phenotype with appropriate existing therapies, or potentially to suggest new therapeutic targets.

Currently, the identification of mutation signatures in cancer genomes has surged ahead of our ability to explain their root causes and mechanisms [1,8]. Of 22 pan-cancer mutation signatures identified to date, only 11 signatures have a proposed or confirmed mechanism [8]. As more data are collected, additional unpublished signatures are also emerging and will require mechanistic explanation [20]. The ability to interpret an observed mutation signature as a process of known origin and mechanism would impact on cancer prevention by conclusively identifying risk factors. Furthermore, ongoing mutational processes in tumors may reflect actionable features, influencing prognosis and treatment. Although more tumor sequencing, and studies linking genotype and phenotype in cancer cells, may suggest new mechanisms for mutation signatures, an alternative proposal gaining popularity has been to use genetically tractable model organisms to generate mutation signatures under controlled settings [8,21–24]. We review the growing number of studies on whole genome mutation analysis in model organisms and their potential convergence with the identification of mutation signatures in human cancer.

Mutation rates and patterns in model organisms

The majority of mutations are deleterious; therefore, there is selective pressure to maintain DNA fidelity, the importance of which is reflected in the multitude of overlapping DNA repair pathways seen across species. However, a small number of acquired mutations can be advantageous to cells, which accounts for the proliferative nature of cancer cells. Determining the causes of variations in mutation rate may be important for understanding the biological significance of such variations and potentially their role in disease. Model organisms, primarily *Escherichia coli*, yeast (*Saccharomyces cerevisiae*), nematodes (*Caenorhabditis elegans*), and flies (*Drosophila melanogaster*), have long been used to assess mutation rates. Estimations of mutation rates conducted in these organisms have used fitness-based evolution experiments or selectable reporter loci [25]. These rates could be compared to human mutation rates determined by analysis of target loci in closely related species separated by a known evolutionary timescale (i.e., human and chimpanzee) [25,26]. More recently, deep-sequencing based estimates of human mutation rates using family trios or founder populations have been created and provide better estimates of human mutation rate for comparison [27,28].

Mutation reporters calculate mutation rate per generation at a single locus under test conditions, whereafter the target locus can be sequenced to identify the spectrum of mutations generated in the experiment. For example, the rate of *CAN1* (arginine permease) loss-of-function mutations in yeast, *unc-22* (uncoordinated 22) inactivation in *C. elegans* [29], or the famous Ames test in which

Box 1. Describing mutation signatures in cancer genomes

From the large-scale studies conducted to date, conventions have emerged as to the information required to describe a mutation signature. The important properties of a signature are reflected in what is usually the most abundant type of mutation – SNVs. SNVs are expressed as one of six possible mutations starting from the pyrimidine moiety of the basepair. In this way C to T (written C>T) and T>C transition mutations, and C>A, C>G, T>A, and T>G transversion mutations, encompass all possible SNVs. The immediate 5′ and 3′ bases also influence mutation signatures, thus each of the six SNV types becomes 16 possible trinucleotide combinations written as NpCpN or NpTpN, where N is any base (see Figure 1 in main text), totaling 96 possible trinucleotide mutations [11]. Mutations can be strand-biased such that a particular mutation type will exhibit preference for the non-transcribed strand [5,6,8,11]. This is taken to indicate the action of transcription-coupled repair processes on the mutation profile. Mutations often occur in a coordinated fashion, and when SNVs occur closer to one another than expected it indicates a unique mutational process. Mutation clustering at specific loci has been linked to various processes including error-prone repair processes related to ssDNA surrounding breakpoints and to hyperactivity of cytidine deaminases [11,14,16,69]. Additional features can also contribute to the uniqueness of mutation signatures. For example, double substitution of two adjacent nucleotides, in particular CC>TT, is observed in signatures that can be associated with UV irradiation [5,8,74]. Similarly, patterns of insertion and deletion (indel) mutations have also been observed in cancers and are proposed to relate to mutations in DNA repair pathways, in particular those that promote microhomology-mediated end joining (MMEJ) [8,11,20]. The lower frequency of detected indels makes strong indel signatures harder to identify. Similarly, there are likely to be characteristic signatures of other types of structural rearrangements (e.g., duplications, inversions, translocations); however, the smaller number of these events in a tumor genome, coupled with the difficulty in detecting them with short-read sequencing technologies, means that they currently contribute less to our understanding of mutation profiles [20].