Signatures of Mutational Processes in Human Cancer

Ludmil B. Alexandrov[1], Serena Nik-Zainal[1], David C. Wedge[1], and Michael R. Stratton[1]

***Short Abstract* — Cancer genomes carry somatic mutations that are the cumulative consequence of the DNA damage and repair processes operative during their cellular lineage. Remarkably, these mutational processes are poorly characterized. Here, we introduce a computational framework for deciphering signatures of mutational processes. We apply it to 7,042 cancers, including most major classes, and extract more than 20 distinct mutational signatures. Certain signatures are associated with known mutagenic exposures or defects in DNA maintenance, but many are of cryptic origin. The results reveal the landscape of mutational processes in human cancer with implications for understanding of cancer etiology, prevention and therapy.**

***Keywords* — Cancer genomics, Mutational processes**

I. Introduction

All cancers are caused by somatic mutations [1]. These may be the consequence of the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA or defective DNA repair. In some cancer types, a substantial proportion of somatic mutations is known to be generated by exposures, *e.g.,* tobacco smoking in lung and ultraviolet light in skin cancers [2].

Different mutational processes often generate different combinations of mutation types, termed "signatures." For example, in smoking-induced lung cancers C:G>A:T transversions predominate, a signature compatible with DNA damage induced by known tobacco carcinogens [2].

Our understanding of the mutational processes that cause somatic mutations in most cancer classes is remarkably limited [1]. Previously, we identified the mutational signatures in 21 breast cancer genomes [3]. Here, we describe our computational approach [4] for deciphering signatures of mutational processes from cancer genomics data and leverage it to survey the mutational signatures and processes operating across the spectrum of human neoplasia.

II. Results

A. Computational Framework

The mutational catalog of a cancer cell, M, can be examined as an approximate linear superposition of the signatures of mutational processes, P, active at some point in the lineage of cells leading to the cancer cell and the intensities of their exposures, E:

$$M \approx P \times E$$

where M, P, and E are nonnegative matrices determined by the examined mutation types. We want to find P and E while only knowing the matrix M. This problem belongs to a class of blind source separation problems and we recently provided an effective theoretical model and computational solution [4]. From a set of mutational catalogues, the algorithm deciphers the minimal set of signatures that optimally explains the proportion of each mutation type found in each catalogue.

B. The landscape of mutational signatures

Applying our approach to somatic mutations derived from 7,042 cancers of 30 different classes revealed 21 distinct patterns of mutational signatures and 3 technology specific sequencing artefacts. Some mutational signatures were characterized by only one type of substitutions, *e.g.*, a T>C signature in liver cancers. We identified mutational signatures with a specific sequencing context, *e.g.*, C>X at TpCpN (mutated base is underlined), strong strand bias, or association with small insertions and deletions. In most cancer classes more than two mutational signatures were observed, with a maximum of six in cancers of the liver, uterus and stomach.

C. Associating cancer etiology and mutational signatures

We were able to propose the underlying causes for 8 of the 21 mutational signatures. We associated these signatures respectively with: patient aging; activity of APOBEC family of cytidine deaminases; mutations in BRCA1/2; IGHV mutations; failure of mismatch repair mechanisms; tobacco smoking; exposure to UV light; and chemotherapy.

D. Case Study: Mutational Signatures in Lung Cancer

We identified 4 mutational signatures in lung adenocarcinomas from 660 patients. Two signatures are induced by cigarette smoking (75% of all mutations), one is due to APOBEC activity (15%), and one to aging (10%).

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Cambridge

References

[1] Stratton MR (2011) Exploring the Genomes of Cancer Cells: Progress and Promise. *Science* **331**, 1553-1558.

[2] Pfeifer GP (2010) Environmental exposures and mutational patterns of cancer genomes. *Genome Med* **2**, 54-58.

[3] Nik-Zainal S, Alexandrov LB et al. (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993.

[4] Alexandrov LB, Nik-Zainal S et al. (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-259