Celebrating 50 years of virus genomics by revisiting the first decade (1976-1985)

Siobain Duffy and J. Steen Hoyer

Department of Ecology, Evolution, and Natural Resources, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, New Jersey, USA

Correspondence to SD: duffy@sebs.rutgers.edu

Abstract

Whole genome sequencing was pioneered in viruses, beginning with the publication of the third and final portion of the levivirus MS2 genome on April 8th, 1976. Fifty years and incalculable viral genomes later, we chose to examine the first decade of viral genomics for posterity. We found that 97 whole genome sequences were published between 1976 and 1985, with some species having as many as seven individual genomes separately sequenced. Early efforts included most genomic architectures of viruses, with roughly equal proportions of RNA (26%), DNA (38%) and retrotranscribing viruses (36%) fully sequenced. However, influenza A was the sole completed -ssRNA viral genome in this decade and no dsRNA viruses were fully sequenced. Nearly two-thirds of sequences (64%) were obtained at least partially using Maxam-Gilbert sequencing, demonstrating that Sanger dideoxy chain termination sequencing did not immediately displace this earlier chemical cleavage technique. As some of these genomes predated the existence of GenBank, many of these genomes were transcribed into GenBank years after publication, and two are still not available in GenBank. This documentation of the initial ten years of what is now a routine part of viral evolution studies should help us celebrate our roots, allow us to order the published genomes as our colleagues working with archaea, bacteria and eukaryotes do, and help correct the misconception that phiX174 was the first genome (or DNA genome) sequenced.

Introduction

From the perspective of 2026, it is nearly impossible to imagine studying viral ecology and evolution without genomic approaches (1). Prior to the advent of reliable and processive DNA sequencing, comparisons between viruses could be based on amino acid content (2) or genome hybridization (3, 4). While hybridization produced qualitatively accurate measures of nucleotide sequence identity, these studies were of limited reuse – if a new strain of a virus was found, one would have to re-do all the hybridization assays with the new strain against all the previously characterized and published relatives (5). The desire to sequence genomes rather than capturing sequence identity by hybridization was strong, but it was a long and thorny path to developing a method of DNA sequencing. RNA sequencing came first, and the most widely used method, developed by Frederick Sanger, required meticulous and tedious 2D gel radiographs, producing short tracts that were better suited for partial gene sequences (6). While many researchers sequenced tRNAs and mRNA to get partial sequences of cellular genes of interest, the very first gene completely sequenced, from mRNA, was from a virus: the +ssRNA levivirus MS2 (7).

The formally published history of genomics began on April 8th 1976, with the complete nucleotide sequence of RNA phage MS2 (8). Related painstaking work by the Fiers lab decoded the rest of the 3,569 base +ssRNA genome in two earlier papers (7, 9). This feat is even more remarkable because more processive methods of sequencing – Maxam-Gilbert (10) and Sanger dideoxy chain termination (11) – had not yet been published; this first genome was decoded on two-dimensional gels of labeled RNA transcripts that had been digested by a nuclease. Only two further genomes were sequenced, at least in part, this way before the modern 1977 methods took over the field: dsDNA polyomavirus SV40, also by the Fiers lab (12) and an insect virus (black beetle virus, (13, 14)). This early history of genomic sequencing has been somewhat undercited and oversimplified in the scientific literature, despite being in the lifetimes of many practicing virologists.

One reason for the murky history of viral genomics is that it was not well curated as the genomic revolution was happening. While colleagues in other areas of Biology have preserved the history of the first 20, 50 or 100 different genomes in their domain were published (e.g., (15)), no similar resource exists for viruses. Lists of fully sequenced bacterial genomes, archaeal genomes and eukaryotic genomes are even available on Wikipedia. We undertook this survey of the literature from 1976-1985 to reconstruct the early history of viral genomics and produce a similar list for viruses. While some of the discoveries in this historical list will confirm the biases of the wider virology community – those working on dsRNA viruses will not be surprised none of their viruses were fully sequenced in this first decade – we imagine that all virologists will be surprised by some aspects of the early history of viral genomics. Our most surprising conclusion is that SV40, not phiX174, was the first complete DNA virus to be completely sequenced – consistent with the writings of Frederick Sanger himself (16).

Results and Discussion

We had a multipronged approach to finding viral genomes published between 1976 and 1985. We used internet searching, going through the tables of contents of relevant journals, reading paper copies of sequences added to GenBank in the mid-1980s and looking within GenBank itself (see Methods, below).

The DNA genome of simian virus 40 (SV40) was completed before that of ssDNA phage PhiX174 (1978)

Despite misconceptions (Figure 1), including in the literature (e.g., (17–19)), phiX174 was not the first viral genome to be fully sequenced (Table 1). The RNA coliphage MS2, using RNA sequencing methods (6), was completed and published in 1976, somewhere between 10 months to two and a half years before that of the ssDNA virus phiX174. The ambiguity on the timing of the complete phiX174 genome's publication is due to the genome appearing in a nearly complete form in February 1977 (20), with the complete, corrected phiX174 genome published in 1978 (21). Many cite the February 1977 Nature paper that had a very good draft sequence of phiX174 as the first DNA genome, but by the authors' own admission that the nearly complete sequence was not perfect (for instance, it lacked an exact nucleotide count, it was just "about 5375 bases" (20). Those who cite that paper for the first genome produced by Sanger sequencing – dideoxy chain termination sequencing – are not correct, as this nearly complete draft sequence was published using a sequencing method the Sanger lab developed and then discarded ("plus and minus" sequencing (22), we recommend this Scientific American article if one wishes to understand this method (23)). A later 1977 paper in PNAS (11) is the correct citation for the well-known Sanger dideoxy chain termination sequencing method [hereafter "Sanger sequencing"], but while the new technique was validated on restriction fragments of the phiX174 genome, that paper also does not contain a full genomic sequence (11). Instead, the complete sequence of the 5386 bases of phiX174's genome was published on October 25, 1978 in the Journal of Molecular Biology (21) and that is the date used in our analyses (Sanger concurred with the 1978 date, (16)). We admit to having made the error of assigning the phiX174 genome to 1977 in the past (24), but if the criterion is a complete genome sequence, then 1978 is the accurate date for phiX174. This makes phiX174 the first complete ssDNA viral genome produced, but the gap in time between the very good draft and the complete sequence allows another DNA virus genome to sneak in – polyomavirus SV40's genome was published by two groups in May 1978 (12, 25). While Sanger sequencing was one of three sequencing methods that contributed to phiX174's complete genome, the first full genome to be sequenced by Sanger methods alone was poliovirus, in 1981 (Table 1). Being the third virus completely sequenced should not diminish phiX174's role as a pioneering virus in the history of biology, it is still the system in which overlapping reading frames were discovered (26), the first template for *in vitro* DNA synthesis (27), the first synthetic genome produced (28) and one of the most frequently sequenced viruses on the planet due to its frequent use as a control in Illumina sequencing (29).

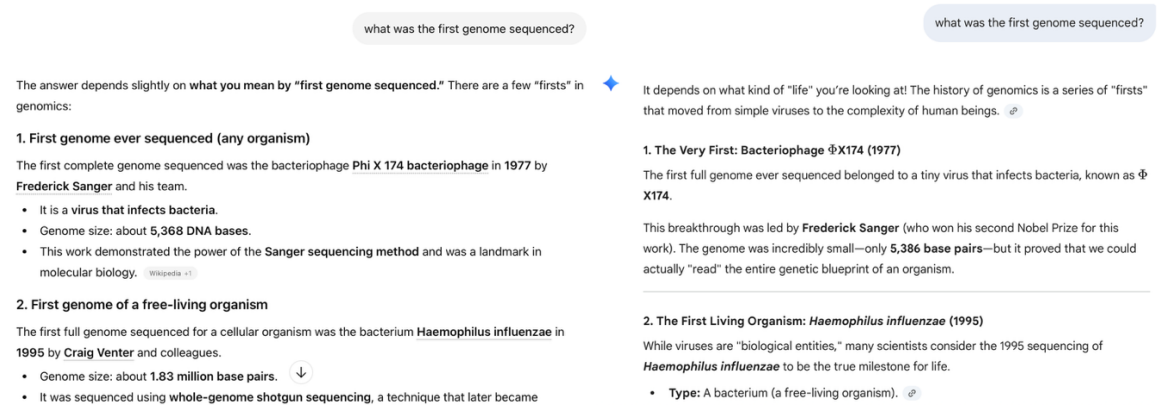

Figure 1: Large language model answers to the prompt "what was the first genome sequenced?" from March 10, 2026. On left, ChatGPT v5.3, on right Gemini 3. Both responses incorrectly identify phiX174 instead of MS2 as the first genome, and both reference the incomplete genome of the 1977 paper (which is currently echoed on Wikipedia). The ChatGPT v5.3 answer has additional incorrect details: the 1977 paper didn't demonstrate the power of the Sanger sequencing method (it was plus and minus) and the genome is 5,386 bases long. The Gemini answer has the incorrect unit of base pairs, as phiX174 is single-stranded.

Table 1. Complete viral genomes published in chronological order. Length is given in nucleotides, the genomic architecture is given by kind of nucleic acid, whether genome is circular (C) or linear (L) and the number of segments. The sequencing methods used are given as Maxam-Gilbert chemical cleavage (MG), Sanger dideoxy chain termination (S), RNA sequencing (RNA), Plus and Minus (±) or forward and back (FB). When a sequence was not associated with any record in GenBank, there is a dash in the "Accession" column.

Year	Month	Name	Length	Genome Type	C/L	Seg #	Sequencing method	Ref	Accession
1976	Apr	MS2	3569	+ssRNA	L	1	RNA	(8)	V00642
1978	May	SV40	5375	dsDNA	C	1	MG	(25)	J02400
1978	May	SV40	5226	dsDNA	C	1	RNA; MG	(12)	J02400
1978	Oct	phiX174	5386	ssDNA	C	1	±; S; MG	(21)	J02482
1978	Nov	G4	5577	ssDNA	C	1	±; S	(30)	V00657
1978	Dec	fd	6408	ssDNA	C	1	MG	(31)	J02451
1979	Oct	hepatitis B virus (ayw)	3182	RT-dsDNA	C	1	MG;S	(32)	V01460
1979	Oct	BK virus(MM strain)	4963	dsDNA	C	1	MG	(33)	V01109
1979	Dec	BK virus (Dunlop strain)	5153	dsDNA	C	1	MG	(34)	V01108
1980	Jan	hepatitis B (adw2)	3221	RT-dsDNA	C	1	MG	(35)	X02763
1980	Jan	polyomavirus A2	5292	dsDNA	C	1	MG; S	(36)	J02288
1980	Feb	polyomavirus strain 3	5295	dsDNA	C	1	MG; S	(37)	J02289
1980	Aug	cauliflower mosaic virus Strasbourg	8024	RT-dsDNA	C	1	MG	(38)	V00141
1980	Oct	M13	6407	ssDNA	C	1	MG	(39)	V00604
1981		Qβ	4220	+ssRNA	L	1	MG	(40)	-
1981	Jun	poliovirus	7433	+ssRNA	L	1	S	(41)	V01148
1981	Jun	cauliflower mosaic virus CM1814	8031	RT-dsDNA	C	1	S	(42)	V00140
1981	Aug	poliovirus	7440	+ssRNA	L	1	MG	(43)	V01149
1981	Oct	Moloney murine leukaemia provirus	8332	RT-ssRNA	L	1	MG	(44)	J02255
1981	Oct	Moloney murine sarcoma virus (replication defective)	5828	RT-ssRNA	L	1	MG	(45)	J02266
1981	Nov	murine sarcoma virus	5833	RT-ssRNA	L	1	MG	(46)	V01185
1981	Dec	f1	6407	ssDNA	C	1	MG	(47)	V00606
1982	Jan	woodchuck hepatitis virus	3308	RT-dsDNA	C	1	MG	(48)	J02442
1982	Feb	human papillomavirus type 1a	7811	dsDNA	C	1	S	(49)	V01116
1982	Mar	influenza virus (A/PR/8/34)	13588	-ssRNA	L	8	FB; MG; S	(50)	V00603, J02151, J02139, V01088, V01084, J02146, V01099, J02150
1982	May	avian sarcoma virus Y73	3718	RT-ssRNA	L	1	MG	(51)	V01170
1982	Oct	tobacco mosaic virus	6395	+ssRNA	L	1	S	(52)	V01408
1982	Oct	f1	6407	ssDNA	C	1	MG	(53)	J02448
1982	Oct	poliovirus (Sabin 1)	7441	+ssRNA	L	1	MG	(54)	V01150
1982	Oct	Fujinami sarcoma virus (defective)	4462	RT-ssRNA	L	1	MG	(55)	J02194
1982	Oct	cauliflower mosaic virus D/H	8016	RT-dsDNA	C	1	MG	(56)	V00139
1982	Oct	bovine papillomavirus type 1	7945	dsDNA	C	1	S	(57)	X02346
1982	Dec	Lambda	48502	dsDNA	C	1	S	(58)	J02459
1983	Jan	parvovirus H-1	5176	ssDNA	L	1	S	(59)	X01457
1983	Jan	cassava latent virus	5503	ssDNA	C	2	MG	(60)	J02057, J02058
1983	Feb	adeno-associated virus 2	4675	ssDNA	L	1	MG	(61)	J01901
1983	Feb	simian sarcoma virus	5779	RT-ssRNA	L	1	MG	(62)	V01201
1983	Mar	FBJ murine osteosarcoma virus	9312	RT-ssRNA	L	1	MG	(63)	V01184
1983	Mar	Rous sarcoma virus	9312	RT-ssRNA	L	1	MG	(64)	J02342
1983	Mar	hepatitis B virus (adr)	3188	RT-dsDNA	C	1	S	(65)	V00867
1983	Mar	hepatitis B virus (adw)	3200	RT-dsDNA	C	1	S	(65)	V00866
1983	Apr	mouse minute virus	5081	ssDNA	L	1	MG; S	(66)	J02275
1983	May	avian myelocytomatosis MC29	3779	RT-ssRNA	L	1	MG	(67)	V01174
1983	Jun	T7	39936	dsDNA	L	1	MG	(68)	V01146
1983	Jun	human adult T-cell leukemia virus	9032	RT-ssRNA	L	1	MG	(69)	J02029
1983	Jun	Abelson murine leukemia virus	5658	RT-ssRNA	L	1	MG	(70)	V01541
1983	Jul	hepatitis B virus (adr)	3214	RT-dsDNA	C	1	MG	(71)	X01587
1983	Aug	poliovirus (3 Leon 12a1b)	7432	+ssRNA	L	1	S	(72)	X00925
1983	Aug	Friend spleen focus-forming provirus	6296	RT-ssRNA	L	1	S	(73)	K00021

1983	Oct	alfalfa mosaic virus	8274	+ssRNA	L	3	MG	(74)	L00163, X01572, K02703
1983	Dec	cowpea mosaic virus	9370	+ssRNA	L	2	MG; S	(75)	X00729, X00206
1983	Dec	human papillomavirus type 6b	7902	dsDNA	C	1	MG	(76)	X00203
1984	Feb	Akv murine leukemia virus	8374	RT-ssRNA	L	1	S	(77)	J01998
1984	Feb	Sindbis virus	11703	+ssRNA	L	1	MG	(78)	J02363
1984	Feb	brome mosaic virus	8200	+ssRNA	L	3	S	(79)	X02380, X01678, J02042
1984	Mar	poliovirus (3 Leon 37)	7431	+ssRNA	L	1	S	(80)	K01392
1984	Mar	duck hepatitis B virus	3021	RT-dsDNA	C	1	MG	(81)	K01834
1984	Apr	Akv murine leukemia virus	8303	RT-ssRNA	L	1	S	(82)	[J01998]
1984	Apr	poliovirus (2 P712, Ch, 2ab)	7439	+ssRNA	L	1	MG	(83)	X00595
1984	Apr	poliovirus (3 Leon 12a1b)	7434	+ssRNA	L	1	MG	(83)	X00925
1984	Jul	Epstein-Barr virus	172282	dsDNA	L	1	S	(84)	V01555
1984	Aug	ground squirrel hepatitis virus	3311	RT-dsDNA	C	1	S	(85)	K02715
1984	Aug	JC virus	5130	dsDNA	C	1	MG	(86)	J02226
1984	Sep	tomato golden mosaic virus	5096	ssDNA	C	2	S	(87)	K02029, K02030
1984	Oct	tobacco mosaic virus (tomato strain)	6384	+ssRNA	L	1	MG	(88)	X02144
1984	Oct	hepatitis B virus (adr)	3215	RT-dsDNA	C	1	MG	(89)	D00630
1984	Oct	maize streak virus	2681	ssDNA	C	1	S	(90)	X01089
1984	Oct	human rhinovirus 14	7205	+ssRNA	L	1	S	(91)	K02121
1984	Nov	adenovirus 2	35937	dsDNA	L	1	S	(92)	J01917
1984	Dec	maize streak virus	2687	ssDNA	C	1	S	(93)	X01633
1985	Jan	AIDS lymphadenopathy-associated virus	9193	RT-ssRNA	L	1	S	(94)	K02013
1985	Jan	IKe	6883	ssDNA	C	1	MG; S	(95)	X02139
1985	Jan	AIDS-associated HTLV-III	9749	RT-ssRNA	L	1	S	(96)	K03455
1985	Feb	AIDS-associated retrovirus ARV-2	9737	RT-ssRNA	L	1	S	(97)	K02007
1985	Feb	bovine leukemia virus	8714	RT-ssRNA	L	1	MG	(98)	K02120
1985	Mar	avian sarcoma virus UR2 (defective)	3165	RT-ssRNA	L	1	MG; S	(99)	M10455
1985	Mar	Shope cottontail rabbit papillomavirus	7868	dsDNA	C	1	S	(100)	K02708
1985	Mar	black beetle virus	4504	+ssRNA	L	2	RNA; MG	(13)	X02396, X00956
1985	Mar	human rhinovirus 2	7102	+ssRNA	L	1	MG; S	(101)	X02316
1985	May	hepatitis A virus	7478	+ssRNA	L	1	S	(102)	K02990
1985	May	hamster papovavirus	5366	dsDNA	C	1	S	(103)	X02449
1985	May	human T-cell leukemia virus II	8952	RT-ssRNA	L	1	MG; S	(104)	M10060
1985	May	mouse minute virus immunosuppressive variant	5087	ssDNA	L	1	S	(105)	X02481
1985	Jun	bean golden mosaic virus	5233	ssDNA	C	2	MG; S	(106)	M10070, M10080
1985	Jun	foot and mouth disease virus A12	7747	+ssRNA	L	1	MG; S	(107)	X00429
1985	Jul	cucumber mosaic virus	8623	+ssRNA	L	3	MG	(108)	X02733, X00985, J02059
1985	Aug	human papillomavirus type 16	7904	dsDNA	C	1	MG; S	(109)	K02718
1985	Aug	duck hepatitis B virus	3021	RT-dsDNA	C	1	MG	(110)	DQ195079
1985	Aug	visna lentivirus	9202	RT-ssRNA	L	1	MG	(111)	M10608
1985	Aug	tobacco mosaic virus (attenuated tomato strain L11A)	5223	+ssRNA	L	1	MG; S	(112)	-
1985	Aug	yellow fever virus	10862	+ssRNA	L	1	MG	(113)	X03700
1985	Sep	wheat dwarf virus	2749	ssDNA	C	1	S	(114)	X02869
1985	Sep	carnation mottle virus	4003	+ssRNA	L	1	MG	(115)	X02986
1985	Oct	deer papillomavirus	8374	dsDNA	C	1	S	(116)	M11910
1985	Oct	Pf3	5833	ssDNA	C	1	MG	(117)	M19377/ M11912
1985	Dec	S13	5386	ssDNA	C	1	MG; S	(118)	M14428
1985	Dec	woodchuck hepatitis virus	3320	RT-dsDNA	C	1	S	(119)	M11082

Importance of the Maxam-Gilbert sequencing method

The other DNA sequencing method published in 1977 that greatly advanced genetics and genomics was Maxam-Gilbert chemical cleavage (10). Both Maxam-Gilbert and Sanger were tremendous advances over the previous techniques that were used to sequence short stretches of RNA, or earlier attempts at DNA sequencing (16). Intriguingly, it was common for both methods to be used to sequence the viruses in Table 1 (46% were produced by Maxam and Gilbert alone, 33% by Sanger alone, 14% by a combination of Maxam-Gilbert and Sanger, with six sequences produced by other methods, often with the help of one or both of these two modern methods). In fact, the corrected, complete phiX174 sequence was produced using some Maxam-Gilbert sequencing (21).

Both modern methods used P^{32} radioactivity to image DNA on X-ray films exposed to acrylamide gels that separated out bands by size, and researchers could read off the sequence of the DNA from four lanes of a gel. The Maxam-Gilbert method involved four separate reactions that allowed determination of which fragments ended with an A, C, G or T. The optimal chemical reactions changed over time (10, 120), becoming easier to perform, but researchers would still have to compare the results of multiple lanes in order to read the sequence. For instance, one lane would show the results of a reaction that would cut after all pyrimidines, and another would have the same reaction at a high salt content, which made it nearly cytosine-specific (Figure 2). Researchers would determine which fragments ended with a C or a T based on whether a fragment that size was present in both lanes (C) or only in one (T). Maxam-Gilbert methods could be used on purified genomic material, usually subjected to restriction digestion, without the polymerization required by Sanger chain termination sequencing.

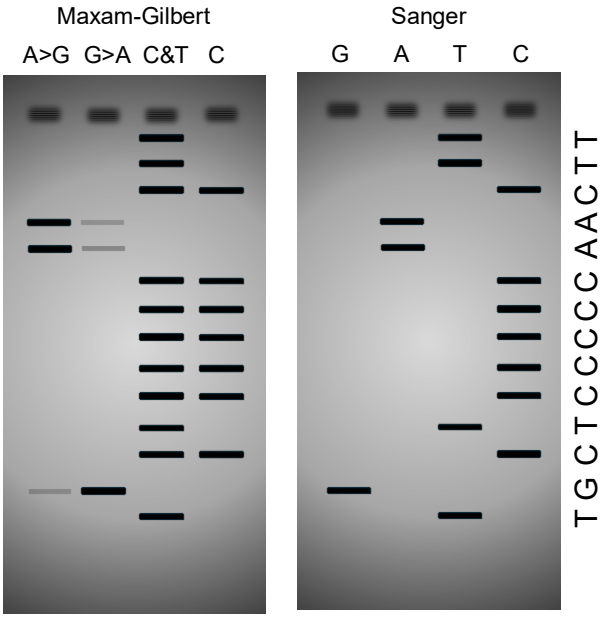

Figure 2. How the phiX174 origin of replication (5' TGCTCCCCCAACTT 3') would appear on films exposed to gels from Maxam-Gilbert (left) and Sanger (right) sequencing, circa 1977. The acrylamide gels allowed size separation, where smaller oligonucleotides run faster and farther

on the gel and larger ones move more slowly, allowing researchers to read the sequence bottom to top of the gel. The Maxam-Gilbert chemical cleavage method uses P32-labeled ssDNA regions of the same length, that are then chemically cleaved using different reactions with varying degrees of specificity for bases (10). Two reactions involving methylation with dimethyl sulfate lead to cleavage after purine bases, but with stronger effects after A (leftmost column) or G (second column). Both reactions for pyrimidines involved hydrazine, and are conducted at two salt concentrations, since higher salt produces higher specificity for cleaving after cytosines (rightmost column on the Maxam-Gilbert side). The reactions for the four columns on the Sanger side are variants of the same polymerization mixture – each column contains ample amounts of three dNTPs and the fourth base – the one that is labeled for each column – is added as a mix of dNTP and dideoxy NTP (ddNTP), the latter of which prevents further polymerization after it is incorporated. When this method was first developed it was not possible to source each dNTP and ddNTP in the same purity or concentration, so the relative amounts used for each column were not identical (11) – this method was refined over time to become more straightforward.

Sanger sequencing became the dominant technique in gene and genome sequencing through two innovations: shotgun cloning into M13 (121, 122) and the innovation of replacing radioactivity with fluorescent tags on ddNTPs and the resulting rise of higher throughput machines (123). Therefore, our boundary of the first decade of viral genomics isn't quite arbitrary – it was much more feasible to sequence viral genomes by automated Sanger sequencing shortly after 1985. However, Maxam-Gilbert methods were still intermittently used in viral sequencing. The last viral genome we could find that partially used Maxam-Gilbert sequencing was published in 2001 (124), though some geminivirus-associated minichromosomes were entirely sequenced by Maxam-Gilbert in a 2003 publication (125), and this method appears to be used for small stretches of sequencing and niche applications into the 2020s (126–128).

After both processive viral sequencing methods were widely adopted and refined, the speed with which genomes were produced increased (Figure 3). Through this decade the genomes of several mitochondria were completely sequenced (129–132), two viral satellites, and more than ten viroids (133) beginning with potato spindle tuber viroid (134). The first complete cellular genome would not be published until 1995 (135).

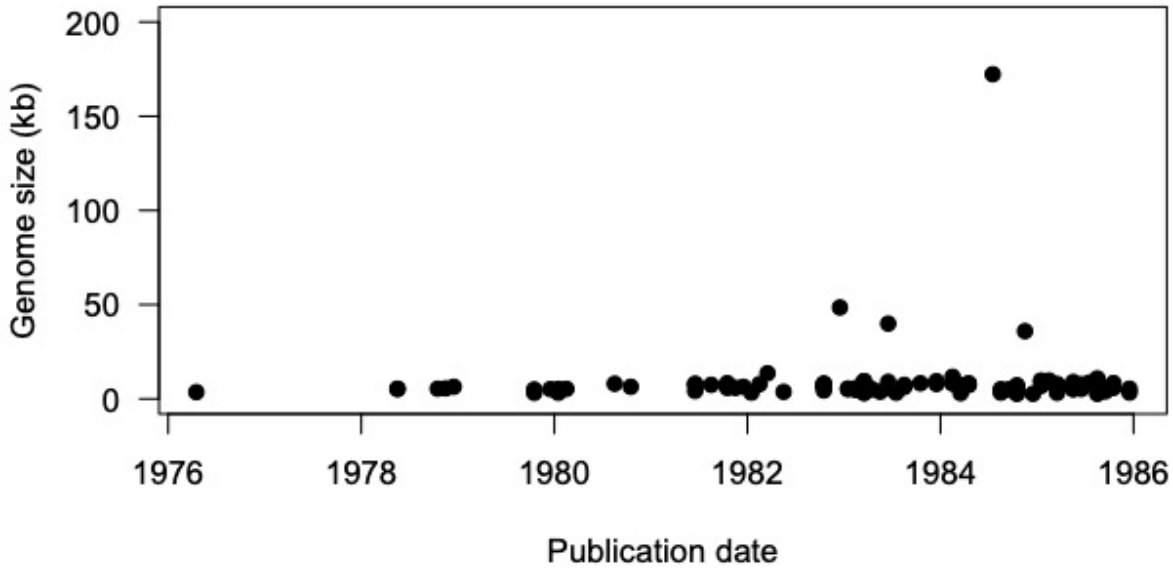

Figure 3. Genome size of sequenced virus genomes over time (by publication date).

Publication trends in this first decade of viral genomics

While many applications of new technologies are quickly shunted from the glamour, high-impact factor journals down to more specialist journals, whole viral genome sequences still held attraction for Cell, Nature and Science through 1985. The most common journals for publishing a full genome in this decade were Nucleic Acid Research (sixteen), Journal of Virology (thirteen), Proceedings of the National Academy of Sciences (thirteen) and Nature (twelve), though genomes were published in a wide range of venues. Far fewer sequencing papers were published in Journal of General Virology or Virology in this time frame.

Even in 1985 it was acceptable to publish a paper on a single gene's sequence from a virus, and there was sufficient novelty in almost any viral genome – it need not be larger or of a more complex genomic architecture. Few large DNA viruses were sequenced in the first 10 years of viral genomics for the obvious reason – it took more time and money than sequencing smaller genomes. The largest genome sequenced in this time frame was Epstein-Barr virus, the only one that was over 100,000 bases (Figure 3). This accomplishment is even more remarkable when one acknowledges that each nucleotide was sequenced 7.3x due to the nature of sequencing cloned sequence fragments (84). Smaller viruses dominated this decade (93% of viruses sequenced had genomes <10kb, 25% had genomes ≤5kb). The most frequently sequenced species was poliovirus (genome size ~7430b), which was sequenced seven times, and hepatitis B virus was sequenced 6 times (Table 1). The first time two genomes were published in a single manuscript was for two hepatitis B sequences (65).

Plant viruses: One of the motivations for this project was that we knew when the first geminivirus genome was sequenced (cassava latent virus, now known as African cassava mosaic virus, formally *Begomovirus manihotis*) but it was not clear whether it was the second, or fourth, or tenth plant virus sequenced. Reading through Table 1, the first plant virus sequenced was the pararetrovirus cauliflower mosaic virus (of which three strains were sequenced independently between 1980-1982) – in 1980 it was the largest genome sequenced to date (38), more than 1500 bases larger than phage fd's 6,408b genome size (31). The second unique plant virus was tobacco mosaic virus (52), and after these four genomes cassava latent virus' sequence was published (60). A total of 17 plant viral genomes were published in this first decade (Table 1)

Bacteriophage genomes and the decline of phage research: The first two phage genomes were MS2 and phiX174, followed by two additional ssDNA phage published in 1978: microvirus G4 (30) and inovirus fd (31). The next phage genomes were inoviruses M13 (39) and f1 (53), then another RNA phage's genome was published only as part of a PhD thesis in 1981 (Qbeta, (40)). The first dsDNA phage's genome was published at the end of 1982: lambda (58), which had a comparatively gargantuan 48,502 bp genome (Figure 3). A total of 13 phage genomes were published in this first decade (Table 1).

In 1976 bacteriophage research was still driving many of the discoveries in biology. By 1985, phage biologists were being driven into other areas of research as there was a pervasive sense that science understood phages and it was time to study more complex organisms (136). Ry Young, who never wavered in his studies of phage, declared "in a real sense, the field of bacteriophage biology died" between the 1970s and 2002 (137). The dominance of animal viruses, and the declining number of phage completely sequenced illustrates this shift (Table 1). A review on the genome of T4 from 1976 correctly states that T4 "is now one of the best understood biological systems," yet its genome, which is comparably sized to Epstein-Barr virus, would not be fully sequenced until the turn of the millennium (and finally published in 2003 (138)). Only ~200 phage were sequenced by the early 2000s (including multiple sequences of the same species), and researchers noted that there was an equal number of prophage that had been inadvertently sequenced in bacterial genomes by that time (139). A contemporaneous estimate of viral genomes (including phage) in GenBank was that over 1,200 separate viral species were represented, confirming the phage were a small minority of sequenced viruses (140).

Animal viruses were the primary focus of genomic sequencing between 1976 and 1985, with 67 genomes published (Table 1). Polyomavirus SV40 was simultaneously sequenced by two groups, and published in 1978 (12, 25). The 7th and 10th genomes were both hepatitis B (32, 35), published three months apart, and two groups independently sequenced and published the genome of BK polyomavirus at the end of 1979 (33, 34). The first +ssRNA animal viruses were sequenced in 1981, when two groups independently sequenced polioviruses (41, 43). What is immediately clear is that there was increased competition for sequencing the first genome of various animal viruses compared to phage and plant viruses: of the first 10 animal virus

genomes, eight were pairs with two groups publishing four species' genomes (Table 1). This didn't abate as sequencing matured – perhaps the best-known publication tie for a viral genome sequence is the three contributions of researchers aiming for the first sequence of what we now know as HIV. Alternately called lymphadenopathy-associated virus (94), AIDS-associated HTLV-III (96) and AIDS-associated retrovirus ARV-2 (97), all three were all published in January and February 1985 – one each in Cell, Nature and Science. One advantage of this competition is that the field was able to study intraspecific sequence polymorphism on the genomic level faster than if researchers had only sequenced a single isolate (105, 141).

Most of the 67 sequenced animal viruses were associated with tumors (polyomaviruses [8], papillomaviruses [6], hepatitis B-like pararetroviruses [12] and retroviruses associated with cancer including early classifications of HIV [20]). The US government emphasized research funding on these viruses through a Special Virus Cancer Program through the end of the 1970s (142); this partially explains the tailwind behind so much sequencing of tumor-causing viruses. Positive-sense ssRNA viruses dominated the non-tumor-related viral sequences. Only one -ssRNA viral genome was completed before 1985 (influenza A), which also stands out as having the highest number of genomic segments sequenced (50). Over 90% of the genomes sequenced in this decade were monopartite (Table 1), and many segmented genomes were only partially sequenced by 1985. Expediency means it was usually the smallest segment that was finished and published, like the S segment of La Crosse virus (143) or the RNA 3 of tobacco streak virus (144).

Comparative genomics

Prior to effective sequencing methods, relative levels of nucleotide sequence identity were often calculated through hybridization. These assays relied on biochemical principles – the degree of hybridization between genomes of related viruses, and the melting temperatures of the hybridized genomes. These hybrids were visualized on Southern blots (DNA), northern blots (RNA) and even with electron microscopy (145). One of the biggest advantages of the era of nucleotide sequencing is the continual reuse of the data obtained by a single lab and shared widely. Restriction fragment length polymorphisms was another way of comparing genomic sequences, which was more re-usable (146), but was much more crude than obtaining the whole genome sequence. The first papers comparing genomic sequences were often published separately from the individually published genomes (147) but the comparisons became more common to publish with new genomes by the end of our ten year window (e.g., (95)).

Towards the end of our studied decade some papers were able to compare the genomes of several related viruses in the same family, but we did not find any phylogenies that used whole genome alignments. The first phylogeny based on most of the sequenced genome of viruses that we could find was of three polyomaviruses (148), which was compared to the phylogeny of their host mammals, as was the traditional approach for viral phylogenies of the time (62). A few papers like Wain-Hobson et al.'s HIV genome used multiple gene genealogies across the lentiviruses (94), but molecular phylogenetic analysis was not a common part of viral genome sequencing papers in this decade. We did find a whole genome phylogenetic analysis based on

restriction fragment polymorphism, of enterovirus 70 in a paper that included a molecular clock analysis (149), but this analysis didn't use sequences. Another molecular clock analysis of influenza A genomic segments did use aligned sequences, but the study looked at segments in isolation from one another, since there was only one complete flu genome at the time (150).

Those who wish to wade into these early genomic papers should gird themselves against the flagrant misuse of "homology." The majority of publications discussed homology not as a Boolean variable but instead incorrectly used it to mean percent sequence identity (151). This is a rarely observed mistake in the current literature.

The path from published sequences to computerized databases

Nucleic acid sequencing emerged in a world that had been curating protein sequences and was thinking ahead to the vastness of protein primary structures that might be obtained. In 1976, Margaret Dayhoff and colleagues said they had a collection of "77,257 amino acid residues in the 767 sequences currently on our sequence data tape" – all empirically obtained, not inferred from the translation of sequenced mRNA/DNA (152). These sequences were distributed both in book form (Atlas of Protein Sequence and Structure) and by mailed magnetic tapes (153), and was the progenitor of SwissProt (154). While labs interested in sequences from a given organism likely curated their own collections of sequences, there was interest in creating a database that could house all of the nucleic acid sequence information being rapidly published. A conference was held at the Rockefeller Institute in 1979, organized by three scientists who were all working with viruses: Carl Anderson (who would sequence polio), Robert Pollack (primarily interested in cancer, but he worked with SV40-transformed cell lines) and Norton Zinder, an exceptional phage biologist who discovered both the process of transduction and the existence of RNA phages (155). This meeting called for the creation of what would become GenBank, but because of difficulties the US NIH encountered for funding the project, Los Alamos National Laboratory (LANL) developed their own database first (154), with other databases developing independently, including that of the Équipe Évolution Moléclaire which published a book form of their group's database of all sequences they had found in the literature in 1981 (156), the National Biomedical Research Foundation's Nucleic Acid Sequence Database in 1981 (157, 158), and the European Molecular Biology Laboratory Data Library, which had its first release of data via magnetic tape in April 1982 (159). "GenBank" rolled out in mid-1982, initially housed by LANL, which used its connections to ARPANET for database distribution (as well as by book, tape and eventual floppy disks). GenBank and EMBL agreed on a shared format for submissions in the mid-1980s and began mirroring each other's submissions (the DNA DataBank of Japan joined in 1986, after the decade we have focused on, (160)).

None of the publications in Table 1 included a GenBank accession number in their text, but many of the published sequences were transcribed into GenBank in its first few years. However, a complication for matching publications with their accessions is that the lengths of sequences often do not match up between these two venues. Sequences, especially in the earlier days of genomics, often went through revisions as confirmatory sequencing reactions produced different results – perhaps best exemplified by the revision and completion of phiX174's genome between 1977 and 1978. These revisions are sometimes trackable as the GenBank

accessions can have notes (e.g., Epstein-Barr's sequence has one fewer bases than the published genome, explained by the deletion of base 359 detailed in the annotation for V01555) or have multiple papers associated with the accession, but not always. For example, the two contemporary papers on the genome of SV40 showed different lengths of 5,375 and 5,226, but both are associated with the same accession number, which has a different length from both of these publications (5,243 bases). Two of the three HIV sequences published in 1985 are associated with accessions of different lengths than the publication (K02013, (94) and K03455 (96)). In the case of Akv murine leukemia virus, both publications from 1984 direct to an entry that corresponds to the first genome (77); the second genome's sequence is not represented on GenBank. The lack of explanation for the different lengths from the publications is confusing, and the fact that all sequences that have been published are not uniquely represented in GenBank is frustrating. It is clear that the focus in the early days of viral genomics was on getting one genome per species or strain more than cataloging diversity within a species. However, we highlight the inovirus Pf3 researchers who sequenced the same clone in two laboratories on either side of the Atlantic; they published only one sequence, but found several nucleotides different between their sequences and have separate accessions for the Dutch and US versions (Table 1).

As frequent users of GenBank are aware, the date information associated with creation and revision of GenBank files should not be taken at face value. Many of the sequences in our table were initially represented by an accession number that was then replaced by another (often a JXXXXX replaced by a VXXXXX). In the current database there are notations that the latter accession replaced the earlier, obsolete version in the early 2000s. Since both the earlier and latter accessions were published with the sequence in the 1985 version of GenBank, the creation of the newer accession was nearly contemporaneous with the earlier version. As these separate accessions may have been to supply more detailed annotation of features, or reflect a sequence correction, this highlights the importance of GenBank adopting version number suffixes for accession numbers (161). This is a much improved way to revise a file without having to create an entirely new accession for what is a nearly identical or identical entry.

Completeness of our dataset

Historical reviews aim to be as accurate as possible, but we were not scientists at any point during the first ten years of viral genomics. A partial validation of our searching was provided by papers reviewing and comparing viral sequences published shortly after our ten-year window. The papillomaviruses (162) and the geminiviruses (163) were two groups that could be validated this way. A 1987 conference presentation on the history of GenBank states "there are on the order of 130 sequences that one might call complete chromosomes or complete genomes" all from "extremely parasitic entities" (153). Our list of 97 complete viral genomes by 1985 (some of which were not in GenBank by the end of 1987) nests neatly into that 1987 value.

Concluding thoughts

It is not hard to see how that world has embraced phiX174 as the first complete DNA genome, even through Frederick Sanger would have disagreed – a Nature paper entitled "Nucleotide sequence of bacteriophage phiX174" has a way of attracting more attention than a later Journal of Molecular Biology paper entitled "The nucleotide sequence of bacteriophage phiX174." Neither of these phiX174 genomics papers are the correct citation for dideoxy chain termination sequencing (11). Despite this confusion about phiX174, it is harder to understand how so much of the scientific world forgot that MS2 was the first genome sequenced. One of those first four databases failed to enter a "complete sequence" for MS2 even though it was published as one (156), but GenBank has always listed it accurately as a complete genome. Perhaps MS2 is ignored because it has an RNA genome, and there's a bias of many biologists towards DNA-based organisms. Post-1988, when Sanger sequencing became the dominant method it is easy to imagine how younger scientists would see genomics exclusively through the lens of Sanger sequencing and erroneously assume that Sanger must have been the first to sequence a viral genome. Scientists raised exclusively on next-generation sequencing techniques may not be too aware of Sanger sequencing (let alone other early sequencing techniques), so their exposure to phiX174 being the first genome may be completely untethered to the shadow of double Nobel prize-winner Frederick Sanger, and most of their familiarity with phiX174 may be as a dsDNA plasmid-like spike-in for illumina sequencing (29). Regardless, SV40 should have a more prominent place in the history of viral genomics as it is more accurately the first completely sequenced DNA virus. It is unfortunate that some of these early genomic sequences are still not represented in GenBank as decisions were made to represent a species with a single representative instead of including all of the sequenced genomic information.

Methods:

Initially we searched (duckduckgo.com, google scholar) for variants of "complete nucleotide sequence" and specific, well-studied virus names or names of viral families. Google scholar searches were bounded between 1976 and 1985. From those scattershot searches, we identified journals that were publishing viral genomes in the appropriate time frame and read through the tables of contents for all issues between 1976 and 1985: Virology, Journal of Virology, Journal of General Virology, The EMBO Journal, Nucleic Acid Research, and Journal of Molecular Biology. We came across an excellent resource compiling full genomic sequences published by 1983 (one of Margaret Dayhoff's last papers, published posthumously), which supplied citations for several viral genomes we had not found in our other searches (133).

For each published study of a complete viral genome we recorded the date of publication. As some monthly journals consider their archived issues as published on the first day of the month (e.g., Journal of Virology) and others did not (e.g., The EMBO Journal), we assumed all monthly journals to be published on the first day of the month. We recorded the virus name, the genome length, details of the genomic architecture (nucleic acid type and polarity, whether the genome was segmented and whether the genome is circular or linear), and what sequencing methods were used to generate the complete viral genome. We included papers that did not

try to estimate the length of the poly-A tails on RNA genomes, but excluded papers that otherwise noted they did not produce a "complete genome." We included proviruses and completely sequenced defective genomes.

GenBank

We sought to connect the papers that contained viral genomic sequences with their accessions in GenBank. Our main tool was the 1985 and 1986/1987 supplements to Nucleic Acids Research from GenBank and EMBL data libraries, where we combed through the phage and virus records for complete genome sequences (164). Careful reading of these supplements provided an additional opportunity to find novel complete sequences. Searching for the title of publications containing sequences in the nucleotide database assisted in filling in remaining gaps. Not all of the sequences in Table 1 are available in GenBank.

Acknowledgements:

SD acknowledges support from NSF 2308503.

The project was assisted by members of the Duffy lab who helped kick start this project during the pandemic lockdown: Abbey Isaac, Alvin Crespo-Bellido, Atila Lima, Taylor Andrews, and Victoria Sharp. We thank Bob Goodman and Brad Hillman for their comments on a preliminary version of Table 1. We thank Janel Mittelstedt for her able and thorough searching assistance. SD would like to thank her first PI, Dr. David Waugh, who ensured her first DNA sequencing experiences were sending out samples to the in-house Sanger capillary sequencer, the undergraduate lab course taught by Dr. David Norris that required her to prepare radiolabeled material for manual Sanger sequencing, and a rotation in graduate school where she prepared the acrylamide slab gels for Dr. Margaret Riley's capillary sequencer. The formatting of Table 1 is to honor the format of Nucleic Acids Research, a leading journal of this era of viral genomics.

References

1. Geoghegan JL, Holmes EC. 2018. Evolutionary Virology at 40. Genetics 210:1151–1162.

2. Erickson AH, Kilbourne ED. 1980. Comparative amino acid analysis of influenza A viral proteins. Virology 100:34–42.

3. Matsuno S, Nakajima K. 1982. RNA of rotavirus: comparison of RNAs of human and animal rotaviruses. J Virol 41:710–714.

4. Cohen M, Rice N, Stephens R, O'Connell C. 1982. DNA sequence relationship of the baboon endogenous virus genome to the genomes of other type C and type D retroviruses. J Virol 41:801–812.

5. Flores J, Perez-Schael I, Boeggeman E, White L, Perez M, Purcell R, Hoshino Y, Midthun K, Chanock RM, Kapikian AZ. 1985. Genetic relatedness among human rotaviruses. J Med Virol 17:135–143.

6. Sanger F, Brownlee GG, Barrell BG. 1965. A two-dimensional fractionation procedure for radioactive nucleotides. J Mol Biol 13:373-IN4.

7. Jou WM, Haegeman G, Ysebaert M, Fiers W. 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature 237:82–88.

8. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A, Volckaert G, Ysebaert M. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. 5551. Nature 260:500–507.

9. Fiers W, Contreras R, Duerinck F, Haegmean G, Merregaert J, Jou WM, Raeymakers A, Volckaert G, Ysebaert M, Van de Kerckhove J, Nolf F, Van Montagu M. 1975. A-protein gene of bacteriophage MS2. Nature 256:273–278.

10. Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. Proc Natl Acad Sci 74:560–564.

11. Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74:5463–5467.

12. Fiers W, Contreras R, Haegeman G, Rogiers R, Van de Voorde A, Van Heuverswyn H, Van Herreweghe J, Volckaert G, Ysebaert M. 1978. Complete nucleotide sequence of SV40 DNA. Nature 273:113–120.

13. Dasmahapatra B, Dasgupta R, Ghosh A, Kaesberg P. 1985. Structure of the black beetle virus genome and its functional implications. J Mol Biol 182:183–189.

14. Dasgupta R, Ghosh A, Dasmahapatra B, Guarino LA, Kaesberg P. 1984. Primary and secondary structure of black beetle virus RNA2, the genomic messenger for BBV coat protein precursor. Nucleic Acids Res 12:7215–7223.

15. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161.

16. Sanger F. 1988. Sequences, sequences, and sequences. Annu Rev Biochem 57:1–29.

17. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. Nature 550:345–353.

18. Clayton J, Dennis C. 2003. The golden years of molecular biology, p. 26–35. *In* Clayton, J, Dennis, C (eds.), 50 Years of DNA. Palgrave Macmillan UK, London.

19. Men AE, Wilson P, Siemering K, Forrest S. 2008. Sanger DNA sequencing, p. 1–11. *In* Next Generation Genome Sequencing. John Wiley & Sons, Ltd.

20. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM, Smith M. 1977. Nucleotide sequence of bacteriophage ϕX174 DNA. Nature 265:687–695.

21. Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchison CA, Slocombe PM, Smith M. 1978. The nucleotide sequence of bacteriophage ϕX174. J Mol Biol 125:225–246.

22. Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441–448.

23. Fiddes JC. 1977. The nucleotide sequence of a viral DNA. Sci Am 237:54–67.

24. Duffy S, Turner PE. 2008. Phage evolutionary biology. Bacteriophage Ecol 147–176.

25. Reddy VB, Thimmappaya B, Dhar R, Subramanian KN, Zain BS, Pan J, Ghosh PK, Celma ML, Weissman SM. 1978. The genome of simian virus 40. Science 200:494–502.

26. Barrell BG, Air GM, Hutchison CA. 1976. Overlapping genes in bacteriophage ϕX174. Nature 264:34–41.

27. Schekman R. 2007. Arthur Kornberg 1918–2007. Cell 131:637–639.

28. Smith HO, Hutchison CA, Pfannkoch C, Venter JC. 2003. Generating a synthetic genome by whole genome assembly: ϕX174 bacteriophage from synthetic oligonucleotides. Proc Natl Acad Sci 100:15440–15445.

29. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. 2015. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. Stand Genomic Sci 10:18.

30. Godson GN, Barrell BG, Staden R, Fiddes JC. 1978. Nucleotide sequence of bacteriophage G4 DNA. Nature 276:236–247.

31. Beck E, Sommer R, Auerswald EA, Kurz Ch, Zink B, Osterburg G, Schaller H, Sugimoto K, Sugisaki H, Okamoto T, Takanami M. 1978. Nucleotide sequence of bacteriophage fd DNA. Nucleic Acids Res 5:4495–4504.

32. Galibert F, Mandart E, Fitoussi F, Tiollais P, Charnay P. 1979. Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in *E. coli*. Nature 281:646–650.

33. Yang RCA, Wu R. 1979. BK virus DNA: complete nucleotide sequence of a human tumor virus. Science 206:456–462.

34. Seif I, Khoury G, Dhar R. 1979. The genome of human papovavirus BKV. Cell 18:963–977.

35. Valenzuela P, Quiroga M, Zaldivar J, Gray P, Rutter WJ. 1980. The nucleotide sequence of the hepatitis B viral genome and the identification of the major viral genes, p. 57–70. *In* Animal Virus Genetics. Elsevier.

36. Soeda E, Arrand JR, Smolar N, Walsh JE, Griffin BE. 1980. Coding potential and regulatory signals of the polyoma virus genome. Nature 283:445–453.

37. Deninger PL, Esty A, LaPorte P, Hsu H, Friedmann T. 1980. The nucleotide sequence and restriction enzyme sites of the polyoma genome. Nucleic Acids Res 8:855–860.

38. Franck A, Guilley H, Jonard G, Richards K, Hirth L. 1980. Nucleotide sequence of cauliflower mosaic virus DNA. Cell 21:285–294.

39. van Wezenbeek PMGF, Hulsebos TJM, Schoenmakers JGG. 1980. Nucleotide sequence of the filamentous bacteriophage M13 DNA genome: comparison with phage fd. Gene 11:129–148.

40. Mekler P. 1981. Determination of nucleotide sequences of the bacteriophage Qβ genome (PhD thesis). Zurich.

41. Kitamura N, Semler BL, Rothberg PG, Larsen GR, Adler CJ, Dorner AJ, Emini EA, Hanecak R, Lee JJ, van der Werf S, Anderson CW, Wimmer E. 1981. Primary structure, gene organization and polypeptide expression of poliovirus RNA. Nature 291:547–553.

42. Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, Messing J. 1981. The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic Acids Res 9:2871–2888.

43. Racaniello VR, Baltimore D. 1981. Molecular cloning of poliovirus cDNA and determination of the complete nucleotide sequence of the viral genome. Proc Natl Acad Sci 78:4887–4891.

44. Shinnick TM, Lerner RA, Sutcliffe JG. 1981. Nucleotide sequence of Moloney murine leukaemia virus. Nature 293:543–548.

45. Reddy EP, Smith MJ, Aaronson SA. 1981. Complete nucleotide sequence and organization of the Moloney murine sarcoma virus genome. Science 214:445–450.

46. Van Beveren C, van Straaten F, Galleshaw JA, Verma IM. 1981. Nucleotide sequence of the genome of a murine sarcoma virus. Cell 27:97–108.

47. Beck E, Zink B. 1981. Nucleotide sequence and genome organisation of filamentous bacteriophages f1 and fd. Gene 16:35–58.

48. Galibert F, Chen TN, Mandart E. 1982. Nucleotide sequence of a cloned woodchuck hepatitis virus genome: comparison with the hepatitis B virus sequence. J Virol 41:51–65.

49. Danos O, Katinka M, Yaniv M. 1982. Human papillomavirus 1a complete DNA sequence: a novel type of genome organization among papovaviridae. EMBO J 1:231–236.

50. Winter G, Fields S. 1982. Nucteotide sequence of human influenza A/PR/8/34 segment 2. Nucleic Acids Res 10:2135–2143.

51. Kitamura N, Kitamura A, Toyoshima K, Hirayama Y, Yoshida M. 1982. Avian sarcoma virus Y73 genome sequence and structural similarity of its transforming gene product to that of Rous sarcoma virus. Nature 297:205–208.

52. Goelet P, Lomonossoff GP, Butler PJ, Akam ME, Gait MJ, Karn J. 1982. Nucleotide sequence of tobacco mosaic virus RNA. Proc Natl Acad Sci 79:5818–5822.

53. Hill DF, Petersen GB. 1982. Nucleotide sequence of bacteriophage f1 DNA. J Virol 44:32–46.

54. Nomoto A, Omata T, Toyoda H, Kuge S, Horie H, Kataoka Y, Genba Y, Nakano Y, Imura N. 1982. Complete nucleotide sequence of the attenuated poliovirus Sabin 1 strain genome. Proc Natl Acad Sci 79:5793–5797.

55. Shibuya M, Hanafusa H. 1982. Nucleotide sequence of Fujinami sarcoma virus: evolutionary relationship of its transforming gene with transforming genes of other sarcoma viruses. Cell 30:787–795.

56. Balázs E, Guilley H, Jonard G, Richards K. 1982. Nucleotide sequence of DNA from an altered-virulence isolate D/H of the cauliflower mosaic virus. Gene 19:239–249.

57. Chen EY, Howley PM, Levinson AD, Seeburg PH. 1982. The primary structure and genetic organization of the bovine papillomavirus type 1 genome. Nature 299:529–534.

58. Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. 1982. Nucleotide sequence of bacteriophage λ DNA. J Mol Biol 162:729–773.

59. Rhode SL, Paradiso PR. 1983. Parvovirus genome: nucleotide sequence of H-1 and mapping of its genes by hybrid-arrested translation. J Virol 45:173–184.

60. Stanley J, Gay MR. 1983. Nucleotide sequence of cassava latent virus DNA. Nature 301:260.

61. Srivastava A, Lusby EW, Berns KI. 1983. Nucleotide sequence and organization of the adeno-associated virus 2 genome. J Virol 45:555–564.

62. Devare SG, Reddy EP, Law JD, Robbins KC, Aaronson SA. 1983. Nucleotide sequence of the simian sarcoma virus genome: demonstration that its acquired cellular sequences encode the transforming gene product p28sis. Proc Natl Acad Sci 80:731–735.

63. Beveren CV, Straaten F van, Curran T, Muller R, Verma IM. 1983. Analysis of FBJ-MuSV provirus and *c-fos* (mouse) gene reveals that viral and cellular *fos* gene products have different carboxy termini. Cell 32:1241–1255.

64. Schwartz DE, Tizard R, Gilbert W. 1983. Nucleotide sequence of Rous sarcoma virus. Cell 32:853–869.

65. Ono Y, Onda H, Sasada R, Igarashi K, Sugino Y, Nishioka K. 1983. The complete nucleotide sequences of the cloned hepatitis B virus DNA; subtype adr and adw. Nucleic Acids Res 11:1747–1757.

66. Astell CR, Thomson M, Merchlinsky M, Ward DC. 1983. The complete DNA sequence of minute virus of mice, an autonomous parvovirus. Nucleic Acids Res 11:999–1018.

67. Dunn JJ, Studier FW, Gottesman M. 1983. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J Mol Biol 166:477–535.

68. Reddy EP, Reynolds RK, Watson DK, Schultz RA, Lautenberger J, Papas TS. 1983. Nucleotide sequence analysis of the proviral genome of avian myelocytomatosis virus (MC29). Proc Natl Acad Sci 80:2500–2504.

69. Seiki M, Hattori S, Hirayama Y, Yoshida M. 1983. Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc Natl Acad Sci 80:3618–3622.

70. Reddy EP, Smith MJ, Srinivasan A. 1983. Nucleotide sequence of Abelson murine leukemia virus genome: structural similarity of its transforming gene product to other onc gene products with tyrosine-specific kinase activity. Proc Natl Acad Sci 80:3623–3627.

71. Fujiyama A, Miyanohara A, Nozaki C, Yoneyama T, Ohtomo N, Matsubara K. 1983. Cloning and structural analyses of hepatitis B virus DNAs, subtype *adr*. Nucleic Acids Res 11:4601–4610.

72. Stanway G, Cann AJ, Hauptmann R, Hughes P, Clarke LD, Mountford RC, Minor PD, Schild GC, Almond JW. 1983. The nucleotide sequence of poliovirus type 3 leon 12 a 1 b: comparison with poliovirus type 1. Nucleic Acids Res 11:5629–5643.

73. Clark SP, Mak TW. 1983. Complete nucleotide sequence of an infectious clone of Friend spleen focus-forming provirus: gp55 is an envelope fusion glycoprotein. Proc Natl Acad Sci 80:5037–5041.

74. Cornelissen BJC, Brederode FTh, Veenerman GH, Van Boorn JH, Bol JF. 1983. Complete nucleotide sequence of alfalfa mosaic virus RNA 2. Nucleic Acids Res 11:3019–3025.

75. Lomonossoff GP, Shanks M. 1983. The nucleotide sequence of cowpea mosaic virus B RNA. EMBO J 2:2253–2258.

76. Schwarz E, Dürst M, Demankowski C, Lattermann O, Zech R, Wolfsperger E, Suhai S, zur Hausen H. 1983. DNA sequence and genome organization of genital human papillomavirus type 6b. EMBO J 2:2341–2348.

77. Herr W. 1984. Nucleotide sequence of AKV murine leukemia virus. J Virol 49:471–478.

78. Strauss EG, Rice CM, Strauss JH. 1984. Complete nucleotide sequence of the genomic RNA of Sindbis virus. Virology 133:92–110.

79. Ahlquist P, Dasgupta R, Kaesberg P. 1984. Nucleotide sequence of the brome mosaic virus genome and its implications for viral replication. J Mol Biol 172:369–383.

80. Stanway G, Hughes PJ, Mountford RC, Reeve P, Minor PD, Schild GC, Almond JW. 1984. Comparison of the complete nucleotide sequences of the genomes of the neurovirulent poliovirus P3/Leon/37 and its attenuated Sabin vaccine derivative P3/Leon 12a1b. Proc Natl Acad Sci 81:1539–1543.

81. Mandart E, Kay A, Galibert F. 1984. Nucleotide sequence of a cloned duck hepatitis B virus genome: comparison with woodchuck and human hepatitis B virus sequences. J Virol 49:782–792.

82. Etzerodt M, Mikkelsen T, Pedersen FS, Kjeldgaard NO, Jørgensen P. 1984. The nucleotide sequence of the Akv murine leukemia virus genome. Virology 134:196–207.

83. Toyoda H, Kohara M, Kataoka Y, Suganuma T, Omata T, Imura N, Nomoto A. 1984. Complete nucleotide sequences of all three poliovirus serotype genomes: Implication for genetic relationship, gene function and antigenic determinants. J Mol Biol 174:561–585.

84. Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, Gibson TJ, Hatfull G, Hudson GS, Satchwell SC, Séguin C, Tuffnell PS, Barrell BG. 1984. DNA sequence and expression of the B95-8 Epstein—Barr virus genome. Nature 310:207–211.

85. Seeger C, Ganem D, Varmus HE. 1984. Nucleotide sequence of an infectious molecularly cloned genome of ground squirrel hepatitis virus. J Virol 51:367–375.

86. Frisque RJ, Bream GL, Cannella MT. 1984. Human polyomavirus JC virus genome. J Virol 51:458–469.

87. Hamilton WDO, Stein VE, Coutts RHA, Buck KW. 1984. Complete nucleotide sequence of the infectious cloned DNA components of tomato golden mosaic virus: potential coding regions and regulatory sequences. EMBO J 3:2197–2205.

88. Ohno T, Aoyagi M, Yamanashi Y, Saito H, Ikawa S, Meshi T, Okada Y. 1984. Nucleotide sequence of the tobacco mosaic virus (tomato strain) genome and comparison with the common strain genome. J Biochem (Tokyo) 96:1915–1923.

89. Midori K, Katsuro K. 1984. Complete nucleotide sequence of hepatitis B virus DNA of subtype adr and its conserved gene organization. Gene 30:227–232.

90. Howell SH. 1984. Physical structure and genetic organisation of the genome of maize streak virus (Kenyan isolate). Nucleic Acids Res 12:7359–7375.

91. Stanway G, Hughes PJ, Mountford RC, Minor PD, Almond JW. 1984. The complete nucleotide sequence of a common cold virus: human rhinovirus 14. Nucleic Acids Res 12:7859–7875.

92. Roberts RJ, O'Neill KE, Yen CT. 1984. DNA sequences from the adenovirus 2 genome. J Biol Chem 259:13968–13975.

93. Mullineaux PM, Donson J, Morris-Krsinich BA, Boulton MI, Davies JW. 1984. The nucleotide sequence of maize streak virus DNA. EMBO J 3:3063–3068.

94. Wain-Hobson S, Sonigo P, Danos O, Cole S, Alizon M. 1985. Nucleotide sequence of the AIDS virus, LAV. Cell 40:9–17.

95. Peeters BPH, Peters RM, Schoenmakers JGG, Konings RNH. 1985. Nucleotide sequence and genetic organization of the genome of the N-specific filamentous bacteriophage IKe: Comparison with the genome of the F-specific filamentous phages M13, fd and f1. J Mol Biol 181:27–39.

96. Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, Doran ER, Rafalski JA, Whitehorn EA, Baumeister K, Ivanoff L, Petteway SR, Pearson ML, Lautenberger JA, Papas TS, Ghrayeb J, Chang NT, Gallo RC, Wong-Staal F. 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature 313:277–284.

97. Sanchez-Pescador R, Power MD, Barr PJ, Steimer KS, Stempien MM, Brown-Shimer SL, Gee WW, Renard A, Randolph A, Levy JA, Dina D, Luciw PA. 1985. Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). Science 227:484–492.

98. Sagata N, Yasunaga T, Tsuzuku-Kawamura J, Ohishi K, Ogawa Y, Ikawa Y. 1985. Complete nucleotide sequence of the genome of bovine leukemia virus: its evolutionary relationship to other retroviruses. Proc Natl Acad Sci 82:677–681.

99. Neckameyer WS, Wang LH. 1985. Nucleotide sequence of avian sarcoma virus UR2 and comparison of its transforming gene with other members of the tyrosine protein kinase oncogene family. J Virol 53:879–884.

100. Giri I, Danos O, Yaniv M. 1985. Genomic structure of the cottontail rabbit (Shope) papillomavirus. Proc Natl Acad Sci 82:1580–1584.

101. Skern T, Sommergruber W, Blaas D, Gruendler P, Fraundorfer F, Pieler C, Fogy I, Kuechler E. 1985. Human rhinovirus 2: complete nucleotide sequence and proteolytic processing signals in the capsid protein region. Nucleic Acids Res 13:2111–2126.

102. Najarian R, Caput D, Gee W, Potter SJ, Renard A, Merryweather J, Van Nest G, Dina D. 1985. Primary structure and gene organization of human hepatitis A virus. Proc Natl Acad Sci 82:2627–2631.

103. Delmas V, Bastien C, Scherneck S, Feunteun J. 1985. A new member of the polyomavirus family: the hamster papovavirus. Complete nucleotide sequence and transformation properties. EMBO J 4:1279–1286.

104. Shimotohno K, Takahashi Y, Shimizu N, Gojobori T, Golde DW, Chen IS, Miwa M, Sugimura T. 1985. Complete nucleotide sequence of an infectious clone of human T-cell leukemia virus type II: an open reading frame for the protease gene. Proc Natl Acad Sci 82:3101–3105.

105. Sahli R, McMaster GK, Hirt B. 1985. DNA sequence comparison between two tissue-specific variants of the autonomous parvovirus, minute virus of mice. Nucleic Acids Res 13:3617–3633.

106. Howarth AJ, Caton J, Bossert M, Goodman RM. 1985. Nucleotide sequence of bean golden mosaic virus and a model for gene regulation in geminiviruses. Proc Natl Acad Sci 82:3572–3576.

107. Robertson BH, Grubman MJ, Weddell GN, Moore DM, Welsh JD, Fischer T, Dowbenko DJ, Yansura DG, Small B, Kleid DG. 1985. Nucleotide and amino acid sequence coding for polypeptides of foot-and-mouth disease virus type A12. J Virol 54:651–660.

108. Rezaian MA, Williams RHV, Gordon KHJ, Gould AR, Symons RH. 1984. Nucleotide sequence of cucumber-mosaic-virus RNA 2 reveals a translation product significantly homologous to corresponding proteins of other viruses. Eur J Biochem 143:277–284.

109. Seedorf K, Krämmer G, Dürst M, Suhai S, Röwekamp WG. 1985. Human papillomavirus type 16 DNA sequence. Virology 145:181–185.

110. Sprengel R, Kuhn C, Will H, Schaller H. 1985. Comparative sequence analysis of duck and human hepatitis B virus genomes. J Med Virol 15:323–333.

111. Sonigo P, Alizon M, Staskus K, Klatzmann D, Cole S, Danos O, Retzel E, Tiollais P, Haase A, Wain-Hobson S. 1985. Nucleotide sequence of the visna lentivirus: relationship to the AIDS virus. Cell 42:369–382.

112. Nishiguchi M, Kikuchi S, Kiho Y, Ohno T, Meshi T, Okada Y. 1985. Molecular basis of plant viral virulence; the complete nucleotide sequence of an attenuated strain of tobacco mosaic virus. Nucleic Acids Res 13:5585–5590.

113. Rice CM, Lenches EM, Eddy SR, Shin SJ, Sheets RL, Strauss JH. 1985. Nucleotide sequence of yellow fever virus: implications for flavivirus gene expression and evolution. Science 229:726–733.

114. MacDowell S w., Macdonald H, Hamilton W d. o., Coutts R h. a., Buck K w. 1985. The nucleotide sequence of cloned wheat dwarf virus DNA. EMBO J 4:2173–2180.

115. Guilley H, Carrington JC, Balàzs E, Jonard G, Richards K, Morris TJ. 1985. Nucleotide sequence and genome organization of carnation mottle virus RNA. Nucleic Acids Res 13:6663–6677.

116. Groff DE, Lancaster WD. 1985. Molecular cloning and nucleotide sequence of deer papillomavirus. J Virol 56:85–91.

117. Luiten RG, Putterman DG, Schoenmakers JG, Konings RN, Day LA. 1985. Nucleotide sequence of the genome of Pf3, an IncP-1 plasmid-specific filamentous bacteriophage of *Pseudomonas aeruginosa*. J Virol 56:268–276.

118. Lau PCK, Spencer JH. 1985. Nucleotide sequence and genome organization of bacteriophage S13 DNA. Gene 40:273–284.

119. Kodama K, Ogasawara N, Yoshikawa H, Murakami S. 1985. Nucleotide sequence of a cloned woodchuck hepatitis virus genome: evolutional relationship between hepadnaviruses. J Virol 56:978–986.

120. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages, p. 499–560. *In* Methods in Enzymology. Academic Press.

121. 1983. New M13 vectors for cloning, p. 20–78. *In* Methods in Enzymology. Academic Press.

122. Sanger F, Coulson AR, Barrell BG, Smith AJH, Roe BA. 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J Mol Biol 143:161–178.

123. Chait E, Page G, Hunkapiller M. 1988. Battle of the DNA sequencers. Nature 333:477–478.

124. Shchelkunov SN, Totmenin AV, Babkin IV, Safronov PF, Ryazankina OI, Petrov NA, Gutorov VV, Uvarova EA, Mikheev MV, Sisler JR, Esposito JJ, Jahrling PB, Moss B, Sandakhchiev LS. 2001. Human monkeypox and smallpox viruses: genomic comparison. FEBS Lett 509:66–70.

125. Pilartz M, Jeske H. 2003. Mapping of Abutilon mosaic geminivirus minichromosomes. J Virol 77:10808–10818.

126. Kerr L, Browning DF, Lemonidis K, Salih T, Hunter IS, Suckling CJ, Tucker NP. 2021. Novel antibiotic mode of action by repression of promoter isomerisation. bioRxiv https://doi.org/10.1101/2020.12.31.424950.

127. Liu X, Zhu TF. 2018. Sequencing mirror-image DNA chemically. Cell Chem Biol 25:1151-1156.e3.

128. Artusi S, Ruggiero E, Nadai M, Tosoni B, Perrone R, Ferino A, Zanin I, Xodo L, Flamand L, Richter SN. 2021. Antiviral activity of the G-quadruplex Ligand TMPyP4 against herpes simplex virus-1. Viruses 13:196.

129. Roe BA, Ma DP, Wilson RK, Wong JF. 1985. The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J Biol Chem 260:9759–9774.

130. Anderson S, de Bruijn MHL, Coulson AR, Eperon IC, Sanger F, Young IG. 1982. Complete sequence of bovine mitochondrial DNA conserved features of the mammalian mitochondrial genome. J Mol Biol 156:683–717.

131. Bibb MJ, Etten RAV, Wright CT, Walberg MW, Clayton DA. 1981. Sequence and gene organization of mouse mitochondrial DNA. Cell 26:167–180.

132. Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457–465.

133. Chen HR, Dayhoff MO, Barker WC, Hunt LT, Yeh L-S, George DG, Orcutt BC. 1983. Nucleic Acid Sequence Database V: completely sequenced genomes. DNA 2:275–280.

134. Gross HJ, Domdey H, Lossow C, Jank P, Raba M, Alberty H, Sänger HL. 1978. Nucleotide sequence and secondary structure of potato spindle tuber viroid. Nature 273:203–208.

135. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh, W, Fields C, Gocayne JD, Scott J, Shirley R, Liu L, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512.

136. Ofir G, Sorek R. 2018. Contemporary phage biology: from classic models to new insights. Cell 172:1260–1270.

137. Young, Ry. 2006. Forward, p. v. *In* Abedon, ST, Calendar, RL (eds.), The Bacteriophages 2nd ed. Oxford University Press, Oxford.

138. Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W. 2003. Bacteriophage T4 genome. Microbiol Mol Biol Rev 67:86–156.

139. Brüssow H. 2006. Prophage genomics, p. 17–25. *In* Abedon, ST, Calendar, RL (eds.), The Bacteriophages 2nd ed. Oxford University Press, Oxford.

140. Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, Rozanov M, Tatusov R, Tatusova T. 2004. National Center for Biotechnology Information Viral Genomes Project. J Virol 78:7291–7298.

141. Okamoto H, Imai M, Shimozaki M, Hoshi Y, Iizuka H, Gotanda T, Tsuda F, Miyakawa Y, Mayumi M. 1986. Nucleotide sequence of a cloned hepatitis B virus genome, subtype *ayr*: comparison with genomes of the other three subtypes. J Gen Virol 67:2305–2314.

142. Lipsick J. 2021. A history of cancer research: tumor viruses. Cold Spring Harb Perspect Biol 13:a035774.

143. Cabradilla CD, Holloway BP, Obijeskit JF. 1983. Molecular cloning and sequencing of the La Crosse virus S RNA. Virology 128:463–468.

144. Cornelissen BJC, Janssen H, Zuidema D, Bol JF. 1984. Complete nucleotide sequence of tobacco streak virus RNA 3. Nucleic Acids Res 12:2427–2437.

145. Banerjee PT, Olson WH, Allison DP, Bates RC, Snyder CE, Mitra S, Brenner S. 1983. Electron microscopic comparison of the sequences of single-stranded genomes of mammalian parvoviruses by heteroduplex mapping. J Mol Biol 166:257–272.

146. Parrish CR, O'Connell PH, Evermann JF, Carmichael LE. 1985. Natural variation of canine parvovirus. Science 230:1046–1048.

147. Danos O, Engel LW, Chen EY, Yaniv M, Howley PM. 1983. Comparative analysis of the human type 1a and bovine type 1 papillomavirus genomes. J Virol 46:557–566.

148. Soeda E, Maruyama T, Arrand JR, Griffin BE. 1980. Host-dependent evolution of three papova viruses. Nature 285:165–167.

149. Tanimura M, Miyamura K, Takeda N. 1985. Construction of a phylogenetic tree of enterovirus 70. Jpn J Genet 60:137–150.

150. Hayashida H, Toh H, Kikuno R, Miyata T. 1985. Evolution of influenza virus genes. Mol Biol Evol 2:289–303.

151. Reeck GR, de Haën C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, Zuckerkandl E. 1987. "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. Cell 50:667.

152. Dayhoff MO, Barker WC, Schwartz RM, Orcutt BC, Hunt LT. 1976. Data base for protein sequences, p. 261–266. *In* Proceedings of the June 7-10, 1976, national computer conference and exposition. Association for Computing Machinery, New York, NY, USA.

153. Burks C. 1988. The GenBank database and the flow of sequence data for the human genome, p. 51–56. *In* Woodhead, AD, Barnhart, BJ, Vivirito, K (eds.), Biotechnology and the Human Genome. Springer US, Boston, MA.

154. Baxevanis AD. 2020. Biological sequence databases, p. 1–18. *In* Baxevanis, AD, Bader, GD, Wishart, DS (eds.), Bioinformatics, 4th ed. John Wiley & Sons.

155. Smith TF. 1990. The history of the genetic sequence databases. Genomics 6:701–707.

156. Gautier C, Gouy M, Jacobzone M, Grantham R. 1981. Nucleic acid sequences handbook. Praeger.

157. Dayhoff MO, Schwartz RM, Chen HR, Barker WC, Hunt LT, Orcutt BC. 1981. Nucleic acid sequence database. DNA 1:51–58.

158. Dayhoff MO, Schwartz, R.M., Chen, H.R., Barker, W.C., Hunt, L.T., Orcutt, B.C. 1981. Nucleic acid sequence database. Washington, D.C.: National Biomedical Research Foundation.

159. Hamm GH, Cameron GN. 1986. The EMBL data library. Nucleic Acids Res 14:5–9.

160. Strasser BJ. 2008. GenBank—natural history in the 21st century? Science 322:537–538.

161. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. Nucleic Acids Res 41:D36–D42.

162. Giri I, Danos O. 1986. Papillomavirus genomes: from sequence data to biological properties. Trends Genet 2:227–232.

163. Howarth AJ, Vandemark GJ. 1989. Phylogeny of geminiviruses. J Gen Virol 70:2717–2727.

164. Armstrong J. 1985. Nucleotide Sequences 1985: A Compilation from the GenBank and EMBL Data Libraries: A Special Supplement to Nucleic Acids Research. IRL Press.