On the stability of sequences inserted into viral genomes

Willemsen, Anouk; Zwart, Mark P

Link to publication in KNAW Research Portal

On the stability of sequences inserted into viral genomes

Anouk Willemsen[1],[*],[†] and Mark P. Zwart[2],[*],[‡]

[1]Laboratory MIVEGEC (UMR CNRS IRD University of Montpellier), Centre National de la Recherche Scientifique (CNRS), 911 Avenue Agropolis, BP 64501, 34394 Montpellier cedex 5, France and [2]Netherlands Institute of Ecology (NIOO-KNAW), Postbus 50, 6700 AB, Wageningen, The Netherlands

[*]Corresponding author: E-mail: willemsenanouk@gmail.com; m.zwart@nioo.knaw.nl

[†]http://orcid.org/0000-0002-8511-3244

[‡]http://orcid.org/0000-0003-4361-7636

Abstract

Viruses are widely used as vectors for heterologous gene expression in cultured cells or natural hosts, and therefore a large number of viruses with exogenous sequences inserted into their genomes have been engineered. Many of these engineered viruses are viable and express heterologous proteins at high levels, but the inserted sequences often prove to be unstable over time and are rapidly lost, limiting heterologous protein expression. Although virologists are aware that inserted sequences can be unstable, processes leading to insert instability are rarely considered from an evolutionary perspective. Here, we review experimental work on the stability of inserted sequences over a broad range of viruses, and we present some theoretical considerations concerning insert stability. Different virus genome organizations strongly impact insert stability, and factors such as the position of insertion can have a strong effect. In addition, we argue that insert stability not only depends on the characteristics of a particular genome, but that it will also depend on the host environment and the demography of a virus population. The interplay between all factors affecting stability is complex, which makes it challenging to develop a general model to predict the stability of genomic insertions. We highlight key questions and future directions, finding that insert stability is a surprisingly complex problem and that there is need for mechanism-based, predictive models. Combining theoretical models with experimental tests for stability under varying conditions can lead to improved engineering of viral modified genomes, which is a valuable tool for understanding genome evolution as well as for biotechnological applications, such as gene therapy.

Key words: recombinant virus; genome evolution; genomic deletion; gene expression; genome stability; gene insertion.

1. Introduction

A large number of virus genomes have been engineered to carry additional sequences for a variety of purposes. Viruses are often used as vectors for heterologous gene expression in cultured cells or the natural host. For example, the baculovirus expression system is widely used for expression work (Chambers et al. 2018), lentiviruses show great promise for gene therapy (Milone and O'Doherty 2018), and phage display allows for selection of desired epitopes (Wu et al. 2016). Marker genes have also been built into

viruses to facilitate tracking infection spread (Dolja, McBride, and Carrington 1992). As viruses evolve rapidly, including the incorporation of genome-rearrangements, it is therefore unsurprising that the insertion of sequences into viral genomes often goes hand in hand with the rapid occurrence of deletions (Koonin, Dolja, and Morris 1993; Pijlman et al. 2001; Zwart et al. 2014). The inserted sequence, and sometimes parts of the viral genome, are then rapidly lost. This genomic instability can have economic ramifications, leading to decreases in heterologous protein expression (Kool et al. 1991; De Gooijer et al. 1992; Scholthof, Scholthof,

and Jackson 1996). It can also introduce limitations and complications to working with marker genes (Dolja et al. 1993; Majer, Daròs, and Zwart 2013). Understanding the stability of inserted sequences therefore has value from an applied perspective, but it could also shed light on basic questions. First, how stable are natural virus genomes, and under what conditions do they become unstable? Second, since horizontal gene transfer (HGT) plays an important role in virus evolution, under what conditions are transferred sequences likely to be retained?

In this review, we consider the stability of inserted sequences and the dynamics of their removal from virus genomes from an evolutionary perspective. First, we provide an overview of empirical results which shed light on insert-sequence stability for viruses, based on the Baltimore classification. Second, we present some conceptual considerations pertaining to sequence stability, identifying important parameters for understanding and potentially predicting stability. We identify theory and experiments that point toward viable strategies for mitigating the rapid loss of inserted genes, and point out key questions that should be addressed in future research. We argue that virus genome organization has a large impact on the stability of inserted sequences, whilst stability is a complex trait that can depend on environmental conditions.

2. Overview of empirical results

We provide an overview of empirical results for the stability of natural and engineered inserted sequences, following the Baltimore classification. Our primary focus is on engineered viruses: studies where gene insertions are an addition to the viral genome (leading to an increase in genome size) and where the subsequent fate of these inserted sequences has been tracked. As inserted sequences can incur a fitness cost, these are often quickly purged from the viral genome. Often these fitness costs are related to a disruption of the viral genome (e.g. gene order). We therefore also consider studies on genome rearrangements in wild viruses and introduce other relevant modifications that shed light on what the impact of genomic inserts can be. We provide an overview of the results and main conclusions of our review in Table 1.

2.1 Group I: double-stranded DNA viruses

Several studies relating to the stability of double-stranded (ds) DNA viruses have been published. The dsDNA viruses have a wide range of genome sizes, from the small *Polyomaviridae* and *Papillomaviridae* ranging from 4.5 to 8.4 kbp, to the relatively

Table 1. We provide an overview of the main conclusions, for all viruses and for the different Baltimore classification groups.

Category	Viruses genera covered in relevant studies	Conclusions of this review
All viruses		• Inserted sequences are often unstable and rapidly lost upon passaging of an engineered virus • The position at which a sequence is integrated in the genome can be important for stability • Sequence stability is not an intrinsic property of genomes because demographic parameters, such as population size and bottleneck size, can have important effects on sequence stability • The multiplicity of cellular infection affects sequence stability, and can in some cases directly affect whether there is selection for deletion variants • Deletions are not the only class of mutations that can reduce the cost of inserted sequences, although they are the most common
I: dsDNA	*Alphabaculovirus*, *Lambdavirus*, *Mastadenovirus*, *Orthopoxvirus*, *T7likevirus*, *Varicellovirus*	• Large genomes that are readily engineered and also highly plastic, as exemplified by the 'genome accordion' in poxviruses • Small insertions can be stable, but larger insertion are rapidly lost • Classic studies with phages exemplify how lower limits to the size of packaged genomes can be used to increase insertion stability
II: ssDNA	*Begomovirus*, *Inovirus*, *Mastrevirus*, *Microvirus*	• Smaller genomes, more restrictions on engineering such as the size of the insertion • Inserted sequence identity has a strong effect on stability
III: dsRNA	*Rotavirus*	• Stoichiometric packaging of segmented genomes may form an impediment to engineering and insert stability
IV: ssRNA(+)	*Alphavirus*, *Betaarterivirus*, *Bromovirus*, *Enterovirus*, *Flavivirus*, *Hepacivirus*, *Potexvirus*, *Potyvirus*, *Tobamovirus*, *Tobravirus*	• Polyproteins are a limitation for engineering insertions, but contribute to stability by restricting which deletions are viable • Insertions are generally unstable • Passage duration matters for stability, with shorter passages leading to a higher observed stability
V: ssRNA(−)	*Morbillivirus*, *Respirovirus*, *Vesiculovirus*	• Easily manipulated due to non-overlapping genes, and insertions are more stable than for ssRNA(+) viruses
VI: ssRNA-RT	*Lentivirus*, *Gammaretrovirus*	• Engineered viruses have plastic genomes, so they are amenable to engineering sequence insertions • Integration in the host genome may foster sequence stability, although this remains to be tested and exploited
VII: dsDNA-RT	*Caulimovirus*	• Longer insertions are unstable and rapidly lost, were as shorter inserts can be reasonably stable

recent discovered giant viruses including the *Mimiviridae* and Pandoravirus with genome sizes ranging from 1,200 to 2,300 kbp. Genomic expansions and reductions are common among dsDNA viruses, indicating that their genomes are flexible and that gene insertions do not necessarily reduce viral fitness. Therefore, one would expect that the selection for genome streamlining in dsDNA viruses might not be as strong as in other viruses, in particular for those viruses that have already large genomes.

2.1.1 Wild viruses

Even when viruses have large DNA genomes, they are not very stable *per se*. A wild non-modified example is vaccinia virus. The inverted terminal repeats of vaccinia virus undergo rapid changes in size due to unequal crossover events leading to stable and unstable forms (Moss, Winters, and Cooper 1981). The diversity in this region is needed for immune evasion and for the colonization of novel hosts and appears to be mainly regulated by recombination events. However, other processes such as mutation leading to accelerated rates of recombination cannot be ruled out.

Poxviruses, such as vaccinia, virus are classified as nucleocytoplasmic large DNA viruses (NCLDVs). These viruses have larger than average genome sizes and the more recently discovered giant viruses are also classified as such. The NCLDVs appear to have undergone a dynamic evolution where gene gain and loss events go in parallel with host-switches between animal and protist hosts (Koonin and Yutin 2018). Interestingly, the phylogenomic reconstructions performed by Koonin and Yutin (2018) suggest that giant viruses (for which the host range appears to be restricted to protists) have evolved from simpler viruses (infecting animals) on many independent occasions. This again suggests that the host plays an important role in genome stability where in animals the pressure for smaller virus genomes is stronger than in protists (Koonin and Yutin 2018).

2.1.2 Engineered viruses

Experimentally it has also been shown that vaccinia virus has a highly plastic genome. After deletion of one host range gene of vaccinia virus, another host range gene increases in copy number (Elde et al. 2012), leading to genomic expansion. The increased gene expression is in itself beneficial, but the high gene copy number also increases the supply of beneficial gain-of-function mutations. Once these gain-of-function mutations are fixed in the population, the other copies of the gene are lost and thus the vaccinia genome size decreased (associated with the cost of an increased genome size) (Elde et al. 2012), leading to accordion-like evolutionary dynamics (Andersson, Slechta, and Roth 1998).

Modified vaccinia virus Ankara (MVA) is used as a viral vector for the development of vaccines against infectious diseases such as malaria, influenza, tuberculosis, HIV/AIDs, and Ebola (Sutter and Staib 2003; Gómez et al. 2012; Gilbert 2013; Stanley et al. 2014) The optimization of poxvirus promoters in this viral vector has proven to be an effective strategy for increasing the stability of antigen (inserted sequence) expression, and therewith the development of MVA-based vaccines (Alharbi 2019).

Although live attenuated vaccines have substantially reduced rabies prevalence after oral-vaccination campaigns were conducted (Lafay et al. 1994; MacInnes et al. 2001), such live vaccines are not efficacious in all rabies vector species. As an alternative, recombinant human adenovirus vaccine vectors expressing the rabies glycoprotein have been developed. The fitness of a replication-competent human adenovirus expressing the rabies glycoprotein was similar to that of the wild-type virus, as tested *in vitro* (Knowles et al. 2009). Moreover, the inserted rabies virus gene was stable during both *in vivo* and *in vitro* passaging (Knowles et al. 2009), demonstrating the potential of this recombinant vaccine vector as an effective alternative.

Non-human adenoviruses can be used as alternative vaccine vectors, providing several advantages such as a limited host range and restricted replication in non-host species. By using bovine adenovirus type 3, a variety of antigens and cytokines were successfully expressed *in vivo* (Ayalew et al. 2015). The stability of bovine adenovirus type 1 was tested by inserting the *EYFP* marker and subsequently passaging the recombinant virus in cell culture (Ren et al. 2018). Although replication of this recombinant virus was less efficient than the wild-type virus, the inserted *EYFP* was stable.

Engineered alphabaculoviruses (infecting arthropods) are widely used as vectors for the expression of heterologous genes in insect cells. Nonetheless, during serial passaging defective interfering (DI) baculoviruses that lack large portions of the genome are rapidly produced, in what appears to be an intrinsic property of baculovirus infection (Pijlman et al. 2001). As a result of having a smaller genome size, these DIs most likely have a replicative advantage (higher fitness). Especially in bioreactor configurations where the cellular multiplicity of infection (MOI, the number of virus particles infecting a cell) is high, faster-replicating DIs can rapidly reach high frequencies (Kool et al. 1991). The rapid generation of DIs involves several recombination steps and prevents the development of stable baculovirus expression vectors, as inserted sequences are then also rapidly lost (Pijlman et al. 2001).

The loss of sequences inserted into baculovirus genomes is not only due to the formation of DIs. When an origin of replication that is enriched in DI genomes was removed, baculovirus genomic stability at high MOIs increased as no DIs were observed. Strikingly, inserted foreign sequences were still rapidly lost (Pijlman, van Schinjndel, and Vlak 2003), showing that rapid DI generation is not the only impediment to the stability of inserted genes. Addition of endogenous viral sequences—homologous repeat regions important for baculovirus replication—to inserted sequences promoted the stability of insertions (Pijlman et al. 2004), highlighting the importance of the genomic context for insert stability.

Another study in which the importance of the genomic context was stressed involved the generation of infectious clones and determination of the stability of Suid herpesvirus 1, the causal agent of Aujeszky's disease. Sequences inserted in infectious clones were genetically stable in *Escherichia coli*. However, for the reconstituted viruses, the insertion at the *gG* locus was highly unstable, whereas the same insert was stable when inserted between the *Us9* and *Us2* genes (Smith and Enquist 1999, 2000). Stability was only determined in a short-term experiment, but these results nevertheless emphasize the importance of the genomic context for stability, even in viruses with relatively large and stable genomes.

Bacteriophages were instrumental in the development of molecular cloning methods. Among dsDNA phages, lambdaviruses of *E.coli* were widely used as cloning vectors, and methods were developed to increase the stability and maximum size of inserts (Chauthaiwale, Therwath, and Deshpande 1992). One interesting approach made use of the fact that there is a minimum genome size for efficient packaging into virus particles. When endogenous genes that are non-essential for the lytic cycle are removed, not only can larger sequences be inserted, but there is also selection for maintaining the inserted sequences

because they increase genome size and enable packaging (Thomas, Cameron, and Davis 1974). Moreover, it has been shown that phage T7 engineered with a biofilm-degrading enzyme (dispersin B) was superior to unmodified phage at clearing short-term biofilms (Lu and Collins 2007). Although providing a 'public' benefit in the form of an exoenzyme that can degrade host defenses, surprisingly this insertion does not have a cost and is therefore stable (Schmerer et al. 2014). Interestingly, the insertion of an endosialidase at the same locus was both beneficial and costly, although in this case evolutionary stability was not determined (Gladstone, Molineux, and Bull 2012).

In summary, engineered dsDNA viruses containing foreign gene insertions are relatively unstable and stability is only reached when the genomic context and demographic conditions (e.g. census population sizes, bottleneck sizes, and population structure) are optimal. Contrarily, in natural conditions dsDNA viruses appear to be highly plastic where increases and decreases in genome size occur on a relatively short evolutionary time scale. In particular, host-switches may play important roles in increased plasticity and stability of dsDNA viral genomes. Even though unstable viral genomes may help increase viral fitness by avoiding the hosts' immune system in natural conditions, it may also prevent the development of stable viral expression vectors in bioreactor configurations.

2.2 Group II: single-stranded DNA viruses

The ssDNA viruses have much smaller genome sizes as compared to the dsDNA viruses (group I), ranging from 1.8 to 2.3 kbp genomes of the *Circoviridae* to the 24.9 kbp genome of the *Spiraviridae*. Judging only by the small range in genome size, one would expect that ssDNA viruses are less plastic compared their dsDNA counterparts, and thus less likely to accept foreign genes in their genomes.

2.2.1 Wild viruses

Although few studies have addressed genomic stability of ssDNA viruses after an insertion, an example in wild viruses of frequent sequence insertions, duplications, and deletions are the *Geminiviridae*, with genomes of about 2.5–3 kbp (monopartite) or 4.8–5.6 kbp (bipartite). During the course of geminivirus infection in plants, shorter subgenomic DNAs often arise. These subgenomic DNAs can range in size and some result in defective DNAs (Stenger et al. 1992; Stanley et al. 1997; Patil et al. 2007), that replicate at the expense of the full-length genome. These subgenomic DNAs can lead to reduced symptom severity in plants and thereby act as modulators of viral pathogenicity. It is speculated that the (sometimes stepwise) deletion process leading to subgenomic DNAs can also be the process leading to the reversion to wild-type full-length DNA molecules with either insertions or deletions that make these bigger or smaller than the wild-type genome (Martin et al. 2011).

2.2.2 Engineered viruses

When inserting sequences into the genome of maize streak virus (MSV, *Geminiviridae*), the infection efficiency decreased as the size of the insert increased (Shen and Hohn 1991). Although, some of the MSV mutants obtained deletions and reverted to the wild-type length, the frequency of the deletion process did not increase linearly with the size of the insert, but rather depended on the nature of the sequence (Shen and Hohn 1991). Deletion mutants of the African cassava mosaic virus (ACMV) have also shown to revert back to the original wild-type genome length through recombination between the two components of the bipartite genome

(Etessami, Watts, and Stanley 1989). The selection pressure on the reversion to wild-type genome length is probably a strong size constraint on encapsidation, where in the case of ACMV the size of encapsidated DNA determines the multiplicity of geminivirus particles (Frischmuth, Ringel, and Kocher 2001).

The *Nanoviridae* family includes ssDNA viruses with a multipartite genome that are composed of six to eight circular segments. Segmented ssDNA viruses present unique challenges when thinking about the stability of inserted sequences, because the frequency of genomic segments is highly plastic for some of these viruses. These viruses might therefore downregulate segments for expression of the inserted sequence, even if downregulation of co-localized homologous genes is costly. A lower frequency would also entail a lower mutational supply, limiting evolvability, the capacity of the virus to generate beneficial variation and subsequently adapt. Segmented viruses might therefore display rapid adaptive responses to inserted sequences, whilst simultaneously limiting their potential for longer-term evolution. To the best of our knowledge, this potentially interesting trade-off has not been shown.

Inserted sequences can be unstable in ssDNA phages, which like their dsDNA counterparts also can have an upper limit to genome size. Inserts of up to 163 bp were stable in X174, despite markedly reducing fitness (Russell and Muller 1984). Genomes with larger insertions were still infectious, although the insert was then rapidly lost. Later, it was shown that short palindromic sequences could be inserted in X174, but that these inserts become more unstable as the number of repeats is increased and when the identity of the repeats is identical (Williams and Müller 1987). In other work, it has been shown that phage display (Wu et al. 2016) can be used to select clones coding for peptides with high affinity for a particular target, although selection for M13 phages with no insert—due to their presumed faster replication—can hamper 'phage panning' (Tur et al. 2001).

Based on the little evidence we obtained there appears to be strong selection for genome streamlining in ssDNA viruses. After a sequence insertion, reversion to the wild-type genome size is observed in both natural and laboratory conditions. Interestingly, the nature of the insert appears to be more important than the size of the insert, indicating that the genomic context also plays an important role in the stability of ssDNA viruses.

2.3 Group III: dsRNA viruses

The dsRNA viruses have a range of genome sizes (3.7–30.5 kb) that is similar to the ssDNA viruses. Most of the dsRNA viruses contain segmented genomes, where during replication, positive-sense ssRNAs are packaged into procapsids and serve as templates for dsRNA synthesis. Thus, the progeny particles contain a complete set of equimolar genome segments. Proper recognition and stoichiometrical packaging of the ssRNAs is indispensable for multi-segmented genome assembly. Although, different dsRNA viruses employ different mechanisms for this assembly, these all rely on proper recognition of the ssRNAs in either specific RNA–protein or RNA–RNA interactions (Borodavka, Desselberger, and Patton 2018). We therefore expect that dsRNA virus genomes are highly streamlined, since most gene insertions will probably disturb the recognition and packaging process of the ssRNAs. Interestingly, for rotaviruses it has been observed that genome segments containing sequence duplications are preferentially packaged into progeny viruses relative to wild-type segments (Troupin et al. 2011), indicating that an increase in genome/segment size may not be a hard constraint. We

hypothesize that few if any gene insertions will lead to viable genomes due to the perturbation of segmented genome assembly into virus particles. If a gene insertion happens to be viable, it will probably be rapidly purged from the viral genome. One exception to this hypothesis could be a gene insertion originating from a closely related virus, for example a virus with similar packaging signals, leading to a fitness advantage, such as increased packaging efficiency.

2.3.1 Engineered viruses

Only a small number of studies that test the stability of inserts in dsRNA viruses are available, and these concern the generation of recombinant rotavirus expressing foreign genes. Group A rotavirus, consisting of eleven segments, has been engineered to express fluorescent proteins (Kanai et al. 2017, 2019; Komoto et al. 2017) such as enhanced green fluorescent protein (eGFP) and mCherry. However, segment 5 in which these genes were introduced is expressed at low levels and is subject to proteasomal degradation (Martínez-Álvarez et al. 2013). By introducing GFP in segment 7 as a fusion protein, Philip et al. (2019) were able to generate a recombinant rotavirus without deleting or interrupting any of the open reading frames (ORFs) of the virus. This virus expressed higher levels of green fluorescent protein (GFP), replicated efficiently, and was genetically stable in vitro, suggesting that this alternative recombinant virus may be a more sensitive tool for studying rotavirus biology (Philip et al. 2019).

Based on limited evidence, we tentatively conclude that stoichiometric packaging of segmented genomes may form an impediment to engineering and insert stability. However, recent work also suggests that careful engineering of dsRNA viruses may lead to stable sequence insertions. The generality of these conclusions for the dsRNA viruses, and their dependence on environmental and demographic conditions, remain to be seen.

2.4 Group IV: single-stranded RNA(+) viruses

The ssRNA(+) viruses range in genome size from 2.3 to 31 kbp. It has been shown that both animal and plant ssRNA(+) viruses can express inserted foreign genes. However, the nature of the ssRNA(+) genomes poses several limitations to efficient expression and maintenance of the insert. Most ssRNA(+) genomes used for the expression of foreign genes code for a polyprotein, a single ORF that is further processed after translation into different mature peptides. The processing occurs through autocatalytic cleaving at specific cleavage sites located between the different proteins to be expressed. Insertions should therefore be carefully engineered, including proper cleavage sites corresponding to the site of insertion. Even when respecting these design rules, inserts may impose restriction on viral replication due to the failure of proper protease cleavage due to conformational constraints. In addition, the genomes of RNA viruses tend to be composed of overlapping genes (Belshaw, Pybus, and Rambaut 2007), which limits their adaptive capacity (Simon-Loriere, Holmes, and Pagán 2013). Overlap can form an impediment to engineering, and perhaps to the likelihood an inserted sequence is maintained, as insertions will often affect multiple genes. We focus exclusively on engineered viruses, given that there are many examples for this virus group.

2.4.1 Engineered viruses

Poliovirus is a good candidate as a live viral vector for the expression of foreign genes, since the attenuated Sabin strains of poliovirus elicit strong protective immune responses without causing disease (Sabin 1957). Insertions of up to 534 bp of the rotavirus VP7 gene into Sabin 3 poliovirus gave rise to infectious viruses that expressed portions of the VP7 outer capsid protein (Mattion et al. 1994). This is promising as antibodies generated to VP7 are able to neutralize the virus. Nevertheless the size of the insert in this construction is limited as only inserts of about 300 bp or smaller were stable upon serial passages in tissue culture, whereas larger insertions failed to produce infectious viruses (Mattion et al. 1994). One of the major limitations appears to be the polyprotein nature of the genome. The recombinant viruses expressing the inserted gene were found to be slower in the assembly of infectious virus particles, and showed smaller plaques and lower virus titers. This is possibly due to slow cleavage at the artificial cleavage sites around the insert (Mattion et al. 1994). Sindbis virus, another ssRNA(+) virus genome that encodes for a single ORF, accepted relatively large inserts of 3.2 kbp in the 11.8 kbp genome (Pugachev et al. 1995). However, the recombinant Sindbis viruses appeared unstable and especially inserts at the 3′ end were rapidly lost during serial passages, suggesting a positional effect.

For members of the Flaviviridae family, such as West Nile virus and hepatitis C virus (HCV), inserted reporter genes appear to be unstable. This instability is related to the size of the insert, and comes about because of the disruption of structural RNA elements required for viral replication (Ruggli and Rice 1999; Pierson et al. 2005). To cope with these issues, recombinant Flaviviridae viruses carrying the split-luciferase gene were generated (Tamura et al. 2018), including dengue virus, Japanese encephalitis virus, HCV and bovine viral diarrhea virus. In vitro, these recombinant viruses appear to be evolutionary stable and propagation was comparable to the wild-type virus, most probably due to the small 11 amino acid insert size. To demonstrate the utility of the split reporter system—to determine in vivo viral dynamics and the efficacy of antiviral reagents—the recombinant HCV was tested in chimeric mice. Chronic infection was established and the luciferase gene was stably maintained in the viral genome (Tamura et al. 2018).

Live attenuated vaccines for porcine reproductive and respiratory syndrome (PRRSV) have failed to provide effective protection due to the genetic diversity of circulating PRRSV strains. To improve the efficacy of PRRSV vaccination, a recombinant virus expressing porcine Interleukin-4 (a regulator of the immune response) was constructed (Zhijun Li et al. 2015). This recombinant virus remained stable upon serial passaging in vitro, and induced higher ratios of Interleukin-4 and CD4+CD8+ double-positive T cells in vivo. Despite the presumably better immune response of the host, the recombinant PRRSV vaccine did not significantly improve protection efficacy (Zhijun Li et al. 2015). In another attempt, granulocyte–macrophage colony-stimulating factor (GM-CSF) was inserted in a PRRSV vaccine strain. The inserted gene was stably expressed upon serial passaging in vitro, and the presence of GM-CSF led to increased surface expression of MHCI+, MHCII+, and CD80/86+ (Yu et al. 2014). Although evaluated solely in vitro, this recombinant strain is expected to elicit stronger immune responses and hereby improve vaccine efficacy against PRRSV infection.

It has been shown that many different plant viruses can express foreign genes, and they have the advantage being able to express these directly in vivo. As an initial strategy to express foreign genes in plants, on many occasions viral genes were replaced with the gene of interest (gene replacement instead of gene insertion). This strategy appeared to be (partially) successful in plant ssDNA viruses (Hayes et al. 1988; Ward, Etessami, and Stanley 1988; Hayes, Coutts, and Buck 1989), as the replaced

coat protein did not appear to play an essential role in virus spread throughout the plant host (Ward, Etessami, and Stanley 1988). However, viral ssRNA(+) genomes seem to be less plastic as the replacement strategy was mostly unsuccessful in plant RNA viruses. Although the RNA viral vectors permitted the expression of replaced genes, either they were only viable in protoplasts and not in whole plants (French, Janda, and Ahlquist 1986; Joshi, Joshi, and Ow 1990), or they were unable to establish systemic infections (Takamatsu et al. 1987; Dawson, Bubrick, and Grantham 1988).

Shortly after, studies showed that gene insertion—rather than gene replacement—was better suited for expressing foreign genes in ssRNA(+) viral genomes (Dawson et al. 1989; Donson et al. 1991; Chapman, Kavanagh, and Baulcombe 1992). The chloramphenicol acetyltransferase (CAT) gene (Dawson et al. 1989), and the dihydrofolate reductase (DHFR) and the neomycin phosphotransferase (NPT) genes (Donson et al. 1991) were successfully expressed in plants by using Tobacco mosaic virus (TMV) as a vector. In addition, the bacterial GUS gene has shown to successfully express when inserted into the viral genome of Potato virus X (PVX) (Chapman, Kavanagh, and Baulcombe 1992). However, in all these cases the presence of a foreign gene leads to genomic instability resulting in the partial deletion of the GUS and NPT genes and a complete deletion of CAT during systemic infection. This instability may result from the presence of the insert leading to lower accumulation levels of the genomic RNA, as well as leading to mRNA instability and/or interfering with synthesis of the viral proteins.

Sequence redundancy due to a promoter duplication can also lead to genomic instability and thus the subsequent deletion of the inserted sequence (Dawson et al. 1989; Chapman, Kavanagh, and Baulcombe 1992). Indeed, for TMV and PVX it has been shown that replacing one of the promoter sequences with that from related viruses (Donson et al. 1991) together with further removal of additional sequence duplications (Dickmeis, Fischer, and Commandeur 2014), leads to increased stability of the insert.

Interestingly, as for the DNA viruses, the site and size of the insert seems to be important for ssRNA(+) viruses. First, the positioning of the CAT gene downstream (instead of upstream) of the TMV coat protein, resulted in a poorly replicating virus that was not able to systematically infect the host plants (Dawson et al. 1989). And second, the DHFR gene (238 bp) inserted in a TMV background appears to be maintained stably through several passages, while the 3.5× larger NPT gene (832 bp) in the same experimental setup was unstable during systemic movement of the virus. This may also be related to the nature of the insert, where sequences with a codon usage similar to that of the viral vector may be retained longer than those that have an opposite codon usage. Interestingly, Chung, Canto, and Palukaitis (2007) generated recombinant plant viruses with inserted genes of unrelated plant viruses and observed instability and variation in the rate of partial or complete loss of the insert depending on the inserted sequence itself, the host used, or the viral vector used (Chung, Canto, and Palukaitis 2007). Also sequences with a high toxicity for the host, are more likely to become deleted faster or to impede viral replication.

In a previous study we reported on experimental evolution of pseudogenization in virus genomes using Tobacco etch virus (TEV) expressing eGFP (Zwart et al. 2014), a gene known to be toxic in many expression systems. In this case eGFP can be considered a non-functional sequence, as it does not add any function to the viral genome. We showed that eGFP has a high

fitness cost in TEV, and the loss of eGFP depended on the passage length, where longer passages led to a faster and assured loss. Similarly, prolonged propagation of TEV and plum pox potyvirus expressing GUS (Dolja, McBride, and Carrington 1992; Dolja et al. 1993; Guo, López-Moya, and García 1998), and TMV expressing GFP (Rabindran and Dawson 2001), led to the appearance of spontaneous deletion variants. Due to the increase in genome size, viruses that carry an insert are unlikely to be as fit as the parental (ancestral) virus, even if they accumulate initially to similar levels. The TEV-eGFP genomes that had lost the insert had a within-host competitive fitness advantage, where the smaller the genome the higher the within-host competitive fitness. Interestingly, although the size of the deletions varied, convergent evolution did occur in terms of fixed point mutations (Zwart et al. 2014). This result also suggests that a demographic 'sweet spot' exists, where heterologous insertions are not immediately lost while evolution can act to integrate them into the viral genome. In summary, in several studies passage duration has an effect on insert stability, with inserts being more stable in shorter passages. We explore these effects in the conceptual section presented at the end of this paper (see also Box 1).

Box 1. A simulation model of the effects of genetic drift on insertion stability.

Here we illustrate how demography can affect the observed stability of an inserted sequence, using a simulation model. This model is based on (Willemsen et al. 2016) and incorporates logistic virus growth, deterministic recombination with a fixed rate, and population bottlenecks after a given number of generations. To describe virus growth and recombination in each generation, two coupled ordinary differential equations are used:

$$\frac{dI}{dt} = \omega_I I\left(1 - \frac{I + \psi_D D}{\kappa}\right) - \rho I, \qquad (1)$$

$$\frac{dD}{dt} = \omega_D D\left(1 - \frac{D + \psi_I I}{\kappa}\right) + \rho I. \qquad (2)$$

Here, I is the number of viruses with the insertion intact, D is the number of viruses with a deletion, ω is initial growth rate of each virus variant, κ is the carrying capacity, ρ is the rate at which I recombines to D, and ψ is a constant for determining the effect of each virus on the others replication, with the effect of D on I being $\psi_D = \omega_D/\omega_I$ and vice versa $\psi_I = 1/\psi_D$. The frequency of the deletion variant is $f_D = D/(D + I)$. At the start of each passage, to simulate the bottleneck we draw the number I from a Binomial distribution with a size α and success probability f_D from the previous time point, and then $D = \alpha - I$. To illustrate the effects of bottlenecks we chose the parameters in Table 2, set the initial f_D to zero, and considered various values of α. The difference in fitness between the virus with insertion and without is large ($\omega_I/\omega_D = 0.8$). The simulation data illustrate how under these conditions narrow bottlenecks can lead to stable inserted sequence (Fig. 2). During each round of passaging the frequency of the deletion variant comes up, but as it does not reach a frequency near $1/\alpha$ this variant is not sampled during the bottleneck. Only when the bottleneck is wider is the probability of sampling the virus variant with a deletion large enough for this to occur regularly. Once a deletion variant has been sampled during the bottleneck, it rapidly goes to fixation as it has a much higher fitness than the full-length virus. Figure 1 provides a simple illustration of the same principle.

When considering host species jumps using the same TEV-eGFP vector, we show that host switches can radically change evolutionary dynamics (Willemsen, Zwart, and Elena 2017).

Table 2. Parameters for the simulation model.

Model parameter	Value
Passages	10
Generations per passage	10
ω_I	1.6
ω_D	2
Ρ	10^{-6}
Κ	10^{6}

After over half a year of evolution in two semi-permissive host species, with a large difference in virus-induced virulence, the *eGFP* insert appears to remain stable. A fitness costs of *eGFP* was only found in the host for which TEV has low virulence. In the hosts for which TEV has high virulence there was no fitness cost and viral adaptation was observed. This contradicts theories that suggest that high virulence could hinder between-host transmission. When considering the evolution of genome architecture, host species jumps might play a very important role, by allowing evolutionary intermediates to be competitive.
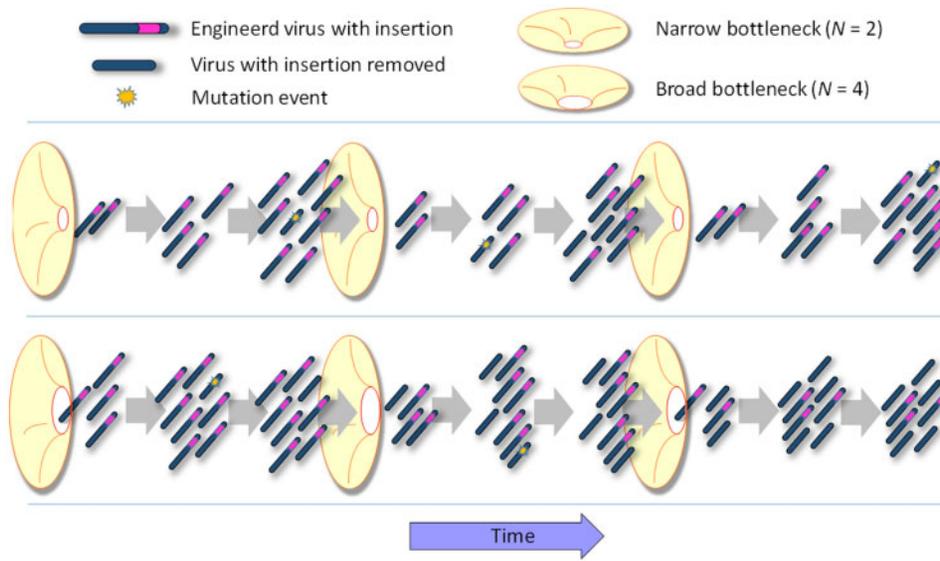

Figure 1. An illustration of how narrow bottlenecks can lead to increase stability of an insertion. A legend for all items in the figure is given on top. The upper sequence of events illustrates serial passage with narrow bottlenecks (N = 2), whereas the lower sequence illustrates passage with a broader bottleneck (N = 4). When bottlenecks are narrow, viral genomes with deletions are less likely to pass through the bottleneck and the population can be 'reset' at the start of every passage. When the bottleneck is broader, viral genomes with deletions are sampled despite the bottleneck and the deletion quickly goes to a high frequency. A more detailed treatment of this issue and a simulation model is presented in Box 1 and Fig. 2.
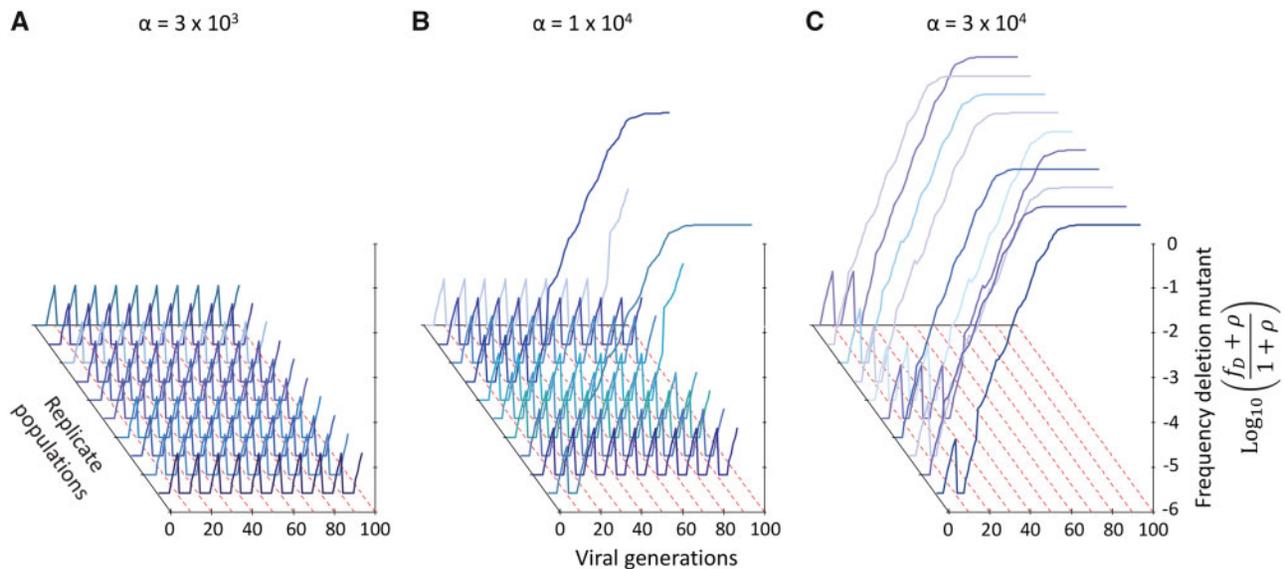

Figure 2. Results generated by the simulation model described in Box 1 are shown, illustrating how demography can affect the stability of inserted sequences. We show how the frequency of a deletion variant (f_D) changes over time when the size of the population bottleneck (α) is changed. For all three panels, viral generations are on the x-axis, the log frequency of the deletion variant which has lost the insert is on the y-axis, and replicate populations are plotted in different randomly chosen shades of blue along the z-axis. We illustrate the bottleneck (i.e. the moment of passaging) with red dotted lines. When the population bottlenecks are narrow (panel A), the frequency of deletion variants is 'reset' during passaging as they are too scarce to be sampled. As the bottleneck is widened (panels B and C), occasionally the deletion variant is sampled.

The stability of an insert could change when considering insertions that might be beneficial for the virus. Using the TEV genome we simulated two HGT events, by separately introducing functional exogenous sequences that are potentially beneficial for the virus (Willemsen et al. 2017). In one case, the insertion was rapidly purged from the viral genome, restoring fitness to wild-type fitness levels. In another case, the inserted gene—the *2b* RNA silencing suppressor from Cucumber mosaic virus—did not seem to have a major impact on viral fitness and was therefore not lost when performing experimental evolution. Interestingly this insertion duplicated the function of RNA silencing suppression function of another gene in the genome. When mutating this functional domain of the TEV gene, the inserted gene provided a replicative advantage. These observations suggest a potentially interesting role for HGT of short functional sequences in improving evolutionary constraints on viruses.

Besides HGT, another mechanism for evolutionary innovation is gene duplication. The effects in the stability on a genetically redundant insert might be variable. On one hand, one would expect the duplicated copy to be rapidly deleted from the genome as it does not confer an additional function. On the other hand, if a duplicated sequence is stable it may act as a stepping stone to the evolution of new biological functions. We have investigated the stability of genetically redundant sequences by generating (TEV) viruses with potentially beneficial gene duplications (Willemsen et al. 2016). All gene duplications resulted in a loss of viability or in a significant reduction in viral fitness. Experimental evolution always led to deletion of the duplicated gene copy and maintenance of the ancestral copy. However, the stability of the different duplicated genes was highly divergent, suggesting that passage duration is not the main factor for determining whether the insert will be stable or unstable. The deletion dynamics of the duplicated genes were associated with the passage duration and the size of the duplicated copy. By developing a mathematical model we showed that the fitness effects alone are not enough to predict genomic stability. A context-dependent recombination rate is also required, with the context being the identity of the insert and its position.

In summary, these experimental observations demonstrate the deleterious nature of gene insertions in ssRNA(+) viruses, where the highly streamlined genomes limit sequence space for the evolution of novel functions, and in turn adaptation to environmental changes.

2.5 Group V: single-stranded RNA(−) viruses

The ssRNA(−) viruses are composed of genomes that range from 10 to 25.2 kbp in size. These viruses are particularly attractive candidates as viral vectors. While in ssRNA(+) viruses inserts are subject to deletion, inserts in their ssRNA(−) counterparts appear more stable (Mebatsion et al.1996; Schnell et al. 1996). One reason for this stability is that in general the genes in the ssRNA(−) viral genomes are non-overlapping and are expressed as separate mRNAs, thus consisting of a modular organization that can be easily manipulated for the insertion of foreign genes. If correctly engineered (e.g. without affecting any regulatory regions), one could expect that gene insertions are more stable in ssRNA(−) viruses as compared to ssRNA(+) viruses, since the complexities surrounding correct processing of a polyprotein are not an issue here. Moreover, if expressed as a separate mRNA, the size of the insert is probably restricted only by the packaging limits of ssRNA(−) viruses. The low rate of homologous recombination in ssRNA(−) viruses can be another explanation for higher genomic stability (Chare, Gould, and Holmes 2003; Han and Worobey 2011). Non-homologous recombination will probably rarely lead to variants with the insert deleted and other regions undisturbed, given it is less constrained than homologous recombination, and hence low homologous recombination rates could be a limiting factor on sequence evolution. However, genomic deletions that disrupt the inserted sequence will be subject to less constraints, as for example they can disrupt the reading frame of the insert without affecting the expression of virus genes.

2.5.1 Engineered viruses

Canine distemper virus (CDV), a species in the *Morbillivirus* genus, is an important pathogen of a variety of animals, including the dog. This virus, however, has shown to be a promising expression vector for the development of vaccines. Although the replicative fitness of a recombinant CDV carrying the rabies virus glycoprotein was slightly lower than the wild-type CDV, the insert was stably expressed during serial passaging *in vitro* and inoculation *in vivo* induced specific neutralizing antibodies against both rabies and CDV (Li et al. 2015).

Similarly, genes expressing foreign antigens can be cloned into recombinant measles virus where measles virus proteins and inserted genes are coexpressed. This relatively small vector can accept large gene insertions, that in most cases are stably expressed (Billeter, Naim, and Udem 2009; Malczyk et al. 2015). For example, for the development of a vaccine against Middle East respiratory syndrome coronavirus (MERS-CoV), it has been shown that a recombinant measles virus expressing the spike glycoprotein of MERS-CoV is genetically stable *in vitro* and induces strong humoral and cellular immunity *in vivo* (Malczyk et al. 2015).

Vesicular stomatitis virus (VSV) is a commonly used vaccine vector that has been engineered to express surface proteins from diverse viruses, including Ebola (Garbutt et al. 2004), human immunodeficiency virus type 1 (HIV-1) (Johnson et al. 1997), and influenza A (Roberts et al. 1999), which can stimulate protective immune responses against these pathogens (Bukreyev et al. 2006). In addition, VSV has shown promise as a candidate for oncolytic virus therapy, as it replicates most efficiently in cells with diminished innate immunity such as cancer cells, which often have impaired production of and/or response to interferon (Barber 2005). Mutations that attenuate VSV growth in healthy immune-competent cells can further enhance the safety of this anti-cancer therapy potential (Barber 2005).

What is particularly interesting about the genome organization of VSV and other ssRNA(−) viruses is that promoter proximal genes are more efficiently expressed than promoter distal ones (Iverson and Rose 1981; Wertz, Perepelitsa, and Ball 1998; Pesko et al. 2015). The efficiency of expression of the inserted gene (and therewith the strength of the immune response) can be controlled (Tokusumi et al. 2002; Roberts et al. 2004). However, inserting a foreign gene close to the promoter also can also reduce the expression of downstream vector genes (Skiadopoulos et al. 2002), which in turn can negatively affect virus transcription and RNA replication (Wertz, Moudy, and Ball 2002; Zhao and Peeters 2003). These empirical observations again show that the site of the insert plays an important role in recombinant vector stability.

When considering the size of the insert and its stability, ssRNA(−) viruses accept relatively large insert without drastically affecting virus replication. Sendai virus, with a genome

size of about 15.3 kbp, can carry and efficiently express gene insertions up to 3.2 kbp (Sakai et al. 1999). However, also here the insert size is limited, where the final virus titers *in vitro* are proportionally reduced as the insert size increases. While *in vivo* no such size-dependent effect was observed, an attenuated replication and pathogenicity were detected (Sakai et al. 1999). Insertions up to 3.9 kbp in the ~15.4 kbp genome of the human parainfluenza virus 3 were viable and replicated efficiently *in vitro* (Skiadopoulos et al. 2000). Nonetheless, the insertions longer than 3,000 bp reduced the robustness to environmental perturbation of the virus, as temperature sensitivity was augmented and replication was restricted to certain sites *in vivo* (Skiadopoulos et al. 2000).

The ssRNA(−) viruses seem promising expression vectors, where one can control gene expression and introduce relatively large inserts that, in many instances, appear to be stable. The constraints imposed on viral gene insertions seem to be the lowest in this group of viruses. Yet, the ideal vector that accepts all types and sizes of foreign gene insertions without decreasing viral replication, has not been identified yet.

2.6 Group VI: retro-transcribing single-stranded RNA(+) viruses

Retro-transcribing ssRNA(+) viruses, or retroviruses, are small viruses varying in genome size from 7 to 11 kbp and are classified in the *Retroviridae* family. After entering a host cell, the retroviral RNA genome is converted into dsDNA by reverse transcription. The viral DNA integrates into the host genome, where viral genes are translated. Therefore, these viruses are often used for gene therapy. Retroviruses frequently undergo genomic rearrangements, including gene insertions and deletions (indels). Moreover, recombination can be common due to the combination of 'diploid' virus particles and high intrinsic recombination rates (Jetzt et al. 2000). Therefore as a general observation this viral group appears to have a highly plastic genome, and should be relatively open to foreign gene insertions. As retroviruses integrate into the host genome, the stability of inserts does not necessarily depend solely on the retrovirus genome configuration and demographic conditions. As host genomes are in general less streamlined than those of viruses, one could expect that gene insertions are stable after integration into the host genome. However, the random integration of retroviruses in the host genome makes it hard to predict genomic stability.

2.6.1 Wild viruses
As a wild example, HIV-1 frequently undergoes genomic rearrangements, where indels are significant source of evolutionary change. These indels appear to have an impact on virus transmission and adaptation as for example indels in the HIV-1 *pol* gene are associated with drug resistance (Rakik et al. 1999), and indels in the *gag* and *vif* genes are associated with disease progression and infectivity (Alexander et al. 2002; Aralaguppe et al. 2017). The HIV-1 surface envelope glycoprotein contains five variable regions (V1–V5) that can tolerate a higher rate of indels than the rest of the genome. Interestingly, indel rate estimates vary significantly among variable regions and subtypes (from different hosts) (Palmer and Poon 2018).

2.6.2 Engineered viruses
When introducing GFP into the five variable regions of HIV-1, certain regions (V4 and V5) were more tolerant to foreign gene insertions than the other variable regions (V1, V2, and V3)

(Nakane, Iwamoto, and Matsuda 2015). In particular, GFP insertions into the V3 region showed lower levels of expression (Nakane, Iwamoto, and Matsuda 2015), which is consistent with V3 having the lowest indel rate (Palmer and Poon 2018), thus having a lower stability after gene insertions. This piece of empirical evidence again shows that the site of insertion plays an important role in determining expression levels and stability.

Retroviruses have a valuable potential as vectors for introducing therapeutic genes into cancer cells. Murine retroviruses are the most commonly used vectors in clinical trials today, and seem promising candidates for human gene therapy as they target dividing cells with a high degree of efficiency and lead to stable gene transfer as they integrate into the chromosomes of the target cell (Edelstein et al. 2004). However, we still have to deal with important safety issues when using retroviruses for gene therapy. The random integration of retroviruses in the host genome poses a risk, as the integration near the LMO2 proto-oncogene promoter can trigger the development of leukemia (Hacein-Bey-Abina et al. 2003).

Besides the risks related to retroviral gene therapy, the limited efficiency of *in vivo* gene transfer poses another obstacle. Replication defective retrovirus vectors are often used in clinical trials but limited since they can only infect a fraction of solid tumor cells (Rainov and Ren 2003). For the delivery of the transgene in all tumor cells, replication-competent retroviral vectors are a promising alternative. The suitability of murine leukemia virus (MLV)-based vectors for cancer gene therapy has been analyzed *in vitro* and *in vivo* by Paar et al. (2007). They found that the choice of the virus strain, the position of the insert, and the host cells used, can influence the replication kinetics, genomic stability, and transgene expression levels (Paar et al. 2007).

Concordantly, the *eGFP* sequence was inserted into MLV under different configurations (i.e. site of insertion and flanking sequence), and the reporter gene was deleted upon extended cell culture (Duch et al. 2004). The stability was improved by decreasing the length of sequence repeats flanking the inserted sequence, however, eventually eGFP was always (partially or completely) deleted (Duch et al. 2004). In another study, transgenes of different sizes (GFP, *hph*, *pac*) were inserted into MLV. Deletions were always observed, where the deletion dynamics depended on the size of the insert and preferred sites of recombination were detected (Logg et al. 2001).

Using retroviral vectors for the expression and transfer of foreign genes is central to the development of gene therapy. An advantage of using retro-transcribing ssRNA(+) viruses is that after reverse transcription a dsDNA molecule stably integrates into the host genome. With careful design, testing, and engineering, the retroviruses are promising vectors for the treatment of diseases, such as cancer.

2.7 Group VII: retro-transcribing dsDNA viruses

The retro-transcribing dsDNA (RT-dsDNA) viruses have small genome sizes varying from 3 to 8.3 kb, and include the viral families *Caulimoviridae* and *Hepadnaviridae*. As the name suggests, the RT-dsDNA viruses replicate through an RNA intermediate, and in some cases the pre-genomic RNA is alternatively spliced. Although genomic rearrangements appear to be frequent in RT-dsDNA viruses, we hypothesize that gene insertions will often be unstable, because 1, they tend to have compact genomes, and 2, insertions can easily disturb a viral regulatory sequence or lead to incorrect processing of the alternative spliced products.

2.7.1 Wild viruses

In contrast to the retroviruses (group VI), the genome replication of the *Caulimoviridae* is entirely episomal. However, fragmented and rearranged endogenous caulimovirus sequences have been found in a wide variety of plant species (Teycheney and Geering 2011). For the *Hepadnaviridae*, the viral genome can be integrated into the host genome, through a process that exploits ds breaks in the host genome. Although this is an infrequent event, the integrated viral DNA often contains deletion, inversions and duplications, often inactivating the virus. In the case of Hepatitis B virus (HBV), integration into the human genome can cause genetic damage and chromosomal instability leading to HBV-induced liver cancer (Shafritz et al.1981; Furuta et al. 2018).

2.7.2 Engineered viruses

Several studies in the 1980s already reported the possibility of inserting foreign DNA into specific sites of the cauliflower mosaic virus (CaMV) genome without greatly affecting viral infectivity or function (Gronenborn et al. 1981; Howell, Walker, and Walden 1981; Dixon, Koenig, and Hohn 1983; Brisson et al. 1984; Lefebvre, Miki, and Laliberté 1987). In two of these studies, functional bacterial genes were introduced into the CaMV genome, where a fragment of the *lac* operator (Gronenborn et al. 1981) and the *DHFR* gene (Brisson et al. 1984) were successfully expressed. In these studies, issues regarding the stability of the insert were raised, where the *lac* operator was lost after five successive transfers and extended growth of the plants, and deletions in the *DHFR* gene started appearing after the second and third transfers. On the contrary, an inserted mammalian *metallothionein* gene appeared to be stable and functional in the CaMV genome (Lefebvre, Miki, and Laliberté 1987). These studies suggest that the differences in stability of inserts in the CaMV genome depend on at least two factors. First, the site of the insert seems to be important as many inserts are lethal for the virus (Gronenborn et al. 1981; Howell, Walker, and Walden 1981; Dixon, Koenig, and Hohn 1983). Second, the size of the insert is important, as CaMV can accept only small foreign genes due to viral encapsidation limits (Gronenborn et al. 1981; Lefebvre, Miki, and Laliberté 1987).

As described along this review, vectors containing the GFP as an insert are often designed to study the infection dynamics of viruses. However, the size of GFP is relatively large (around 700 nt) and often leads to instability of vectors (Zwart et al. 2014; Nakane, Iwamoto, and Matsuda 2015). To cope with the size limitation a split GFP system has been engineered (Cabantous, Terwilliger, and Waldo 2005), where only a small part of GFP is introduced in the viral vector and the other part is expressed using a transgenic host. When the two GFP fragments are together, spontaneous association leads the formation of a fluorescent molecule. In the CaMV genome this system allowed to track a CaMV protein *in vivo* (Dáder et al. 2019). The partial GFP insertion was stable for ten or four serial passages, depending on the host plant species used, suggesting that the demographic conditions such as the host play an important role in stability.

Although the number of studies on insert stability in RT-dsDNA viruses is limited, we reason that several constraints limit insert stability in these viruses. Although small inserts will allow to track viral infection dynamics, the use of RT-dsDNA viruses for gene therapy does not seem practicable as integration into the host genome is a rare event for these viruses.

3. Some conceptual considerations

Sequence loss is inherently an evolutionary process, at a minimum involving mutation and selection, and therefore needs to be framed in an evolutionary context. Here, we consider how theory might help to better understand and ultimately predict this process. First, inspired by empirical results we consider the effects of virus population and bottleneck sizes on sequence loss. Second, we consider whether there are different evolutionary trajectories that lead to a restoration of fitness following insertion of a sequence, and their implications for sequence stability.

3.1 Effects of demography on sequence stability

We understand demography to be a description of the size and structure of virus populations over time. In this discussion we will consider virus populations that are divided into demes at the host or cell level. Theory suggests that demography could have major implications for the loss of inserted sequences, with small population sizes, narrow bottlenecks, and short time intervals between bottlenecks resulting in high sequence stability. Hence, the stability of the inserted sequence cannot be viewed solely as a property of a genome, rather it is a phenotype and therefore depends on the environment. In this section, we motivate this argument and present a simulation model that highlights the effects of demography on the deletion of inserted sequences.

At its core, the stability of genomic insertions in viral genomes depends on two key factors. First, the supply of mutations removing the insertion is crucial, because selection can only act on existing heritable variation. Second, selection then acts to fix variants with the inserted sequence removed. All other things equal, the larger the supply of mutations that remove the insertion and the larger the selection coefficients of variants with the insert removed, the less stable the insertion will be. The interplay between mutation and selection will govern the stability of genomic inserts, and in many cases demography has an important role in shaping this interplay. For example, low fitness can lead to small population sizes, which in turn will limit the mutation supply (Chao 1990; Lynch and Gabriel 1990). A high-cost inserted sequence might therefore limit viral evolvability, thereby promoting its own stability.

Genetic drift can also play an important role in determining the stability of inserted sequences, as inserts can have high stability if a viral population regularly passes through population bottlenecks. This idea is inspired by the empirical observation that a group IV plant virus appears to be stable when short-duration passages are used, but not in long-duration passages (Dolja et al. 1993; Zwart et al. 2014; Willemsen et al. 2016, 2018). Viruses pass through bottlenecks at many points during infection, *in vitro* and *in vivo* (Zwart and Elena 2015), it is therefore important to consider these effects. Even if there is a large supply of deletions and strong selection for the deletion mutant, if deletion mutants fail to reach a frequency $\geq /\alpha$, where α is the bottleneck size, they are unlikely to pass through the bottleneck (Willemsen et al. 2016, 2018). This leads to a 'resetting' of the virus population by each bottleneck event (Fig. 1), effectively resulting in high stability of the inserted sequences (Box 1, see also Fig. 2). Short passages shorten the time for deletion mutants to reach the frequency $1/\alpha$, making it more difficult for these variants to pass through bottlenecks and hereby promoting insert stability. It is important to remember that assays for detecting deletion mutants, such as deep sequencing or the

polymerase chain reaction, do have limited sensitivity. Deletions may therefore also be detected more readily in longer passages, whilst low frequency mutations that will be purged by bottlenecks may not be detected (Bull, Nuismer, and Antia 2019).

Demography can also modulate the strength of selection itself. The MOI (cellular multiplicity of infection) is a key demographic parameter at the cellular level, as it describes the number of virus particles infecting a cell. If an inserted sequence affects viral fitness *in trans* at the within cell level—for example by being toxic—then the MOI will determine whether there can be selection (Miyashita and Kishino 2010). At high MOIs there will be no selection, because the toxin is produced in all cells and affects the replication of both producers and non-producers of the toxin (Fig. 3). An interesting conundrum is that high MOIs also tend to promote the evolution of DI viruses
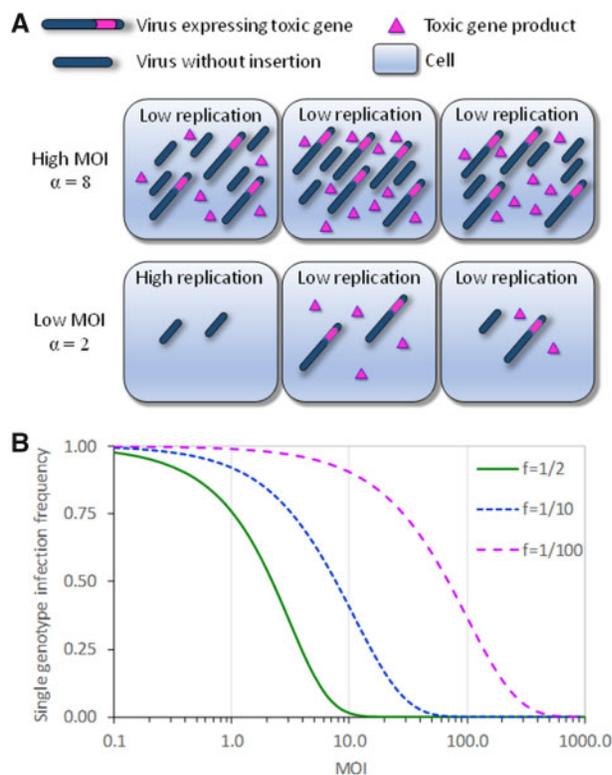

Figure 3. In panel A, we illustrate how the cellular MOI can have a direct effect on selection strength. Consider a virus that expresses a product that is toxic and acts *in trans* within cells to lower replication levels, but deletions can remove the gene coding this gene. If there is a mixed virus population with variants with the insertion intact and deleted, at high MOI all cells will be infected with both variants and the toxin will lower replication. The ubiquitousness of the toxin will limit selection against the virus variant with the deletion. When MOI is low, due to genetic drift at the cellular not all cells will contain both variants, and virus variant with the deletion is selected because those cells infected only with this variant have higher replication. In panel B, the relationship between the cellular MOI (ordinate) and the frequency of single-genotype infection (abscissa) for a virus population with genotypes a and b is given, for different frequencies of the two virus genotypes in the population (f_a shown, $f_b = 1 - f_a$). Note that the frequency of single-genotype infections is given as the proportion of infected cells in which only virus genotypes a or b are present. As the MOI increases, the frequency of single-genotype infections decreases, although it depends on the frequency of the two virus genotypes in the population. If genotype a expresses a gene that has fitness costs that act *in trans* (e.g. toxicity), then selection can only act against this genotype when there is an appreciable number of single-genotype infections.

(Huang 1973) due to within-cell selection, and hence these two effects must be weighed accordingly.

In this review, we considered only a few cases in which inserted sequences potentially could have beneficial effects on a virus (Thomas, Cameron, and Davis 1974; Gladstone, Molineux, and Bull 2012; Schmerer et al. 2014; Willemsen et al. 2017). Beneficial effects could promote insertion stability and are therefore interesting from a bioengineering perspective, but demography can once again play a role in determining sequence stability. Heterologous expression of endosialidase, an exoenzyme that degrades a key biofilm component after phage-induced cell lysis, lead to increased amplification of phage T7 in capsulated *E.coli* (Gladstone, Molineux, and Bull 2012). However, a phage that did not express the dispersin outcompeted the engineered virus, as it could reap the benefits of dispersin production whilst not bearing costs. This tragedy of the commons is a reversal of the situation sketched above for high MOIs (Fig. 3). One proposed strategy to increase stability would be setting up culture conditions such that phages are growing in isolation or spatially structured environments (Gladstone, Molineux, and Bull 2012), other examples of demography-based approaches to increasing insert stability.

It will certainly not always be possible to address issues of insert stability through demographic changes, but theory suggests this can be an interesting approach. Some experimental protocols already exploit some of these principles, in particular strict adherence to low MOI (Fitzgerald et al. 2006). One should caution against naive applications of evolutionary theory, as the details of each real-world system matter (Schmerer et al. 2014).

3.2 Effects of alternative mutational trajectories on sequence loss

There are multiple, non-mutually exclusive mechanisms by which an inserted sequence can be costly for a virus. Consequently, deletion of the inserted sequence may not be the only class of mutation that ameliorates the insert's effects on fitness, a possibility we explore in this section. We argue that alternative trajectories may sometimes play a role, but that due to mutation supply of different types of mutations, deletion of the inserted sequence is the most likely trajectory.

A cost of the insert can arise because of the attributes of the inserted sequence (i.e. metabolic costs of expressing extra genes, toxicity of gene products), reorganization of the genome due to the insertion (i.e. disruption of the regulation of gene expression, polyprotein processing, and subgenomic RNAs), or limitations on genome size imposed by virus-particle packaging. Deletion of the inserted sequence is therefore not the only plausible class of mutation that can restore viral fitness, as other mutations can also affect fitness. These mutation types are 1, regulatory mutations (i.e. promoter mutations) that downregulate gene expression (van Opijnen, Boerlijst, and Berkhout 2006), 2, removal of immunogenic sequence motifs (Fros et al. 2017), 3, alteration of unfavorable secondary RNA structures (McFadden et al. 2013), and 4, adopting a more favorable codon usage (Carrasco, de la Iglesia, and Elena 2007; Agashe et al. 2013; Cladel et al. 2013), as synonymous mutations can have marked effects on virus fitness.

These different mutation classes are likely to have different mutation rates, and mutation bias might therefore drive the evolutionary route that is followed (Stoltzfus and McCandlish 2017). For example, consider that recombination rates are high for many viruses (Tromas and Elena 2010), and there are many

possible recombination events that partially remove an insertion. In contrast, probably only a small fraction of point mutations will be beneficial (Sanjuán, Moya, and Elena 2004; Carrasco, de la Iglesia, and Elena 2007), e.g. in this case by lowering expression of the inserted gene or leading to more favorable codon usage. We therefore conjecture that mutation supply is likely to favor the evolution of deletions in the transgene over beneficial point mutations that affect fitness cost. Consider the 'genomic accordion' observed in poxviruses (Elde et al. 2012): beneficial point mutations typically occur long after gene amplification by copy number variation. Likewise we expect deletions that remove an insertion to be fixed before point mutations that also lessen its impact occur. Nevertheless, the occurrence of alternative evolutionary trajectories could, depending on the exact mutation supply and effect sizes for different classes of mutations, contribute to making stability of genomic inserts less repeatable and predictable in some cases (De Visser and Krug 2014; Bolnick et al. 2018).

4. Concluding remarks

Whereas some sequences inserted into viral genomes are stable, others are clearly not. Although there are some factors that appear to explain these differences, at the end of the day there is still a great deal about the relatively simple question of stability that we do not understand. In contrast, these different outcomes are encouraging, because they suggest that if we understand the process well enough, we can design more stable insertions. For most viruses, strong selective constraints appear to exist against increasing genome size. In natural conditions, this is an impediment for evolutionary innovation by gene duplication or HGT. In laboratory conditions, this is an impediment for expressing a gene of interest by using engineered viral vectors.

When stratifying by viral groups, we observe that the stability of viral genomes partially depends on the nature of the genome. Viral genomes with separately expressed non-overlapping ORFs (group V: ssRNA−) appear to have less constraints imposed on sequence insertions as compared to genomes with genes encoded in one single ORF (group IV: ssRNA+). Although the dsDNA (group I) virus genomes are extremely plastic in natural conditions, this observation is not a good predictor for stability of engineered viral genomes as inserts are generally lost. In the case of ssDNA (group II) viruses, the varying frequency of genomic segments might lead to rapid adaptive responses to inserted sequences. While in the case of segmented dsRNA (group III) viruses, sequence insertions probably perturb segmented genome assembly. When comparing the retro-transcribing viruses, the RT-ssRNA(+) (group VI) viruses appear to successfully express sequences of interest after stable integration into the host genome, whilst the RT-dsDNA (group VII) viruses are less stable and only rarely integrate into the host genome. Multipartite viruses, represented in various groups, also present unique challenges when thinking about the stability of inserted sequences.

When comparing all viral genome architectures, we conclude that that genomic stability is not a fixed, intrinsic property. Although we show that insert stability depends on the nature of the genome, the site and size of the insert and the recombination rate, the host species and demographic conditions (i.e. population and bottleneck size) can radically change viral evolutionary dynamics. We have illustrated this idea with a simple simulation model that considers the effect of genetic bottlenecks (Box 1), where the observed stability of the viral genome decreases as the bottleneck is widened. The interplay between all factors affecting insert stability appears to be complex and unexpectedly sensitive to the exact conditions under which a virus population evolves.

Given these complexities, we think it may be challenging to develop predictive models of insert stability, for different types of virus genomes under different conditions. We hope to see developments in this area, possibly linked to resurging interest in preventing and exploiting DI viruses. However, we think that experimental tests of the stability of viral constructs will remain important in the foreseeable future. Experimental evolution can detect design problems in engineered genomes by looking at fitness and evolutionary stability (Springman et al.). As Springman and collaborators suggest, experimental evolution may also prove useful for optimizing the stability of expression vectors by ameliorating constraints for which solutions are hard to predict because we lack a mechanistic understanding, such as codon usage (Carrasco, de la Iglesia, and Elena 2007; Agashe et al. 2013). This approach can lead to improved engineering of viral genomes, which is also of interest for designing vectors with tags to follow viral infection, and for the use of viral vectors for gene therapy as well as for vaccine vectors. Finally, for real-world applications it can be useful to determine quantitatively the impact of the loss of inserted sequences on the desired output. For example, models suggest that deletions in vector vaccines may not have a large impact on eliciting the desired immune response (Bull, Nuismer, and Antia 2019).

We have noticed that a surprisingly large number of studies draw conclusions on the stability of inserted sequences in viral genomes based on experiments with either no or low replication. We cannot stress enough the importance of replication in studying genomic stability, in the first place because mutation is a stochastic process. Moreover, as illustrated by our simple simulations—in which mutation is deterministic—bottlenecks and population dynamics can also introduce further stochastic effects that influence stability (Fig. 1). Furthermore, empirical studies with high levels of replication show the extent to which observed stability does vary between replicates (Zwart et al. 2014).

References

Agashe, D. et al. (2013) 'Good Codons, Bad Transcript: Large Reductions in Gene Expression and Fitness Arising from Synonymous Mutations in a Key Enzyme', *Molecular Biology and Evolution*, 30: 549–60.

Alexander, L. et al. (2002) 'Inhibition of Human Immunodeficiency Virus Type 1 (HIV-1) Replication by a Two-amino-acid Insertion in HIV-1 Vif from a Nonprogressing Mother and Child', *Journal of Virology*, 76: 10533–9.

Alharbi, N. K. (2019) 'Poxviral Promoters for Improving the Immunogenicity of MVA Delivered Vaccines', *Human Vaccines & Immunotherapeutics*, 15: 203–9.

Andersson, D. I., Slechta, E. S., and Roth, J. R. (1998) 'Evidence That Gene Amplification Underlies Adaptive Mutability of the Bacterial *Lac* Operon', *Science*, 282: 1133–5.

Aralaguppe, S. G. et al. (2017) 'Increased Replication Capacity Following Evolution of PYxE Insertion in Gag-p6 is Associated with Enhanced Virulence in HIV-1 Subtype C from East Africa', *Journal of Medical Virology*, 89: 106–11.

Ayalew, L. E. et al. (2015) 'Bovine Adenovirus-3 as a Vaccine Delivery Vehicle', *Vaccine*, 33: 493–9.

Barber, G. N. (2005) 'VSV-tumor Selective Replication and Protein Translation', *Oncogene*, 24: 7710–9.

Belshaw, R., Pybus, O. G., and Rambaut, A. (2007) 'The Evolution of Genome Compression and Genomic Novelty in RNA Viruses', *Genome Research*, 17: 1496–504.

Billeter, M. A., Naim, H. Y., and Udem, S. A. (2009) 'Reverse Genetics of Measles Virus and Resulting Multivalent Recombinant Vaccines: Applications of Recombinant Measles Viruses', *Current Topics in Microbiology and Immunology*, 329: 129–62.

Bolnick, D. I. et al. (2018) '(Non)parallel Evolution', *Annual Review of Ecology, Evolution, and Systematics*, 49: 303–30.

Borodavka, A., Desselberger, U., and Patton, J. T. (2018) 'Genome Packaging in Multi-segmented dsRNA Viruses: Distinct Mechanisms with Similar Outcomes', *Current Opinion in Virology*, 33: 106–12.

Brisson, N. et al. (1984) 'Expression of a Bacterial Gene in Plants by Using a Viral Vector', *Nature*, 310: 511–4.

Bukreyev, A. et al. (2006) 'Nonsegmented Negative-strand Viruses as Vaccine Vectors', *Journal of Virology*, 80: 10293–306.

Bull, J. J., Nuismer, S. L., and Antia, R. (2019) 'Recombinant Vector Vaccine Evolution', *PLoS Computational Biology*, 15: e1006857.

Cabantous, S., Terwilliger, T. C., and Waldo, G. S. (2005) 'Protein Tagging and Detection with Engineered Self-assembling Fragments of Green Fluorescent Protein', *Nature Biotechnology*, 23: 102–7.

Carrasco, P., de la Iglesia, F., and Elena, S. F. (2007) 'Distribution of Fitness and Virulence Effects Caused by Single-nucleotide Substitutions in Tobacco Etch Virus', *Journal of Virology*, 81: 12979–84.

Chambers, A. C. et al. (2018) 'Overview of the Baculovirus Expression System', *Current Protocols in Protein Science*, 91: 5.4.1–6.

Chao, L. (1990) 'Fitness of RNA Virus Decreased by Muller's Ratchet', *Nature*, 348: 454–5.

Chapman, S., Kavanagh, T., and Baulcombe, D. (1992) 'Potato Virus X as a Vector for Gene Expression in Plants', *The Plant Journal*, 2: 549–57.

Chauthaiwale, V. M., Therwath, A., and Deshpande, V. V. (1992) 'Bacteriophage Lambda as a Cloning Vector', *Microbiological Reviews*, 56: 577–91.

Chung, B. N., Canto, T., and Palukaitis, P. (2007) 'Stability of Recombinant Plant Viruses Containing Genes of Unrelated Plant Viruses', *Journal of General Virology*, 88: 1347–55.

Cladel, N. M. et al. (2013) 'Synonymous Codon Changes in the Oncogenes of the Cottontail Rabbit Papillomavirus Lead to Increased Oncogenicity and Immunogenicity of the Virus', *Virology*, 438: 70–83.

Dáder, B. et al. (2019) 'Split Green Fluorescent Protein as a Tool to Study Infection with a Plant Pathogen, Cauliflower Mosaic Virus (V Citovsky, ed.)', *PLoS One*, 14: e0213087.

Dawson, W. O., Bubrick, P., and Grantham, G. L. (1988) 'Modifications of the Tobacco Mosaic Virus Coat Protein Gene Affecting Replication, Movement, and Symptomatology', *Phytopathology*, 78: 783–9.

—— et al. (1989) 'A Tobacco Mosaic Virus-hybrid Expresses and Loses an Added Gene', *Virology*, 172: 285–92.

De Gooijer, C. D. et al. (1992) 'A Structured Dynamic Model for the Baculovirus Infection Process in Insect-cell Reactor Configurations', *Biotechnology and Bioengineering*, 40: 537–48.

De Visser, J., and Krug, J. (2014) 'Empirical Fitness Landscapes and the Predictability of Evolution', *Nature Reviews Genetics*, 15: 480–90.

Dickmeis, C., Fischer, R., and Commandeur, U. (2014) 'Potato Virus X-based Expression Vectors are Stabilized for Long-term Production of Proteins and Larger Inserts', *Biotechnology Journal*, 9: 1369–79.

Dixon, L. K., Koenig, I., and Hohn, T. (1983) 'Mutagenesis of Cauliflower Mosaic Virus', *Gene*, 25: 189–99.

Dolja, V. V., McBride, H. J., and Carrington, J. C. (1992) 'Tagging of Plant Potyvirus Replication and Movement by Insertion of Beta-glucuronidase into the Viral Polyprotein', *Proceedings of the National Academy of Sciences*, 89: 10208–12.

—— et al. (1993) 'Spontaneous Mutagenesis of a Plant Potyvirus Genome After Insertion of a Foreign Gene', *Journal of Virology*, 67: 5968–75.

Donson, J. et al. (1991) 'Systemic Expression of a Bacterial Gene by a Tobacco Mosaic Virus-based Vector', *Proceedings of the National Academy of Sciences of the United States of America*, 88: 7204–8.

Duch, M. et al. (2004) 'Transgene Stability for Three Replication-competent Murine Leukemia Virus Vectors', *Gene*, 329: 61–9.

Edelstein, M. L. et al. (2004) 'Gene Therapy Clinical Trials Worldwide 1989–2004—An Overview', *The Journal of Gene Medicine*, 6: 597–602.

Elde, N. C. et al. (2012) 'Poxviruses Deploy Genomic Accordions to Adapt Rapidly Against Host Antiviral Defenses', *Cell*, 150: 831–41.

Etessami, P., Watts, J., and Stanley, J. (1989) 'Size Reversion of African Cassava Mosaic Virus Coat Protein Gene Deletion Mutants During Infection of *Nicotiana benthamiana*', *Journal of General Virology*, 70(Pt 2): 277–89.

Fitzgerald, D. J. et al. (2006) 'Protein Complex Expression by Using Multigene Baculoviral Vectors', *Nature Methods*, 3: 1021–32.

French, R., Janda, M., and Ahlquist, P. (1986) 'Bacterial Gene Inserted in an Engineered RNA Virus: Efficient Expression in Monocotyledonous Plant Cells', *Science*, 231: 1294–7.

Frischmuth, T., Ringel, M., and Kocher, C. (2001) 'The Size of Encapsidated Single-stranded DNA Determines the Multiplicity of African Cassava Mosaic Virus Particles', *Journal of General Virology*, 82: 673–6.

Fros, J. J. et al. (2017) 'CpG and UpA Dinucleotides in Both Coding and Non-coding Regions of Echovirus 7 Inhibit Replication Initiation Post-entry', *eLife*, 6: e29112.

Furuta, M. et al. (2018) 'Correction: Characterization of HBV Integration Patterns and Timing in Liver Cancer and HBV-infected Livers', *Oncotarget*, 9: 25075–88.

Garbutt, M. et al. (2004) 'Properties of Replication-competent Vesicular Stomatitis Virus Vectors Expressing Glycoproteins of Filoviruses and Arenaviruses', *Journal of Virology*, 78: 5458–65.

Gilbert, S. C. (2013) 'Clinical Development of Modified Vaccinia Virus Ankara Vaccines', *Vaccine*, 31: 4241–6.

Gladstone, E. G., Molineux, I. J., and Bull, J. J. (2012) 'Evolutionary Principles and Synthetic Biology: Avoiding a Molecular Tragedy of the Commons with an Engineered Phage', *Journal of Biological Engineering*, 6: 13.

Gómez, C. E. et al. (2012) 'Poxvirus Vectors as HIV/AIDS Vaccines in Humans', *Human Vaccines & Immunotherapeutics*, 8: 1192–207.

Gronenborn, B. et al. (1981) 'Propagation of Foreign DNA in Plants Using Cauliflower Mosaic Virus as Vector', *Nature*, 294: 773–6.

Guo, H. S., López-Moya, J. J., and García, J. A. (1998) 'Susceptibility to Recombination Rearrangements of a Chimeric Plum Pox Potyvirus Genome After Insertion of a Foreign Gene', *Virus Research*, 57: 183–95.

Hacein-Bey-Abina, S. et al. (2003) 'LMO2-Associated Clonal T Cell Proliferation in Two Patients After Gene Therapy for SCID-X1', *Science*, 302: 415–9.

Han, G. Z., and Worobey, M. (2011) 'Homologous Recombination in Negative Sense RNA Viruses', *Viruses*, 3: 1358–73.

Hayes, R. J. et al. (1988) 'Gene Amplification and Expression in Plants by a Replicating Geminivirus Vector', *Nature*, 334: 179–82.

——, Coutts, R. H. A., and Buck, K. W. (1989) 'Stability and Expression of Bacterial Genes in Replicating Geminivirus Vectors in Plants', *Nucleic Acids Research*, 17: 2391–403.

Howell, S. H., Walker, L. L., and Walden, R. M. (1981) 'Rescue of In Vitro Generated Mutants of Cloned Cauliflower Mosaic Virus Genome in Infected Plants', *Nature*, 293: 483–6.

Huang, A. S. (1973) 'Defective Interfering Viruses', *Annual Review of Microbiology*, 27: 101–18.

Iverson, L. E., and Rose, J. K. (1981) 'Localized Attenuation and Discontinuous Synthesis During Vesicular Stomatitis Virus Transcription', *Cell*, 23: 477–84.

Jetzt, A. E. et al. (2000) 'High Rate of Recombination Throughout the Human Immunodeficiency Virus Type 1 Genome', *Journal of Virology*, 74: 1234–40.

Johnson, J. E. et al. (1997) 'Specific Targeting to CD4+ Cells of Recombinant Vesicular Stomatitis Viruses Encoding Human Immunodeficiency Virus Envelope Proteins', *Journal of Virology*, 71: 5060–8.

Joshi, R. L., Joshi, V., and Ow, D. W. (1990) 'BSMV Genome Mediated Expression of a Foreign Gene in Dicot and Monocot Plant Cells', *The EMBO Journal*, 9: 2663–9.

Kanai, Y. et al. (2017) 'Entirely Plasmid-based Reverse Genetics System for Rotaviruses', *Proceedings of the National Academy of Sciences of the United States of America*, 114: 2349–54.

—— et al. (2019) 'Development of Stable Rotavirus Reporter Expression Systems', *Journal of Virology*, 93: e01774–18.

Knowles, M. K. et al. (2009) 'In Vitro and In Vivo Genetic Stability Studies of a Human Adenovirus Type 5 Recombinant Rabies Glycoprotein Vaccine (ONRAB)', *Vaccine*, 27: 2662–8.

Komoto, S. et al. (2017) 'Reverse Genetics System Demonstrates That Rotavirus Nonstructural Protein NSP6 is Not Essential for Viral Replication in Cell Culture', *Journal of Virology*, 91: e00695–17.

Kool, M. et al. (1991) 'Detection and Analysis of Autographa californica Nuclear Polyhedrosis Virus Mutants with Defective Interfering Properties', *Virology*, 183: 739–46.

Koonin, E. V., Dolja, V. V., and Morris, T. J. (1993) 'Evolution and Taxonomy of Positive-strand RNA Viruses: Implications of Comparative Analysis of Amino Acid Sequences', *Critical Reviews in Biochemistry and Molecular Biology*, 28: 375–430.

——, and Yutin, N. (2018) 'Multiple Evolutionary Origins of Giant Viruses', *F1000Research*, 7: 1840.

Lafay, F. et al. (1994) 'Vaccination Against Rabies: Construction and Characterization of SAG2, a Double Avirulent Derivative of SADBern', *Vaccine*, 12: 317–20.

Lefebvre, D. D., Miki, B. L., and Laliberté, J.-F. (1987) 'Mammalian Metallothionein Functions in Plants', *Nature Biotechnology*, 5: 1053–6.

Li, et al. (2015) 'A Recombinant Canine Distemper Virus Expressing a Modified Rabies Virus Glycoprotein Induces Immune Responses in Mice', *Virus Genes*, 50: 434–41.

Li, Zhijun et al. (2015) 'Rescue and Evaluation of a Recombinant PRRSV Expressing Porcine Interleukin-4', *Virology Journal*, 12: 185.

Logg, C. R. et al. (2001) 'Genomic Stability of Murine Leukemia Viruses Containing Insertions at the Env 3' Untranslated Region Boundary', *Journal of Virology*, 75: 6989–98.

Lu, T. K., and Collins, J. J. (2007) 'Dispersing Biofilms with Engineered Enzymatic Bacteriophage', *Proceedings of the National Academy of Sciences of the United States of America*, 104: 11197–202.

Lynch, M., and Gabriel, W. (1990) 'Mutation Load and the Survival of Small Populations', *Evolution*, 44: 1725–37.

MacInnes, C. D. et al. (2001) 'Elimination of Rabies from Red Foxes in Eastern Ontario', *Journal of Wildlife Diseases*, 37: 119–32.

Majer, E., Daròs, J.-A., and Zwart, M. (2013) 'Stability and Fitness Impact of the Visually Discernible Rosea1 Marker in the Tobacco Etch Virus Genome', *Viruses*, 5: 2153–68.

Malczyk, A. H. et al. (2015) 'A Highly Immunogenic and Protective Middle East Respiratory Syndrome Coronavirus Vaccine Based on a Recombinant Measles Virus Vaccine Platform', *Journal of Virology*, 89: 11654–67.

Martin, D. P. et al. (2011) 'Recombination in Eukaryotic Single Stranded DNA Viruses', *Viruses*, 3: 1699–738.

Martínez-Álvarez, L. et al. (2013) 'The Shift from Low to High Non-structural Protein 1 Expression in Rotavirus-infected MA-104 Cells', *Memórias do Instituto Oswaldo Cruz*, 108: 421–8.

Mattion, N. M. et al. (1994) 'Attenuated Poliovirus Strain as a Live Vector: Expression of Regions of Rotavirus Outer Capsid Protein VP7 by Using Recombinant Sabin 3 Viruses', *Journal of Virology*, 68: 3925–33.

McFadden, N. et al. (2013) 'Influence of Genome-scale RNA Structure Disruption on the Replication of Murine Norovirus—Similar Replication Kinetics in Cell Culture But Attenuation of Viral Fitness In Vivo', *Nucleic Acids Research*, 41: 6316–31.

Mebatsion, T. et al. (1996) 'Highly Stable Expression of a Foreign Gene from Rabies Virus Vectors', *Proceedings of the National Academy of Sciences*, 93: 7310–14.

Milone, M. C., and O'Doherty, U. (2018) 'Clinical Use of Lentiviral Vectors', *Leukemia*, 32: 1529–41.

Miyashita, S., and Kishino, H. (2010) 'Estimation of the Size of Genetic Bottlenecks in Cell-to-cell Movement of Soil-borne Wheat Mosaic Virus and the Possible Role of the Bottlenecks in Speeding Up Selection of Variations in Trans-acting Genes or Elements', *Journal of Virology*, 84: 1828–37.

Moss, B., Winters, E., and Cooper, N. (1981) 'Instability and Reiteration of DNA Sequences Within the Vaccinia Virus Genome', *Proceedings of the National Academy of Sciences*, 78: 1614–8.

Nakane, S., Iwamoto, A., and Matsuda, Z. (2015) 'The V4 and V5 Variable Loops of HIV-1 Envelope Glycoprotein are Tolerant to Insertion of Green Fluorescent Protein and are Useful Targets for Labeling', *Journal of Biological Chemistry*, 290: 15279–91.

Paar, M. et al. (2007) 'Effects of Viral Strain, Transgene Position, and Target Cell Type on Replication Kinetics, Genomic Stability, and Transgene Expression of Replication-competent Murine Leukemia Virus-based Vectors', *Journal of Virology*, 81: 6973–83.

Palmer, J., and Poon, A. F. (2018) 'Phylogenetic Measures of Indel Rate Variation among the HIV-1 Group M Subtypes', *Virus Evolution*, 5: vez022.

Patil, B. L. et al. (2007) 'Deletion and Recombination Events Between the DNA-A and DNA-B Components of Indian Cassava-infecting Geminiviruses Generate Defective Molecules in Nicotiana benthamiana', *Virus Research*, 124: 59–67.

Pesko, K. et al. (2015) 'Genome Rearrangement Affects RNA Virus Adaptability on Prostate Cancer Cells', *Frontiers in Genetics*, 6: 121.

Philip, A. A. et al. (2019) 'Generation of Recombinant Rotavirus Expressing NSP3-UnaG Fusion Protein by a Simplified Reverse Genetics System', *Journal of Virology*, doi: 10.1128/JVI.01616-19.

Pierson, T. C. et al. (2005) 'An Infectious West Nile Virus That Expresses a GFP Reporter Gene', *Virology*, 334: 28–40.

Pijlman, G. P. et al. (2001) '*Autographa californica* Baculoviruses with Large Genomic Deletions are Rapidly Generated in Infected Insect Cells', *Virology*, 283: 132–8.

——, van Schinjndel, J. E., and Vlak, J. M. (2003) 'Spontaneous Excision of BAC Vector Sequences from Bacmid-derived Baculovirus Expression Vectors upon Passage in Insect Cells', *Journal of General Virology*, 84: 2669–78.

—— et al. (2004) 'Evaluation of Baculovirus Expression Vectors with Enhanced Stability in Continuous Cascaded Insect-cell Bioreactors', *Biotechnology and Bioengineering*, 87: 743–53.

Pugachev, K. V. et al. (1995) 'Double-subgenomic Sindbis Virus Recombinants Expressing Immunogenic Proteins of Japanese Encephalitis Virus Induce Significant Protection in Mice Against Lethal JEV Infection', *Virology*, 212: 587–94.

Rabindran, S., and Dawson, W. O. (2001) 'Assessment of Recombinants That Arise from the Use of a TMV-based Transient Expression Vector', *Virology*, 284: 182–9.

Rainov, N. G., and Ren, H. (2003) 'Clinical Trials with Retrovirus Mediated Gene Therapy: What Have we Learned?', *Journal of Neurooncology*, 65: 227–36.

Rakik, A. et al. (1999) 'A Novel Genotype Encoding a Single Amino Acid Insertion and Five Other Substitutions Between Residues 64 and 74 of the HIV-1 Reverse Transcriptase Confers High-level Cross-resistance to Nucleoside Reverse Transcriptase Inhibitors', *Journal of Acquired Immune Deficiency Syndromes*, 22: 139–45.

Ren, J. et al. (2018) 'Generation of Infectious Clone of Bovine Adenovirus Type I Expressing a Visible Marker Gene', *Journal of Virological Methods*, 261: 139–46.

Roberts, A. et al. (1999) 'Attenuated Vesicular Stomatitis Viruses as Vaccine Vectors', *Journal of Virology*, 73: 3723–32.

—— et al. (2004) 'Complete Protection from Papillomavirus Challenge After a Single Vaccination with a Vesicular Stomatitis Virus Vector Expressing High Levels of L1 Protein', *Journal of Virology*, 78: 3196–9.

Ruggli, N., and Rice, C. M. (1999) 'Functional cDNA Clones of the Flaviviridae: Strategies and Applications', *Advances in Virus Research*, 53: 183–207.

Russell, P. W., and Muller, U. (1984) 'Construction of Bacteriophage phiX174 Mutants with Maximum Genome Sizes', *Journal of Virology*, 52: 822–7.

Sabin, A. B. (1957) 'Properties and Behavior of Orally Administered Attenuated Poliovirus Vaccine', *Journal of the American Medical Association*, 164: 1216–23.

Sakai, Y. et al. (1999) 'Accommodation of Foreign Genes into the Sendai Virus Genome: Sizes of Inserted Genes and Viral Replication', *FEBS Letters*, 456: 221–6.

Sanjuán, R., Moya, A., and Elena, S. F. (2004) 'The Distribution of Fitness Effects Caused by Single-nucleotide Substitutions in an RNA Virus', *Proceedings of the National Academy of Sciences*, 101: 8396–401.

Schmerer, M. et al. (2014) 'Challenges in Predicting the Evolutionary Maintenance of a Phage Transgene', *Journal of Biological Engineering*, 8: 21.

Schnell, M. J. et al. (1996) 'The Minimal Conserved Transcription Stop-start Signal Promotes Stable Expression of a Foreign Gene in Vesicular Stomatitis Virus', *Journal of Virology*, 70: 2318–23.

Scholthof, H. B., Scholthof, K.-B., and Jackson, A. O. (1996) 'Plant Virus Gene Vectors for Transient Expression of Foreign Proteins in Plants', *Annual Review of Phytopathology*, 34: 299–323.

Shafritz, D. A. et al. (1981) 'Integration of Hepatitis B Virus DNA into the Genome of Liver Cells in Chronic Liver Disease and Hepatocellular Carcinoma', *New England Journal of Medicine*, 305: 1067–73.

Shen, W. H., and Hohn, B. (1991) 'Mutational Analysis of the Small Intergenic Region of Maize Streak Virus', *Virology*, 183: 721–30.

Simon-Loriere, E., Holmes, E. C., and Pagán, I. (2013) 'The Effect of Gene Overlapping on the Rate of RNA Virus Evolution', *Molecular Biology and Evolution*, 8: 1916–28.

Skiadopoulos, M. H. et al. (2000) 'Long Nucleotide Insertions Between the HN and L Protein Coding Regions of Human Parainfluenza Virus Type 3 Yield Viruses with Temperature-sensitive and Attenuation Phenotypes', *Virology*, 272: 225–34.

—— et al. (2002) 'Evaluation of the Replication and Immunogenicity of Recombinant Human Parainfluenza Virus Type 3 Vectors Expressing up to Three Foreign Glycoproteins', *Virology*, 297: 136–52.

Smith, G. A., and Enquist, L. W. (1999) 'Construction and Transposon Mutagenesis in *Escherichia coli* of a Full-length Infectious Clone of Pseudorabies Virus, an Alphaherpesvirus', *Journal of Virology*, 73: 6405–14.

——, and —— (2000) 'A Self-recombining Bacterial Artificial Chromosome and Its Application for Analysis of Herpesvirus Pathogenesis', *Proceedings of the National Academy of Sciences*, 97: 4873–8.

Springman, R. et al. (2012) 'Evolutionary Stability of a Refactored Phage Genome', *ACS Synthetic Biology*, 1: 425–30.

Stanley, D. A. et al. (2014) 'Chimpanzee Adenovirus Vaccine Generates Acute and Durable Protective Immunity Against Ebolavirus Challenge', *Nature Medicine*, 20: 1126–9.

Stanley, J. et al. (1997) 'Novel Defective Interfering DNAs Associated with Ageratum Yellow Vein Geminivirus Infection of *Ageratum conyzoides*', *Virology*, 239: 87–96.

Stenger, D. C. et al. (1992) 'A Number of Subgenomic DNAs are Produced following Agroinoculation of Plants with Beet Curly Top Virus', *Journal of General Virology*, 73: 237–42.

Stoltzfus, A., and McCandlish, D. M. (2017) 'Mutational Biases Influence Parallel Adaptation', *Molecular Biology and Evolution*, 34: 2163–72.

Sutter, G., and Staib, C. (2003) 'Vaccinia Vectors as Candidate Vaccines: The Development of Modified Vaccinia Virus Ankara for Antigen Delivery', *Current Drug Target Infectious Disorders*, 3: 263–71.

Takamatsu, N. et al. (1987) 'Expression of Bacterial Chloramphenicol Acetyltransferase Gene in Tobacco Plants Mediated by TMV-RNA', *The EMBO Journal*, 6: 307–11.

Tamura, T. et al. (2018) 'Characterization of Recombinant Flaviviridae Viruses Possessing a Small Reporter Tag', *Journal of Virology*, 92: e01582–17.

Teycheney, P.-Y., and Geering, A. D. W. (2011) 'Endogenous Viral Sequences in Plant Genomes', in Caranta, C., Aranda, M. A., Tepfer, M., and Lopez-Moya, J. J. (eds) *Recent Advances in Plant Virology,* pp. 343–62. Norfolk, UK: Caister Academic Press.

Thomas, M., Cameron, J. R., and Davis, R. W. (1974) 'Viable Molecular Hybrids of Bacteriophage Lambda and Eukaryotic DNA', *Proceedings of the National Academy of Sciences*, 71: 4579–83.

Tokusumi, T. et al. (2002) 'Recombinant Sendai Viruses Expressing Different Levels of a Foreign Reporter Gene', *Virus Research*, 86: 33–8.

Tromas, N., and Elena, S. F. (2010) 'The Rate and Spectrum of Spontaneous Mutations in a Plant RNA Virus', *Genetics*, 185: 983–9.

Troupin, C. et al. (2011) 'Rotavirus Rearranged Genomic RNA Segments are Preferentially Packaged into Viruses Despite Not Conferring Selective Growth Advantage to Viruses (DX Liu, ed.)', *PLoS One*, 6: e20080.

Tur, M. K. et al. (2001) 'Selection of scFv Phages on Intact Cells Under Low pH Conditions Leads to a Significant Loss of Insert-free Phages', *Biotechniques*, 30: 404–13.

van Opijnen, T., Boerlijst, M. C., and Berkhout, B. (2006) 'Effects of Random Mutations in the Human Immunodeficiency Virus Type 1 Transcriptional Promoter on Viral Fitness in Different Host Cell Environments', *Journal of Virology*, 80: 6678–85.

Ward, A., Etessami, P., and Stanley, J. (1988) 'Expression of a Bacterial Gene in Plants Mediated by Infectious Geminivirus DNA', *The EMBO Journal*, 7: 1583–7.

Wertz, G. W., Moudy, R., and Ball, L. A. (2002) 'Adding Genes to the RNA Genome of Vesicular Stomatitis Virus: Positional Effects on Stability of Expression', *Journal of Virology*, 76: 7642–50.

——, Perepelitsa, V. P., and —— (1998) 'Gene Rearrangement Attenuates Expression and Lethality of a Nonsegmented Negative Strand RNA Virus', *Proceedings of the National Academy of Sciences of the United States of America*, 95: 3501–6.

Willemsen, A. et al. (2016) 'Predicting the Stability of Homologous Gene Duplications in a Plant RNA Virus', *Genome Biology and Evolution*, 8: 3065–82.

—— et al. (2017) '2b or Not 2b: Experimental Evolution of Functional Exogenous Sequences in a Plant RNA Virus', *Genome Biology Evolution*, 9: 297–310.

——, Zwart, M. P., and Elena, S. F. (2017) 'High Virulence Does Not Necessarily Impede Viral Adaptation to a New Host: A Case Study Using a Plant RNA Virus', *BMC Evolutionary Biology*, 17: 25.

—— et al. (2018) 'Going, Going, Gone: Predicting the Fate of Genomic Insertions in Plant RNA Viruses', *Heredity*, 121: 499–509.

Williams, W., and Müller, U. (1987) 'Effects of Palindrome Size and Sequence on Genetic Stability in the Bacteriophage ϕX174 Genome', *Journal of Molecular Biology*, 196: 743–55.

Wu, C.-H. et al. (2016) 'Advancement and Applications of Peptide Phage Display Technology in Biomedical Science', *Journal of Biomedical Science*, 23: 8.

Yu, L. et al. (2014) 'Construction and *In Vitro* Evaluation of a Recombinant Live Attenuated PRRSV Expressing GM-CSF', *Virology Journal*, 11: 201.

Zhao, H., and Peeters, B. P. H. (2003) 'Recombinant Newcastle Disease Virus as a Viral Vector: Effect of Genomic Location of Foreign Gene on Gene Expression and Virus Replication', *Journal of General Virology*, 84: 781–8.

Zwart, M. P. et al. (2014) 'Experimental Evolution of Pseudogenization and Gene Loss in a Plant RNA Virus', *Molecular Biology and Evolution*, 31: 121–34.

——, and Elena, S. F. (2015) 'Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution', *Annual Review of Virology*, 2: 161–79.