YALE JOURNAL OF BIOLOGY AND MEDICINE 79 (2006), pp.179-185.

ESSAY

Thirty Years into the Genomics Era: Tumor Viruses Led the Way

Daniel DiMaio[a]* and George Miller[b]

[a]Departments of Genetics and Therapeutic Radiology, Yale University School of Medicine, New Haven, Connecticut; and [b]Departments of Pediatrics, Epidemiology and Public Health, and Molecular Biophysics and Biochemistry, Yale University School of Medicine, New Haven, Connecticut

This year marks the 30th anniversary of an important milestone in the history of biology: the dawn of the genomics era with the first report of the complete DNA sequence of a viral genome in 1977.

Viruses are submicroscopic micro-organisms that are dwarfed in size and complexity by all free-living forms of life. Thus, when rudimentary methods for determining short stretches of DNA sequence were developed in the 1970s, a few visionary laboratories realized that viral genomes should be the target of the first large-scale sequencing efforts. The results of this endeavor were eagerly awaited. An entire page of the 1970 edition of Albert Lehninger's *Biochemistry*, then the standard text in the field, was devoted to the *imaginary* DNA sequence of a bacterial virus, the bacteriophage ϕX174.

At the time, the prospect of sequencing the vastly larger human genome seemed impossible. Nevertheless, the sequencing of viral genomes provided seminal insights not only into viral growth, but also into animal and human carcinogenesis. These studies resulted in the discovery of cellular oncogenes and led directly to the identification of viruses as an important cause of human cancer and the idea that non-viral cancer is, in large part, a genetic disease. In addition, the sequence analysis of viral genomes laid the intellectual and technical framework that enabled the eventual sequencing and analysis of cellular genomes, an effort that in turn is unlocking many of the secrets of cancer and providing new opportunities for more effective cancer treatments.

SETTING THE STAGE

The breakthrough that allowed the sequencing of viral genomes (and later, larger cloned genomes) was the use of restriction endonucleases to manipulate and analyze DNA. Bacterial cells synthesize restriction enzymes to defend themselves against invading foreign DNA, and these enzymes were first identified and characterized on the basis of their ability to cleave bacteriophage DNA.

The value of restriction enzymes was established by a series of pioneering studies by Daniel Nathans at Johns Hopkins and Paul Berg at Stanford on the genome of SV40, a

*To whom all correspondence should be addressed: Daniel DiMaio, MD, PhD, Department of Genetics, Yale University School of Medicine, Sterling Hall of Medicine, I-141, 333 Cedar Street, New Haven, CT 06510; Tel: 203-785-2684; Fax: 203-785-6765; E-mail: daniel.dimaio@yale.edu.

small virus that normally infects monkeys. SV40 was discovered in tissue cultures of monkey kidney cells in which poliovirus vaccine was produced, and many vaccinated individuals were inadvertently inoculated with SV40 during early poliovirus vaccination programs.

Enhanced interest in SV40 lay in the fact that it and the closely related mouse polyomavirus caused tumors in rodents. It was hoped that detailed analysis of tumor viruses would shed light on the genesis of common human cancers and suggest new approaches to prevent and treat cancer.

This hope has been realized many times over. SV40 has been implicated in some human cancers, including childhood brain tumors, mesotheliomas, and non-Hodgkins lymphoma, but an etiologic role of SV40 in these tumors has not been established unequivocally [1]. Nevertheless, the early development of molecular biology techniques to analyze SV40 DNA led to the first restriction map and the first localization of a genetic signal, the viral origin of DNA replication, in eukaryotic cells [2-4].

The first recombinant DNA molecule constructed *in vitro* consisted of SV40 linked to bacteriophage lambda DNA [5].

This work set the stage for much of modern genetics and molecular biology and won Nobel prizes for Nathans and Berg. Studies on SV40 also played a major role in the genetic engineering revolution and the birth of the biotechnology industry.

THE FIRST SEQUENCING SUCCESSES

Prior to the use of restriction endonucleases, sequencing studies were largely limited to the handful of bases at the very ends of viral genomes (a striking limitation in the case of circular genomes). Now the entire viral genome was accessible.

The initial large-scale sequencing efforts bore fruit in 1977 with the publication of the complete DNA sequence of ϕX174 by a consortium headed by Frederick Sanger (who already had won a Nobel Prize for developing methods to sequence proteins and later won a second Nobel for DNA sequencing) [6].

The next year, the DNA sequence of the tumor virus SV40 was solved by groups headed by Sherman Weissman at Yale University and Walter Fiers in Belgium [7-8].

ϕX174 and SV40 are relatively simple viruses, each with only about 5,000 DNA base pairs (compared to the 6 billion base pairs in each of our cells), but with the primitive methods available at that time, this sequence analysis required many years and prodigious effort. With today's technology, the genomes of these small viruses can be knocked off in a busy afternoon.

VIRUSES TAKE CENTER STAGE

With the support of the "War on Cancer" declared by the National Cancer Institute in the 1970s, tumor virologists rapidly adopted sequencing as a primary method to analyze viral genomes, and larger DNA tumor viruses soon fell to sequencing efforts: papillomaviruses, the cause of warts and cervical and some other human cancer, with 8,000 base pairs, in 1982; and adenoviruses, which cause respiratory tract disease in humans and tumors in experimental animals, with 36,000 base pairs, in 1984. The following year, 1985, saw the completion of the 172,000 base pair sequence of the DNA of Epstein-Barr virus, a herpesvirus that infects virtually all humans and causes certain infected individuals to develop Hodgkin's disease, Burkitt's lymphoma, or nasopharyngeal carcinoma. Since the sequencing was done manually, this heroic feat required eight postdoctoral scientists working full time for several years. The poxviruses, a virus group that includes the dreaded smallpox virus and related viruses that can cause tumors in animals, weighed in at 191,636 base pairs in 1990.

The necessity of sequencing and analyzing ever-larger viral genomes spurred improvements in technology. The first sequences were derived from RNA copied from the viral DNA, but researchers soon moved to sequencing the DNA itself. A key breakthrough was the development of rapid enzymatic DNA sequencing methods by Sanger and colleagues at Cambridge University [9], replacing more cumbersome chemical methods invented by Maxam and Gilbert at Harvard University [10]. The Cambridge group also developed “shot-gun” methods for sequencing the genome of Epstein-Barr virus, in which the sequence of random fragments of viral DNA was determined, and the sequences of overlapping DNA segments were then aligned by computer. This method turned out to be essential for solving the sequences of the much larger bacterial and cellular genomes, and eventually the postdoctoral fellows were replaced by robots for the cloning operations and by automated DNA sequencing machines. Our ability to record and analyze long DNA sequences was also sorely tested, a predicament that helped lead to the development of methods of computational analysis required for the study of bacterial and cellular genomes.

VIRAL GENOMES AS TARGETS AND TEACHERS ABOUT CANCER AND OTHER DISEASES

The sequences of viral genomes revealed the great promise and limitations of large-scale sequence analysis. For the first time, the entire genetic legacy of an organism was manifest. Genes and regulatory signals were identified, protein structure and function was deduced, and overall genetic organization was laid bare. It seemed possible that a life form could be reduced to a set of simple instructions, thus revealing its most intimate secrets.

Novel human viruses were discovered based on DNA sequencing; genes that viruses captured from cellular DNA were identified; and pathogenic mechanisms and the genetic and evolutionary relationships between different viruses were elucidated. Again, studies of tumor viruses led the way.

Kaposi Sarcoma herpesvirus, which commonly causes cancer in AIDS victims, was discovered solely on the presence of novel viral DNA sequences, related to Epstein-Barr virus in a tumor [11], and hepatitis C virus was defined based on sequence information [12,13]. Hepatitis C virus and its compatriot hepatitis B virus are responsible for the development of hepatocellular carcinoma, the third most common cancer in the world [14,15]. Analysis of oncogenic retroviruses isolated from animal tumors revealed that these viruses carried genes responsible for cancer and that these genes were actually altered versions of cellular oncogenes [16], a discovery that won the Nobel Prize for Michael Bishop and Harold Varmus at the University of California at San Franscisco. Extensive sequencing also resulted in the identification of multiple different types of human papillomaviruses and the discovery that certain high-risk HPV types cause important human cancers [17,18].

These studies ultimately led to development of prophylactic vaccines designed to prevent virus infection and the resulting human cancers. Vaccines based on the major capsid protein of the high-risk human papillomaviruses will prevent much cervical cancer, and the hepatitis B virus surface antigen vaccine already is causing a reduction in the incidence of hepatocellular cancer [14,19]. These cancer vaccines are produced by using recombinant DNA technology with defined viral DNA segments because of the difficulty of propagating these human tumor viruses in the laboratory. These two vaccines alone have the potential to eliminate greater than 10 percent of the world-wide cancer burden. The prevention of cancers by vaccination is the culmination of a century-long effort to tie infectious agents to cancer and validates this entire enterprise [20]. In addition, ongoing work at Yale and elsewhere demonstrates that viral gene expression is required to maintain the survival and growth of certain cancers, suggesting that anti-viral drugs may be useful in treating these cancers once they develop [21,22,23]. Vaccination against viral non-structural antigens expressed by cancer cells also has emerged as a potential therapeutic approach.

Viruses responsible for acute infections have been subjected to intense sequencing efforts. Complete sequence information enabled the laboratory synthesis of infectious poliovirus, a small virus that has so far eluded sustained efforts for eradication [24]. In a remarkable example of molecular archaeology, sequencing of archived samples of the 1918 pandemic influenza virus, which caused the most deadly acute epidemic in history, allowed the resurrection of this virus [25]. Studies based on this work will have a profound effect on our understanding of these important human pathogens and the genetic basis for future influenza pandemics, but this work also inspired much discussion about the potential misuse of sequence information and scientific knowledge. The very rapid identification of a novel coronavirus as the causative agent of SARS was a stunning achievement based entirely on sequencing [26,27]. Recent sequencing studies also revealed that the environment harbors a vast number of uncharacterized bacteriophage and other viruses [28,29]. Remarkably, the genes contained in these viruses often do not resemble previously known genes, hinting at immense genetic diversity of viruses.

SURPRISES AND COMPLICATIONS

It was anticipated that genomic sequence information would lead to the rapid elucidation of all the details of virus reproduction and myriad interactions between viruses and cells and suggest new rational approaches to combat virus infections and virally induced tumors. But there also were complications. Viruses were more clever, versatile, and enigmatic than we appreciated, and not all their secrets were revealed by simple sequence analysis alone. RNA splicing, first discovered in oncogenic adenoviruses and widespread in cellular genes as well [30,31], results in the synthesis of proteins assembled from segments encoded by DNA scattered about the genome, complicating efforts to deduce protein sequence from DNA sequence.

In other cases, viruses selectively ignored signals that should tell the cellular machinery where genes started and stopped, and RNA editing or ribosomal frameshifting confounded the definition of a gene (e.g., [32,33]). One of the first viral sequences solved revealed the startling situation that the same piece of DNA could encode two proteins with entirely different amino acid sequences [34]. Thus, it became clear that complete understanding of an organism, even one as simple as a virus, cannot be deduced from sequence information alone, but the services of experimental biologists were still needed.

MOVING BEYOND VIRUSES

The great value of sequencing was immediately clear to the scientific community and inspired successful attempts to decipher larger cellular genomes.

The first complete DNA sequence of a bacterial genome, that of *Mycoplasma genitalium*, was reported in 1995 [35], and now hundreds of bacterial genomes have been decoded, providing valuable information about the lifestyle, evolution, and pathogenesis of these organisms. And in 2001, the entire human genome sequence was completed.

In parallel, the genomes of several other pathogens and organisms useful as experimental models, including the malaria parasite, insects, worms, plants, and other mammals, have been sequenced. These sequences uncovered a treasure trove of information that is revolutionizing our understanding of evolution, gene organization and function, cellular function, development, and disease. These studies also revealed the disquieting fact that a substantial fraction of our own genomes is derived from remnants of viruses related to oncogenic retroviruses. It has even been postulated that the transfer of viral genes into primordial RNA-based cells gave rise to the three branches of cellular life on earth: bacteria, archea, and eukaryotes [36].

MORE INSIGHTS INTO CANCER

DNA sequencing has had a particularly profound effect on our understanding of carcinogenesis and on efforts to control this disease. Sequencing of the cellular versions of the oncogenes first identified in tumor viruses revealed newly arising mutations in many sporadically occurring human cancers [37], helping to establish the notion that much of cancer has a genetic basis.

Many of these oncogenes encode growth factors, growth factor receptors, signal transduction proteins, and transcription factors, which are crucial components of signaling pathways that determine how cells respond to their environment, grow, and die. Based on these and related studies, it is now possible in some cases to predict a patient's prognosis or response to chemotherapy based on the sequence of cellular proteins involved in cell signaling and growth [38].

The elucidation of the human genome sequence also allowed the construction of microarrays that have been used to assess the transcriptional profile of normal and cancer cells. These studies revealed the existence of subsets of various tumors with markedly different prognosis, suggesting underlying biological differences that can be exploited in developing improved approaches for managing these diseases [39].

In addition, advances in rapid sequencing and computational analysis allowed comprehensive comparison of the sequences of normal and cancer cell genomes and the identification of crucial shared mutations in independently arising cancers [40]. The genes identified in this manner are likely to play important roles in carcinogenesis and provide new targets for therapy and diagnosis.

Sequencing of DNA isolated from cancers also revealed the frequent occurrence of mutations in genes encoding tumor suppressor proteins, particularly p53, which was first identified as a binding partner of SV40 large T antigen, the viral DNA replication protein [41,42]. Large T antigen also binds the retinoblastoma tumor suppressor protein, a central component of another tumor suppressor pathway often inactivated by mutation in human cancers [43]. Strikingly, three unrelated groups of DNA tumor viruses, SV40, human papillomaviruses, and adenoviruses, all inactivate these same tumor suppressor proteins, implying the existence of a limited number of regulatory nodes essential for viral replication and cellular carcinogenesis.

These important cellular genomics efforts clearly had their origins in studies of viral genomes. Francis Collins, who heads the National Human Genome Research Institute and directed the United States' effort to sequence the human genome, was trained in the Weissman laboratory at Yale that sequenced SV40. Two of the prime movers of the landmark 1995 *M. genitalium* sequence were Hamilton Smith and Craig Venter. Smith and his Johns Hopkins' colleague Nathans supplied essential information and materials for the SV40 sequencing project, and a year before the appearance of the first complete bacterial sequence, Venter published the sequence of smallpox virus DNA.

The ability to decode bacterial and cellular genomes is a watershed achievement in biology, one that is transforming our concepts of microbial pathogenesis, cellular function, evolution, and human disease, particularly cancer. This remarkable accomplishment is the logical extension of pioneering studies begun a quarter of a century earlier on viruses, the simplest organisms at the border between chemistry and life.

REFERENCES

1. Shah KV. Simian Virus 40 and human disease. J Infect Dis. 2004;190:2061-64.
2. Danna KJ, Nathans D. Specific cleavage of simian virus 40 DNA by restriction endonuclease of *Hemophilus influenzae*. Proc Natl Acad Sci USA. 1971;68:2913-17.
3. Danna KJ, Nathans D. Bidirectional replication of simian virus-40 DNA.Proc Natl Acad Sci USA. 1972;69:3097-3102.
4. Danna KJ, Sack GH, Nathans D. Studies of simian virus-40 DNA. Cleavage map of SV40 genome. J Mol Biol. 1973;78:363-76.
5. Jackson DA, Symons RH, Berg P. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: Circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. Proc Natl Acad Sci USA. 1972;69:2904-09.
6. Sanger F., Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977;265:687-95.
7. Fiers W, Contreras R, Haegemann G, Rogiers R, Van de Voorde A, Van Heuverswyn H, et al. Complete nucleotide sequence of SV40 DNA. Nature. 1978;273:113-120.

8. Reddy VB, Thimmappaya B, Dhar R, Subramanian N, Zain BS, Pan J, et al.The genome of simian virus 40. Science. 1978;200:494-502.
9. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977;74:5463-67.
10. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA. 1977;74:560-4.
11. Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, Moore PS. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. Science. 1994;266:1865-69.
12. Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science. 1989;244:359-62.
13. Miller RH, Purcell RH. Hepatitis C virus shares amino acid sequence similarity with pestiviruses and flaviviruses as well as members of two plant virus supergroups. Proc Natl Acad Sci USA. 1990;87:2057-61.
14. Beasley RP. Hepatitis B virus. The major etiology of hepatocellular carcinoma. Cancer. 1988;61:1942-56.
15. DiBisceglie AM. Hepatitis C and hepatocellular carcinoma. Hepatology. 1997;26:34S-38S.
16. Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature. 1976;260:170-3.
17. Bosch FX, Lorincz A, Munoz N, Meijer CJ, Shah KV. The causal relation between human papillomavirus and cervical cancer. J Clin Pathol. 2002;55:244-65.
18. Durst M, Gissmann L, Ikenberg H, zur Hausen H. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. Proc Natl Acad Sci USA. 1983;80:3812-15.
19. Koutsky LA, Ault KA, Wheeler CM, Brown DR, Barr E, Alvarez B, et al. A controlled trial of a human papillomavirus type 16 vaccine. N Engl J Med. 2002;347:1645-51.
20. Klein G, DiMaio D. Principles of Human Tumor Virology. In: Garcea R, DiMaio D, editors. *The Papillomaviruses*. New York: Springer; 2007. p. 19-29.
21. Hwang ES, Riese DJ, Settleman J, Nilson LA, Honig J, Flynn S, DiMaio D. Inhibition of cervical carcinoma cell line proliferation by the introduction of a bovine papillomavirus regulatory gene. J Virol. 1993;67:3720-29.
22. Kennedy G, Komano J, Sugden B. Epstein-Barr virus provides a survival factor to Burkitt's lymphomas. Proc Natl Acad Sci USA. 2003;100:14269-74.
23. von Knebel Doeberitz M, Oltersdorf T, Schwarz E, Gissmann L. Correlation of modified human papilloma virus early gene expression with altered growth properties in C4-1 cervical carcinoma cells. Cancer Res. 1988;48:3780-86.
24. Cello J, Paul AV, Wimmer E. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. Science. 2002;297:1016-18.
25. Tumpey TM, Basler CF, Aguilar V, Zeng H, Solorzano A, Swayne DE, et al. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. Science. 2005;310:77-80.
26. Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med. 2003;348:1967-76.
27. Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emergy S, et al. A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med. 2003;348:1953-66.
28. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, et al. The marine viromes of four oceanic regions. PLoS Biology. 2006;4:2121-31.
29. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304:66-74.
30. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA. 1977;74:3171-3175.
31. Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell. 1977;12:1-8.
32. Jacks T, Varmus HE. Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. Science. 1985;230:1237-42.
33. Sanchez A, Trappier SG, Mahy BW, Peters CJ, Nichol ST. The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. Proc Natl Acad Sci USA. 1996;93:3602-07.
34. Barrell BG, Air GM, Hutchison CA. Overlapping genes in bacteriophage phiX174. Nature. 1976;264:34-41.
35. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of Mycoplasma genitalium. Science. 1995;270:397-403.
36. Koonin EV, Senkevich TG, Dolja VV. The ancient Virus World and evolution of cells. Biol Direct. 2006;1:29.
37. Tabin CJ, Bradley SM, Bargmann CI, Weinberg RA, Papageorge AG, Scolnick EM, et al. Mechanism of activation of a human oncogene. Nature. 1982;300:143-9.
38. Lynch, TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al.

Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 204;350:2129-39.

39. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286:531-7.
40. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006;314:268-74.
41. Lane DP, Crawford LV. T antigen is bound to a host protein in SV40-transformed cells. Nature. 1979;278:261-3.
42. Linzer DI, Levine AJ. Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. Cell. 1979;17:43-52.
43. DeCaprio JA, Ludlow JW, Figge J, Shew JY, Huang CM, Lee WH, et al. Large tumor antigen forms a specific complex with the product of the retinoblastoma susceptibility gene. Cell. 1988;54:275-83.