# Mechanisms of DNA Virus Evolution

**Moriah L Szpara,** Pennsylvania State University, University Park, United States
**Koenraad Van Doorslaer,** University of Arizona, Tucson, AZ, United States



## Glossary

**Adaptation vs. Evolution** Here we use the term "adaptation" to refer to events that occur within a host, and "evolution" to refer to those that occur over much longer spans of time. This allows us to highlight that local adaptation within a host is due to different selective pressures than those that impact transmission to new hosts, or that act across multiple generations of hosts.

**Consensus genome** This term refers to a genome derived by selecting the most commonly observed allele detected at each genomic position in a sequencing-based analysis of a virus sample. Compare this term to "minor variant" below.

**Fossilized or endogenized viral genome** When viral DNA becomes integrated into the host's germline, these endogenous viruses can be vertically transmitted. These endogenized viruses represent a molecular fossil record of past viral invasions.

**Horizontal gene transfer (HGT)** Refers to the movement of a fragment of genetic material between unrelated species. Viral HGT can occur between host and virus, between two viruses, or between a virus and a coincident species that enters the same host cell. Viruses are thought to be major mediators or vectors of HGT, due to their ability to introduce genetic material into new host cells and to infect multiple closely-related host species.

***In vivo* vs. *in vitro*** These terms are used here to distinguish between experiments conducted within a complex host organism (*in vivo*), vs. within cells in culture (*in vitro*).

**Latency** A phase where a herpesvirus is present as an episome in a host cell nucleus, mostly quiescent, and not producing any lytic viral progeny.

**Lysogeny** A phase where bacteriophage or archaeal viruses integrate into their host genome and are propagated along with the host genome as the cell divides.

**Minor variant** A sequence variant which is not the most common allele in a given virus population (e.g., within an infected host). Compare this term to "consensus genome" above.

**Persistent or chronic infection** This term is used to refer to a long-lasting viral infection, i.e., one that exceeds the time frame of an acute infection for that virus species.

**Recombination (homologous vs. non-homologous)** This term refers to the joining of DNA segments after a break. Homologous recombination encompasses several mechanisms such as strand invasion, single-strand annealing, and microhomology-mediated end-joining. Non-homologous recombination involves end-joining without any homology required.

**Single nucleotide polymorphism (SNP)** This term is used here to denote a single nucleotide difference (allele), which is observed when comparing sequenced isolates of a given viral species.

**Standing variation** This refers to a viral population that contains more than one allele or variant at a given locus, or at multiple loci in the genome (e.g., within a single infected host or within a group of hosts). See also the term "minor variant" above.

**Tandem repeats** Short repetitive elements found in any nucleotide sequences. These are categorized based on the length of their repeating unit, *n*, as follows: homopolymers ($n = 1$ base pair, bp), microsatellites ($n < 10$ bp), macrosatellites ($n \geq 10$ bp), minisatellites ($n \geq 100$ bp).

**Transposable elements (TEs)** Transposons are segments of DNA that can move, as a unit, from one location in the genome to another.

## Introduction

A historical view of viral evolution might suggest that the evolutionary processes of RNA and DNA viruses adhere to distinct and non-overlapping rules. RNA virus evolution, as covered elsewhere in this volume, involves error-prone polymerases, an inability to perform error-correction (except in rare cases such as the coronaviruses), the existence of viral quasispecies, and a constant interplay of mutation and fitness-based selection. In contrast, DNA virus evolution is often discussed in more sweeping historical terms, with a focus on how evolution has led to speciation through the slow accumulation of genetic drift and relatively rare fixation of recombination-based genetic shifts. However, there is actually much in common between the mechanisms of evolution for both RNA and DNA viruses. For instance, while the polymerases used by DNA viruses are less error-prone and can perform error-correction, the larger size of many DNA virus genomes still leaves room for the accumulation of genetic variation in every round of viral replication. Furthermore, evidence from multiple DNA viruses suggests that rather than being rare, recombination between DNA virus genomes is rampant. The progeny of these genetic exchanges go unnoticed when recombination occurs between identical or highly similar genomes, or if the progeny do not survive fitness-based selection. Host-linked evolution or co-divergence may also contribute to the apparent low mutation rates in DNA viruses. Understanding the factors that determine the rate at which viral genomes generate and fix mutations provides essential insights into their evolutionary mechanisms. We

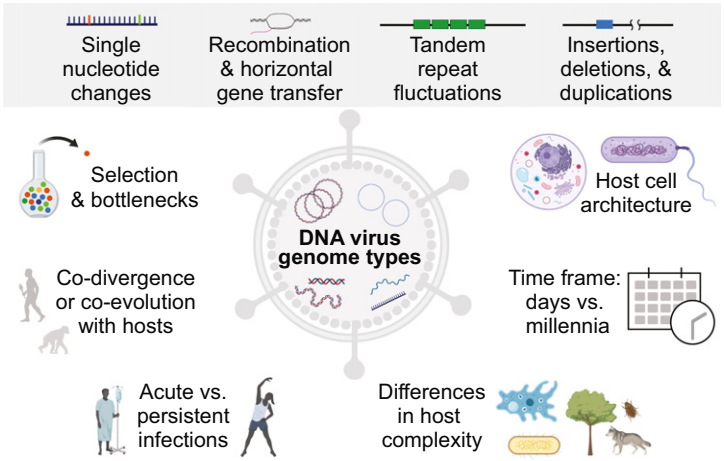


**Fig. 1** DNA virus evolution relies on molecular mechanisms (top, shaded gray) which are impacted by host biology (arrayed below). DNA viruses exist in a range of genome formats (center) and sizes, each of which has a different propensity to evolve via these mechanisms. Viral genome formats include circular and linear DNA that is either single- or double-stranded, with lengths ranging from ~2 to >2000 kbp. The molecular mechanisms that underlie DNA virus evolution include single nucleotide changes, recombination and horizontal gene transfer, fluctuations in tandem repeat length, and sequence gain or loss through insertions, deletions, and segment duplications. Host impacts on DNA virus evolution (listed clockwise) include host cell architecture (e.g., nucleated vs. non-nucleated host cells), the time frame being considered (e.g., one round of infection or many generations), the host complexity (single-cells vs. complex organisms), an acute vs. long-term persistent duration of host infection, selective pressures and bottlenecks that act on each virus population, and co-divergence with host species over millennia. Image created using BioRender.com and Adobe Illustrator.

cover these topics in greater detail below, after introducing a number of additional considerations to the discussion of how DNA viruses evolve (see Fig. 1 for summary).

## Diversity of DNA Virus Genome Types

A simplistic division of evolutionary mechanisms for viruses is generally split based on whether the genome being considered is RNA or DNA. While a single-stranded RNA virus and a double-stranded DNA virus might be considered typical exemplars of each group, these are by no means the only genome types — there are numerous variations on these themes. The prototypic double-stranded DNA (dsDNA) viruses exist in both linear and circular forms. These dsDNA viruses run the gamut in terms of size, from tiny (~5–8 kilobase pairs, kbp) papillomaviruses and polyomaviruses, to large bacteriophage, adenovirus, and herpesvirus genomes (ranging from ~30–250 kbp), to the over-sized nucleocytoplasmic large DNA viruses (NCLDVs) such as poxviruses and phycodnaviruses (~130–400 kbp), and finally the giant viruses found in algae and amoeba (upwards of ~1–2 megabase pairs, Mbp). There are also unusual genome formats among these dsDNA viruses, for instance, the covalently-closed ends of linear poxvirus genomes, or the partially gapped circular dsDNA genome of hepadnaviruses (e.g., hepatitis B virus). Also, there are abundant examples of single-stranded DNA (ssDNA) viruses, which include both linear (e.g., parvovirus and densovirus) and more numerous circular forms (e.g., circovirus, nanovirus, and geminivirus; these are also known as Circular Rep-Encoding Single-Stranded or CRESS DNA viruses). In each case, these genome formats lead to particular constraints and opportunities for the evolutionary mechanisms discussed here. We describe the evolutionary mechanisms below in light of the most common dsDNA virus examples, and where possible, we note those areas where other DNA virus genome formats may differ.

## Host Cell Biology and Availability of Host Enzymes Constrains Virus Evolution

It is possible – though not advisable – to discuss the mechanisms of DNA virus evolution without considering host cell biology. This simplification is enabled by the fact that all known hosts for these viruses are DNA-based life forms, with the concomitant presence of the requisite machinery of a DNA polymerase for replication, RNA polymerase for transcription, and ribosomes for translation. The most apparent distinctions among potential hosts for DNA viruses fall along the known bifurcations of the tree of life – namely bacteria, archaea, and the major groups of eukaryotes (i.e., plants, animals, fungi, and protists). In bacterial and archaeal hosts, the absence of a nucleus removes any distinction in where DNA virus replication occurs. However, in eukaryotes, many host enzymes are constrained to the nucleus, including host DNA and RNA polymerases as well as the RNA splicing machinery, whereas translation is limited to the cytoplasm. Viruses that utilize the host DNA polymerase to copy their genomes, such as members of the *Polyomaviridae* and *Papillomaviridae*, must therefore replicate in the nucleus. Likewise, while the *Herpesviridae* and *Adenoviridae* encode their own DNA polymerase, they use the host RNA polymerase and splicing functions, restricting their replication to the nucleus. In contrast, members of the *Poxviridae* and *Mimiviridae* that replicate in the cytoplasm

encode their own DNA and RNA polymerases, whose fidelity can, therefore, evolve on a separate trajectory from that of the host. Finally, while ssDNA viruses use host DNA polymerases, their observed mutation rate far exceeds that detected in their host-cell genomes or in dsDNA viruses, suggesting that other sources of mutation such as oxidative damage and/or lack of DNA repair may be at play. For these reasons, knowledge of the host cell biology and the usage of host enzymes by a given virus species is a requirement for understanding the constraints on viral evolution.

## Time Frames: Viral Adaptation Within a Host vs. Evolution Over Multiple Generations

Any discussion of the mechanisms of virus evolution needs to begin by defining the time scale under consideration. At the shortest end of this spectrum lies the time frame of a single round of viral infection. As noted below, the first infected cell may be anything from a single-celled organism to the first cellular entry point into a complex human host. From a clinical perspective, viral infection and disease are often considered on the time frame of a single individual's infection – often a human or animal subject. As described below, the virus population within a given host may undergo adaptation within the relatively short time frame of the host's infection. Mechanisms that enable diversification or speciation of a given virus usually require thousands of viral replication cycles, encompassing multiple host generations. At the grandest scale, the origins of viruses and specific lineages thereof spans the history of life on earth. The origins of viruses as we know them are covered elsewhere in this volume, so here we focus solely on the mechanisms that form the foundation of all viral adaptation and evolution. As such, we focus mostly on the time scale of an individual cell and/or host infection, which can include the contributions of virus populations that are more diverse and/or less fit than those which we see preserved over longer sweeps of evolutionary time.

## DNA Virus Hosts Vary From Single Cells to Complex Multi-Cellular Organisms

An understanding of DNA virus adaptation and evolution requires a consideration of the host as a single-cell versus a complex multi-cellular organism. A basic theoretical model of viral replication would include productive viral replication in a single cell, followed by spread to nearby uninfected cells, potentially over multiple generations. This model may well apply to bacterial and archaeal cells, and to single-celled eukaryotic species such as marine alga or amoeba. However in most cases, more complex eukaryotic organisms, from plants to animals and humans, require a complicated series of steps for successful virus propagation and spread. These steps include entry via an accessible portal of the organism, dissemination within the organism to reach susceptible cells, evasion of host defensive responses (including innate and adaptive immunity), and egress to allow for potential spread to new hosts. There is ample evidence that evolution acts within a single host, although for the sake of clarity we will refer to these intra-host events as "adaptation" rather than evolution. Using these terms allows us to highlight the distinction that local adaptation within a host is due to selective pressures that differ from those that impact transmission to new hosts, or that act across multiple generations of hosts. Also, the virus population within a complex organism may partition into distinct environmental niches within the host. For instance, the genomic diversity of human cytomegalovirus (HCMV) in patient samples is often analyzed from blood samples, and yet this viral population does not directly represent a common source of natural virus transmission between hosts (e.g., saliva). Studies of virus evolution need to carefully consider the source material used in examinations of viral diversity, and how this choice may influence the resulting observations of evolutionary fitness.

## The Contributions of DNA Virus Persistence and Chronic Infections

We referred above to a theoretical model of DNA virus replication that involved productive replication in a single cell and spread into nearby uninfected cells, across multiple viral generations. An underlying assumption in such a model is that multiple rounds of productive infection occur sequentially. However, the lifecycle of many if not most DNA viruses exhibit other phases of existence, namely through persistence and chronic infections. For many bacteriophage and archaeal viruses, a common strategy is the well-known cycle of lysis versus lysogeny. For these viruses, the productive and often cell-destructive strategy of lytic replication is interleaved with phases of lysogeny, when the viral genome integrates into the host genome and is propagated as part of the host genome during cell division. A similar strategy exists for the large family of herpesviruses that infect most animal species and humans, with the long-term non-lytic phase being termed latency instead of lysogeny. An important distinction is that with a few notable exceptions, integration into the host genome is not a normal part of herpesvirus latency. Instead, these herpesviruses remain episomal in the host nucleus during lifelong latency. At the molecular level latency can be defined by the absence of significant viral replication and limited viral gene expression. Herpesvirus episomes can undergo sporadic reactivation to produce new viral progeny, which is followed by additional cycles of latency and reactivation. Similar to herpesviruses, certain members of the *Adenoviridae* can progress from a lytic infection of epithelial cells to a latent infection in T-lymphocytes of the tonsils and other adenoid tissues. The ability to establish a long-term infection is thus a vital part of the viral lifecycle of many DNA viruses, which contrasts with the acute infectious period of many RNA viruses (e.g., influenza virus or rotavirus). Persistence and chronic infections motivate the need to explore the contributions of within-host variation and adaptation to the evolutionary mechanisms of DNA viruses.

In addition to latency and lysogeny, virus persistence or chronic infection includes a whole class of DNA virus infections where viral replication is readily detected in the host, but the infection is not cleared for a significant length of time. Many smaller DNA

viruses such as papillomaviruses, polyomaviruses, and certain members of the *Circoviridae* use this "low-and-slow" approach. These viruses replicate in actively dividing cells, but have evolved to avoid detection by the host immune system. Interestingly, many of these viruses appear to be pathogenic only if the virus persists for an extraordinarily long time. For example, in most cases, the host immune system will eventually clear human papillomavirus infections. This process typically spans several months if not years. However, a long-term infection ($>$ 2 years) dramatically increases the risk of virus-induced cancer. Similarly, while polyomavirus infections in humans are typically asymptomatic, long-term persistence of JC polyomavirus causes complications in immunocompromised hosts. In these hosts, the otherwise benign infection can spread into the nervous system, where the viral infection can then induce significant damage (as discussed further below). The duration of animal lifespans, as opposed to single-celled hosts, means that long-term persistent viruses of animal cells have evolved to have significantly more interactions with their host's immune system during lifelong latency, than are observed during bacteriophage or archaeal virus lysogeny. Recent advances in high-throughput sequencing technology are now enabling researchers to interrogate whether mutations in viral genomes are specifically correlated with disease progression in these chronic infection settings.

## Co-Divergence With Hosts as a Driver of DNA Diversification

A common perception is that RNA viruses mutate rapidly while DNA viruses are slow and stable. This may stem from the view that the diversity of many DNA viruses can be explained by co-divergence with host species, thus placing viral evolution on a timescale of millions of years. Long-term co-divergence and consequently low rates of nucleotide substitution have been supported in some DNA viruses; however, this is likely only part of the equation. The development of new sequencing technologies and the ability to include temporal information into molecular clock models allows us to estimate the rate and timescale of virus evolution independent of the (strong) assumption of co-divergence. Indeed, many DNA viruses show evolutionary rates close to those of RNA viruses, which themselves span a range of mutation rates. It is important to note that time-structured sequence data spanning years or decades often contain short-lived polymorphisms. Researchers should thus use caution when comparing mutation rates at such distinct evolutionary scales.

Nonetheless, for many viruses, it is essential to acknowledge that both short and long timescales may provide valuable information. While there is strong evidence supporting co-divergence of the *Polyomaviridae* with their hosts, recent studies have demonstrated the need to account for faster evolution within this virus family. In immunocompromised patients, mutations in the JC polyomavirus capsid protein allow it to escape neutralizing antibodies and invade the central nervous system, causing an opportunistic brain disease called progressive multifocal leukoencephalopathy (PML). The ability to evade the immune system – while remaining extraordinarily stable over longer timeframes – suggests that the *Polyomaviridae* evolve at two distinct rates. In the case of ssDNA parvoviruses, researchers seeking to understand the determinants of host range variation have tested the outcome of culturing several closely related viruses ($>$98% nucleotide identity) in cells derived from phylogenetically distinct hosts. The authors found that canine parvovirus (CPV-2) underwent extensive mutation during passage in non-native host cells, while no mutations arose in cells from the native host. These data indicate that the virus was well-adapted to its current host species, but that multiple mutations in its surface protein were needed for it to infect diverse host species efficiently. These data illustrate how long-term host dependency may constrain evolutionary rates in many DNA viruses.

## Single Nucleotide Differences as a Measure of Evolutionary Change

Specific mutations such as single nucleotide polymorphisms (SNPs), insertions, and deletions (together termed in/dels) are likely to experience different selection dynamics, which impact the chances that these variations become fixed in the population. However, unlike for nucleotide substitutions (i.e., SNPs), the methods for measuring the evolutionary rate of insertions and deletions (in/dels) are not well developed. Because of this limitation, our understanding of viral evolution is primarily based on measuring the accumulation of SNPs over time, which ignores the potentially critical influence of other sources of variation, such as in/dels, tandem repeat fluctuations, and recombination (discussed further below). Recent studies have also provided evidence that viral evolutionary rate estimates decrease as their measurement timescales increase. This is evident in the field of paleovirology. For example, "fossilized" hepadnavirus DNA integrated into bird genomes suggests that these viruses are at least 19 million years old. In turn, this implies a significantly slower evolutionary rate than what was predicted based solely on extant viruses.

Early studies on the mutation rate of DNA viruses using single-gene or single-locus analyzes estimated a mutation rate on the order of $1 \times 10^{-7}$ to $1 \times 10^{-8}$ substitutions/site/year. These values have been further supported by genome-wide comparisons for a handful of large DNA viruses. For instance, a recent study used a high-fidelity high-throughput sequencing (HTSeq) technique called duplex sequencing to detect spontaneous mutations in clonal lineages of human adenovirus 5, and the authors found that these occurred at a rate of $1.3 \times 10^{-7}$ per base, per infection cycle. This rate matches well to a genome-wide estimate of the *in vitro* and *in vivo* mutation rates for murine CMV, which was obtained by shotgun Sanger approaches just before the development of HTSeq ($\sim 1 \times 10^{-7}$ mutations per bp per day). These low mutation rates are often cited by those wishing to contrast DNA virus stability with RNA virus diversity. However, data from both modeling and newer HTSeq-based comparative genomics studies have indicated that large DNA viruses may have mutation rates closer to $1 \times 10^{-5}$ or $1 \times 10^{-6}$. In our comparisons of sub-clones generated from a parental population of herpes simplex virus 1 (HSV-1), we observed 3%–4% variation between sub-clones,

genome-wide. Other studies of HSV have shown that antiviral drug resistance mutations can be selected from a naïve virus population in just one round of viral passage *in vitro*. These data suggest that at least under certain circumstances, standing variation is maintained in DNA virus populations. An alternative or additional theory is that *de novo* mutations may occur at specific genomic regions more often than others (e.g., hot spots). The wider application of genome-wide measurements of viral variation will help to elucidate these possibilities.

### *In Vivo* Observations of Within-Host Diversity and Adaptation of DNA Viruses

Recent advances in high-throughput sequencing have now enabled the detection of minor variants within a single viral isolate or patient. These minor alleles can manifest as a new dominant allele or genotype after population bottlenecks or selective pressures such as antiviral therapy. Evidence of sequential takeover by distinct HCMV strains has been observed in immunocompromised adult patients, demonstrating both the existence of co-infections as well as the opportunities for recombination and/or subsequent selection. Studies of vaccine-associated rashes for varicella-zoster virus (VZV), and of congenital infections by HCMV, have demonstrated the potential for niche-specific adaptation or segregation of viral variants within specific body sites of infected hosts. For human papillomavirus 16 (HPV16), a recent study of several thousand women used a combination of PCR and Illumina-based HTSeq to reveal an unexpectedly high level of viral genetic variability. Of note, there was higher HPV16 genetic variability between patients than within a single patient, suggesting that many of the identified sequence differences were specific to each patient. Interestingly, women with pre-cancerous lesions had significantly less variation than women with a productive (early stage) HPV16 infection, confirming that cellular transformation by HPV represents a genetic bottleneck. This high level of inter-patient variability demonstrates that, at least within some settings, the mutation rate for HPV must be significantly higher than the previously estimated $2 \times 10^{-8}$ nucleotide substitutions/site/year for the viral coding genome. Importantly, the higher-than-expected rate of inter-host evolution argues against the notion that a subset of (oncogenic) human papillomaviruses were acquired by archaic hominins during their migration out of Africa. Together these data indicate that many DNA virus populations may contain and/or generate standing variation following infection. It also appears that this variation is not often transmitted to a new host. The lack of successful transmission of these minor variants suggests that the standing variation in viral populations only becomes phenotypically apparent after population bottlenecks or selection. Importantly, these studies provide corroborating evidence that the molecular mechanisms of DNA virus evolution which have been demonstrated *in vitro,* also operate *in vivo*.

### Fluctuations in Tandem Repeat Copy Number as a Mechanism of Evolution

Changes in the length or copy number of tandem repeats (TRs) provide another mechanism of virus evolution. Short TRs are usually categorized into three groups: homopolymers, which are sequential repeats of a single base (e.g., 5 or more C's in a row); microsatellites, which have a repeating unit of $<10$ base pairs (bp); and mini- or macrosatellites, which include repeating units of 10–500 bp. The mechanisms of repeat expansion or contraction vary by the repeat size. Homopolymer-based length variants are presumed to arise primarily through polymerase slippage, whereas larger TRs may arise either by template looping during polymerase progression or through recombination as discussed below. The repeating units of TRs may be perfect copies or include minor imperfections in the repeating sequence, and these repeats can occur in both coding and non-coding regions. In coding sequences, repeated elements may contribute to structural units of protein folding (e.g., turns of an alpha-helix) or provide variable lengths of unstructured regions within a multi-domain protein. Noncoding repeats have been shown to include promoter elements, chromatin or insulator binding motifs, as well as secondary structural elements such as quadruplexes and other motifs.

For many tandem repeats, the only viral data available is their conservation of position in the genome of a given species, and perhaps data on the degree to which a given TR varies in length across different virus isolates of the species. Functional roles have been demonstrated for select TRs in just a handful of DNA viruses. In the few herpesvirus species that have been shown to integrate into a host genome, there are viral telomeric repeats that function in their integration into the host. In other non-integrating herpesviruses, length variations at homopolymeric tracts in the thymidine kinase (TK) and polymerase genes are a common route of viral escape from the antiviral drug acyclovir. Ribosomal frameshifting of defective transcripts in these drug-resistant genomes allows the translation of a low level of functional TK or polymerase, enabling viral survival even in the face of an otherwise disabling mutation. Fluctuations in TR lengths have also been described for JC polyomavirus populations in patients. In this case a predominant polyomavirus genotype, or archetype, is shed in the urine of most infected individuals, while rearranged forms with deletions and TR variations are found in the brains of patients with PML disease. For poxviruses and other large DNA viruses, restriction fragment length polymorphisms (RFLPs) have often been used to track changes in the dominant virus genotypes and TRs over time. In a recent study of myxoma virus (a *Poxviridae* member), the predominant RFLP type was observed to change each year. Expansion of the inverted terminal repeat boundaries appears to provide myxoma virus with an opportunity for evolution. Likewise, the genome of the vaccinia poxvirus shows similar heterogeneity of the terminal repeats. Repeated plaque based purifications have shown that heterogeneity in the terminal repeats can evolve rapidly from the DNA of a single vaccinia virion. As technologies to track fluctuations in the length of TRs improve, it will no doubt become easier to examine these changes and gain a better understanding of their contribution to virus adaptation and evolution.

## Large DNA Viruses Undergo Frequent Recombination

Recombination can serve as a driving force for evolutionary shifts in DNA viruses, akin to the genetic shifts that result from reassortment in segmented RNA viruses. Recombination can be classified as homologous recombination – between like sequences – or as illegitimate or non-homologous recombination. For most large DNA viruses, the potential of the viral genome to recombine has been studied by analyzing phylogenetic relationships between naturally circulating viral genomes. Among the adenoviruses, which include seven species (human adenovirus A-G) and multiple serotypes, recent studies applying HTSeq-based comparative genomics have demonstrated both intra-species and interspecies recombinants – often in association with pathogenic infections. For instance, a naturally circulating intratypic recombinant of human adenovirus subtype C was found to be the etiologic agent of severe acute respiratory infections in children in China. There are also examples of both historical and recent isolates of pathogenic adenoviruses that appear to have arisen from zoonotic transmission and recombination between simian and human adenoviruses. For the beta-herpesvirus HCMV, multiple studies have demonstrated a history of rampant recombination between the genomes of different isolates. Particular sections or islands of the HCMV genome appear to have co-segregated, while widespread recombination between strains has created a mixture of alleles elsewhere in the genome. It is thought that genes in these islands are co-dependent, thus placing a fitness cost on any recombination events that occur inside these regions. Similar levels of within-species recombination have been shown for most herpesviruses with sufficient genome sequence availability to make these comparisons. Recently, data supporting potential inter-species recombination among these viruses have been observed as well, with HSV-1-like DNA detected in several loci of the HSV-2 genome. Likewise, a virulent avian herpesvirus that created an outbreak in Australian poultry was revealed to be a spontaneous recombinant derived from two live-attenuated vaccines in use in the area. For large DNA viruses such as herpesviruses and poxviruses, laboratory co-infection studies and analysis of recombinant progeny by HTSeq have further defined the genome-wide potential for recombination and begun to define hot spots or regions with a higher propensity to recombine. Together these data demonstrate the extensive role of recombination in the evolution of both nuclear- and cytoplasmic-replicating large DNA virus genomes.

## Recombination at Different Frequencies for Small DNA Virus Genomes

Large DNA viruses appear to recombine more readily than the small dsDNA viruses of the *Papillomaviridae* and *Polyomaviridae*. Even under controlled experimental conditions, no conclusive evidence for recombination within these two virus families has been described. One theory for this lack of observable recombination is that smaller viruses have fully optimized the usage of their genomic real-estate, such that recombination events would be highly likely to interrupt co-dependent genes or regulatory sequences – and thus carry too high a fitness cost to survive. However, phylogenetic analyses have identified evidence for several recombination events within the *Papillomaviridae*. As in HCMV, it appears that ancient recombination has segregated functional regions of the viral genome, separating the genes coding for non-structural proteins from the structural genes. Recombination does not appear to play a significant role in the short-term adaptation of the papillomaviruses, implying that recombined daughter viruses are not as fit as the parental genomes. Supporting this hypothesis, even when evidence of HPV16 recombination was detected within a single patient, these recombinant genomes were incapable of sustained replication within the host. Similarly, while phylogenetic analysis can detect evidence for ancient recombination near the root of the *Polyomaviridae* phylogenetic tree, recombination does not appear to be a significant component of ongoing polyomavirus evolution. However rare recombination events can and do contribute to virus evolution. For instance, conservation efforts to prevent the extinction of the western barred bandicoot have been hampered by an outbreak of the bandicoot papillomatosis carcinomatosis virus type 1 (BPCV1), a recombinant between an ancestral papillomavirus and polyomavirus. This virus is a hybrid that appears to have recombined the structural genes of the *Papillomaviridae* with the non-structural genes of the *Polyomaviridae*. These examples illustrate how rare and unusual recombination events can enable the dramatic expansion of viral evolutionary sequence space.

Despite being roughly the same size as the *Polyomaviridae*, single-stranded DNA viruses recombine relatively efficiently. Single-stranded parvoviruses have shown an ability to jump to new hosts rapidly, and recombination along with a relatively high mutation rate has been hypothesized to underlie this ability. Parvoviruses have also been demonstrated to readily recombine in cell culture. Although the mechanism of parvovirus recombination is not known, a role for viral secondary structure has been proposed. Indeed, the parvovirus origin of replication forms a hairpin structure that is a recombination hot spot, potentially due to stalling of DNA polymerase at this secondary structure. Template swapping before re-initialization of replication could then result in the formation of a chimeric genome. Alternatively, parvovirus replication may create intermediate concatemers. Resolving these concatemers may activate DNA repair enzymes, leading to the creation of mosaic viruses through the homologous recombination repair system. While recombination appears to play an essential role in the evolution of ssDNA viruses, these viruses appear to have adapted to minimize combinations of incompatible regulatory elements. For example, the gene encoding the replication protein (Rep) and the cis-acting elements that interact with the replication protein are usually within 100 nucleotides of one another. This ensures that the replication machinery is highly likely to remain together and compatible following any recombination events. A detailed comparison of recombination patterns within ssDNA viruses also found that breakpoints tend to fall outside of known genes. These observations imply that viruses expressing recombinant proteins are not usually tolerated.

### Duplication and Deletions of Genes and Genome Segments

The outcome of recombination within identical or highly similar genomes is rarely noticed, except for occasions where this event leads to gene duplication or loss. Evidence of gene duplication and subsequent divergence is prevalent in adenovirus genomes. Ancient incidents of gene capture presumably produced those adenoviral gene products with similarity to host genes or those of other viruses, which are found across many adenoviral genera. Other more evolutionarily-recent duplications are found in smaller subsets of adenoviral species. The phenomenon of gene loss has been well-documented in herpesviruses, where across the diverse alpha-, beta-, and gamma-subfamilies of the *Herpesviridae*, many examples of gene loss have been found during viral propagation *in vitro*. The phenomena of genetic drift and gene loss were first detected in laboratory-passaged strains of the beta-herpesvirus HCMV, where the gene regions lost *in vitro* were later found to have functions associated with cell tropism and immune evasion *in vivo*. The extremely large mimivirus dsDNA genome has also been shown to undergo gene loss from both its termini during repeated passage in an amoebal host. In mimivirus, this gene loss was associated with a phenotypic change in virions, which was visible as a loss of fibrils on the virion surface.

In contrast to gene loss, the duplication of genetic segments – a gene accordion – has been best demonstrated by a series of elegant studies in poxviruses grown *in vitro*. These studies showed that expansion of gene copy number could provide functional fitness recovery after deletion of a core viral gene, by driving higher expression of a less-efficient gene version. This expansion also enabled the adaptation and eventual evolution of improved function, via mutations that occurred in the redundant copies of this gene. Whether or not this type of gene accordion occurs for DNA viruses that replicate in the nucleus remains to be determined. The segregated nature of nuclear replication and transcription, followed by translation in the cytoplasm, means that nuclear-replicating viruses will complement defects in co-replicating genomes *in trans*, since proteins made in the cytoplasm can be utilized by all progeny genomes.

Among the small DNA viruses, a subset of human papillomaviruses is associated with recurrent respiratory papillomatosis (RRP). Interestingly, these RRP-associated viruses are not typically considered as oncogenic viruses. However while RRP is considered a benign neoplasm of the larynx, involvement of the lungs is almost invariably fatal. Whole genome sequencing efforts have implicated a duplication of the viral promoter and a subset of viral genes in the RRP progression towards lung invasion. While the expansion of these loci in the papillomaviral genome is likely not important during a normal viral lifecycle, these data illustrate how duplications can provide a powerful adaptation mechanism for otherwise slow-evolving viruses.

### Host-Virus Exchange via Horizontal Gene Transfer and Transposable Elements

Horizontal gene transfer (HGT) provides another avenue for evolutionary adaptation of both viruses and their hosts. HGT has been well-documented between bacterial and archaeal host species, often vectored by large DNA bacteriophages or archaeal viruses. Recent data have demonstrated that HGT may also take place between eukaryotic hosts and their viruses. For example, transposable elements (TEs) found in the moth genome have also been detected in the genomes of baculoviruses that infect these moths. Since this baculovirus infects several species of sympatric, co-occurring moths, it may well be the historical vector that moved TEs among these different host species. Other host-derived sequences were also detected in about 5% of progeny baculovirus genomes, although the co-opted host DNA was not carried beyond a few cycles of viral replication. Most of the integrated host sequences were TEs, but others appeared to result from recombination at sites of microhomology between the host and viral genomes. Most large DNA viruses are not known to integrate into the host genome as part of their overall replication strategy. Select herpesviruses of the alpha- and gamma- subfamilies do integrate into the host genome, although for these viruses it appears to be a reversible process that can lead to later excision and non-integrative replication. Marek's disease virus, an alpha-herpesvirus of poultry, and human herpesvirus (HHV) 6A and 6B, two gamma-herpesviruses of humans, integrate into host telomeres as a central part of their lifecycle. The germline or chromosomal integration of human herpesviruses (ciHHV), usually HHV6A, is detected in about 1% of the human population, although the clinical consequences of ciHHV are as yet unknown. These examples recommend the use of genome-wide HTSeq of viral populations as a means to detect horizontal gene transfer in action.

For the small DNA polyoma- and papillomaviruses, integration of all or a fragment of the viral genome into the host cell DNA is an evolutionary dead end, with an outcome that is nonetheless well-known for having the potential to induce dramatic outcomes of dysregulated cell division and tumor formation. In a recent study of HTSeq data from HPV-positive head and neck cancers, evidence was found to suggest that the HPV genome can replicate as an independent viral–human hybrid mini-chromosome, at least in some instances. These data implied that following an integration event, the viral genome may be excised from the human chromosome, creating a viral–human hybrid circular episome. Under particular circumstances, these hybrid genomes could theoretically get packaged into infectious virions. However, considering the tight regulation of papillomavirus replication, it appears unlikely that these hybrid genomes would be able to establish an infection in the next host.

## Conclusions

Much remains to be resolved about the dichotomy between the measurably low rate of polymerase error in most DNA viruses, and their ability to undergo rapid genetic change in the face of intense selective pressures. However as discussed here, the multiple mechanisms of DNA virus evolution beyond single nucleotide substitutions likely provide the resources to confer this level of

evolutionary adaptability. Researchers have long agreed that ancient events of recombination and horizontal gene transfer, as well as gene duplications and subsequent divergence, could explain many aspects of virus origins. The breadth of new insights offered by high-throughput and deep viral sequencing, as well as by virus discovery and metagenomic approaches, have begun to broaden and clarify this picture. Deep sequencing has revealed the level and ubiquity of standing variation in virus populations, which provides fodder for future adaptation and selection. Metagenomic approaches and viral discovery have allowed researchers to detect novel viruses and recombinants that would have been missed using prior methods, which tracked viral presence using single-point genetic markers. These data provide ample assurance that the textbook explanation of mechanisms of virus evolution will need continued revision in the years to come, as more examples are brought forward and we expand our knowledge of how viral diversity arises and fuels virus evolution.

## Acknowledgments

We appreciate the contributions of Molly Rathbun and other members of the Szpara and Van Doorslaer labs for their helpful input, as well as our many colleagues whose research and insights have been incorporated into this article. M.L.S. acknowledges support from the Eberly College of Science and the Huck Institutes of the Life Sciences at Pennsylvania State University, the Pennsylvania Department of Health Commonwealth Universal Research Enhancement (CURE) Program, as well as from NIH grants R01 AI132692, R21AI130676, and R21 AI140443. KVD is supported by a State of Arizona Improving Health TRIF, and by a USDA Hatch grant NC229.

## Further Reading

Allison, A.B., Kohler, D.J., Ortega, A., *et al.*, 2014. Host-specific parvovirus evolution in nature is recapitulated by *in vitro* adaptation to different carnivore species. PLoS Pathogens 10, e1004475.
Buck, C.B., Van Doorslaer, K., Peretti, A., *et al.*, 2016. The ancient evolutionary history of polyomaviruses. PLoS Pathogens 12, e1005574.
Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: Patterns and determinants. Nature Reviews Genetics 9, 267–276.
Elde, N.C., Child, S.J., Eickbush, M.T., *et al.*, 2012. Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell 150, 831–841.
Firth, C., Kitchen, A., Shapiro, B., *et al.*, 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. Molecular Biology and Evolution 27, 2038–2051.
Gilbert, C., Feschotte, C., 2018. Horizontal acquisition of transposable elements and viral sequences: Patterns and consequences. Current Opinion in Genetics & Development 49, 15–24.
Greenbaum, B.D., Ghedin, E., 2015. Viral evolution: Beyond drift and shift. Current Opinion in Microbiology 26, 109–115.
Houldcroft, C.J., Beale, M.A., Breuer, J., 2017. Clinical and biological insights from viral genome sequencing. Nature Reviews Microbiology 15, 183–192.
Ismail, A.M., Cui, T., Dommaraju, K., *et al.*, 2018. Genomic analysis of a large set of currently–and historically–important human adenovirus pathogens. Emerging Microbes & Infections 7, 1–22.
Jansen, A., Gemayel, R., Verstrepen, K.J., 2012. Unstable microsatellite repeats facilitate rapid evolution of coding and regulatory sequences. Genome Dynamics 7, 108–125.
Koonin, E.V., Yutin, N., 2019. Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. Advances in Virus Research 103, 167–202.
Mirabello, L., Yeager, M., Yu, K., *et al.*, 2017. HPV16 E7 genetic conservation is critical to carcinogenesis. Cell 170, 1164–1174.
Renner, D.W., Szpara, M.L., 2018. The impacts of genome-wide analyses on our understanding of human herpesvirus diversity and evolution. Journal of Virology 92, e00908-17.
Renzette, N., Pfeifer, S.P., Matuszewski, S., Kowalik, T.F., Jensen, J.D., 2017. On the analysis of intrahost and interhost viral populations: Human cytomegalovirus as a case study of pitfalls and expectations. Journal of Virology 91, e01976-16.
Stedman, K., 2013. Mechanisms for RNA capture by ssDNA viruses: Grand theft RNA. Journal of Molecular Evolution 76, 359–364.
Van Doorslaer, K., 2013. Evolution of the papillomaviridae. Virology 445, 11–20.
Zhao, L., Rosario, K., Breitbart, M., Duffy, S., 2019. Eukaryotic circular rep-encoding single-stranded dna (CRESS DNA) viruses: Ubiquitous viruses with small genomes and a diverse host range. Advances in Virus Research 103, 71–133.