ESSAY

SCIENCE & SCILIFELAB PRIZE

Understanding the origins of human cancer

Analyzing the DNA sequences of more than 12,000 cancer patients revealed signatures of mutational processes

By **Ludmil B. Alexandrov**

All cancers originate from a single cell that starts to behave abnormally, to divide uncontrollably, and, eventually, to invade adjacent tissues (*1*). The aberrant behavior of this single cell is due to somatic mutations—changes in the genomic DNA produced by the activity of different mutational processes (*1*). These mutational processes include exposure to exogenous or endogenous mutagens, abnormal DNA editing, the incomplete fidelity of DNA polymerases, and failure of DNA repair mechanisms (*2*). Early studies that sequenced *TP53*, the most commonly mutated gene in human cancer, provided evidence that mutational processes leave distinct imprints of somatic mutations on the genome of a cancer cell (*3*). For example, C:G>A:T transversions predominate in smoking-associated lung cancer, whereas C:G>T:A transitions occurring mainly at dipyrimidines and CC:GG>TT:AA double-nucleotide substitutions are common in ultraviolet light–associated skin cancers. These patterns of mutations matched the ones induced experimentally by tobacco mutagens and ultraviolet light, respectively, the major, known, exogenous carcinogenic influences in these cancer types, and demonstrated that examining patterns of mutations in cancer genomes can yield information about the mutational processes that cause human cancer (*4*).

When I started my Ph.D. at Mike Stratton's lab at the Wellcome Trust Sanger Institute, large-scale global initiatives, such as the International Cancer Genome Consortium, had started performing molecular characterization of thousands of cancer patients around the world (*5*). However, at that time, there had only been limited characterization of patterns of mutations imprinted by mutational processes. During my Ph.D. studies, I explored the possibility of leveraging the available cancer genomics data to elucidate the mutational processes operative in human cancer. I started by conceptualizing the problem and developing a mathematical model that describes the interconnection between the activity of mutational processes in cancer cells and the mutational catalogs generated by next-generation sequencing of cancer genomes (*6*). The mathematical model was subsequently used to develop a computational approach (*6*), which I later applied to thousands of sequenced human cancers (*7*).

Biologically, the somatic mutations in a cancer genome are the cumulative result of the mutational processes that have been active since the very first division of the fertilized egg from which the cancer cell was derived (*2*). Different mutational processes often generate unique combinations of mutation types, and we termed these patterns "mutational signatures." Multiple distinct mutational signatures may be recorded on the genome of a single cancer cell and, as such, an individual cancer genome is insufficient for identifying all imprinted mutational signatures. However, the availability of thousands of samples in which mutational signatures are present with different frequencies makes it possible to decipher their patterns. Mathematically, a set of mutational catalogs of cancer genomes could be examined as a linear mixture of unknown numbers of mutational signatures. The mutational catalogs of these cancer genomes are known from DNA sequencing, and the aim is to identify the patterns of the mutational signatures as well as the number of mutations attributed to each signature in each sample. This problem belongs to a well-known class of blind source separation (BSS) problems, in which mixtures of recordings need to be separated with very little information about the underlying mixing process. To solve this cancer-specific BSS problem in a practical way, I developed a computational framework that uses the previously established multiplicative update algorithm for non-negative matrix factorization (*8*). The framework was extensively evaluated with simulated and real data, demonstrating that it allows one to accurately identify mutational signatures both from whole-genome and whole-exome sequenced samples (*6*).

Initially, I applied the developed computational framework to the somatic mutations found in 21 whole-genome sequenced breast cancers (*9, 10*). Analysis revealed the existence of multiple distinct mutational signatures (*9*), and we were able to explore the activity of these signatures over time (*10*). This initial application of the developed computational framework was followed by a comprehensive global analysis of mutational signatures across the spectrum of human neoplasia (*7*). I curated the majority of publicly available data and compiled a data set encompassing ~5 million somatic mutations from the mutational catalogs of 7042 primary cancers of 30 different classes. These data revealed the existence of 21 distinct mutational signatures in human cancer. Some were present in many cancer types, notably a signature attributed to the APOBEC family of cytidine deaminases (*7,*

*Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos, NM 87545, USA. *E-mail: lba@lanl.gov*

PHOTO: CREDIT GOES HERE AS SHOWN; CREDIT GOES HERE AS SHOWN

CATEGORY WINNER: GENOMICS AND PROTEOMICS

Ludmil Alexandrov

Ludmil Alexandrov is an Oppenheimer Fellow in the Theoretical Biology and Biophysics Group at Los Alamos National Laboratory. He earned his Bachelor of Science degree in Computer Science from Neumont University and received his Master's of Philosophy in Computational Biology as well as his Ph.D. in Cancer Genetics from the University of Cambridge. He is a recipient of the 2015 Weintraub Award for Graduate Research and, in 2013, he was listed by Forbes magazine as one of the "30 brightest stars under the age of 30" in the field of Science and Healthcare. His work is focused on understanding the mutational processes responsible for human cancer and human aging. In 2015, his research was highlighted by the American Society of Clinical Oncology as an important step forward in the fight against cancer.

11); others were confined to a single cancer class. For some of these processes, the underlying biological mechanism is still unknown. However, some of the identified mutational signatures were associated with age of cancer diagnosis, tobacco smoking, exposure to ultraviolet light, treatment with anticancer drugs, presence of *BRCA1* or *BRCA2* mutations, activity of polymerase η, activity of polymerase ε, and inactivation of mismatch repair genes.

The performed comprehensive pan-cancer analysis was complemented by a plethora of studies focusing on individual cancer types. In the last year of my Ph.D. studies, I contributed to further elaborating the understanding of mutational signatures in breast cancer (*12*), prostate cancer (*13*–*15*), liver cancer (*16*), renal cancer (*17*), B cell lymphoma (*18*), a diverse set of childhood cancers (*19*), multiple myeloma (*20*), and acute lymphoblastic leukemia (*21*). Additionally, I participated in mapping the signatures of the somatic mutational processes in human mitochondria (*22*) as well as in understanding the mutational processes operative in normal somatic cells (*23, 24*). Overall, the pan-cancer analysis and the hitherto mentioned research resulted in identifying 30 distinct signatures of somatic

Signatures of mutational processes in human cancer

Detailed patterns of the mutational signatures as well as most up-to-date information could be found at our website (*25*).

SIGNATURE NUMBER	CHARACTERISTIC MUTATIONAL PATTERN	MOST COMMON CANCER TYPES	PROPOSED ETIOLOGY	ETIOLOGY PROPOSED BASED ON:
Signature 1	C>T at CpG	All cancer types	Deamination of 5-methylcytosine	Similarity of the mutational pattern
Signature 2	C>T at TpC	Twenty-two different cancer types	*APOBEC1*, *APOBEC3A*, or *APOBEC3B*	Similarity of the mutational pattern
Signature 3	Uniform mutational signature	Breast, ovarian, and pancreatic cancer	Defective repair of DNA double-strand breaks based on homologous recombination	Statistical association with mutations in *BRCA1* or *BRCA2*
Signature 4	C>A mutations with strand bias	Lung, head and neck, and liver cancer	Tobacco smoking	Similarity of the mutational pattern and statistical association
Signature 5	Mostly uniform mutational signature with some peaks of T>C mutations at ApT	All cancer types	Unknown etiology	N/A
Signature 6	C>A mutations and C>T at GpC mutations	Seventeen different cancer types but most prevalent in colorectal and uterine cancers	Defective DNA mismatch repair	Similarity of the mutational pattern and statistical association
Signature 7	C>T at dipyrimidines	Malignant melanoma and lip cancers	Ultraviolet light	Similarity of the mutational pattern
Signature 8	C>A mutations with a moderate strand bias	Breast cancer and medulloblastoma	Unknown etiology	N/A
Signature 9	T>G transversions at ApT and TpT	Chronic lymphocytic leukemias and B-cell lymphomas	Polymerase η	Similarity of the mutational pattern and statistical association
Signature 10	C>A at TpCpT and C>T at TpCpG	Colorectal and uterine cancers	Polymerase ε	Statistical association
Signature 11	C>T substitutions	Malignant melanoma and glioblastoma multiforme	Treatment with temozolomide	Similarity of the mutational pattern and statistical association
Signature 12	T>C substitutions with strand bias	Liver and uterine cancer	Unknown	N/A
Signature 13	C>A and C>G at TpC	Twenty-two different cancer types	*APOBEC1*, *APOBEC3A*, or *APOBEC3B* and *REV1*	Similarity of the mutational pattern
Signature 14	C>A mutations and C>T at GpC mutations	Low grade glioma and uterine cancer	Unknown etiology	N/A
Signature 15	C>T at GpC mutations	Stomach and lung cancer	Defective DNA mismatch repair	Similarity of the mutational pattern
Signature 16	T>C mutations at ApT with extremely strong strand-bias	Liver cancer	Unknown etiology	N/A
Signature 17	T>G at TpT and T>C at CpT	Esophagus cancer, liver cancer, stomach cancer, and B-cell lymphoma	Unknown etiology	N/A
Signature 18	C>A mutations	Neuroblastoma	Unknown etiology	N/A
Signature 19	C>T mutations	Pilocytic astrocytoma	Unknown etiology	N/A
Signature 20	C>A and C>T mutations	Stomach cancer	Defective DNA mismatch repair	Similarity of the mutational pattern

Signatures of mutational processes in human cancer (continued)

Detailed patterns of the mutational signatures as well as most up-to-date information could be found at our website (*25*).

SIGNATURE NUMBER	CHARACTERISTIC MUTATIONAL PATTERN	MOST COMMON CANCER TYPES	PROPOSED ETIOLOGY	ETIOLOGY PROPOSED BASED ON:
Signature 21	T>C mutations	Stomach cancer	Unknown etiology	N/A
Signature 22	T>A mutations	Urothelial (renal pelvis) carcinoma and liver cancers	Exposure to aristolochic acid	Similarity of the mutational pattern and statistical association
Signature 23	C>T mutations	Liver cancer	Unknown etiology	N/A
Signature 24	C>A mutations with strand bias	Liver cancers	Exposures to aflatoxin	Similarity of the mutational pattern and statistical association
Signature 25	T>A mutations with strand bias	Hodgkin lymphomas	Unknown etiology	N/A
Signature 26	T>C mutations	Breast, cervical, stomach and uterine cancer	Defective DNA mismatch repair	Statistical association
Signature 27	T>A mutations with strand bias	Kidney cancer	Unknown etiology	N/A
Signature 28	T>G mutations	Stomach cancer	Unknown etiology	N/A
Signature 29	C>A mutations with strand bias	Gingivo-buccal oral squamous cell carcinoma	Tobacco chewing	Statistical association
Signature 30	C>T mutations	Breast cancers	Unknown etiology	N/A

mutational processes, most of which were previously unknown.

These 30 mutational signatures are briefly summarized in the table.

In summary, my Ph.D. thesis provided a basis for deciphering mutational signatures from cancer genomics data and developed the first comprehensive census of mutational signatures in human cancer. The results reveal the diversity of mutational processes underlying the development of cancer and have far-reaching implications for understanding cancer etiology, as well as for developing cancer prevention strategies and novel targeted cancer therapies. ■

REFERENCES AND NOTES

1. M. R. Stratton, *Science* **331**, 1553 (2011).
2. L. B. Alexandrov, M. R. Stratton, *Curr. Opin. Genet. Dev.* **24**, 52 (2014).
3. M. Hollstein *et al.*, *Mutat. Res.* **431**, 199 (1999).
4. B. Vogelstein, K. W. Kinzler, *Nature* **355**, 209 (1992).
5. T. J. Hudson *et al.*; International Cancer Genome Consortium, *Nature* **464**, 993 (2010).
6. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, M. R. Stratton, *Cell Reports* **3**, 246 (2013).
7. L. B. Alexandrov *et al.*; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, *Nature* **500**, 415 (2013).
8. D. D. Lee, H. S. Seung, *Nature* **401**, 788 (1999).
9. S. Nik-Zainal *et al.*; Breast Cancer Working Group of the International Cancer Genome Consortium, *Cell* **149**, 979 (2012).
10. S. Nik-Zainal *et al.*; Breast Cancer Working Group of the International Cancer Genome Consortium, *Cell* **149**, 994 (2012).
11. S. Nik-Zainal *et al.*, *Nat. Genet.* **46**, 487 (2014).
12. L. R. Yates *et al.*, *Nat. Med.* **21**, 751 (2015).
13. G. Gundem *et al.*; ICGC Prostate UK Group, *Nature* **520**, 353 (2015).
14. M. K. Hong *et al.*, *Nat. Commun.* **6**, 6605 (2015).
15. C. S. Cooper *et al.*; ICGC Prostate Group, *Nat. Genet.* **47**, 367 (2015).
16. K. Schulze *et al.*, *Nat. Genet.* **47**, 505 (2015).
17. N. Kanu *et al.*, *Oncogene* **34**, 5699 (2015).
18. R. Wagener *et al.*, *Leukemia* **29**, 1612 (2015).
19. A. Shlien *et al.*; Biallelic Mismatch Repair Deficiency Consortium, *Nat. Genet.* **47**, 257 (2015).
20. N. Bolli *et al.*, *Nat. Commun.* **5**, 2997 (2014).
21. E. Papaemmanuil *et al.*, *Nat. Genet.* **46**, 116 (2014).
22. Y. S. Ju *et al.*, *eLife* **3**, (2014).
23. I. Martincorena *et al.*, *Science* **348**, 880 (2015).
24. S. Behjati *et al.*, *Nature* **513**, 422 (2014).
25. Mutational Signatures, The Cancer Genome Project; http://cancer.sanger.ac.uk/cosmic/signatures.

10.1126/science.aad7363

Understanding the origins of human cancer

Ludmil B. Alexandrov

Science **350** (6265), 1175-1177.
DOI: 10.1126/science.aad7363