Experimental evolution of genome architecture and complexity in an RNA virus

Anouk Willemsen

Programa Oficial de Postgrado en Biotecnología

Directores:

Prof. Santiago F. Elena Fito

Dr. Mark P. Zwart

ibmcp
Instituto de Biología Molecular y Celular de Plantas

CSIC
Consejo Superior de Investigaciones Científicas

Universitat Politècnica de València

MINISTERIO
DE ECONOMÍA
Y COMPETITIVIDAD

CSIC
Consejo Superior de Investigaciones Científicas
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
INSTITUTO DE BIOLOGIA MOLECULAR Y CELULAR DE PLANTAS

Santiago F. Elena Fito, doctor en Ciencias Biológicas y Profesor de Investigación del Consejo Superior de Investigaciones Científicas (CSIC) en el Instituto de Biología Molecular y Celular de Plantas *Primo Yúfera* (IBMCP), centro mixto del CSIC y de la Universidad Politécnica de Valencia y **Mark P. Zwart**, Investigador Doctor del Instituto de Física Teórica de la Universidad de Colonia (Colonia, Alemania).

CERTIFICAN:

Que Doña **Anouk Willemsen**, máster en Biología Molecular, Celular y Genética por la Universitat de València, ha realizado bajo su supervisión la Tesis Doctoral titulada "*Experimental Evolution of Genome Architecture and Complexity in an RNA Virus*".

Para que así conste, en cumplimiento de la legislación vigente, firman el presente certificado en Valencia a 8 de Marzo de 2016.

Fdo. Santiago F. Elena Fito

Fdo. Mark P. Zwart

Prof. Santiago F. Elena, PhD
santiago.elena@csic.es
http://bioxeon.ibmcp.upv.es/EvolSysVir
Twitter: @SFElenaLab

Campus UPV, CPI 8E
Ingeniero Fausto Elío, s/n
46022 València
Tel: 963 877 895
Fax: 963 877 859

Table of Contents

Acknowledgements

A PhD is no one-man job. Therefore, I would like to express my gratitude to all those that have made this thesis possible. First of all, I would like to thank both of my directors Santiago an Mark for guiding me throughout my PhD project. Santiago, thank you for receiving me in your lab, for your continuous support, your advices, and for giving me the independence to work on this very exciting project. Mark, thank you for your patience, support, enthusiasm, and for valuable input. I very much appreciate all the time that you dedicated transmitting to me your knowledge on every single aspect: from the laboratory, to statistics, R programming, critical thinking, and scientific writing. Then, there are two persons that deserve a special thank you: Paqui and Paula. Thank you for giving me excellent technical support and helping me out with the tremendous amount of work in the lab. I would also like to thank Nicolas, José Luis, Silvia, and José-Antonio for their collaboration and valuable advices on this project, as well as Julia, Susana and José Manuel, for their interest and support. Outside the lab there are also several persons I want to thank. My parents and grandparents, thank you for giving me the opportunity to go to University, which was not taken for granted back in their time, and respecting me in my decision to continue my studies abroad. Nacho, thank you for recommending me for this PhD position and Luis David for convincing me. Manuel, thank you for the reviving hikes in the mountains. And last but not least, Alejandro, thank you for always being there when I needed to practice a presentation, for the bioinformatic skills you taught me, for your advises, for being my friend and partner, and for going through this important and sometimes stressful period with me.

Abstract

The evolution of genome architecture – the dimensions and organization of an organism's hereditary material – is poorly understood. There are clear differences in genome architecture over organisms, and it is clear that this complexity can be altered on relatively short time scales. Here, three specific processes of the evolution of genome architecture in viruses are studied: (*i*) the reshuffling of existing elements, (*ii*) the decrease of genome complexity through loss of redundant or unnecessary genetic material, (*iii*) and the increase of genome complexity through the acquisition of new genes. These topics were addressed *in vivo* using *Tobacco etch virus*, a plant RNA virus, as a model organism. Important changes in the viral genome were generated – from changes in gene order, to duplications of existing genes, to the introduction of exogenous sequences – followed by experimental evolution to observe how these changes were accommodated. The evolved and ancestral lineages were compared by next-generation sequencing and measurements of virulence, viral accumulation and within-host competitive fitness. First, we identified multiple barriers to the evolution of alternative gene orders. Second, we observed differences in the deletion dynamics of genetically and functionally redundant sequences and we developed a model to predict the stability of gene insertions. Third, we found an exogenous sequence that was evolutionary stable in the *Tobacco etch virus* genome that does not appear to affect viral fitness and can act as a backup in case of failure of the viral host defense mechanism. Lastly, we observed that a host species jump can be a game changer for evolutionary dynamics, allowing unstable viruses to be competitive in alternative hosts.

General Introduction

1. Introduction to viruses

1.1. Discovery of viruses

The first discovery of a pathogenic agent, smaller than any known bacteria, was done in 1892 by the Russian scientist Dimitrii Ivanovsky. It was the causative agent of tobacco mosaic disease in plants, that was not retained by the filters used to remove bacteria from extracts and culture media. However, the concept of this pathogen being a distinctive agent, was introduced six years later by Martinus Beijerinck. In the same year (1898), Koch, Friedrich, Loeffler and Frosch studied the causative agent of foot-and-mouth disease and discovered that these infectious filterable agents consisted of small particles. These particles, that passed through filters that retain bacteria, were eventually called viruses, the Latin word for poison.

1.2. Characteristics of viruses

One characteristic that defines a virus is their absolute dependence on a living host for reproduction. A virus cannot replicate on its own, and needs a host cell to be able to reproduce, evolve and metabolize. Therefore, viruses are often defined as infectious, obligate intracellular parasites. However this definition is incomplete as many bacteria and viriods follow this definition as well. To make the definition of viruses more appropriate, it can be added that viruses are parasites of the translational machinery. This distinguishes viruses from viriods

as these are parasites of both the transcriptional and translational machinery. Nevertheless, this definition is neither fully satisfactory as some DNA viruses, *e.g.* begomoviruses, also use cellular DNA polymerases for their replication.

Another characteristic of viruses is that their genome comprises either DNA or RNA, which contains information to initiate and complete an infectious cycle. Viruses package their genome in a particle, or virion, which is formed by *de novo* self-assembly of newly synthesized components. Virions transmit the viral genome from host to host, where disassembly initiates the next infectious cycle. Virions are small in size, which allows them to enter a host cell. For example, the virion of poliovirus is about 30 nm in diameter (Dales *et al.* 1965; Flint *et al.* 2009; Racaniello 2013), similar to the size of a ribosome (Palade 1955). One of the biggest viruses known, to date, belong to the genus *Mimivirus* (family: *Mimiviridae*), which has a diameter of 750-800 nm (La Scola *et al.* 2003, 2008; Xiao *et al.* 2005). The size of a virus can be seen as an indicator on how much a virus depends on its host. Small viruses can not carry many genes, so they depend more on the cell they infect. If viruses are larger, they can carry many genes that are not in other viruses, and can even code for their own DNA replication machinery. Thereby, a large virus becomes less of a parasite and less dependent on the host cell.

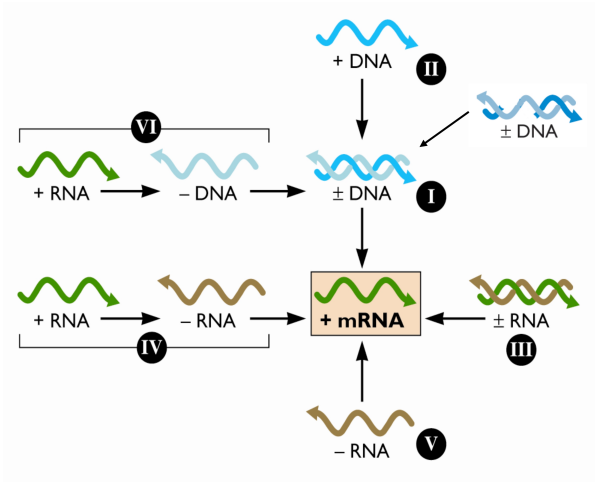

Figure I.1. The Baltimore classification. Adapted from *Principles of Virology* (Flint *et al.* 2015).

1.3. Virus classification

Viruses come in many different shapes, sizes and compositions and infect a wide range of hosts. To catalog viruses, two types of viral classification systems exist: (*i*) the classical hierarchical system and (*ii*) the Baltimore system. One one hand, the classical hierarchical system groups viruses according to their shared properties (Lwoff *et al.* 1962). Four main characteristics are used: (*i*) the nature and sequence of nucleic acid in the virion (RNA or DNA), (*ii*) the symmetry of the protein shell, or capsid, (*iii*) the presence or absence of a lipid membrane, or envelope, and (*iv*) the dimensions of the virion and capsid. The classical hierarchical system goes down from Kingdom, Phylum, Class, Order (-virales), Family (-viridae), Genus (-virus), to Species. So far, virus isolates from bacteria, plants and animals have been classified into 7 orders, 104 families, 505 genera and 3,186 species (http://ictvonline.org/). On the other hand, the Baltimore system (**Fig I.1**), bases its classification on the structure and composition of the viral genome (Baltimore 1971). This scheme is a reflection of the method of replication and packaging, and is based on the fact that every viral genome hast to make mRNA so it can be translated by the hosts ribosomes. Seven different viral genomes exist: dsDNA, gapped dsDNA, ssDNA, dsRNA, ss(+)RNA, ss(-)RNA, ss(+)RNA with DNA intermediate. The gapped dsDNA genome was originally not in the scheme as it was identified later. It is not clear why viruses have all these configurations and if there is any selective advantage of having a certain structure. It can be speculated that a ss(+)RNA genome has the highest advantage for fast replication, since it is an mRNA, and as soon as it enters the cell it can be translated. On the other hand, dsRNA and DNA are much more stable than ssRNA, also offering potential advantages.

1.4. Origin of viruses

To date, it is not clear where viruses came from. Three main hypotheses exist. One hypothesis is that viruses may have arisen from mobile genetic elements, retrotransposons, that are present in most eukaryotic genomes. The movement of viral-like retrotransposons within a genome is similar to the replication process of retroviruses (Hurst and Werren 2001; Pommier *et al.* 2005). Therefore, it can be speculated that retrotransposons gained the ability to exit one cell and enter another, and thereby becoming an infectious agent. Another hypothesis is that viruses may have originated from more complex (free-living) organisms that adapted a parasitic replication strategy. During this adaptation process, this organism underwent genome reduction by losing essential genes, and eventually losing the ability to replicate independently. This hypothesis is best illustrated by the nucleocytoplasmic large DNA viruses (NCLDVs) (Iyer *et al.* 2001), proposed as the new *Megavirales* order (Colson *et al.* 2012), which have larger and more complex genomes than other viruses. This might indicate that they are descendants of more complex ancestors. A last hypothesis supports the idea that viruses may have been the first replicating entities, predating the last universal cellular ancestor (LUCA) (Nasir *et al.* 2012). This is supported by the similarities in virion architecture and coat protein topology of viruses infecting hosts in the three domains of life (Bamford *et al.* 2005). However, none of these three hypothesis may be entirely correct or complete. Viruses could also have evolved multiple times using multiple mechanisms, or using mechanisms yet to be discovered.

2. RNA virus evolution

2.1. Evolutionary plasticity

Virus evolution is the result of continuing interaction between viral and host cell genes and selection for survival of the fittest. RNA viruses are probably one of the most abundant parasitic life forms. They infect a wide range of hosts in all domains of life, including other parasites. The key to their ability to infect many different hosts, results from the evolutionary plasticity of RNA virus genomes. This plasticity favors adaptation to environmental changes and results in RNA viruses being some of the most important emerging pathogens. Mechanisms that contribute to a high plasticity are short generation times, large population sizes, and high mutation rates (Domingo 2000; Elena and Sanjuán 2007; Holmes 2009). Additionally, even the robustness against genetic or environmental perturbations can result in a higher plasticity (Elena 2012).

2.2. Mutation rate

Although the estimation of viral mutation rate is a complex issue, RNA viruses have been estimated to generate between 10^{-6} and 10^{-4} errors per nucleotide (Sanjuán *et al.* 2010). This means that there is approximately one mutation per genome per replication cycle. The mutation rate in RNA viruses is much higher than those reported for DNA viruses, which is estimated to range from 10^{-8} to 10^{-6} (Sanjuán *et al.* 2010). One reason for this difference is that, unlike DNA-dependent DNA polymerases, many RNA-dependent RNA polymerases do not have a proofreading mechanism, and are therefore more error prone. The low fidelity of the RNA polymerase seems to be crucial for viral survival, as high-

fidelity mutant viruses negatively affect viral fitness (Pfeiffer and Kirkegaard 2005; Vignuzzi *et al.* 2006; Coffey *et al.* 2011). For RNA viruses not to reach the error threshold, *i.e.* the number of mutations within a population at which a viruses can no longer be propagated, selection and survival must balance genetic fidelity and the mutation rate.

2.3. Quasispecies, population size and fitness

RNA virus populations exist as dynamic distributions of nonidentical but related and interactive replicons, termed mutant swarms or quasispecies (Eigen and Schuster 1977; Domingo and Holland 1997). The composition of viral quasispecies can be altered by (*i*) mutation, (*ii*) homologous and nonhomologous recombination, (*iii*) reassortment of genome segments, and (*iv*) genome segmentation. Viral quasispecies constitute huge reservoirs of variants where the input of beneficial mutations allows for escape and survival in the face of an environmental change (*e.g.,* drug-resistant variants). Despite the huge amount of variation in a virus population, the average consensus sequence may remain invariant for many generations. (Domingo *et al.* 1978; Steinhauer *et al.* 1989). Only when a quasispecies is poorly adapted to the environment or is adapting to a new environment, newly arising mutants will have a higher probability to be more fit and increase in frequency in the population.

Changes in population size have important effects on genetic variation and viral fitness (Domingo *et al.* 2012). Serial passaging of large virus populations in a constant environment generally result in a fitness increase (Novella *et al.* 1995; Escarmís *et al.* 1999; Lorenzo-Redondo *et al.* 2011), which may or may not result in a modification of the consensus sequence. Genetic bottlenecks

involving drastic reductions in population size have been commonly associated with decreases in genetic variation, the accumulation of mutations which are reflected in the consensus sequence, and hence decreases in viral fitness (**Fig I.2**) (Duarte *et al.* 1992; Clarke *et al.* 1993; Novella 2004; Li and Roossinck 2004; de la Iglesia and Elena 2007; Escarmís *et al.* 2009). This is in line with the "Muller's ratchet" concept, describing that asexual organisms with a small

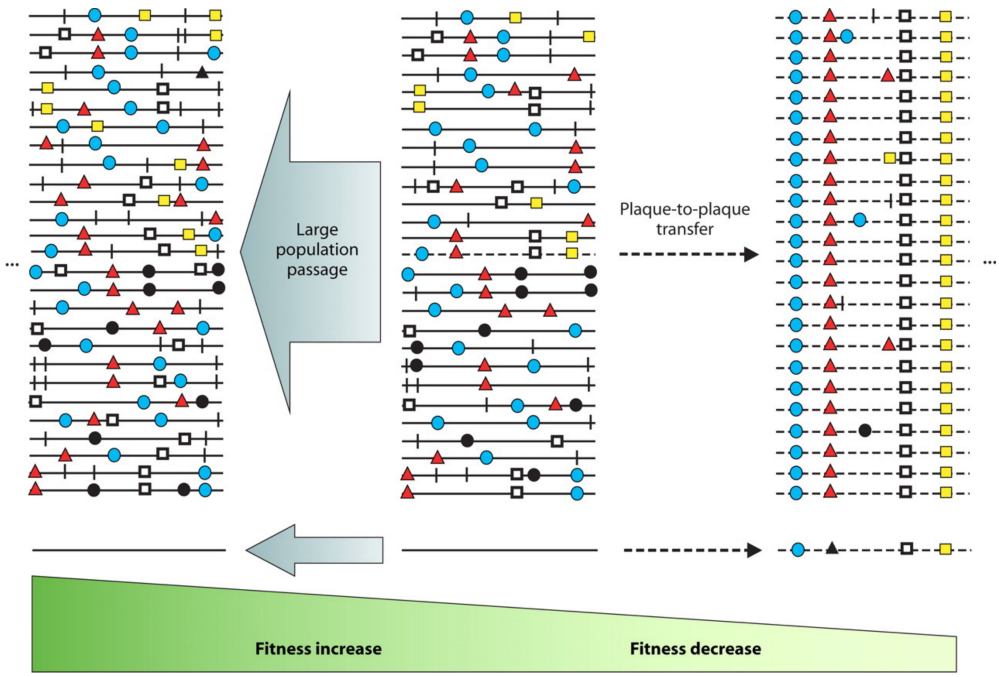

Figure I.2. The effect RNA virus population size on viral fitness. Mutant distributions are represented as horizontal lines and colored symbols. Large population passages in a constant environment generally result in a fitness increase (bottom trapezoid), which may or may not result in a modification of the consensus sequence (lines below the distribution). In contrast, genetic bottlenecks (plaque-to-plaque transfers; discontinuous genome in the central mutant distribution and discontinuous arrow) result in an accumulation of mutations, also reflected in the consensus sequence, and fitness decline. From Domingo *et al*. (2012). Reprinted with permission from the *American Society for Microbiology*.

population size and a high mutation rate will tend to incorporate deleterious mutations in an irreversible, ratchet-like manner unless mechanisms like recombination can compensate for such damage (Muller 1964).

Genetic bottlenecks appear to be frequent during the life cycle of RNA viruses. The main sources for bottlenecks are between-host transmission, either horizontal (Zwart *et al.* 2011) or vertical (Fabre *et al.* 2014), as well as bottlenecks during the spread of infection within the host (Tromas *et al.* 2014a). Nevertheless, genetic bottlenecks are not the only means to drive quasispecies to low fitness, drastic changes in the environment can do so as well. Adaptation of a virus to another environment can lead to perturbation of the mutant spectrum.

2.4. Mutant spectrum

The mutant spectrum – the ensemble of genomes that constitute a viral quasispecies – is not always at an equilibrium. Population equilibrium refers to viral populations that maintain a constant consensus sequence, whereas disequilibrium refers to mutations thereof. Positive selection or stochastic events can lead to disequilibrium, where subsets of the spectrum increase in frequency and replace the previous distribution. In addition to disequilibrium, the mutant spectrum of a virus can display evolutionary stasis in one environment, while in another environment it displays rapid evolution. One example is avian Influenza virus that shows stasis in its natural avian hosts, but causes disease in alternative avian hosts or in mammalian hosts (Gorman *et al.* 1992). High mutation rates permit but are not necessarily involved in rapid evolution.

2.5. Genetic robustness

Most mutations occurring in virus populations are deleterious or even lethal mutations (Sanjuán 2010), which can have a negative impact on viral fitness. Even though individual molecules in a population may be very brittle to these mutational effects, the average effect of mutation on the whole population may be small (Elena 2012). RNA viruses can be robust against mutational or environmental perturbations. Potential mechanisms that provide viral robustness are: (*i*) large population sizes, (*ii*) recombination, (*iii*) reassortment, and (*iv*) complementation in the case of co-infection of two or more viruses (Elena *et al.* 2006; Elena 2012; Lauring *et al.* 2013).

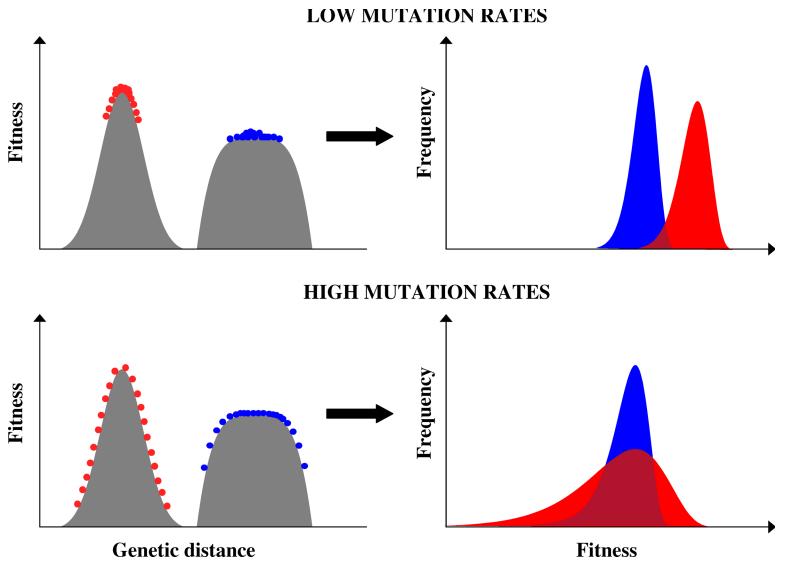

Figure I.3. Schematic representation survival of the flattest effect. Left panel: Dots represent individuals located on each peak at low and high mutation rates. Fitter but more brittle populations are indicated in red, whereas more robust populations are indicated in blue. Right panel: Qualitatively expected distribution of individual fitness values for the two populations. The red individuals benefit from the highest fitness, although the blue populations show less variance in fitness due to the higher mutational robustness. Reprinted from Sanjuán *et al* (2007).

Recent studies have demonstrated that more robust viral populations can be superior competitors, despite having lower replication rates (Codoñer *et al.* 2006; Sanjuán *et al.* 2007). The phenomenon of robust genotypes producing equally fit phenotypes is called "survival of the flattest" (**Fig I.3**), where mutations are neutral. Contrarily, non-robust genotypes suffer from mutational fitness effects.

3. The evolution of genome architecture

3.1. Genome architecture across organisms

There is astounding variation in genome architecture, *i.e.* genome organization, over organisms. Viruses tend to have small genomes, with minimal intergenic sequences and typically with overlapping genes (Lynch 2006; Belshaw *et al.* 2007). Gene overlaps are though to be a form of genome compression, allowing the virus to increase its number of proteins without increasing its genome length (Barrell *et al.* 1976; Scherbakov and Garber 2000; Lillo and Krakauer 2007; Chung *et al.* 2008). Prokaryotes have compact genomes, from less than 200 kb to over 13 Mb (Rocha 2008), but with larger intergenic sequences than viruses, and long overlapping genes are rare (Lillo and Krakauer 2007). Eukaryotes have a wide range of genome sizes, from within the prokaryotic range to orders of magnitude larger, and are marked by their intron-exon organization (Hawkin 1988; Roy and Gilbert 2006). Intronic and intergenic sequences contribute to the large eukaryotic genome sizes and their genomes can contain up to 60% of non-coding DNA, while viruses and prokaryotes consist mostly (> 85%) of coding DNA (Lynch 2006).

There is a notable increase in genome complexity from viruses to prokaryotes to eukaryotes, in terms of genome size, gene number, mobile-element number, intron number, intron size, and complexity of regulatory regions (Lynch and Conery 2003). Even so, the differences in genome architecture over organisms are not defined by clear boundaries. The discovery of the giant viruses (La Scola *et al.* 2003; Raoult *et al.* 2004) and, conversely, of bacterial symbionts with tiny genomes (Bennett and Moran 2013) eliminated the separation of cellular and viral genomes by size. Therefore, it has been suggested that the primary forces driving the divergence in genome architecture evolution are population genetic mechanisms, rather than different lifestyles, cell structures or physiologies of organisms (Lynch 2006).

3.2. Mechanisms shaping viral genome architecture

Population genetic mechanisms like the mutation rate and effective population size could play a role in shaping the divergent genome architectures. The so-called Drake's Law (Drake 1991; Drake *et al.* 1998), describes how the genome-wide mutation rate is more or less constant across organisms. Therefore, a negative correlation between the mutation rate and genome size exists (Drake 1991; Drake *et al.* 1998; Gago *et al.* 2009; Lynch 2010; Sanjuán *et al.* 2010; Sung *et al.* 2012). This is also a consequence of the Eigen's error threshold paradox (Eigen 1971): the length of a genome is limited by the mutation rate. Whenever the mutation rate is below the error threshold, a population will be maintained by mutation-selection balance.

Viruses, with small genomes compared to multicellular species, have a higher mutation rate (**Fig I.4**). In particular the high mutation rates observed for RNA

viruses is several orders of magnitude higher than those in most DNA based organisms (Duffy *et al.* 2008; Sanjuán *et al.* 2010). Even so, the boundaries of mutation rates between viruses with different genome architectures are not clear, and the high mutation rates of RNA viruses are matched by some DNA viruses (Duffy *et al.* 2008). Therefore it has been suggested that other aspects, like genome architecture and replication speeds rather than low RNA polymerase fidelity alone, better explain the differences in mutation rates in viruses.

The transitions from simple to more complex organisms are also found to be associated with large reductions in population size (Lynch and Conery 2003). In small population sizes, the power of random genetic drift increases which provides space for the production of various genomic features that would otherwise be eliminated by purifying selection. Therefore, it has been proposed that a large effective population size – like those of RNA viruses – may be an important barrier to the evolution of complex genomes (Lynch and Conery 2003).

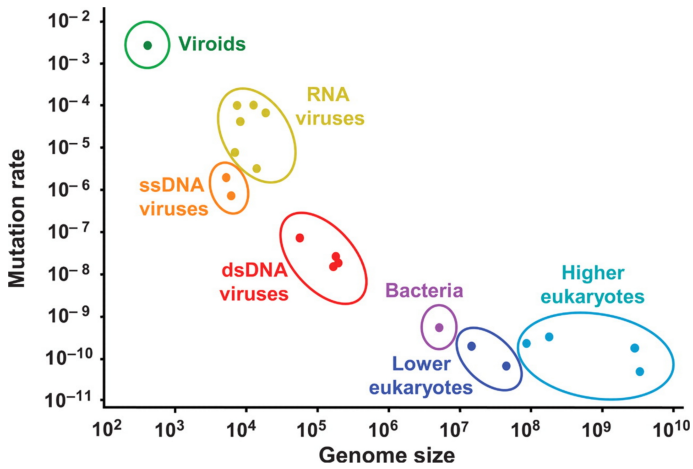

Figure I.4. The relationship between mutation rate per nucleotide site and genome size for different organisms. From Gago *et al.* (2009). Reprinted with permission from *AAAS*.

3.3. RNA virus genome architecture

When comparing the nucleotide sequences of RNA viral genomes, blocks of genes that encode proteins with similar functions can be identified. The (–)RNA viruses carry a limited number of genes, ranging from 4 to not more than 13, while this number ranges from 3 to more than 12 in (+)RNA viruses (Flint *et al.* 2015). The (+)RNA viruses represent the largest group of viruses (Francki *et al.* 1991) and are classified into three tribes: picorna-, alpha- and flavi-like (Koonin and Dolja 1993; Fauquet *et al.* 2005). These viruses are characterized by conserved gene clusters (**Fig I.5**), and specially the helicase-polymerase arrangement, where the helicase gene is typically located upstream of the polymerase gene (Koonin and Dolja 1993). In particular, the picorna-like tribe is identified by the partial conservation of core genes that consists of the RNA-dependent RNA polymerase (RdRp), a chymotrypsin-like protease (3CPro), a superfamily 3 helicase (S3H), and a genome-linked protein (VPg) (Koonin and Dolja 1993; Fauquet *et al.* 2005; Koonin *et al.* 2008). Moreover, core genes

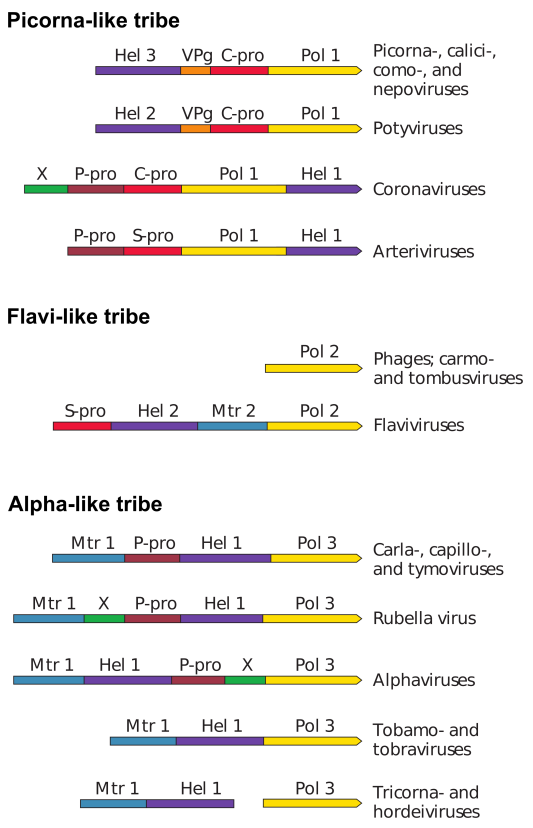

Figure I.5. Organization of postive-strand RNA virus genomes. Adapted from *Principles Of Virology* (Flint *et al.* 2015).

tend to form ordered arrays, whereas non-core genes are responsible for genome reorganization and recombination between distant groups of viruses from all three tribes (Koonin and Dolja 1993; Fauquet *et al.* 2005).

4. The model system: a plant RNA virus

4.1. Tobacco etch virus characteristics

To study the evolution of genome architecture, we chose to work with *Tobacco etch virus* (TEV). This plant pathogen has a single-stranded (+)RNA genome, classified within the genus *Potyvirus* of the *Potyviridae* family. Virions in this genus are non-enveloped, flexuous filaments of approximately 680-900 nm long and 11-13 nm in diameter (King *et al.* 2011), and a helical symmetry (**Fig I.6**).

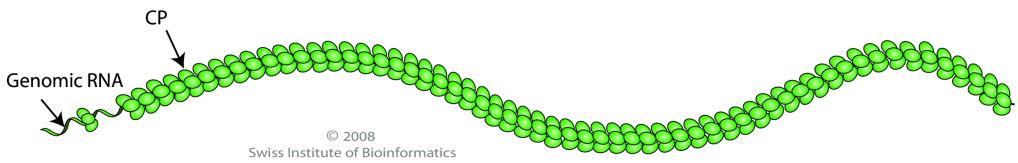

Figure I.6. Schematic representation of a potyvirus virion. From ViralZone (http://viralzone.expasy.org/). Reprinted with permission from SIB, *Swiss Institute of Bioinformatics*.

To date, 190 viral species have been assigned into the *Potyviridae* family (http://www.ictvonline.org/virusTaxonomy.asp). These species are divided into eight different genera; *Brambyvirus* (1), *Bymovirus* (6), *Ipomovirus* (6), *Macluravirus* (6), *Poacevirus* (2), *Potyvirus* (158), *Rymovirus* (3), *Tritimovirus* (6), and two species which are still unassigned. The different genera are defined

based on genome composition and structure, sequence similarity and vector organisms responsible of plant-to-plant transmission. All viruses in these genera have a monopartite genome, except for viruses in the *Bymovirus* genus which have a bipartite genome (**Fig I.7**). Interestingly, the gene order of the different members within the *Potyviridae* family is very well conserved.

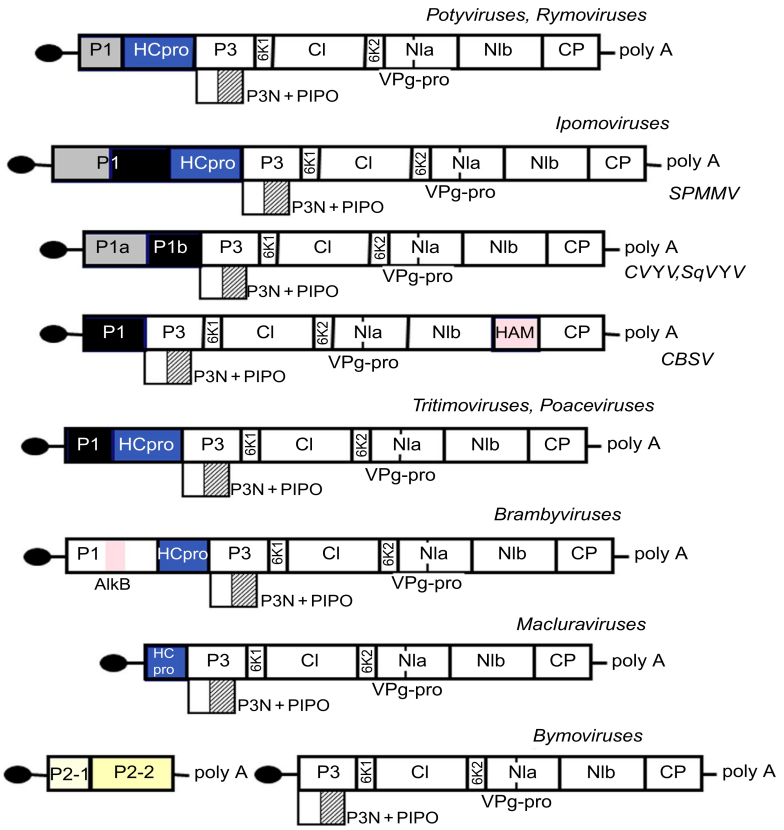

Figure I.7. The genomic organization of the eight genera in the *Potyviridae* family. Reprinted from Revers and García (2015), with permission from *Elsevier*.

The monopartite linear genome of TEV is about 9.5 kb in length. It codes for a single ORF (*i.e.,* polyprotein) that is further processed after translation into ten mature peptides (**Fig I.8**), in addition P3N-PIPO is translated at a +2 frameshift within the P3 cistron. The P1 protein is a serine protease that cleaves at its own C-terminus (Verchot *et al.* 1991). P1 acts as an accessory factor for virus amplification (Verchot and Carrington 1995a) and is likely to be implicated in host range determination (Shi *et al.* 2007). HC-Pro is a cysteine protease that cleaves itself at its C-terminus. HC-Pro is a multifunctional protein, however, it was named to its first discovered function: Helper Component (HC) for aphid transmission (Carrington *et al.* 1989). Other functions of HC-Pro are RNA silencing suppression (Kasschau and Carrington 2001; Mallory *et al.* 2002; Soitamo *et al.* 2011) and the enhancement of the yield of virus particles (Valli *et al.* 2014). The P3 protein is required for viral replication (Klein *et al.* 1994) and is suggested to be involved in host adaptation (Lin *et al.* 2011). P3N-PIPO, located within the P3 protein, is involved in virus movement from cell-to-cell (Wen and Hajimorad 2010; Vijayapalani *et al.* 2012). The role of 6K1 is not know, it is though that this protein regulates the activity of P3 (Riechmann *et al.* 1995). Like HC-Pro, CI is a multifunctional protein with ATPase and RNA helicase activities which are required for replication (Fernandez 1997; Sorel *et al.* 2014). The CI protein forms cylindrical inclusion bodies in the form of pinwheels in the cytoplasm of infected cells (Edwardson and Christie 1996). It has also been suggested that CI helps virus movement through plasmodesmata (Gabrenaite-Verkhovskaya *et al.* 2008), the symplasmic tunnels between cells. 6K2 is associated with the VPg-NIaPro protein complex in endoplasmatic reticulum (ER) derived membranes, forming cytoplasmatic vesicles (Léonard *et al.* 2004; Beauchemin *et al.* 2007). The VPg-NIaPro complex is formed by the

processing of the largest protein NIa. VPg is a 5'-end genome-linked protein, which serves as a primer for RNA replication. NIa-Pro is responsible for the proteolytic processing of the central and C-terminal regions of the potyvirus polyprotein (Adams *et al.* 2005b). In addition, NIa-Pro has DNase activity, which may have a regulatory role in host gene expression relevant for viral infection (Anindya and Savithri 2004). NIb is the RNA-dependent RNA polymerase, responsible for potyviral genome replication (Hong and Hunt 1996). CP is the coat protein, responsible for the encapsidation of the viral genome.

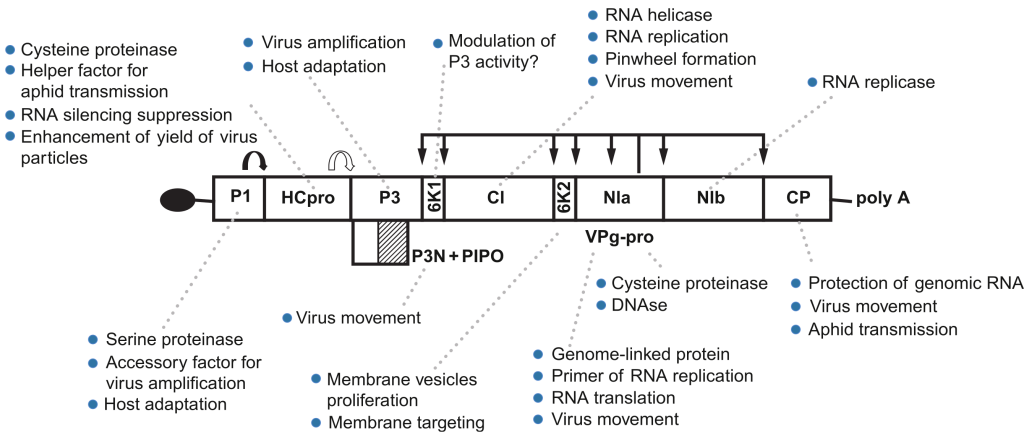

Figure I.8. Schematic representation of a potyvirus genome. Relevant features and proposed functions are indicated. Reprinted from Revers and García (2015), with permission from *Elsevier*.

4.2. Host plants and transmission

TEV has a moderate host range, infecting mainly plants in the *Solanaceae* family. Examples of its natural hosts are: *Nicotiana tabacum*, *Datura*

stramonium, *Capsicum annuum*, and *Nicotiana benthamiana*. Infection of TEV induces mottling, necrotic etching and leaf distortion in most of its hosts. **Fig I.9** displays the differences between a healthy *N. tabacum* plant (**A**) and a plant infected with TEV (**B**).

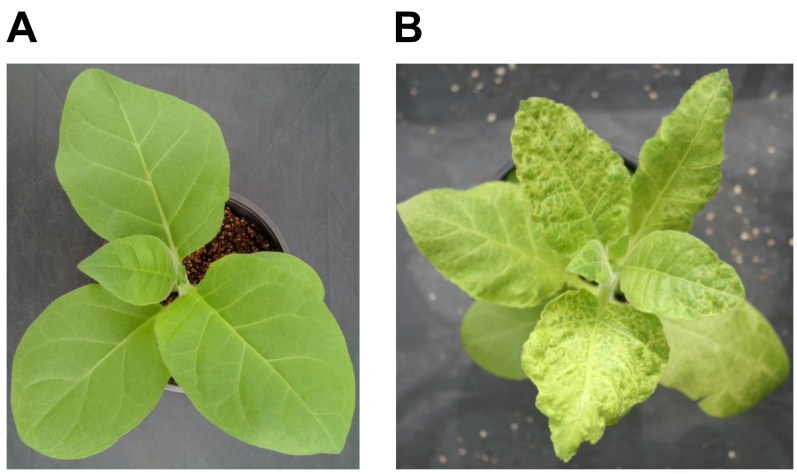

Figure I.9. Symptomatology. (**A**) Healthy *N. tabacum* plant. (**B**) *N. tabacum* plant infected with TEV.

TEV is transmitted by aphids in a non-persistent and non-circulative manner. Unlike circulative viruses, non-circulative viruses have a more superficial and transient relationship with the vector and only enter into the stylet of the aphid (**Fig I.10**). The majority of the aphid vectors that transmit plant viruses belong the the subfamily *Aphidinae* (Blackman and Eastop 2000). The piercing-sucking mouthparts of the aphid facilitate the delivery of virions into plant cells without causing unrepairable damage. TEV does not replicate in the aphid vector, therefore it can be rapidly acquired and transmitted (NG and Perry 2004). The virus is hosted in the anterior food canal of aphids (**Fig I.10**), where

the HC-Pro and CP proteins reversibly bind the virus particles to putative receptor sites (Moreno *et al.* 2012). In addition to transmission by aphids, TEV can also be transmitted mechanically.

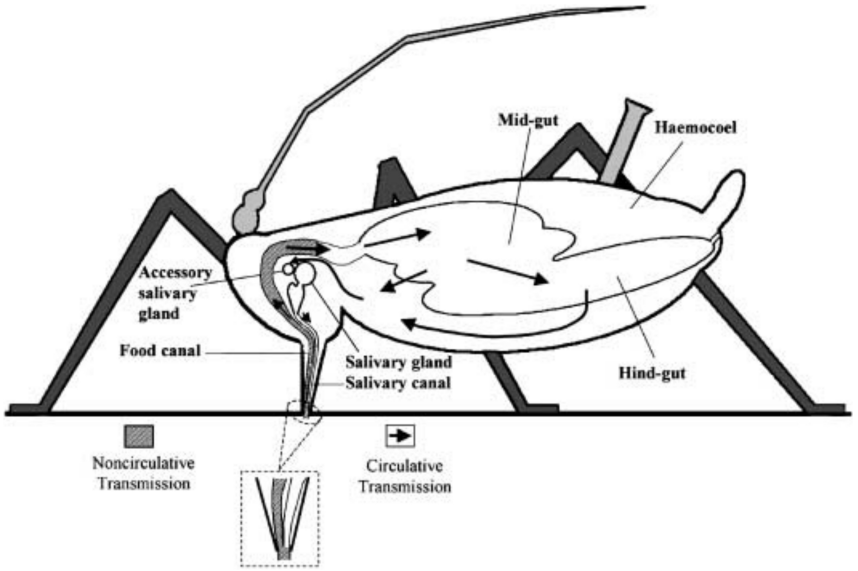

Figure I.10. Transmission of plant viruses by aphid vectors. The food canal and foregut (hatched regions) are retention sites for non-circulative transmitted viruses. Arrows represent the passage of circulative transmitted viruses. From Ng and Perry (2004). Reprinted with permission from *John Wiley and Son*s.

4.3. Replication

Once entered a plant cell, either by vector or mechanical inoculation, TEV replicates in the cytoplasm of infected cells (**Fig I.11**). The genomic RNA is released after a poorly understood step of virion disassembly. Translation can start directly, as it is a (+)RNA genome. The replication complex is formed thanks to the production of viral proteins. Then, complementary negative-strand (–)RNA copies are formed, that serve as templates for the synthesis of new

genomic RNAs in a stamping-machine manner (Martínez *et al.* 2011), which are involved in new replication steps, translated or encapsidated. The replication takes place in vesicles that are produced by host endomembrane recruitment (Grangeon *et al.* 2012). These vesicles are derived from the ER and are found distributed throughout the cortical and perinuclear ER membrane systems (Beauchemin *et al.* 2007; Beauchemin and Laliberté 2007). However, vesicles have been found to target chloroplasts and it is suggested that chloroplasts are the main location for genome replication, whereas the ER is the site where initiation of replication takes place (Wei *et al.* 2010).

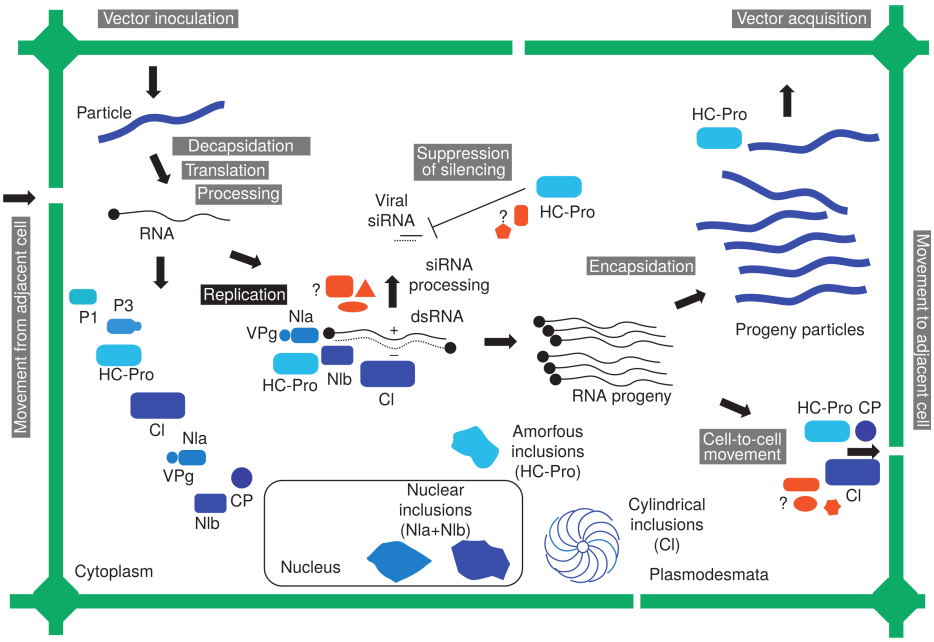

Figure I.11. Schematic representation of different events during a potyvirus infection of a plant cell. Reprinted from *Encyclopedia of Virology* (López-Moya and García 2008).

4.4. Infection dynamics

TEV expands within its host plant by two different manners: by (*i*) cell-to-cell movement, where virions and replication complexes are transferred from cell-to-cell via plasmodesmata (Cruz *et al.* 1998; Gabrenaite-Verkhovskaya *et al.* 2008), and by (*ii*) systemic movement, where the virus accesses the vascular tissue and is transported in the phloem sap within and between leaves (Dolja *et al.* 1992). The timing of the onset of systemic infection depends on the speed of cell-to-cell movement and the number of primary infected foci in the inoculated leaf, which is quicker when both processes increase (Lafforgue *et al.* 2012; Rodrigo *et al.* 2014). However, model predictions suggest that there is a limited range for the number of primary infection foci, as the onset of systemic infection will be limited by the latency period (Rodrigo *et al.* 2014).

As TEV achieves local spread by cell-to-cell movement, infected cells are likely to be found together forming aggregates. The cellular contagion rate, *i.e.* the number of secondary infections per infected cell per day, was estimated to range from 2.43 to values close to 0 (Tromas *et al.* 2014a). So even though TEV has a high replication rate within the cell, the cell-to-cell movement is much slower. The aggregation of virus-infected cells play a role in the low contagion rate, as only cells at the edge of an aggregate can contribute to virus expansion. Additionally, host immune responses, in particular RNA-silencing, are likely to play a role in the slow virus expansion between cells. The cellular multiplicity of infection (MOI) for TEV, in plants, has an estimated mean of 1.14 virions per infected cel (Tromas *et al.* 2014a). At this low MOI, a small fraction of cells will still be infected by 2 or more virions, allowing for limited cellular co-infection.

At the plant level, there is a low overall frequency of cellular infection with TEV (Tromas *et al.* 2014a). When TEV infects a plant, the virus moves up in through the different leaves as the plant grows. The infection dynamics vary greatly between the different leaves of an infected plant (**Fig I.12**). In the inoculated leaf, a low level of cellular infection is detected, compared to leaves located higher in the plant (Tromas *et al.* 2014a). Interestingly, in this experimental setup, which is similar the one used for this thesis, no infection can be detected in the leaf above the inoculated leaf (**Fig I.12**; leaf 4), which is probably related to the maturity of the leaf. At the time of the inoculation the leaf above is already mature and a sink-source transition (Turgeon 1989) – producing and excess of photoassimilate (*e.g.,* energy storing monosaccharides) – has probably already taken place in this leaf. It is likely that phloem-transported virions cannot traverse the sink-source boundary.

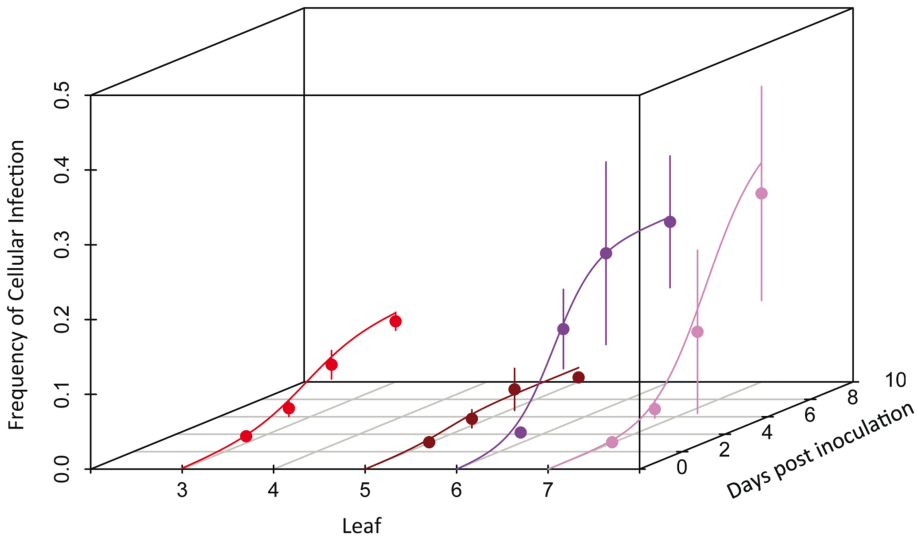

Figure I.12. The frequency of infected cells at different times, measured in the different leaves. The inoculated leaf (leaf 3) and the above leaves are shown (leaves 4, 5, 6 and 7). Adapted from Tromas *et al.* (2014a).

4.5. The effective population size

The effective population size (N_e) for viruses is defined as the number of horizontally transmitted virions. N_e determines how strong genetic drift acts on viral populations (Moya *et al.* 2000) and the frequency of co-infection by multiple virus genotypes (Zwart *et al.* 2009). The independent-action hypothesis (IAH), proposes that each virion has a non-zero probability of infection and that the action of virions is independent of the presence of other conspecific virions. Under this hypothesis, viral population size is linked to the level of host infection. The IAH was first described in plant viruses (Bald 1937), and was later confirmed in other plant RNA viruses, (Furumoto and Mickey 1967), including TEV (Zwart *et al.* 2011). Thus, the number of virions needed for systemic infection of TEV can be very small, starting from one virion, although the probability of infection per virion is very small as well and a large inoculum dose of virions is therefore typically needed (Zwart *et al.* 2011). For TEV, N_e is dose dependent as the number of primary infection of foci increased linearly with dose, whilst for a given dose the number of primary infections approximately follows Poisson distribution (Zwart *et al.* 2011).

4.6. Mutation rate

The upper limit of the mutation rate of TEV is estimated to be 3×10^{-5} per nucleotide per replication cycle, which is in the lower range accepted for RNA viruses (Sanjuán *et al.* 2009). However, this estimate was done on a highly constrained region and was therefore under selective pressure. The spontaneous mutation rate for TEV is estimated to be in the range of 10^{-6} and 10^{-5} mutations per nucleotide per replication cycle (Tromas and Elena 2010). This estimate was

done on a neutral region in TEV. The low mutation rate appears to be a general observation in plant RNA viruses (Malpica *et al.* 2002; Sanjuán *et al.* 2009; Tromas and Elena 2010), supporting the idea that plant RNA viruses may have lower mutation rates than their animal counterparts.

4.7. TEV as a model for experimental evolution of genome architecture

TEV is well studied and a commonly used model to study many aspects of the molecular biology of plant RNA viruses. More recently, TEV is also used as a model organism for experimental evolution of plant RNA viruses. One advantage of this system is that a full length clone of this virus is available (pTEV7DA, GenBank: DQ986288), which allows the *in vitro* production of infectious transcripts. More stable versions of this clone have now been generated as well (Bedoya and Daròs 2010). The genome of TEV is relatively small, and therefore facilitates the generation of recombinant variants of this virus. Another advantage of this experimental system is that many potyviral genomes are available, allowing for comparisons amongst them. The genome architecture of TEV appears to be very rigid. The gene order has been very well conserved in all eight genera within the *Potviridae* family (**Fig I.7**). However, it is unclear why this architecture has evolved and whether this is the optimal architecture for a potyvirus.

Objectives

The evolution of genome architecture in viruses can be grossly divided into three sorts of processes. First, the decrease of genome complexity, for example, the deletion of a redundant gene or regulatory sequence, which results in a reduction of genome size. Second, the increase in genome complexity, *e.g.* horizontal gene transfer (HGT), gene duplication, or *de novo* evolution of genes, which result in an increase in genome size. Third, the reshuffling of existing elements without any duplication events, which does not result in a change of genome size, but does result in a change of gene order. In this thesis, I investigated these three processes of genome architecture using TEV as a model system. There are four main objectives of this study that correspond to the four chapters in this thesis:

The first objective is to better understand the conservation of gene order in virus orders and families. Using TEV as an *in vivo* model, the evolutionary trajectories to alternative gene orders are explored. The results will shed light on the factors that constrain or promote gene-order conservation.

The second objective is to better understand the stability of genetic redundancy and how viruses evolve smaller genomes by removing this redundancy. The stability and fitness costs of genetic redundancy are measured, by experimentally evolving TEV variants containing potentially beneficial gene duplications. Based on the experimental data, a model that can predict the stability of genetic redundancy will be developed, contributing to our understanding of which biological features constrain the likelihood of maintenance of duplicate genes.

The third objective is to explore the evolutionary fate of an increase in genome size, in the context of HGT. Two different exogenous sequences are introduced in the TEV genome, that simulate the acquisition of a new function and the acquisition of an existing function. These viruses are experimentally evolved, to determine the stability of the exogenous sequences. The results will indicate if HGT is possible in a present-day potyvirus.

The fourth objective is to better understand the effects of virulence and transmission on evolution. It is considered whether the evolutionary patterns observed for viruses with an altered genome architecture are similar in alternative hosts. By using hosts for which TEV has a large difference in virulence, it is explored how virulence affect adaptability to a new host. The result will demonstrate if adaptive evolution is predictable in alternative hosts.

Methods

1. Generation and infection of altered TEV clones

1.1. Virus clones

The TEV genome used to generate all virus clones, was originally isolated from *N. tabacum* plants (Carrington *et al.* 1993). In this study, 10 different variants of TEV were generated, namely as the following virus genotypes: TEV-NIb_1-ΔNIb_9, TEV-NIb_2-ΔNIb_9, TEV-NIb_1-NIb_9, TEV-NIb_2-NIb_9, TEV-$HCPro_2$-$HCPro_3$, TEV-$NIaPro_2$-$NIaPro_8$, TEV-CP_{10}-CP_{11}, TEV-2b, TEV-AlkB, and TEV-eGFP.

TEV-NIb_1-ΔNIb_9, TEV-NIb_2-ΔNIb_9, TEV-NIb_1-NIb_9 and TEV-NIb_2-NIb_9 were generated from cDNA clones constructed using plasmid pGTEVa, which consists of a TEV infectious cDNA (GenBank: DQ986288, including two silent mutations, G273A and A1119G) flanked by *Cauliflower mosaic virus* (CaMV) 35S promoter and terminator in a binary vector derived from pCLEAN-G181 (Thole *et al.* 2007). TEV-$HCPro_2$-$HCPro_3$, TEV-$NIaPro_2$-$NIaPro_8$, TEV-CP_{10}-CP_{11}, TEV-2b and TEV-AlkB were generated from cDNA clones constructed using plasmid pMTEVa, which like pGTEVa consist of TEV infectious cDNA (accession: DQ986288, including two silent mutations, G273A and A1119G) flanked by SP6 phage RNA promoter derived from pTEV7DA (GenBank: DQ986288), however pMTEVa contains a minimal transcription cassette to ensure a high plasmid stability (Bedoya and Daròs 2010). TEV-eGFP was constructed in a previous study based on pMTEV (Zwart *et al.* 2011). Clones were constructed using standard molecular biology techniques, including PCR

amplification of cDNAs with the high-fidelity Phusion DNA polymerase (Thermo Scientific), DNA digestion with *Eco*31I (Thermo Scientific) for assembly of DNA fragments (Engler *et al.* 2009), DNA ligation with T4 DNA ligase (Thermo Scientific) and transformation of *E. coli* DH5α by electroporation. For generation of TEV-2b, the primers used for amplification were phosphorylated, prior to the ligation step, due to the presence of an *Eco*31I site within the *2b* sequence. Sanger sequencing confirmed the sequences of the resulting plasmids.

1.2. Virus stocks

For TEV-NIb$_1$-ΔNIb$_9$, TEV-NIb$_2$-ΔNIb$_9$, TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$, the ancestral and resulting binary plasmids were transformed in *Agrobacterium tumefaciens* C58C1, harboring helper plasmid pCLEAN-S48 (Thole *et al.* 2007). *N. tabacum* L cv. Xanthi (NN) plants were agroinoculated with *A. tumefaciens* cultures (Bedoya and Daròs 2010) and symptomatic tissue collected 7 days post-inoculation (dpi). To generate large virus stocks, the collected tissue was homogenized, ground into fine powder using liquid nitrogen and a mortar, and resuspended 1:1 in phosphate buffer (50 mM KH_2PO_4, pH 7.0, 3% polyethylene glycol 6000). The third true leaf of 4-week-old *N. tabacum* plants was mechanically inoculated with 50 µl of the TEV genotypes. All systemically infected tissues were harvested 7 dpi and stored at −80 ºC. For TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-CP$_{10}$-CP$_{11}$, TEV-2b, TEV-AlkB and TEV-eGFP, the plasmids were linearized by digestion with *Bgl*II prior to *in vitro* RNA synthesis using the mMESSAGE mMACHINE® SP6 Transciption Kit (Ambion), as described in Carrasco *et al*.

(2007). The third true leaf of 4-week-old *N. tabacum* plants was mechanically inoculated with varying amounts (5 μg - 30 μg) of transcribed RNA. All symptomatic tissue was collected 7 dpi.

2. Experimental setup

2.1. Plants

Most experiments were performed in *N. tabacum* plants, from which the wild-type TEV in our experiments was originally isolated. Alternative hosts used for the experimental evolution of TEV-eGFP are *N. benthamiana* and *D. stramonium.* All plants were kept in a BSL-2 greenhouse at 24º C with 16 h light.

2.2. Serial passages

For the serial passage experiments, 500 mg homogenized stock tissue was ground into fine powder and diluted in 500 μl phosphate buffer. From this mixture, 20 μl were then mechanically inoculated on the third true leaf of 4-week old *N. tabacum* plants (**Fig M.1**). At least five independent replicates were used of each virus variant. At the end of the designated passage duration (3 or 9 weeks) all leaves above the inoculated leaf were collected and stored at −80 ºC. For subsequent passages the frozen tissue was homogenized and a sample of the homogenized tissue was ground and resuspended with an equal amount of phosphate buffer (Zwart *et al.* 2014). Then, new *N. tabacum* plants were mechanically inoculated as described above. The serial passages of TEV-

eGFP in *N. benthamiana* and *D. stramonium* were done similarly. However TEV has a high virulence for *N. benthamiana*, and therefore the long 9-week passages had to be restricted to 6-weeks passages. Additionally, the sixth true leaf of *N. benthamiana* was inoculated instead of the third true leaf inoculated in *N. tabacum* and *D. stramonium*.

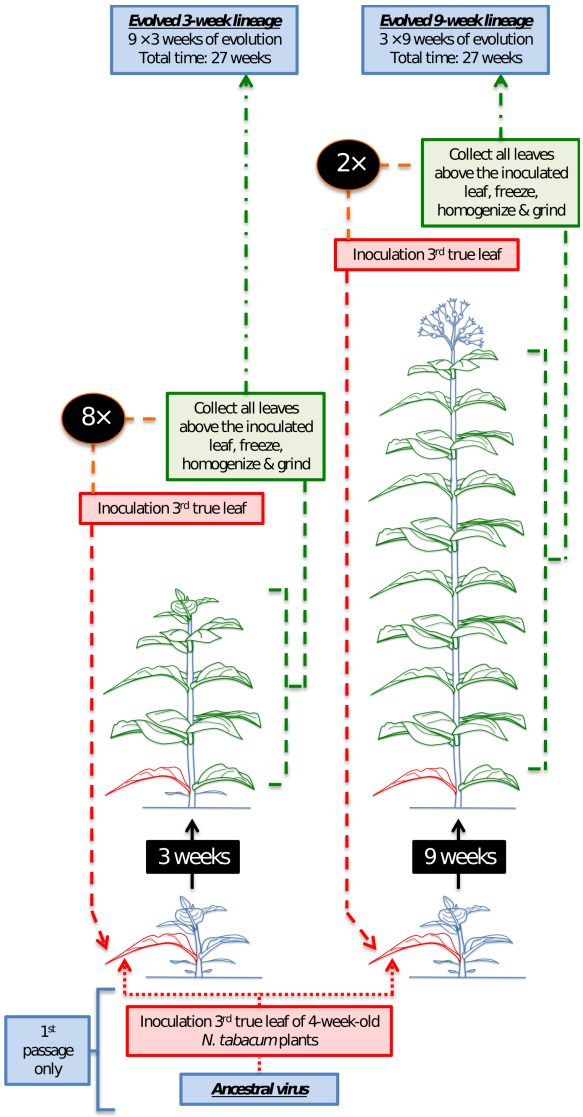

Figure M.1. Schematic representation of the serial passages. At the start of the serial passage experiment, 4-week-old *N. tabacum* plants were inoculated with TEV or one of its altered genotypes in the third true leaf (indicated in red). At the end of the designated passage duration (3 or 9 weeks) all leaves above the inoculated leaf (indicated in green) were collected and stored at −80 ºC. The frozen tissue was then homogenized, and a sample of the homogenized tissue was ground to a fine powder. For inoculation of subsequent passages, powder was resuspended in phosphate buffer and new *N. tabacum* plants were inoculated. Adapted from Zwart *et al*. (2014).

3. Virus detection

3.1. Symptoms

The wild-type TEV produces systemic symptoms in all three host plants used in this study. However, some of the altered genotypes show few or no symptoms and virus infection had to be confirmed by RT-PCR.

3.2. Reverse transcription polymerase chain reaction (RT-PCR)

To determine the stability of the rearranged, inserted or duplicated genes, RNA was extracted from 100 mg homogenized infected tissue using the InviTrap Spin Plant RNA Mini Kit (Stratec Molecular). Reverse transcription (RT) was performed using M-MuLV reverse transcriptase (Thermo Scientific) and a reverse primer located in the 3' UTR of the TEV genome. PCR was then performed with Taq DNA polymerase (Roche) and primers flanking the regions of interest. As some of the viruses showed weak to no symptoms, especially at the start of the experiment, infection was confirmed by amplifying a region outside the altered region, where deletions were unlikely to occur. PCR products were resolved by electrophoresis on 1% agarose gels. For the viruses where we observed deletions during the evolution experiment, we estimated the genome size based on the amplicon size and the genome size of the ancestral viruses.

3.3. RT-qPCR

The standard curves for measuring accumulation and within-host competitive

fitness were prepared using the pMTEV, pMTEV-eGFP, and pMTEV-mCherry plasmids (Bedoya and Daròs 2010; Zwart *et al.* 2011). These plasmids were linearized by digestion with *Bgl*II prior to *in vitro* RNA synthesis using the mMESSAGE mMACHINE® SP6 Transciption Kit (Ambion), as described in Carrasco *et al*. (2007). After synthesis, the RNA was diluted to a concentration of 50 ng/µl. The genome equivalents (*i.e.,* viral copy numbers) per µl can then be calculated based on the length of each genome and the molar mass (**Table M.1**).

Table M1. Genome equivalents per µl.

	TEV	TEV-eGFP	TEV-mCherry
measured ng/µl RNA	50	50	50
length genome (nt)	9539	10301	10271
g/mole per nt RNA	340	340	340
weight of one mole of genome (g)	3.243×10^{6}	3.502×10^{6}	3.492×10^{6}
Avogadro molecules / mole	6.022×10^{23}	6.022×10^{23}	6.022×10^{23}
weight of one genome (g)	5.386×10^{-18}	5.816×10^{-18}	5.799×10^{-18}
measured RNA (g)	5.000×10^{-8}	5.000×10^{-8}	5.000×10^{-8}
genome equivalents/µl	9.284×10^{9}	8.597×10^{9}	8.622×10^{9}

For quantification of a standard TEV accumulation or competition assay, six standard curve (SC) 5× serial dilutions are used starting with a concentration of 5×10^{8} genome equivalents/µl. To normalize the standard curve to the total RNA extracted from the infected plants, the SC dilutions are mixed 1:1 with total RNA extracted from a healthy plant. Real-time quantitative RT-PCR (RT-qPCR) was performed using the One Step SYBR PrimeScript RT-PCR Kit II (Takara), in accordance with manufacturer instructions, in a StepOnePlus Real-Time PCR

System (Applied Biosystems). The StepOne Software v.2.2.2 (Applied Biosystems) was used to analyze the data.

4. Fitness assays

4.1. Sample normalization

Prior to performing assays, the genome equivalents per 100 mg of tissue of the ancestral virus stocks and all evolved lineages were determined for subsequent assays. The InviTrap Spin Plant RNA Mini Kit (Stratec Molecular) was used to isolate total RNA of 100 mg homogenized infected tissue. Specific primers for the *CP* (coat protein) gene were used. The concentration of genome equivalents per 100 mg of tissue was then normalized to that of the sample with the lowest concentration, using phosphate buffer.

4.2. Accumulation assay

For the accumulation assays, 4-week-old *N. tabacum* plants were inoculated with 50 µl of the normalized dilutions of ground tissue. For each ancestral and evolved lineage, at least three independent plant replicates were used. Plant height was measured and leaf tissue was harvested 7 dpi. For TEV-eGFP in *N. benthamiana* and *D. stramonium*, leaf tissue was harvested 10 dpi. Total RNA was extracted from 100 mg of homogenized tissue. Virus accumulation was then determined by means of RT-qPCR for the *CP* of the ancestral and the evolved lineages. For each of the harvested plants, at least three technical replicates were used in the RT-qPCR.

4.3. Within-host competitive fitness assay

To measure within-host competitive fitness, we used TEV carrying an enhanced green fluorescent protein (TEV-eGFP) as a common competitor. TEV-eGFP has proven to be stable up to six weeks (using 1- and 3-week serial passages) in *N. tabacum* (Zwart *et al.* 2014), and is therefore not subjected to eGFP loss in our 1-week long competition experiments. To measure within-host competitive fitness for TEV-eGFP itself, we used TEV carrying a red fluorescent protein: TEV-mCherry (Zwart *et al.* 2011). All ancestral and evolved viral lineages were again normalized to the sample with the lowest concentration, and 1:1 mixtures of viral genome equivalents were made with TEV-eGFP or TEV-mCherry. As the TEV-NIb_1-ΔNIb_9 and TEV-NIb_2-ΔNIb_9 have a low infectivity, the ratio was 10 genome equivalents set to 1 of TEV-eGFP. The mixture was mechanically inoculated on the same species of host plant on which it had been evolved, using three independent plant replicates per viral lineage. The plant leaves were collected at 7 dpi, and stored at −80 ºC. The plant leaves of TEV-eGFP evolved in *N. benthamiana* and *D. stramonium* were harvested 10 dpi. Total RNA was extracted from 100 mg homogenized tissue. RT-qPCR for the CP was used to determine total viral accumulation, and independent RT-qPCR reactions were also performed for the eGFP or mCherry sequence using specific primers. The ratio of the evolved and ancestral lineages to TEV-eGFP (*R*) is then $R=(n_{CP}-n_{eGFP}/n_{eGFP})$, where n_{CP} and n_{eGFP} are the RT-qPCR measured copy numbers of *CP* and *eGFP*, respectively. To calculate ratio of the evolved an ancestral lineages to TEV-mCherry, n_{eGFP} is replaced for $n_{mCherry}$ in the formula. Then we can estimate the within-host competitive fitness as $W=\sqrt[t]{R_t/R_0}$, where R_0 is the ratio at the start of the experiment and R_t the

ratio after *t* days of competition (Carrasco *et al.* 2007). Note that the method for determining *R* only works well when the frequency of the common is below ~ 0.75. This limitation was not problematic though, since in these experiments the fitness of the evolved virus populations remained the same or increased.

5. Sequencing

5.1. Sanger

For those evolved virus populations in which deletions were detected by RT-PCR, the exact positions of these deletions were determined. The genomes were partly sequenced by Sanger's method. RT was performed using AccuScript Hi-Fi (Agilent Technologies) reverse transcriptase and a reverse primer outside the region to be PCR-amplified for sequencing. PCR was then performed with Phusion DNA polymerase (Thermo Scientific) and primers flanking the deletions. Sanger sequencing was performed at GenoScreen (Lille, France: www.genoscreen.com) with an ABI3730XL DNA analyzer. Several sequencing reaction were done to obtain a coverage of at least 2 for each of the amplicons.

5.2. Illumina

For Illumina next-generation sequencing (NGS) of the evolved and ancestral lineages, the viral genomes were amplified by RT-PCR using AccuScript Hi-Fi (Agilent Technologies) reverse transcriptase and Phusion DNA polymerase (Thermo Scientific), with six independent replicates that were pooled. Each virus was amplified using three primer sets, generating three amplicons of

similar size. Equimolar mixtures of the three PCR products were made. Sequencing was performed at GenoScreen. Illumina HiSeq2500 2×100bp paired-end libraries with dual-index adaptors were prepared along with an internal PhiX control. Libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina Inc.). Sequencing quality control was performed by GenoScreen, based on PhiX error rate and Q30 values.

6. Bioinformatic analyses

6.1. Statistical analyses

The statistical analyses were performed using R (R Core Team 2014). The nonlinear regression analysis and the generalized linear model (GLM) comparisons were done using IBM SPSS Statistics version 23.

6.2. Sanger reads assembly

Sequences were assembled using Geneious v.8.0.3 (www.geneious.com) and the start and end positions of the deletions were determined. Based on the ancestral reference sequences, new reference sequences were constructed for each of the evolved lineages.

6.3. Illumina reads mapping, variant and SNP calling

Read artifact filtering and quality trimming (3' minimum Q28 and minimum read length of 50 bp) was done using FASTX-Toolkit v.0.0.14

(http://hannonlab.cshl.edu/fastx_toolkit/index.html). De-replication of the reads and 5' quality trimming requiring a minimum of Q28 was done using PRINSEQ-lite v.0.20.4 (Schmieder and Edwards 2011). Reads containing undefined nucleotides (N) were discarded. As an initial mapping step, the evolved sequences were mapped using Bowtie v.2.2.6 (Langmead and Salzberg 2012) against their corresponding ancestral sequences. Subsequently, mutations were detected using SAMtools' mpileup (Li *et al.* 2009) in the evolved lineages as compared to their ancestral lineage. At this point, we were only interested in mutations at a frequency > 10%. Therefore we present frequencies as reported by SAMtools, which has a low sensitivity for detecting low-frequency variants (Spencer *et al.* 2014).

After the initial pre-mapping step, error correction was done using Polisher v2.0.8 (available for academic use from the Joint Genome Institute) and consensus sequences were defined for every lineage. Subsequently, the cleaned reads were remapped using Bowtie v.2.2.6 against the corresponding consensus sequence for every lineage. For each new consensus, SNPs within each virus population were identified using SAMtools' mpileup and VarScan v.2.3.9 (Koboldt *et al.* 2012). For SNP calling maximum coverage was set to 40000 and SNPs with a frequency < 1% were discarded.

6.4. Mapping large genomic deletions

For the genomes were large genomics deletions were detected by RT-PCR and/or at the pre-mapping step of the Illumina reads, the most common deletions observed were defined manually, and for every lineage, a new reference sequence was constructed masking each position of the defined

deletion with the symbol N. These new reference genomes, together with the cleaned reads, were used as input for the program GapFiller v1.9 (Boetzer and Pirovano 2012), which reliably closes gaps within preassembled scaffolds using paired reads. GapFiller fills the gap from each edge in an iterative manner. In our case, it partially closed the gaps, base by base, until it could not extend any further given the difference between the a priori estimated deletion size and the actual size encountered. At both sides, overlapping sequences were manually identified and the ends were joined to reconstruct the new consensus sequences.

7. Modeling the stability of gene insertions

The model as described in equations 1 and 2 (**Page 108**) was implemented in R version 3.1.0 in order to estimate the parameter Δ, the recombination rate. The model describes how a population composed initially of only virus variants with a gene duplication (variant *A*) acquires and eventually fixes a new variant that only retains the original copy of the duplicated gene (*B*). In this section, those methodological details which are not described in the results section are covered. Note that the model includes a genetic bottleneck at the start of each round of passaging (*i.e.,* the initiation of infection in the inoculated leaf), with a fixed total number of founders λ and binomially distributed number of founders for variants containing the gene duplication. Following this genetic bottleneck, there is deterministic growth of both variants as well as deterministic recombination of *A* into *B*.

All model parameters except Δ could be estimated *a priori* for each virus with a gene duplication (**Table M.2**). To fit the model to the data, we considered model

predictions of the frequency of three kinds of virus populations over time: (*i*) those populations containing only the full-length ancestral virus with a gene duplication (X_1), (*ii*) those populations containing only variants with a genomic deletion removing the artificially introduced second gene copy (X_2), and (*iii*) those populations containing a mixture of both variants (X_3). Model predictions for the frequency of the different classes were obtained by iterating the model one-thousand times for a given set of passaging conditions. As we used a PCR-based method with limited sensitivity to characterize experimental populations, we assumed that the predicted frequency of *A* must be greater than 0.1 and less than 0.9 to be considered a mixture. We then compared model predictions for the frequency of the three different kinds of virus populations with the data by means of the multinomial likelihood. The likelihood of the number of occurrences of these three stochastic variables denoting observations of a particular kind of population (X_1, X_2, X_3) follows a multinomial distribution with probabilities p_1, p_2 and p_3 ($\sum_{i=1}^{3} p_i = 1$). The multinomial probability of a particular realization (x_1, x_2, x_3) is given by:

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = (\sum_{i=1}^{3} x_i)! \prod_{i=1}^{3} p_i^{x_i} / \prod_{i=1}^{3} x_i!$$

The negative log likelihood (NLL) was minimized by means of grid searches, considering all values of log(Δ) between −20 and −0.1, with intervals of 0.1. We first fitted the model with a single value of Δ to all the data (Model 1; 1 parameter). Next, we fitted the model with a virus-dependent value of Δ, but one which is independent of passage duration (Model 2; 4 parameters). We then fitted the model with Δ value dependent on passage duration, but the same for

each virus (Model 3; 2 parameters). Finally, we fit the model to each experimental treatment separately (Model 4; 8 parameters). For all these different model fittings, 95% fiducial estimates of Δ were obtained by fitting the model to 1000 bootstrapped datasets.

Table M2. Model parameters

Parameter	Value	Explanation
λ	500	Number of founders of infection (Zwart *et al.* 2014).
$\kappa_{t = 9\ weeks}$	4×10^9	Final value time-varying carrying capacity (9 weeks post-infection), weight of leaves multiplied by carrying capacity as estimated (Zwart *et al.* 2012).
s	1.344	Initial growth rate (per generation) for virus with single gene copy (Zwart *et al.* 2012).
r	φs	Initial growth rate for virus with a duplicated gene, where φ is the relative fitness of the virus with duplications compared to the virus with a single gene copy (see results)
g	2.91	Generations per day (Martínez *et al.* 2011)
β	s/r	The effect of *A* the replication of *B*
α	r/s	The effect of *B* the replication of *A*

Chapter 1: Multiple barriers to the evolution of alternative gene orders

1. Introduction

The organization of genes within a genome can vary greatly between phylogenetically distant species. Several comparative studies of bacterial, archaeal and eukaryotic genomes have concluded that in general gene order is not conserved (Watanabe *et al.* 1997; Himmelreich *et al.* 1997; Kolstø 1997; Siefert *et al.* 1997; Koonin and Galperin 1997; Dandekar *et al.* 1998; Rocha 2008). In stark contrast, gene order within virus orders and families is often conserved. Viral genomes tend to be smaller, with minimal intergenic sequences and in some cases overlapping genes (Lynch 2006; Belshaw *et al.* 2007; Koonin 2009). The reasons why a particular gene order supports the required patterns of virus-gene expression and virus replication have in many cases also been elucidated. For example, different expression levels for viral gene products can arise through the generation of subgenomic RNAs (de Haan *et al.* 2003), frameshifts (Chung *et al.* 2008), stuttering of the RNA polymerase in intergenic regions (Wertz *et al.* 1998), having multiple genome segments with different regulatory elements (Sullivan and Ahlquist 1997), or varying the frequency of different genome segments (Sicard *et al.* 2013). Altering gene order in viral genomes can therefore be associated with great fitness costs (Novella *et al.* 2004; Springman *et al.* 2005), and rearrangement of essential genes is not always reversible (Wertz *et al.* 1998). Nevertheless, it is not always obvious why, in viruses, gene order has been so well conserved.

Phylogenetic approaches have helped unveil interesting patterns in gene-order evolution. An intriguing example is the endornaviruses, found in plants, fungi and protists (Valverde *et al.* 1990; Wakarchuk and Hamilton 1990; Fukuhara *et al.* 2006), which have acquired domains with similar functions from these different hosts (Song *et al.* 2013). Despite their distinct origins, these domains have a strict functional order within the *Endornavirus* genus (Roossinck *et al.* 2011; Song *et al.* 2013), even though they highly vary as to presence or absence. Whereas phylogenetic approaches can identify patterns in gene-order evolution, experimental evolution could potentially shed light on the short-term dynamics and underlying mechanisms. The evolution of gene order has been experimentally explored for phage T7 (Springman *et al.* 2005) and *Vesicular stomatitis virus* (VSV) (Pesko *et al.* 2015).

T7 has a double-stranded DNA genome of ~40 kb. The T7 genome contains three promoters for the *Escherichia coli* RNA polymerase, and host-mediated transcription draws the first part of the T7 genome into the cell. Once the T7 RNA polymerase protein in this early region is expressed, it initiates transcription for the rest of the genome from its associated promoters, internalizing the remaining part of the T7 genome and achieving a high level of transcription of the late genes. The artificial repositioning of the T7 RNA polymerase downstream of its normal location resulted in a delay of the phage life cycle and had severe impacts on viral fitness (Endy *et al.* 2000; Springman *et al.* 2005). Subsequent experimental evolution only led to a modest recovery in fitness (Springman *et al.* 2005). In one evolved line the RNA polymerase was restored to the wild-type position, but at the same time other genes in the T7 genome were relocated, and a full regain of fitness was not observed. This study demonstrates that gene order is important for fitness, and that the wild-type

levels of fitness are not rapidly re-evolved after reorganizing the genome.

VSV is a non-segmented negative-strand RNA virus, with a genome size of ~11.2 kb, encoding five proteins. Transcription of VSV is regulated by a single promoter located at the 3' end of the genome. The stuttering of the VSV RNA polymerase causes greater mRNA production in upstream genes, which is a strategy to regulate gene expression. Gene order in VSV was altered by moving the nucleocapsid (*N*) gene, located at the 3' end, sequentially downstream in the genome (Wertz *et al.* 1998). This led to a stepwise decrease in *N* mRNA production and protein expression (Wertz *et al.* 1998). The initial fitness of the VSV variants was low (Novella *et al.* 2004), however, fitness gains were observed in evolutionary time and fitness improved the most for the variant with the lowest initial fitness (Pesko *et al.* 2015). Nevertheless, the variant with the wild-type gene order still grew better than the other variants.

T7 and VSV are different in nature, gene content, structure, and both use different replication strategies. Despite these differences, gene order is important for the regulation of gene expression in both viruses. Moreover, the different constraints on gene-order evolution observed in these two studies raise the question of their general applicability. Do most viruses and viral genome architectures suffer from the major constraints, as has been observed for T7? What about viruses that do not use promoters for the transcription of mRNA, like the positive-strand RNA viruses? Many emerging viruses with large societal impacts, as well as viral model systems with great relevance to fundamental research, are positive-strand RNA viruses, making it relevant to address these questions.

To study the evolution of alternative gene orders in positive-strand RNA virus,

in the context of a real multicellular-host infection, we used the picorna-like *Tobacco etch virus* (TEV; genus *Potyvirus*, family *Potyviridae*). TEV has a 9.5 kb genome that codes for a single polyprotein that is further processed into eleven mature peptides (**Fig C1.1A**). As it is composed of positive-strand RNA, TEV genome can immediately be translated upon entering a cell. Unlike the bacteriophage T7 and VSV, replication of TEV is not regulated by a promoter, but by the VPg protein linked to the 5' end of the genome which helps initiate RNA replication (Puustinen and Mäkinen 2004). Then, the viral NIa-Pro protease is responsible for processing the polyprotein at most of its proteolytic cleavage sites (Revers and García 2015), except for the processing of the first two proteins, P1 serine protease and HC-Pro cysteine protease, which are self-cleaving. Therefore, the rate of synthesis of the mature proteins depends on three factors: the amount of positive-strand RNAs accessible to ribosomes, the rate and effectiveness of translation into the polyprotein and the efficiency of its proteolysis by the viral proteases. Within the *Potyviridae* family, gene order has been strictly conserved, including the *Bymovirus* genus that has evolved a segmented bipartite genome (Revers and García 2015).

Given the need for correct polyprotein processing, a polyprotein-mediated gene-expression strategy is likely to impose constraints on gene order evolution. Rearranged viral genomes must conserve proteolytic cleavage sites, and as a consequence most recombination events are likely to disrupt polyprotein processing. However, even if they do not, would the resulting viruses be viable? The RdRp of TEV is coded by the *NIb* gene, and deletion thereof leads to virus variants that cannot replicate on their own (Li and Carrington 1995). In a previous study we considered whether virus infectivity was maintained when the *NIb* was relocated to all possible intergenic positions in the TEV genome,

without maintaining the original *NIb* copy (Majer *et al.* 2014). Only 2/9 viruses with reordered genomes were viable: the genotypes with the *NIb* relocated to the first two intergenic positions (**Fig C1.1B**). A variant with the *NIb* relocated to the third intergenic position was not infectious in wild-type plants, whilst it could cause infection in transgenic plants expressing NIb *in trans*, albeit at a lower frequency (Majer *et al.* 2014). Moreover, in these cases, we always found an exact deletion of the *NIb* and therefore we do not consider this a viable virus for reordering.

In this study, we used experimental evolution to better understand the dynamics of genome-architecture evolution and gene order conservation. Why has gene order been conserved within positive-strand RNA virus orders and families? Are there accessible evolutionary trajectories to alternative orders, or is a lack of accessible trajectories an important impediment in present-day viruses? Here we use the two viable reordered TEV variants to address these issues. We first explore whether there are accessible evolutionary trajectories that result in these two viable reordered viruses. To consider whether such natural trajectories exist, genomes containing duplications of the *NIb* were generated, tested for viability, and evolved in plants. We consistently found that the *NIb* at the alternative position was rapidly lost due to the occurrence of large genomic deletions. Finally, we explored the evolutionary potential of viruses with reordered genomes, by evolving viruses with a single *NIb* in an alternative position. We then measured virus accumulation and fitness and used next generation sequencing to identify genomic changes. Although we found evidence for adaptation of these reordered viruses in terms of increasing virus accumulation, they were still less fit than the wild-type virus. This study therefore revealed multiple barriers to the evolution of alternative gene orders.

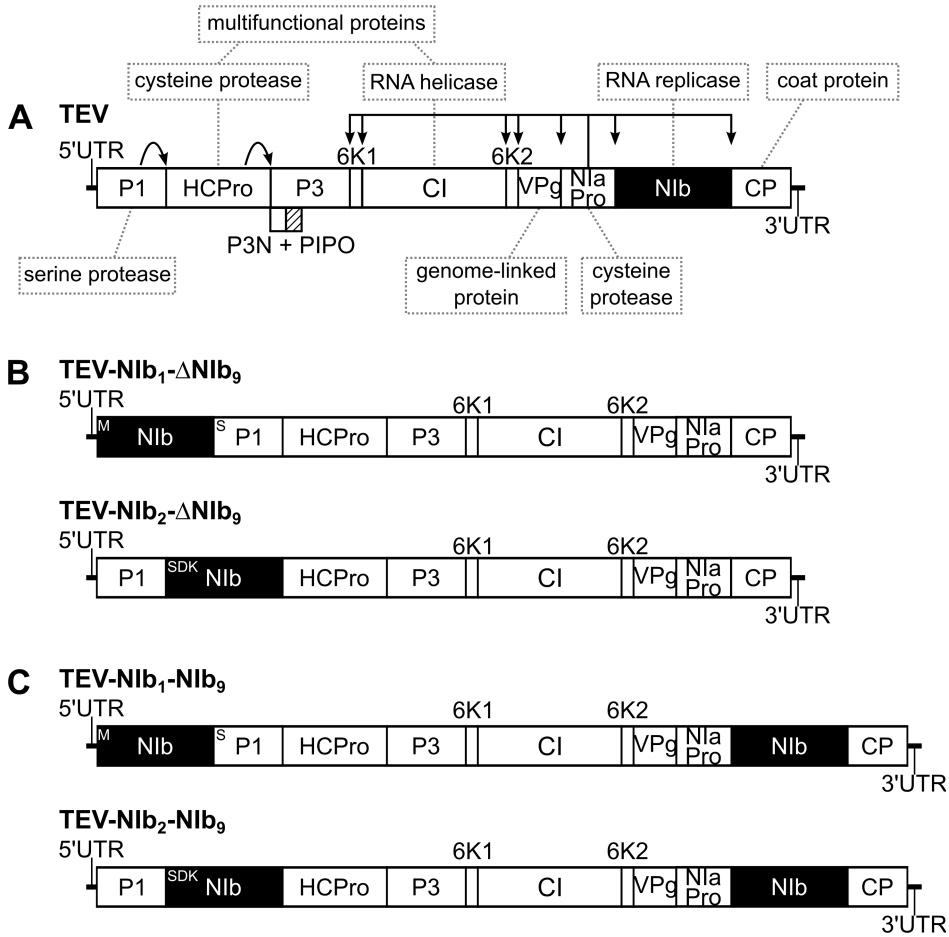

Figure C1.1. Schematic representations of the different *Tobacco etch virus* genotypes used in this study. (**A**) The wild-type TEV. (**B**) For constructing the viral genotypes with the NIb replicase in an alternative position, a copy of the *NIb* was moved to the first and the second position within the TEV genome; removing *NIb* from its original position. (**C**) For construction the viral genotypes with a duplication of the *NIb*, a second copy of the *NIb* was introduced into the first and the second position within the TEV genome; preserving the original *NIb* and hereby generating a duplication event. For simplification P3N-PIPO is only drawn at the wild-type TEV.

2. Results

2.1. Study framework: plausible evolutionary trajectories to alternative gene orders

For the rearrangement of genes in a viral genome, there are many potential evolutionary trajectories that can be envisioned. Here, we use "plausible" to denote trajectories that could be traversed by the virus in terms of consistently maintaining replication (*i.e.,* virological perspective), whereas we use the term "accessible" to denote trajectories that could be traversed if the fitness and stability of intermediate steps are considered (*i.e.,* evolutionary perspective). In the potyvirus model we are considering, a number of constraints conspire to effectively limit the number of plausible trajectories to one. First, we are considering repositioning of the essential *NIb* replicase, which must be present in every cell that will contribute to infection and between-host transmission. When plants are inoculated with multiple potyvirus genotypes, the observed rate of cellular co-infection is typically very low, with the main exception being early infection prior to systemic movement (Dietrich and Maiss 2003; Zwart *et al.* 2011; Tromas *et al.* 2014a; Gutiérrez *et al.* 2015). It is therefore not surprising that whilst TEV missing the *NIb* (TEV-ΔNIb) can autonomously infect plants expressing NIb (Li and Carrington 1995), it cannot co-infect wild-type tobacco plants when co-inoculated with a wild-type virus (Tromas *et al.* 2014b). Since each intermediate along a reordering trajectory must be capable of autonomous replication, a plausible trajectory for an essential gene – like the *NIb* – will necessarily involve a gene duplication event. Second, although higher gene expression as a consequence of gene duplication may have benefits, the increase in genome size and the possible disruption of the expression of

other genes could significantly reduce viral fitness. Therefore, the most plausible trajectory to a new gene order is gene duplication, followed by the deletion of the ancestral copy (**Fig C1.2**). A complication in this process is that variants with a deletion of the new gene copy may very well be favored, due to the fine-tuning of polyprotein processing and expression levels at this position. A deletion of the new gene copy would, however, be a *cul-de-sac* along this evolutionary trajectory (**Fig C1.2**, variant *c*). Alternatively, viruses with only the new copy may be fixed (**Fig C1.2**, variant *d*), successively followed by the evolutionary fine-tuning (*e.g.,* adaptive mutations) of the relocated gene (**Fig C1.2**, variant *e*).

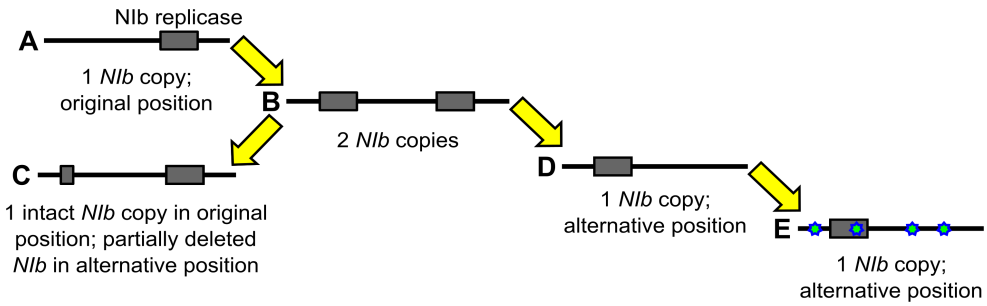

Figure C1.2. Plausible evolutionary trajectory for reordering of an essential gene in a potyvirus genome. First the essential gene must be duplicated (variant *a* → variant *b*), to ensure the virus can replicate autonomously at all times. Subsequently, we anticipate that one copy of the duplicate gene will be lost; either the second new copy is deleted (variant *c*), or the alternative gene copy is deleted (variant *d*). In the latter event, further evolution will take place to accommodate this reordered genome (variant *e*).

Given that biological constraints strongly suggest that an evolutionary trajectory through gene duplication is the most plausible route to gene reordering, we decided to study this route in detail for the repositioning of *NIb* in the TEV

genome. To get a complete picture of the likelihood that a new gene order can evolve by means of this route, we decided to consider the following five key steps: (*i*) the fitness of viruses with gene duplications, rendering an indication of how long such variants can persist. (*ii*) The evolutionary potential of viruses with gene duplications: focusing on the stability of the new gene copy, as this will show whether they can act as a bridge to the evolution of a new gene order. (*iii*) The fitness of viruses with a single *NIb* copy in an alternative position: to determine the likely fate of such viruses in the background of its direct ancestor; the corresponding double *NIb* genotype. (*iv*) The evolutionary potential of viruses with a single *NIb* copy in an alternative position: since should such a virus occur and have low fitness, we can infer whether – following a period of reproductive isolation – they could eventually be competitive with wild-type viruses.

From the outset, however, we are faced with a barrier to the evolution of alternative gene orders for TEV: there are only two alternative positions to place the *NIb* replicase for which viruses appear to be viable (Majer *et al.* 2014): (*i*) before the *P1* serine protease gene and (*ii*) between *P1* and the *HC-Pro* cysteine protease genes (**Fig C1.1B**). A first barrier to the evolution of alternative gene order is therefore the number of potentially viable intergenic sites, which is limited to only two out of nine for TEV. Recombination events leading to the movement of a gene, as well as conservation of the reading frame and polyprotein processing will be rare, and in addition, all other things being equal 7/9 of these events will lead down trajectories that are ultimately *cul-de-sacs*. The number of viable alternative positions therefore means the effective supply of first-step recombinants leading to alternative gene order is almost 10-fold smaller than suggested by the mutational supply alone.

We therefore focus on those evolutionary pathways to reach the two viable alternative gene orders, and consequently, four different TEV genotypes were constructed. The *NIb* was inserted at the first and the second positions in the TEV genome, whilst preserving *NIb* at the original position (**Fig C1.1C**). Henceforth we refer to these viruses with a duplication of *NIb* as TEV-NIb_1-NIb_9 and TEV-NIb_2-NIb_9 (**Table C1.1**), with subscripts denoting the intergenic position of *NIb*. Note that NIb_9 is referring to *NIb* at its original position. We also generated viruses in which *NIb* was moved to alternative positions and the original gene was deleted (**Fig C1.1B**). We refer to these viruses with a single *NIb* copy at alternative positions as TEV-NIb_1-ΔNIb_9 and TEV-NIb_2-ΔNIb_9 (**Table C1.1**). For all viruses generated, the termini of the relocated NIb were adjusted such that this protein is properly translated and sites for cleavage from the viral polyprotein are provided (**Fig C1.1**), similar to the original proteolytic cleavage sites at the corresponding positions. For these four key intermediate viruses along an evolutionary trajectory to altered gene order, we could then measure virulence, viral accumulation, competitive fitness, and study their evolutionary potential.

Table C1.1. TEV genotypes constructed for this study

Viral genotype	Original *NIb* gene present?	Alternative position of *NIb* gene
TEV-NIb_1-ΔNIb_9	No	1
TEV-NIb_2-ΔNIb_9	No	2
TEV-NIb_1-NIb_9	Yes	1
TEV-NIb_2-NIb_9	Yes	2

2.2. Viruses with a duplication of the NIb have reduced fitness and accumulation

TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ were reconstituted from infectious clones and inoculated in *Nicotiana tabacum* L. cv. Xanthi nc (NN) plants. The collected tissue from these plants served as the starting material for all succeeding experiments. Subsequently, we measured within-host competitive fitness (*W*), by individually competing the two viruses with *NIb* duplications against the wild-type virus carrying a GFP marker (TEV-eGFP), and viral accumulation, by measuring the number of virions (genome equivalents) present in the host plant, in the absence of a competing virus. After normalizing the number of virions for each viral genotype to the same concentration, both competition and accumulation experiments were performed for a total duration of one week. For both ancestral viruses TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$, we observed statistically significant decreases – as compared to the ancestral wild-type virus – in competitive fitness (**Fig C1.3A**; TEV-NIb$_1$-NIb$_9$: $t_4 = 6.379$, $P = 0.003$; TEV-NIb$_2$-NIb$_9$: $t_4 = 8.348$, $P = 0.001$), and accumulation (**Fig C1.4A**; TEV-NIb$_1$-NIb$_9$: $t_4 = 45.097$, $P < 0.001$; TEV-NIb$_2$-NIb$_9$: $t_4 = 8.650$, $P < 0.001$). Compare the light grey bars labeled “ancestral” in panel A of both **Fig C1.3** and **Fig C1.4**, to see how the duplicated viruses performed compared to the wild-type virus. Due to the increase in genome size, we were not surprised that duplication of the *NIb* leads to a virus with reductions in these fitness components. Nevertheless, this observation has an important implication; the first step that must be taken along the plausible evolutionary route we have suggested is already unlikely. Viruses with *NIb* duplications cannot be maintained in virus populations for long periods of time. These viruses must therefore establish a bridgehead in viral populations by means of genetic drift,

from whence they can continue along the evolutionary trajectory to alternative gene orders. Consequently, the low fitness of viruses with *NIb* duplications constitutes a second barrier to reordering.

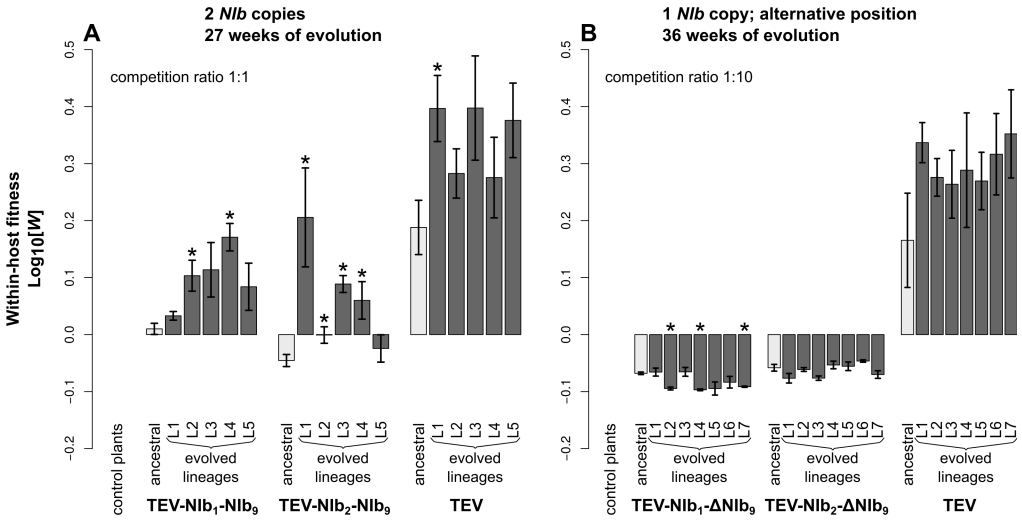

Figure C1.3. Within-host competitive fitness. Fitness (*W*), as determined by competition experiments and RT-qPCR, of the different viral genotypes with: a duplication of the *NIb* (**A**), the *NIb* moved to an alternative position (**B**), the wild-type TEV and healthy control plants, with respect to a common competitor; TEV-eGFP. In panel A, the competition experiment was started with an inoculum containing a 1:1 mixture of both competitors, while a 1:10 mixture (common competitor : virus of interest) was used in panel B. The ancestral lineages are indicated with light-gray bars and the evolved lineages with dark-gray bars. The viruses in panel A were evolved using five replicate lineages each (L1-L5), for a total of 27 weeks, and viruses in panel B were evolved using seven replicate lineages each (L1-L7) for a total of 36 weeks. Evolved lineages that tested significantly different compared to their ancestral lineage are indicated with an asterisk (*t*-test with Holm-Bonferroni correction for multiple tests).

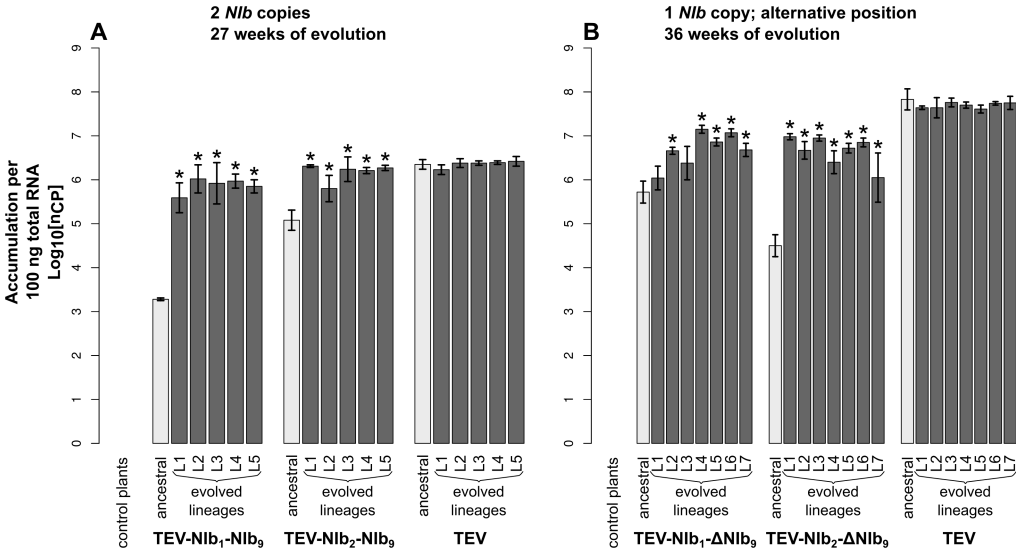

Figure C1.4. Virus accumulation of the evolved and ancestral lineages. Virus accumulation, as measured by RT-qPCR, of the different viral genotypes with: a duplication of the *NIb* (A), the *NIb* moved to an alternative position (B), the wild-type TEV and healthy control plants. The ancestral lineages are indicated with light-gray bars and the evolved lineages with dark-gray bars. The viruses in panel A were evolved using five replicate lineages each (L1-L5), for a total of 27 weeks, and viruses in panel B were evolved using seven replicate lineages each (L1-L7) for a total of 36 weeks. Evolved lineages that tested significantly different compared to their ancestral lineage are indicated with an asterisk (*t*-test with Holm-Bonferroni correction for multiple tests).

2.3. An evolutionary cul-de-sac: after duplication, the NIb is pervasively deleted from an alternative position

The next key stage in the evolution of alternative gene order is to determine the evolutionary potential of viruses with gene duplications, TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$. The genotypes containing two *NIb* copies (Fig **C1.1C**) were therefore evolved in *N. tabacum* plants for a total of 27 weeks, using both nine 3-week passages and three 9-week passages. Our choice for passage duration

was based on a previous study where we showed that selection appears to act more strongly for longer-duration passages, whilst non-functional sequences are more stable during shorter passages (Zwart *et al.* 2014). Hence, these conditions may fulfill requirements for further evolution of duplicated viruses, by (*i*) retaining the duplicated gene copy, whilst (*ii*) still allowing for selection – and not mainly genetic drift – to act on these virus populations. At the start of the evolution experiment the TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ viruses had a reduced infectivity and showed little or no symptoms, consisting of < 7 sparse small chlorotic spots per leaf. These symptoms are much weaker than typical symptoms observed for wild-type TEV, which consist of vein clearing, mosaic mottling, chlorosis, and stunting of the tobacco leaves together with reduced plant growth (Velasquez *et al.* 2014; Revers and García 2015). However, during the evolution experiment these symptoms changed to wild-type like symptoms, indicated by the green lines in Fig 5 In the first 3- and 9-week passages the reduced symptoms turned into mild wild-type like symptoms for the TEV-NIb$_1$-NIb$_9$ lineages (**Fig C1.5**). For the TEV-NIb$_2$-NIb$_9$ lineages, mild wild-type symptoms appear later; for the 3-week lineages in passage 3 (9 weeks on the *x*-axis in **Fig C1.5**) and for the 9-week lineages in passage 2 (18 weeks on the *x*-axis in **Fig C1.5**). At the end of the evolution experiment (27 weeks) all lineages of both viruses showed wild-type like symptoms in tobacco plants.

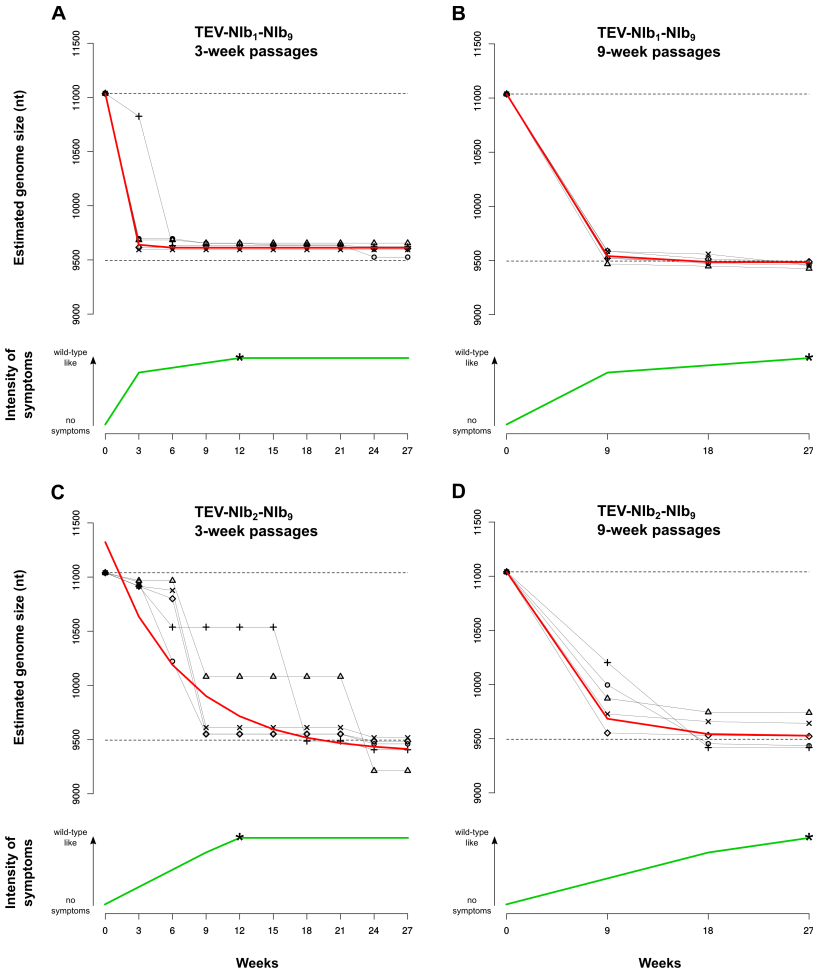

Figure C1.5. Symptomatology and genome size over time of genotypes with a duplication of NIb. The different panels display how the genome size and the symptomatology of the TEV-NIb$_1$-NIb$_9$ (**A** and **B**) and TEV-NIb$_2$-NIb$_9$ (**C** and **D**) genotypes change along the evolution experiments, performing 3-week (**A** and **C**) and 9-week (**C** and **D**) serial passages. The green lines plot the observed severity of symptoms. The arrows on the left side of the plots indicates the gradient from no symptoms to wild-type like symptoms. The asterisk on each green line indicates the passage where we first observed wild-type like symptoms. The black symbols connected by the continuous black lines represent the estimated genome size of the different viral lineages along the evolution experiment. The upper dashes lines represent the genome size of the ancestral viruses and the lower dashes lines represent the genome size of the wild-type virus. The red line plots the median estimates of a non-linear regression model. Note that the reduction in genome size appears to be negatively correlated to the severity of symptoms developed.

Through RT-PCR, deletions were detected in the second alternative copy of the *NIb*, but never in the original one. Genome size for all the evolved lineages was estimated at every passage (**Fig C1.5**). As deletions occur, the viral genomes reduce in size to one that is similar to that of the wild-type TEV. An exponential decay model ($S(t)=S(0)+a\exp(-bt)$) was fitted to the estimated genome size (*S*) over time (*t*) for every lineage (red line in **Fig C1.5**) and the rates of size change, *b*, were compared by means of a generalized linear model (GLM) using a gamma probability distribution. Whereas *PASSAGE DURATION* did not have a significant effect on genome size, both virus *GENOTYPE* and the interaction *GENOTYPE* × *PASSAGE DURATION* were significant (**Table C1.2**).

Table C1.2. GLM analysis of the genome size data

Term	LRT	df	*P*
Intercept	28.420	1	< 0.001
GENOTYPE	10.280	1	0.001
PASSAGE DURATION	0.005	1	0.943
GENOTYPE × *PASSAGE DURATION*	4.707	1	0.030

The alternative *NIb* copy was deleted more quickly in the TEV-NIb_1-NIb_9 lineages as compared to the lineages from TEV-NIb_2-NIb_9 (**Fig C1.5**; compare panels A and B to panels C and D), whereas the interaction term suggests the effect of genotype is particularly strong for the 3-week passages (**Table C1.2**). Additionally, in the TEV-NIb_2-NIb_9 lineages there appears to be more variation in the time points at which the second *NIb* copy was deleted along the evolution experiment (**Fig C1.5**; panels C and D). The alternative *NIb* copy from TEV-

NIb_2-NIb_9 is therefore more stable, suggesting that the second position in the TEV genome is a more accessible trajectory for duplication of a gene and subsequent reorganization within the genome. The decrease in genome size also appears to correlate with the appearance of stronger symptoms (**Fig C1.5**). At the end of the evolution experiment all lineages had a genome size very similar to that of the wild-type. All TEV-NIb_1-NIb_9 3-week lineages and 4/5 9-week lineages even evolved to a genome size smaller than the wild-type virus by deleting a part of the 5' UTR. For TEV-NIb_2-NIb_9, 4/5 3-week lineages and 3/5 9-week lineages evolved to a smaller genome size by deleting part of the HC-Pro cysteine protease.

For all the evolved lineages, we then measured within-host competitive fitness and accumulation (**Fig C1.3A** and **C1.4A**). Compare bars labeled "ancestral" and "evolved lineages" in **Fig C1.3A**. When pairwise comparisons were made between the ancestral virus and the evolved lineages (*t*-test with Holm-Bonferroni correction), significant increases in within-host competitive fitness were found for 2/5 TEV-NIb_1-NIb_9 lineages and for 4/5 TEV-NIb_2-NIb_9 lineages (**Fig C1.3A**; asterisks). However, fitness never reaches the fitness of the wild-type, and therefore we did find a significant effect of treatment (ANOVA with *post hoc* Tukey HSD) comparing the evolved TEV-NIb_1-NIb_9 and TEV-NIb_2-NIb_9 to the evolved TEV lineages (**Table C1.3** and **C1.4**). Now compare bars labeled "ancestral" and "evolved lineages" in **Fig C1.4A**. Whereas no significant increases in accumulation were found comparing the ancestral virus and the evolved lineages of the wild-type TEV, we did find significant increases in viral accumulation for all the evolved lineages of both TEV-NIb_1-NIb_9 and TEV-NIb_2-NIb_9. And these lineages reached similar accumulation levels as the evolved wild-type lineages. However, we did find a significant effect of

treatment on accumulation comparing the evolved lineages of the viral genotypes, due to the differences between the evolved TEV-NIb$_1$-NIb$_9$ and the wild-type TEV lineages (**Table C1.3** and **C1.4**).

Table C1.3. Nested ANOVAs on accumulation and within-host fitness of evolved lineages

	Trait	Source of variation	df	*SS*	*F*	*P*
2 *NIb* copies	Accumulation	Treatment	2	1.818	19.476	< 0.001
		Lineage within treatment	12	0.922	1.646	0.131
		Error	30	1.400		
	Within-host fitness	Treatment	2	0.665	117.296	< 0.001
		Lineage within treatment	12	0.208	6.124	< 0.001
		Error	32	0.091		
1 *NIb* copy; alternative position	Accumulation	Treatment	2	14.451	18.236	< 0.001
		Lineage within treatment	18	4.778	6.699	< 0.001
		Error	42	1.664		
	Within-host fitness	Treatment	2	1.966	686.952	< 0.001
		Lineage within treatment	18	0.028	1.078	0.405
		Error	42	0.060		

The evolution of viruses with two *NIb* copies results in an increase in fitness related to the reduction in genome size. However, we consistently observed the deletion of the *NIb* at an alternative position, leading back to the ancestral wild-type virus, and we never observed the deletion of the *NIb* at its ancestral position. This evolutionary *cul-de-sac* therefore represents a third barrier to reordering.

Table C1.4. *Post hoc* Tukey HSD test on nested ANOVAs

	Trait	Treatments compared		*P*
2 *NIb* copies	Accumulation	TEV-NIb$_1$-NIb$_9$	TEV-NIb$_2$-NIb$_9$	0.002
		TEV-NIb$_1$-NIb$_9$	TEV	< 0.001
		TEV-NIb$_2$-NIb$_9$	TEV	0.053
	Within-host fitness	TEV-NIb$_1$-NIb$_9$	TEV-NIb$_2$-NIb$_9$	0.593
		TEV-NIb$_1$-NIb$_9$	TEV	< 0.001
		TEV-NIb$_2$-NIb$_9$	TEV	< 0.001
1 *NIb* copy; alternative position	Accumulation	TEV-NIb$_1$-NIb$_9$	TEV-NIb$_2$-NIb$_9$	0.832
		TEV-NIb$_1$-NIb$_9$	TEV	< 0.001
		TEV-NIb$_2$-NIb$_9$	TEV	< 0.001
	Within-host fitness	TEV-NIb$_1$-NIb$_9$	TEV-NIb$_2$-NIb$_9$	0.161
		TEV-NIb$_1$-NIb$_9$	TEV	< 0.001
		TEV-NIb$_2$-NIb$_9$	TEV	< 0.001

2.4. Whole genome sequences of evolved lineages of viruses with an NIb duplication

All evolved and ancestral lineages described in this study have been fully sequenced using the Illumina technology. The sequences of the ancestral lineages were used as an initial reference for the evolved lineages. Furthermore, for the genotypes that originally had two *NIb* copies, parts of the genome were sequenced by Sanger to determine the exact deletion sites, which have been previously detected by RT-PCR (**Fig C1.5**). The majority deletion variants were used to construct new reference sequences for each of the evolved TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ lineages. After the initial mapping step, mutations were detected in the evolved lineages as compared to their corresponding ancestor.

At the sequence level, the main changes in TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ were large genomic deletions in the second *NIb* copy. In other words, for both viruses we consistently observed pseudogenization of the duplicated *NIb* copy, in accordance with RT-PCR results. In TEV-NIb$_1$-NIb$_9$ none of these deletions included the N-terminus of P1, while for 9/10 lineages these deletions included the 3' end of the 5' UTR (**Fig C1.6**). Only in half of the TEV-NIb$_1$-NIb$_9$ lineages the reading frame is maintained after pseudogenization. As the deletion occurs at the start of the genome, it is not necessary to maintain the reading frame, as long as the original second methionine in P1 is preserved. In TEV-NIb$_2$-NIb$_9$ none of the deletions included the C-terminus of P1, but for 7/10 lineages the deletions included the N-terminal region of the HC-Pro (**Fig C1.6**), similar to results obtained by previous studies (Dolja *et al.* 1993; Zwart *et al.* 2014). HC-Pro is a multifunctional protein, the N-terminal region of HC-Pro is implicated in transmission by aphids (Thornbury *et al.* 1990; Atreya *et al.* 1992) and is not essential for replication and movement (Dolja *et al.* 1993; Cronin *et al.* 1995). In the TEV-NIb$_2$-NIb$_9$ genotype, the reading frame was maintained in the sequence of all lineages after pseudogenization. This could be simply explained by the fact that these lineages depend on only one methionine codon at the beginning of the polyprotein, at the start of the coding region in P1.

On the other hand, for these lineages little evidence for adaptive evolution was found at the level of single-nucleotide mutations (**Fig C1.6**). We determined mutations present in the evolved lineages with respect to their corresponding ancestral sequences. In the TEV-NIb$_1$-NIb$_9$ 3-week lineages we detected three high frequency (> 10%) convergent nonsynonymous mutations occurring in 2/5 lineages; located in the pseudogenized alternative *NIb* copy (A1643U), at the 3' end of *CI* (U7066C), and in *VPg* (U7703A). For TEV-NIb$_1$-NIb$_9$ 9-week

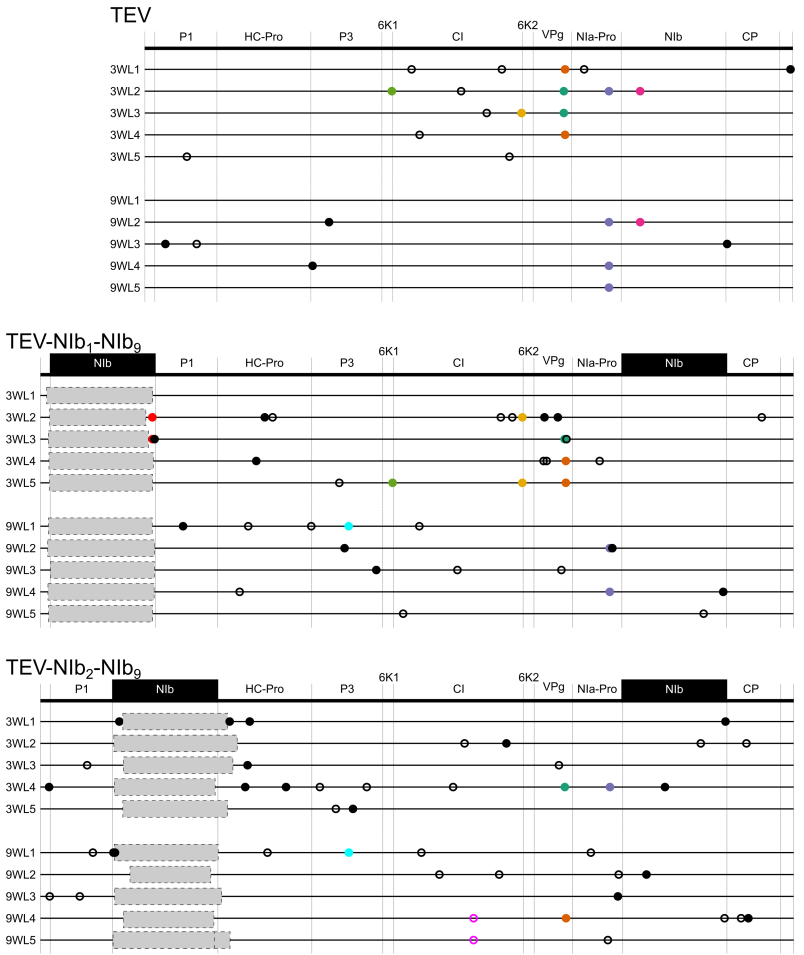

Figure C1.6. Genomes of the evolved lineages with a duplication of *NIb*. Mutations were detected using NGS data of the evolved TEV-NIb_1-NIb_9 and TEV-NIb_2-NIb_9 lineages as compared to their ancestral lineages. The wild-type TEV is given for comparative purposes. The names on the left identify lineages (*e.g.*, 3WL1 is the final population of 3-week passages, lineage 1). No NGS data was available for TEV-NIb_1-NIb_9 3WL1. Full circles and open circles represent nonsynonymous and synonymous substitutions, respectively. Black substitutions occur in only one lineage, whereas color-coded substitutions are repeated in two or more lineages. Note that the mutations are present at different frequencies as reported by SAMtools (> 10%). Grey boxes indicate genomic deletions in the majority variant, as detected by Sanger sequencing technology.

lineages we observed one convergent nonsynonymous mutation that occurred in 2/5 lineages in *NIa-Pro* (A8347G). The same mutation was also found in TEV wild-type 3- and 9-week lineages. For the TEV-NIb_2-NIb_9 3-week lineages, no repeated mutations were found and in the 9-week lineages one convergent synonymous mutation was found in *CI* (C6351U) for 2/5 lineages. After remapping the cleaned reads against a new defined consensus sequence for each lineage, we looked at the variation within each lineage. Single nucleotide polymorphisms (SNPs) were detected from a frequency as low as 1%. In the evolved TEV-NIb_1-NIb_9 lineages a total of 301 SNPs were detected, with a median of 36 (27-45) per lineage. In the evolved TEV-NIb_2-NIb_9 lineages a total of 220 SNPs were detected, with a median of 23.5 (4-44) per lineage. In both virus genotypes, most of the SNPs were present at low frequency, with a higher percentage of synonymous (TEV-NIb_1-NIb_9: 66.4%, TEV-NIb_2-NIb_9: 64.5%) versus nonsynonymous changes (**Fig C1.7**). However, the difference in the distribution of synonymous versus nonsynonymous SNP frequency was not significant (Kolmogorov–Smirnov test; TEV-NIb_1-NIb_9: $D = 0.146$, $P = 0.073$; TEV-NIb_2-NIb_9: $D = 0.144$, $P = 0.217$).

The results from the whole-genome data are congruent with RT-PCR results and phenotypic assays: the main change in the evolved lineages of the viruses with *NIb* duplications is the deletion of the second copy, which turns into a virus which is in all respects similar to the ancestral wild-type. Although there are some convergent single-nucleotide mutations, these occur only in a small fraction of lineages and moreover are often shared between the TEV, TEV-NIb_1-NIb_9 and TEV-NIb_2-NIb_9 lineages. These mutations therefore appear to represent general adaptations, without a strong link to the transient presence of the second *NIb* copy.

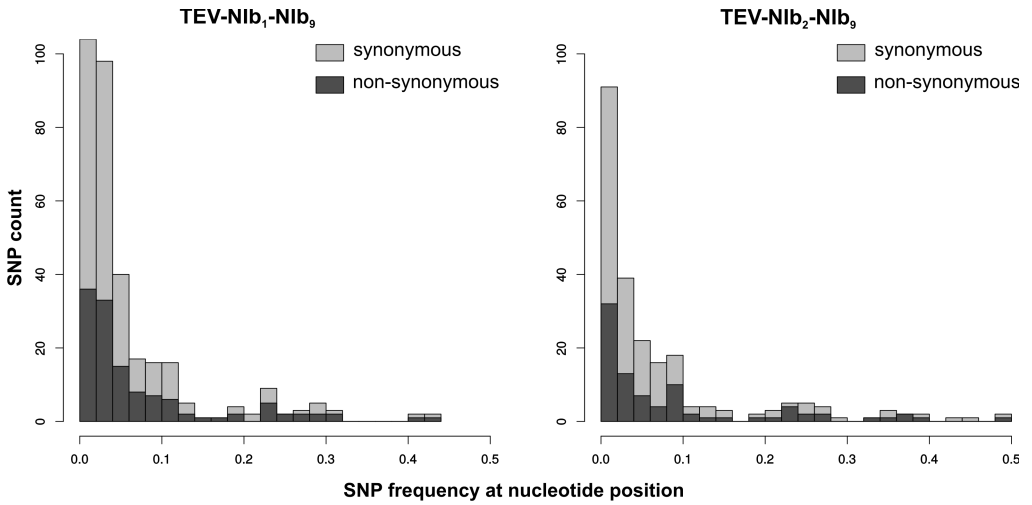

Figure C1.7. The distribution of SNP frequencies in the evolved TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ lineages.

2.5. Viruses with NIb moved to an alternative position have further reductions in fitness and viral accumulation

The viruses for which *NIb* was moved to an alternative position without conserving the original *NIb* copy, TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$, were reconstituted from infectious clones. Subsequently we measured their within-host competitive fitness and viral accumulation. Compare bars labeled "ancestral" in both **Fig C1.3B** and **Fig C1.4B**. For both TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$, we observed significant decreases – as compared to the wild-type virus – in competitive fitness (**Fig C1.3B**; TEV-NIb$_1$-ΔNIb$_9$: $t_4 = 4.897$, $P = 0.008$; TEV-NIb$_2$-ΔNIb$_9$: $t_4 = 4.692$, $P = 0.009$) and accumulation (**Fig C1.4B**; TEV-NIb$_1$-ΔNIb$_9$: $t_4 = 10.463$, $P < 0.001$; TEV-NIb$_2$-ΔNIb$_9$: $t_4 = 16.453$, $P < 0.001$). Deletion of the original *NIb* copy therefore leads to further reductions in

viral fitness, suggesting that TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$ will be rapidly outcompeted by both their direct ancestors (TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$, respectively) as well as viruses with deletions in the new copy of *NIb* (fitness similar to the wild-type virus). Therefore, we are confronted with a fourth evolutionary barrier: viruses with a single copy of *NIb* in an alternative position cannot outcompete the duplicated virus, meaning that they must be maintained, or probably fixed, by genetic drift to have the opportunity to undergo further evolution.

2.6. Limited short-term evolutionary potential of viruses with NIb moved to an alternative position

In the evolutionary trajectory we have postulated, the final step is the evolution of a virus with *NIb* only in an alternative position. If these viruses managed to occur through a series of chance events and could exist in isolation from their ancestral viruses for a period of time, what would their evolutionary fate be? Could these viral populations readily converge on a fitness peak that allowed them to be comparable or superior to the ancestral TEV?

To address these questions, TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$ were evolved in tobacco plants for a total of 36 weeks, using four 9-week serial passages. We did not perform 3-week passages, as we expected these genomes to be stable and therefore considered only a condition with maximal selection and minimal drift by intermittent bottlenecks. Both reordered viruses have very low infectivity, and serial passage was successfully completed for only seven out of 10 lineages. For both TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$, tobacco plants had very weak or no symptoms of infection, and these symptoms did not

become more severe over time. Therefore, infection had to be confirmed by RT-PCR. Two different regions of the reordered genomes were amplified using primers flanking the region of the new *NIb* position as well as primers flanking the original position of the removed *NIb*. No evidence was found of any insertions or deletions within these amplified sites, indicating that reordered viruses were stable over time and there was no presence of wild-type like viruses.

For all the evolved lineages, we measured within-host competitive fitness and accumulation (**Fig C1.3B** and Fig **C1.4B**). Compare bars labeled "ancestral" and "evolved lineages" in **Fig C1.3B**. Pairwise comparisons between the ancestral and evolved lineages showed there were no significant increases in within-host competitive fitness (**Fig C1.3B**), whilst in 3/5 TEV-NIb$_1$-NIb$_9$ lineages we even found significant decreases. Now compare bars labeled "ancestral" and "evolved lineages" in **Fig C1.4B**. Accumulation levels of the wild-type virus did not change significantly as compared to the ancestral TEV, while for 5/7 of the evolved TEV-NIb$_1$-ΔNIb$_9$ lineages and all TEV-NIb$_2$-ΔNIb$_9$ lineages accumulation increased significantly (**Fig C1.4B**; asterisks). However, these accumulation levels never reached the same levels as the wild-type virus. Comparing only the evolved lineages, there was a significant effect of treatment (ANOVA with *post hoc* Tukey HSD; **Table C1.3** and **C1.4**) on viral accumulation and within-host competitive fitness, indicating that the wild-type TEV outperforms the two reordered viruses for both fitness components.

2.7. Whole genome sequences of evolved lineages of viruses with a single NIb copy at an alternative position

We found evidence of adaptive convergent evolution comparing the evolved and ancestral lineages containing one reordered *NIb* copy (**Fig C1.8**). Mutations in the TEV-NIb$_1$-ΔNIb$_9$ lineages were found in (*i*) the reordered *NIb* gene at the first position (U428C), (*ii*) in *P1* around the proteolytic cleavage site of NIb and P1 (U1688C, U1697C and U1697A), and (*iii*) in *NIa-Pro* (U8210C and A8347G). Mutation U1688C modifies the start codon of *P1* (M563T). This is explained by the introduction of an additional start codon at the first position of the reordered genome that makes the original methionine redundant. Mutations in the TEV-NIb$_2$-ΔNIb$_9$ lineages were found (*i*) in the reordered *NIb* gene at the second position (G1066A, G1090A, G1264A, and U1346C), (*ii*) in *HC-Pro* (G3213U, A3632G and U3803C), (*iii*) in *P3* (A4016G), and (*iv*) in *NIa-Pro* (U8285C and A8350G). The former mutation in *NIa-Pro* (U8285C) was also found in one lineage of TEV-NIb$_1$-ΔNIb$_9$, and the latter mutation (A8350G) was also found in the evolved lineages of TEV-NIb$_1$-ΔNIb$_9$ and the wild-type TEV. Not a single mutation was detected in *VPg*, which is putatively involved in translation and replication.

As for the within population variation, in the evolved TEV-NIb$_1$-ΔNIb$_9$ lineages we detected 137 SNPs, with a median of 22 (7-37) per lineage. In the evolved TEV-NIb$_2$-ΔNIb$_9$ lineages 155 SNPs were detected, with a median of 22 (18-48) per lineage. In both virus genotypes, most of the SNPs were low frequency SNPs, with a higher percentage of synonymous changes (58.4%) in the TEV-NIb$_1$-ΔNIb$_9$ lineages, while in the TEV-NIb$_2$-ΔNIb$_9$ lineages the percentage of nonsynonymous changes was higher (63.9%, see **Fig C1.9**). However, no

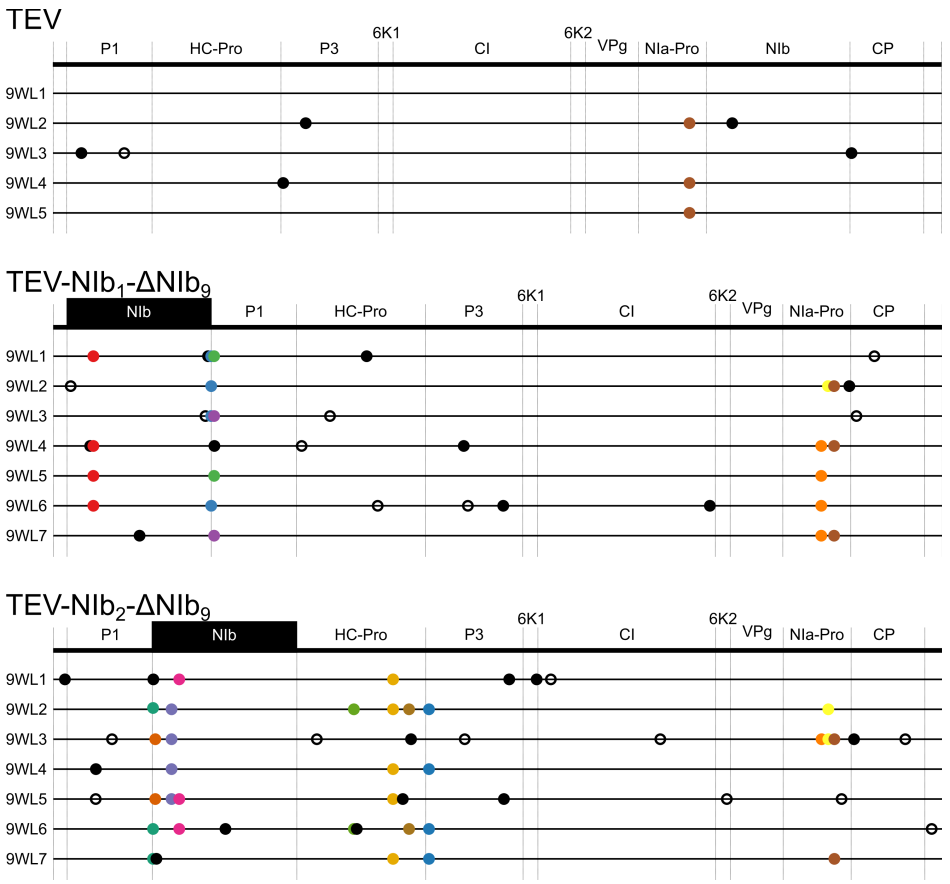

Figure C1.8. Genomes of the evolved lineages with the *NIb* moved to an alternative position. Mutations were detected using NGS data of the evolved TEV-NIb_1-ΔNIb_9 and TEV-NIb_2-ΔNIb_9 lineages as compared against their ancestral lineages. The wild-type TEV is given for comparative purposes. The names on the left identify lineages (*e.g.*, 9WL1 is the final population of 9-week passages, lineage 1). Full circles and open circles represent nonsynonymous and synonymous substitutions, respectively. Black substitutions occur in only one lineage, whereas color-coded substitutions are repeated in two or more lineages. Note that the mutations are present at different frequencies as reported by SAMtools (> 10%).

significant difference was found in the distribution of synonymous versus nonsynonymous SNP frequencies (Kolmogorov–Smirnov test; TEV-NIb$_1$-ΔNIb$_9$: $D = 0.203$, $P = 0.078$; TEV-NIb$_2$-ΔNIb$_9$: $D = 0.152$, $P = 0.318$).

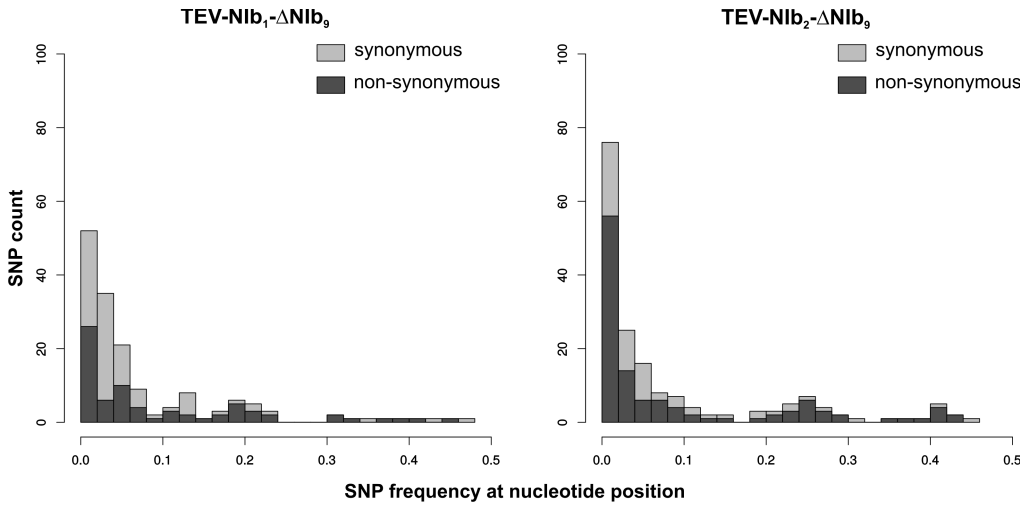

Figure C1.9. The distribution of SNP frequencies in the evolved TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$ lineages.

The mutations appeared to be contingent upon the ancestral genotype; most of the convergent mutations that were found in the lineages of one reordered genotype were not found in the other genotype. All convergent mutations found were nonsynonymous. Note that the convergent mutations are present at different frequencies and none of these convergent mutations were fixed in all the replicate lineages.

For TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$ we therefore see markedly different patterns of genome evolution than for TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$. In

agreement with RT-PCR results, TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$ show no signs of major genomic rearrangements, such as the large deletions seen in TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$. Given the strong selection, which would undoubtedly occur for a variant with wild-type gene order, these results suggest that the mutation supply and unviability of viruses without *NIb* is a major limiting factor. This result provides support for our conjecture that gene order evolution involving essential genes must occur through gene duplications in potyviruses. On the other hand, the observed convergent single-nucleotide mutations are congruent with the observed improvement of virus accumulation, suggesting that selection is acting on these virus populations.

Studying the evolutionary potential of the reordered viruses TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$, although demonstrating clear signatures of adaptive evolution, therefore also shows further barriers to the evolution of alternative gene orders. First, increased accumulation and sequence-level convergent evolution illustrate that adaptive evolution occurred, meaning that the reordered viruses, despite initially having low levels of accumulation, are still somewhat evolvable and show marked improvement in key fitness components. Conversely, accumulation was still significantly lower than the wild-type virus. Furthermore, within-host competitive fitness remained similar to the ancestral TEV-NIb$_1$-ΔNIb$_9$ or TEV-NIb$_2$-ΔNIb$_9$, or strikingly, even decreased in some lineages. These observations suggest that even after 36 weeks of evolution under conditions that optimize selection, these reordered viruses remain grossly inferior competitors to the wild-type virus. Therefore, even if a virus population should overcome all these previous four barriers, a final evolutionary barrier to the reordering of potyvirus genomes remains.

3. Discussion

In this study we have explored whether the most plausible evolutionary trajectory for the rearrangement of gene order in a positive-strand RNA potyvirus is accessible. Overall, we have identified five barriers to the evolution of viruses with the essential *NIb* replicase gene moved to an alternative position. First, only 2/9 viruses with *NIb* moved to an alternative position are viable. Second, the fitness of viruses with *NIb* duplications was low, meaning that such viruses would be quickly displaced from populations if they arose (**Fig C1.10** shows a summary of fitness data). Third, for viruses with gene duplications, the new *NIb* copy was lost in all lineages whereas the original copy was maintained. This propensity represents a second *cul-de-sac*, since the loss of the new *NIb* copy entails a trajectory that leads back to the ancestral gene order. Fourth, viruses with only a single *NIb* copy at an alternative position had a low fitness and accumulation, notably lower than the viruses with duplications (**Fig C1.10**). Therefore, to reach a virus with a single *NIb* copy in an alternative position, two rare recombination events must occur within a small time window, as the intermediate step is unstable. Moreover, the low-fitness recombinants would need to be maintained or fixed by genetic drift, as they would be outcompeted otherwise. Fifth, even if this unlikely sequence of events occurs and the resulting virus becomes reproductively isolated, after more than half a year of evolution under optimal conditions this virus would still not stand a chance in head-to-head competition with its ancestor (**Fig C1.10**). We therefore conclude that – under the conditions we have considered – the evolution of alternative gene orders for TEV is highly unlikely, because the evolutionary trajectory to alternative gene order we have studied is not

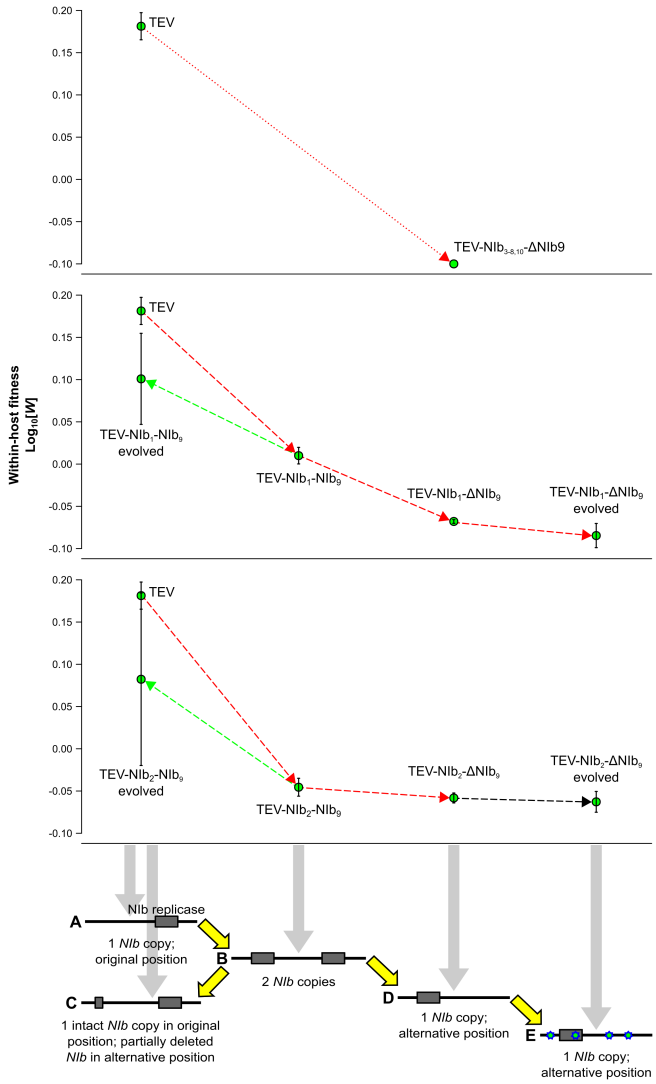

Figure C1.10. Within-host competitive fitness of all trajectories explored. For all the trajectories proposed (*x*-axis) the fitness (*W* at the *y*-axis) was plotted. The viruses with the *NIb* replicase gene moved to intergenic positions 3-8 and 10 are unviable genotypes (upper panel; $W < 0$). The trajectories of viruses with *NIb* moved to intergenic positions 1 and 2 are indicated in the middle and lower panels, respectively. Red lines indicate a decrease in fitness, black lines indicate no significant change in fitness and green lines indicate an increase in fitness. Note how the ancestral and evolved TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$ lineages have an inferior within-host competitive fitness as compared to their ancestral TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ viruses (respectively) and the wild-type TEV.

accessible. These results suggest that one reason gene order has been conserved in potyviruses is therefore the lack of accessible trajectories to alternative gene orders.

The observation that the evolved virus lineages with an alternative gene order have improved accumulation, while having unchanged or deteriorated within-host fitness, is noteworthy. Our serial passage experiment was conducted in single host plants, and we would therefore expect within-host competitive fitness to improve. In other words, the virus variant that is present at highest frequency in the final population has the highest probability of carrying over to the next round of infection, irrespective of the level of accumulation. We hypothesize that unchanged or lowered competitive fitness is probably due to low infection levels during serial passaging of TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$. These low infection levels would result in limited direct competition for space within the host, with adaptation occurring by means of more prudent use of cellular resources and improved repression of host immune responses, for example. Evolution during low-level infections may therefore not improve performance in direct competition, or even lower it due to antagonistic pleiotropy, given that accumulation and competitive fitness do not correlate for TEV (Zwart *et al.* 2014). Since the common competitor used in the fitness assays is a strain derived from TEV, high-level infections will occur in these assays. For these reasons, fitness improvements in lineages of TEV-NIb$_1$-ΔNIb$_9$ and TEV-NIb$_2$-ΔNIb$_9$ may very well not be conspicuous in direct competitions with the ancestral TEV, and can only be detected systemically by measuring virus accumulation.

The first step in our proposed model, gene duplication, is a rare event in the

recent history of RNA viruses (Simon-Loriere and Holmes 2013). The few cases describing gene duplication in RNA viruses seem to occur through either homologous or non-homologous recombination (Forss and Schaller 1982; Tristem *et al.* 1990; Boyko *et al.* 1992; Walker *et al.* 1992; Wang and Walker 1993; Karasev *et al.* 1995; LaPierre *et al.* 1999; Peng *et al.* 2001; Valli *et al.* 2007; Simon-Loriere and Holmes 2013). In our experiments, the gene duplications are engineered artificially, whilst the loss of the duplicated gene generates shorter genomes that are efficiently selected for, probably by means of slippage of the RdRp. We also observed deletions in the N-terminal region of the HC-Pro cysteine protease, since there is no selection to maintain the intact HC-Pro. However, these specific deletions would probably not perform well in nature, because the resulting truncated protein will preclude aphid-mediated transmission.

We speculate that the rearrangement of genes within the TEV genome could lead to differences in efficiency of cleavage at the proteolytic sites. Due to the lower conservation of gene order and sequence homology at the 5' end in potyviruses, repositioning of genes to this side of the genome would least disturb polyprotein processing. Therefore, it is not surprising that the two positions available for relocation of the *NIb* replicase gene are located at the N-terminus, before and between the *P1* and *HC-Pro* genes. Within the *Potyviridae* family, both the P1 and the multifunctional HC-Pro have the lowest sequence conservation (Adams *et al.* 2005a; Revers and García 2015). P1 is not an essential gene (Verchot and Carrington 1995a), however, the cleavage that separates P1 and HC-Pro is required (Verchot and Carrington 1995b). The introduction of *NIb* before or between *P1* and *HC-Pro* could result in a delay in this separation or even make P1 inactive, which in turn results in a low

accumulation of the virus (Verchot and Carrington 1995a). We do observe that the viral genomes with an alternative gene order have low accumulation levels, however, the accumulation levels are able to improve over time with the same gene order. This suggests that the adaptive mutations alone are responsible for the increase in accumulation levels. Nevertheless, these adaptive mutations could not compensate for the rearrangements done, since within-host competitive fitness remains low throughout evolutionary time.

It can be postulated that the early expression of the *NIb* gene would have advantages for viral replication. However, the position of NIb in potyviruses' polyprotein seems to be optimized for its interaction with VPg, for initiation of replication. Furthermore, it is thought that the interaction with both VPg and NIa-Pro targets NIb to the membranous structures where viral RNA replication takes place (Dufresne *et al.* 2008). So the movement of NIb away from these interacting proteins might result in a delay in the interaction, replication and gene expression. This is a difficult, maybe even impossible, burden to overcome for a virus.

The relocation of the *NIb* in our model system did lead to the evolution of adaptive or compensatory mutations that allowed for an increase in viral accumulation. Interestingly, convergent mutations occur in both the relocated *NIb* and two regions that are responsible for proteolytic activity: the C-terminal region of *HC-Pro* and the main viral protease *NIa-Pro*. When relocating the *NIb* replicase to the first position in the genome, and therewith introducing a second methionine codon before *NIb* in the viral genome, we observe a mutation that changes the original methionine (in P1) to a threonine. Therefore, a reversion of *NIb* to its original position is not a subsequent step we would expect to see in

this viral genotype.

Comparing our study to the experimental evolution done on the repositioning of the T7 RNA polymerase (Springman *et al.* 2005), similar results were obtained despite of T7 having a different genome composition, architecture and replication and gene expression strategies. In T7 studies, fitness is measured as the rate of population growth, which is comparable to our accumulation measurements. As in our study, the fitness of the rearranged phages improved but never reached the wild-type level. Additionally, as in our study, the reordered T7 genomes are stable during a long period of time. Concerning these similarities, the lack of accessible evolutionary trajectories to alternative gene orders cannot be entirely explained by the most obvious impediment: that potyviruses with reordered genomes would most likely be unviable due to improper autocatalytic processing of the polyprotein. In our study, experimental evolution of the re-ordered viruses illustrates other important barriers to gene order evolution.

In VSV, the evolution of alternative gene orders is more plausible than for TEV, as the movement of the *N* gene does not affect viability of the virus (Wertz *et al.* 1998). The evolvability of variants of VSV with the *N* in alternative positions, was higher in a cell line from an alternative host (Pesko *et al.* 2015). Our results for the evolution of TEV-NIb_1-ΔNIb_9 and TEV-NIb_2-ΔNIb_9 are comparable, although we considered evolution in a permissive host and as such expect to see little adaptation of the ancestral TEV. Pesko *et al*. (2015) did not consider the evolutionary trajectory leading up to the formation of a virus with a rearranged genome. However, at least in the cell-culture environment where MOI can be high for VSV, this may not be as big an impediment as for TEV. A

computational study suggests that point mutations in VSV intergenic regions preceded or co-evolved with the fixation of the wild-type gene order, resulting in a sub-optimal genome organization (Lim and Yin 2009).

For our experiments, we have considered only one present-day potyvirus, as well as variants with rearranged gene orders derived from it. Whether our results will extend to other present-day potyviruses is a valid question, but it is equally important to consider that these results may have only a limited bearing on the potential for gene-order evolution in ancestral potyviruses. As a result of epistasis, adaptive evolution can limit accessible evolutionary trajectories (Salverda *et al.* 2011) whilst purifying selection can result in entrenchment (Shah *et al.* 2015). It is therefore plausible that evolutionary trajectories to alternative gene orders may have been more accessible to ancestral viruses. We therefore do not rule out that other factors may have been important in conserving potyvirus gene order, in particular at early time points in their evolution.

What conditions could make the evolutionary trajectory we have studied accessible for present-day viruses, or could open up alternative evolutionary trajectories? We have explored evolvability of a single potyvirus genotype at different points along the trajectory to alternative gene orders. It is possible that certain genotypes could be less constrained, although given the low fitness of the intermediates considered and the instability of the new *NIb* copy, this potentiating variation would probably have to preexist prior to the first recombination step leading to gene duplication. Prime candidates for mutations that may mitigate such constraints are the convergent mutations found in the evolved lineages of TEV-NIb_1-ΔNIb_9 and TEV-NIb_2-ΔNIb_9. An alternative host

species in which high MOI occurs also could open up new evolutionary trajectories, by allowing complementation between viral genomes and hereby widening the set of plausible trajectories beyond only those involving gene duplication. We think it is unlikely that such hosts exist, however, given that (*i*) in general MOI estimates for plant RNA viruses tend to be low (Zwart *et al.* 2013), and (*ii*) alternative hosts would tend to be semi-permissive, we would therefore intuitively expect lower infection levels and MOI in such a host. We therefore think the most promising avenue for further research on alternative gene orders is to consider the impact of potentiating mutations.

By showing these different barriers to alternative gene orders in viruses, we expect to drive further research on the diversity of gene order over different organisms. Our results serve as a roadmap for testing which factors constrain or promote gene order conservation across different viruses and could be compared to the great diversity of gene order in other taxa.

Chapter 2: Predicting the stability of homologous gene duplications

1. Introduction

Gene duplication results in genetic redundancy. In other words, the existence of genetic elements that encode for the same function. It is a powerful process to regulate gene expression, to increase the genetic and environmental robustness of organisms, and can be a stepping stone to the evolution of new biological functions. Therefore, it is not surprising that gene duplication is a frequent phenomenon in many organisms (Zhang 2003; Andersson and Hughes 2009), especially in eukaryotes.

There are few examples of genetic redundancy in viral genomes. In general, viral genomes tend to be highly streamlined, with limited intergenic sequences and in many cases overlapping open reading frames (ORFs), suggesting genome size is under strong selection (Lynch 2006). RNA viruses typically have smaller genomes than DNA viruses, and consequently there is an extreme low prevalence of gene duplication in RNA viruses (Simon-Loriere and Holmes 2013). For the reverse-transcribing viruses, three different gene duplication events have been reported within the *Retroviridae* family (Tristem *et al.* 1990; LaPierre *et al.* 1999; Kambol *et al.* 2003). This low prevalence of gene duplication in retroviruses is surprising, since repeated sequence elements of endogenous retroviruses are thought to mediate genomic rearrangements, including gene duplication (Hughes and Coffin 2001). For the ss(-)RNA viruses, two different tandem gene duplications have been reported (Walker *et*

al. 1992; Gubala *et al.* 2010; Blasdell *et al.* 2012; Simon-Loriere and Holmes 2013) within the *Rhabdoviridae* (infecting vertebrates, invertebrates and plants). For the ss(+)RNA viruses, single duplication events have been reported for three different domains: (*i*) a tandem duplication of the coat protein gene (*CP*) within the *Closteroviridae* (infecting plants) (Boyko *et al.* 1992; Fazeli and Rezaian 2000; Kreuze *et al.* 2002; Tzanetakis *et al.* 2005; Tzanetakis and Martin 2007; Simon-Loriere and Holmes 2013); (*ii*) a tandem duplication of the genome-linked protein gene (*VPg*) in *Foot-and-mouth disease virus* from the *Picornaviridae* (infecting vertebrates) (Forss and Schaller 1982); and (*iii*) a duplication of the third segment generating an additional segment in *Beet necrotic yellow vein viru*s from the *Benyviridae* (Simon-Loriere and Holmes 2013). To date, no cases of gene duplication in dsRNA viruses have been reported.

The variation in genome sizes and structures indicates that gene duplication must have played a role in the early diversification of virus genomes. However, the rapid evolution of RNA viruses and the potential fitness costs associated with harboring an extra chunk of genetic material probably makes it unlikely to detect viruses with duplications, or even the signatures of recent duplication events in present day viruses. Strong selective constraints against increasing genome sizes are thought to play a role in the lack of gene duplications that we nowadays observe in RNA viruses (Holmes 2003). One of these constraints is the high mutation rates of RNA viruses, which is approximately one mutation per genome replication (Sanjuán *et al.* 2010). This limits the probability of copying without errors a genome above the length limit imposed by Eigen's error threshold (Eigen 1971): the inverse of the per site mutation rate. Another constraint is the need for fast replication due to strong within-cell and within-

host competition (Turner and Chao 1998). Together, an increase in genome size is likely to increase the number of deleterious mutations that occur during each round of replication, and to slow down the replication process. However, the small and streamlined RNA virus genomes also limit sequence space for the evolution of novel functions, and in turn adaptation to environmental changes.

Here we therefore consider experimentally the evolutionary fate of RNA viruses with gene duplications, in terms of their effects on fitness, the stability of the duplicated gene and the evolvability of these viruses. We experimentally explore four cases of homologous duplication of genes within the *Tobacco etch virus* (TEV) genome: (*i*) the multi-functional protein (HC-Pro) involved in aphid transmission, polyprotein cleavage, genome amplification and suppression of RNA silencing, (*ii*) the main viral protease (NIa-Pro), (*iii*) the viral RNA-dependent RNA polymerase (NIb), and (*iv*) the coat protein (CP) (Revers and García 2015). Potyviruses are a particularly interesting system for studying the evolution of gene duplications, as they encode a single polyprotein that is auto catalytically processed into the mature gene products. For each complete positive sense RNA, as well as frame-shifted transcripts where translation terminates at P3-PIPO, there will be isostoichiometric expression of all genes. Assuming there are no unknown mechanisms that regulate gene expression, the scope for the regulation of gene expression in potyviruses could therefore be very limited. Gene duplication may represent a way to bypass these constraints and achieve higher expression of specific genes.

We speculated that the duplication of these four proteins might have widely different impacts in TEV fitness. As HC-Pro is a multifunctional protein, two copies of HC-Pro could lead to improvement of one or more of its functions.

This improvement could possibly be caused by two mechanisms. First, by simply producing more protein there could be an immediate benefit in one of HC-Pro functions. Second, there could be improvement of protein function when the duplicated virus is evolved, because the two gene copies can each specialize on different functions. Higher levels of NIa-Pro may result in a more efficient processing of the polyprotein, making more mature viral proteins available faster for the replication process. As potyviruses have only a limited number of post-translational mechanisms for regulating gene expression levels, we predicted that the overproduction of NIa-Pro will alter the equilibrium concentrations of all the different mature peptides and thus have a major impact in TEV fitness. Higher levels of NIb may result in faster replication of the virus and this could lead to higher levels of accumulation and potentially the within-host spread of infection by a greater number of virions. The cellular multiplicity of infection (MOI), which has been estimated to be as low as 1.14 virions per infected cell for TEV (Tromas *et al.* 2014a), might even increase. Higher levels of CP expression could allow for the encapsidation of more genomic RNA molecules without affecting the accumulation of all other mature peptides. However, completion of the infectious cycle would still depend on the cytoplasmic amount of other limiting proteins (*e.g.,* replicase NIb or silencing suppressor HC-Pro).

The duplication events that we explore here could therefore conceivably have beneficial effects on TEV replication, perhaps offsetting the costs inherent to a larger genome and increasing overall fitness. Moreover, especially in the case of HC-Pro, they could perhaps lead to the evolution of greatly improved or novel functions. However, given the scarcity of gene duplications in RNA viruses, we expected that the fitness costs of duplication where likely to be high, and that

one of the two gene copies would be rapidly lost. If further mutations could potentially help accommodate the duplicated gene, then this could lead to interesting evolutionary dynamics: will the duplicated gene be lost or will beneficial mutations that lead to stable maintenance of the gene occur first? Moreover, as they could potentially disrupt correct processing of the polyprotein, the possibility that some of the duplications would not be viable in the first place could also be discounted (Majer *et al.* 2014). To address these issues we have constructed four viruses with gene duplications and tested their viability. We subsequently evolved these viruses and determined the stability of the duplicated gene, as well as looking for signals of accommodation of the duplicated gene.

2. Results and Discussion

2.1. Genetic redundant constructs and the viability of the resulting viruses

To simulate the occurrence of duplication events within the TEV genome (**Fig C2.1A**), different TEV genotypes were constructed using four genes of interest **Fig C2.1**). Each of these genotypes therefore represents a single gene duplication event. Where necessary, the termini of the duplicated gene copies were adjusted, such that the proteins can be properly translated. Cleavage sites are provided, similar to the original proteolytic cleavage sites at the corresponding positions. A description of every duplication event will be given in the same order as these genes occur within the TEV genome.

First, for duplication of the multifunctional *HC-Pro* cysteine protease gene, a second copy of *HC-Pro* was inserted in the second position within the TEV

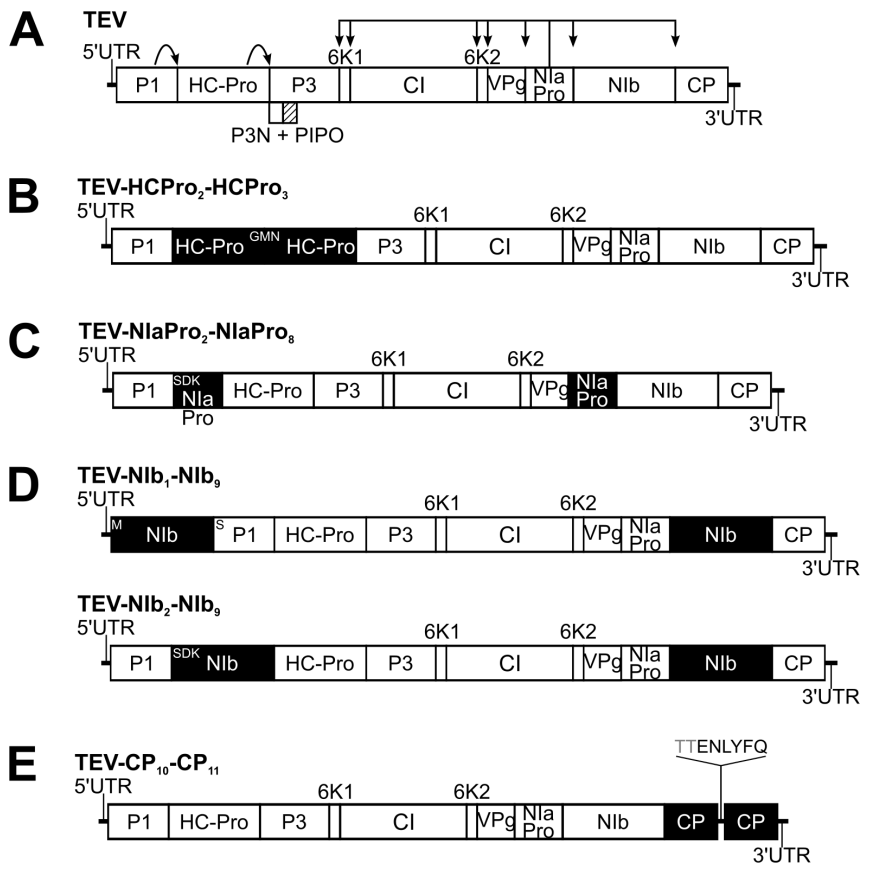

Figure C2.1. Schematic representation of the different *Tobacco etch virus* genotypes containing gene duplications. The wild-type TEV (**A**) codes for 11 mature peptides, including P3N-PIPO embedded within the P3 protein at a +2 frameshift. Five different viral genotypes containing a single gene duplication were constructed. Second copies of *HC-Pro* (**B**), *NIa-Pro* (**C**) and *NIb* (**D**) were introduced between *P1* and *HC-Pro*. A second copy of *NIb* was also introduced before *P1* (**D**). And a second copy of *CP* was introduced between *NIb* and *CP* (**E**). For simplification P3N-PIPO is only drawn at the wild-type TEV.

genome, between the *P1* serine protease gene and the original gene copy, generating a tandem duplication (**Fig C2.1B**). This position is a common site for the cloning of heterologous genes (Zwart *et al.* 2011). Second, a copy of the

NIa-Pro main viral protease gene was introduced between *P1* and *HC-Pro* (Fig **C2.1C**). Third, two genotypes containing a duplication of the *NIb* replicase gene were generated, see also Chapter 1, where a copy of the *NIb* gene was inserted at the first position (before P1) and the second position in the TEV genome (**Fig C2.1D**). Fourth, for duplication of the *CP* we introduced a second copy at the tenth position between *NIb* and *CP,* generating a tandem duplication (**Fig C2.1E**). Henceforth we refer to these five genetic redundant viruses as TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, TEV-NIb$_2$-NIb$_9$ and TEV-CP$_{10}$-CP$_{11}$, respectively, with subscripts denoting the intergenic positions of the duplicated gene in question.

The viability of these viruses was tested in *N. tabacum* L. cv. Xanthi (NN) plants, by inoculating plants with approximately 5 µg *in vitro* generated transcripts. TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, and TEV-NIb$_2$-NIb$_9$, were found to infect *N. tabacum* plants, as determined by RT-PCR on total RNA extracted from these plants. After performing multiple viability tests, TEV-CP$_{10}$-CP$_{11}$ demonstrated to have a very low infectivity and high amounts (> 20 µg) of RNA are needed for infection to occur. Performing RT-PCR of the region containing two CP copies, we detected either (*i*) a band corresponding to the wild-type virus (**Fig C2.2A**; plant inoculated with 20 µg of RNA), indicating that upon infection with RNA the second CP copy is deleted immediately, or (*ii*) a band that indicates the two CP copies are present (**Fig C2.2A**; plant inoculated with 30 µg of RNA). Taking the latter as a starting population for experimental evolution, within the first passage, we detect a band corresponding to the wild-type virus, in six out of eighth lineages, and we did not detect any infection in the remaining two lineages. When sequencing the region containing the deletions in the different lineages, using Sanger

technology, exact deletions of the second CP copy were observed. We discontinued further experiments on this virus due to the extreme instability of the second CP copy.

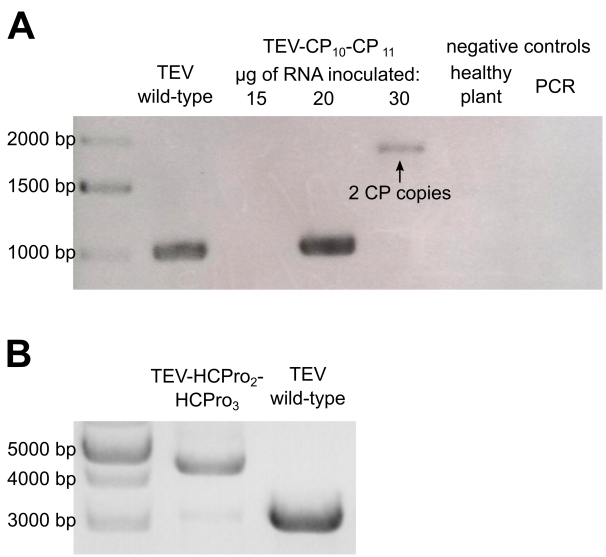

Figure C2.2. Ancestral TEV-CP_{10}-CP_{11} and TEV-$HCPro_2$-$HCPro_3$ populations. (**A**) An agarose gel with RT-PCR products of the duplicated CP region. Plants were inoculated with 15 µg, 20 µg and 30 µg RNA. When inoculated with 20 µg a band corresponding to the wild-type virus was detected. When inoculated with 30 µg, a band corresponding to a virus with two CP copies was detected. Healthy plants and the PCR mix were used as negative controls. (**B**) An agarose gel with RT-PCR products of the duplicated HC-Pro region. The ancestral TEV-$HCPro_2$-$HCPro_3$ population, contains a low frequency deletion variant corresponding to the wild-type

Remarkable is that within the superfamily of the (+)RNA viruses, the coding region of CP is usually located at the 5´-terminal to ensure high level of expression. However, this is not the case for the *Potyviridae,* where the *CP* occupies the 3´ end terminal position and appears not to be subjected to expression regulation.

2.2. Evolution of genetic redundant viruses

After reconstitution of TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, and TEV-NIb$_2$-NIb$_9$ from infectious clones, these viruses containing gene duplications were evolved in *N. tabacum* plants. All viruses were evolved for a total of 27 weeks, using nine 3-week passages and three 9-week passages with at least five independent lineages for each passage length. In the starting population of TEV-HCPro$_2$-HCPro$_3$ we observed a mild symptomatology, however these symptoms rapidly turned into wild-type symptoms in lineages from the first 3- and 9-week passages. At the start of the evolution experiment, TEV-NIaPro$_2$-NIaPro$_8$ also displayed altered symptomatology: symptoms were milder and appeared to expand slower compared to the wild-type TEV. However, in the first 9-week passage symptoms became stronger, similar to the wild-type virus, as the virus expanded through the plant. This symptomatology was also displayed in the subsequent 9-week passages. Increases in symptomatology were also observed for the TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ viruses (see Chapter 1).

Partial and complete deletions of the duplicated gene copy were detected by RT-PCR (**Fig C2.3**), but never in the ancestral gene. Deletion of the duplicated gene copy of the TEV-HCPro$_2$-HCPro$_3$ variant occurred rapidly after infection of the plants; after one passage no virions carrying the gene duplication could be detected by RT-PCR **(Fig C2.3A**). By performing several RT-PCRs, in which annealing temperature and template input concentration were varied, we were able to detect a deletion in the ancestral TEV-HCPro$_2$-HCPro$_3$ population (**Fig C2.2B**). Nevertheless, the RT-PCR results show that this deletion variant is present only at low frequency. This estimate is moreover conservative, as the

shorter PCR template representative of the deletion variant will be more efficiently amplified than the intact duplication variant. No deletions were detected in the TEV-NIaPro$_2$-NIaPro$_8$ lineages using the shorter 3-week passages (**Fig C2.3B**). Deletions were not detected in the first 9-week passage either, but in the second passage partial or complete deletions did occur. Mixed populations that contain virions with a deletion together with virions that have maintained the intact duplicated copy, are mainly present in the TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ lineages. However, TEV-NIb$_1$-NIb$_9$ loses the duplicated copy faster (**Fig C2.3C**; second 3-week passage, and first 9-week passage) than the TEV-NIb$_2$-NIb$_9$ virus (**Fig C2.3D**; third 3-week passage, and second 9-week passage).

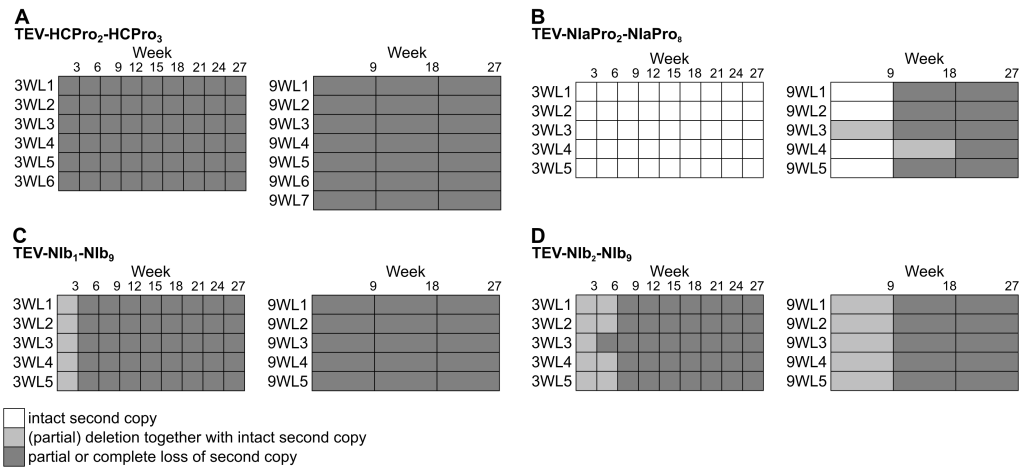

Figure C2.3. Deletion detection along the evolution experiments. RT-PCR was performed on the region containing a duplication in the viral genotypes. Either an intact duplicated copy (white boxes), a deletion together with an intact duplicated copy (light-grey boxes), or a partial or complete loss of the duplicated copy (dark-grey boxes) were detected.

Based on the majority deletion variants observed by RT-PCR, genome size was estimated for every passage (**Fig C2.4**). Comparing the different viral genotypes, there are clear differences in the time until the duplicated gene copy is deleted. The faster occurrence of deletion variants during longer-duration passages is congruent with results from a previous study (Zwart *et al.* 2014),

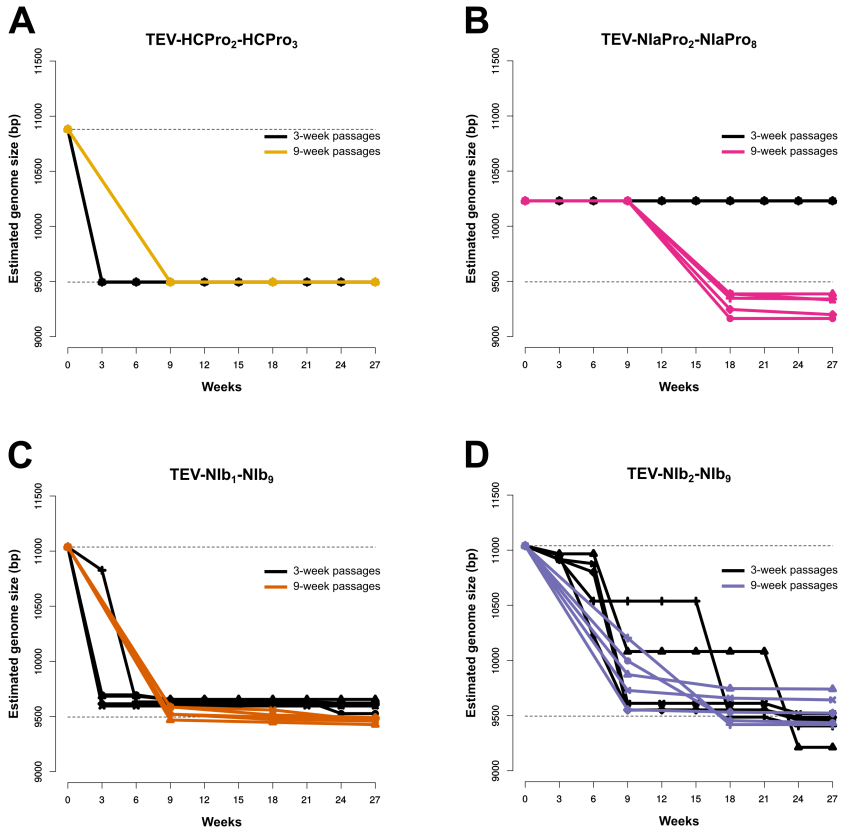

Figure C2.4. The reduction in genome size over time. The different panels display how the genome size of the different viral genotypes containing duplications changes along the evolution experiments. The dashed black lines indicate genome sizes of the wild-type virus (below) and the ancestral viruses (above). The genome sizes of the 3-week lineages are drawn with black lines, and those of the 9-week lineages are drawn with colored lines.

where deletions of eGFP marker inserted in the TEV genome were usually observed after a single 9-week passage, but were rarely spotted even after nine 3-week passages. On the other hand, the highly diverging results for genome stability obtained here for the different viruses, suggests that passage duration is not the principle factor determining whether gene duplication will be stable. Therefore, the size of the duplicated gene, the nature of the gene, and/or the position for duplication could play a role in the stability of genomes with a duplication.

2.3. Viruses with a gene duplication have reduced fitness which cannot always be restored after deletion

For both the ancestral and evolved virus populations, we measured within-host competitive fitness (**Fig C2.5**) and viral accumulation (**Fig C2.6**). Comparing the ancestral viruses containing a gene duplication to the ancestral wild-type virus (filled circles in **Fig C2.5** and **C2.6**), we observed statistically significant decreases in competitive fitness (**Fig C2.5A**; TEV-HCPro$_2$-HCPro$_3$: $t_4 = 8.398$, $P = 0.001$; **Fig C2.5B**; TEV-NIaPro$_2$-NIaPro$_8$: $t_4 = 12.776$, $P < 0.001$ **Fig C2.5C**; TEV-NIb$_1$-NIb$_9$: $t_4 = 6.379$, $P = 0.003$; TEV-NIb$_2$-NIb$_9$: $t_4 = 8.348$, $P = 0.001$). Statistically significant decreases in accumulation levels were also observed for TEV-HCPro$_2$-HCPro$_3$, TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ (**Fig C2.6A**; TEV-HCPro$_2$-HCPro$_3$: $t_4 = 3.491$, $P = 0.0251$; **Fig C2.6C**; TEV-NIb$_1$-NIb$_9$: $t_4 = 45.097$, $P < 0.001$; TEV-NIb$_2$-NIb$_9$: $t_4 = 8.650$, $P < 0.001$). However, there was no difference in accumulation for the virus with a duplication of NIa-Pro (**Fig C2.6B**; TEV-NIaPro$_2$-NIaPro$_8$: $t_4 = 2.099$, $P = 0.104$).

By evolving the viruses with gene duplications using three 9-week passages, we

observe an increase in within-host competitive fitness in all four virus genotypes (open circles in **Fig C2.5**), comparing these to their corresponding ancestral virus (**Fig C2.5**; asterisks indicate a significant increase, *t*-test with Holm-Bonferroni correction). Within-host fitness levels similar to the evolved wild-type TEV, were reached by both evolved TEV-HCPro$_2$-HCPro$_3$ (Fig **C2.5A**; Mann-Whitney $U = 23$, $P = 0.432$) and TEV-NIaPro$_2$-NIaPro$_8$ (Fig **C2.5B**; Mann-Whitney $U = 20$, $P = 0.151$) lineages. On the contrary, the evolved TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ lineages, did not reach wild-type virus within-host fitness levels (**Fig C2.5C**; TEV-NIb$_1$-NIb$_9$: Mann-Whitney $U = 0$, $P = 0.008$; TEV-NIb$_2$-NIb$_9$: Mann-Whitney $U = 0$, $P = 0.008$). Together with within-host fitness, virus accumulation also increased significantly for the evolved TEV-NIb$_1$-NIb$_9$ and TEV-NIb$_2$-NIb$_9$ genotypes (**Fig C2.6C**; asterisks), comparing these to their corresponding ancestral virus. However, accumulation levels did not increase significantly for most of the evolved lineages of the TEV-HCPro$_2$-HCPro$_3$ and TEV-NIaPro$_2$-NIaPro$_8$ genotypes. Nevertheless, these two genotypes have much higher initial accumulation levels than the genotypes with a duplication of Nib. When comparing the accumulation levels of the evolved lineages to those of the wild-type, TEV-HCPro$_2$-HCPro$_3$ (**Fig C2.6A**; Mann-Whitney $U = 20$, $P = 0.755$), TEV-NIaPro$_2$-NIaPro$_8$ (**Fig C2.6B**; Mann-Whitney $U = 11$, $P = 0.841$), and TEV-NIb$_2$-NIb$_9$ (**Fig C2.6C**; Mann-Whitney $U = 3$, $P = 0.056$) do reach wild-type accumulation levels, whilst TEV-NIb$_1$-NIb$_9$ does not (**Fig C2.6C**; Mann-Whitney $U = 0$, $P = 0.008$).

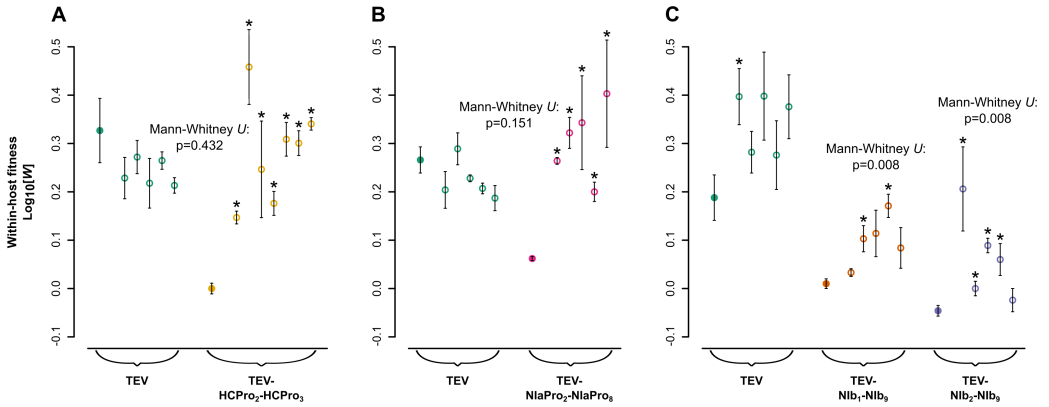

Figure C2.5. Within-host competitive fitness of the evolved and ancestral lineages Fitness (*W*), as determined by competition experiments and RT-qPCR of the different viral genotypes with respect to a common competitor; TEV-eGFP. The ancestral lineages are indicated by filled circles and the evolved lineages by open circles. The asterisks indicate statistical significant differences of the evolved lineages as compared to their corresponding ancestral lineages (*t*-test with Holm-Bonferroni correction). The *P*-values indicate the significance of the differences between the evolved variant lineages as compared to the evolved wild-type lineages (Mann-Whitney *U* test).

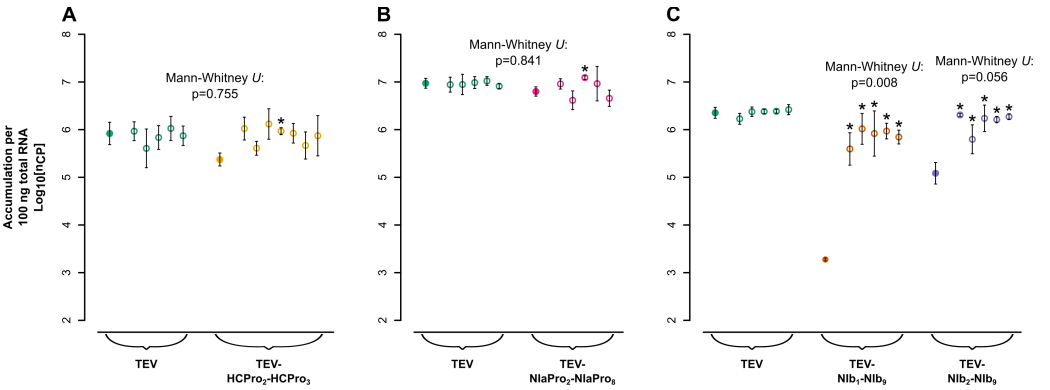

Figure C2.6. Virus accumulation of the evolved and ancestral lineages. Virus accumulation, as determined by accumulation experiments and RT-qPCR at 7 dpi of the different viral genotypes The ancestral lineages are indicated by filled circles and the evolved lineages by open circles. The asterisks indicate statistical significant differences of the evolved lineages as compared to their corresponding ancestral lineages (*t*-test with Holm-Bonferroni correction). The *P*-values indicate the significance of the differences between the evolved variant lineages as compared to the evolved wild-type lineages (Mann-Whitney *U* test).

When comparing the within-host competitive fitness of the evolved TEV-NIaPro$_2$-NIaPro$_8$ 9-week lineages to the 3-week lineages, we found that there is a linear relationship between genome size and within-host competitive fitness (**Fig C2.7**; Spearman's rank correlation ρ = -0.795, 10 d.f., P = 0.006). Moreover, the evolved 9-week lineages, that contain genomic deletions, have a significant higher within-host competitive fitness (Mann-Whitney U = 4.5, P < 0.001) than the evolved 3-week lineages without deletions.

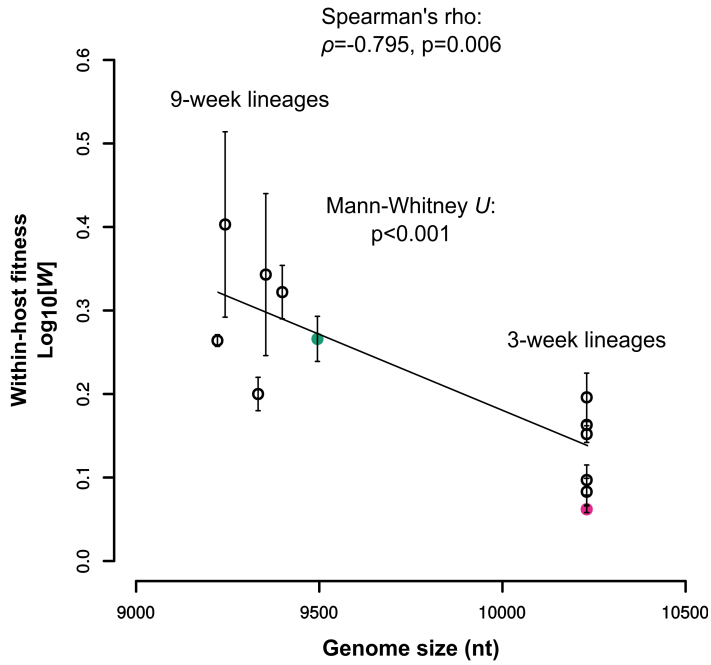

Figure C2.7. The relationship between genome size and within-host competitive fitness. The pink filled circle represents the within-host competitive fitness of the ancestral TEV-NIaPro$_2$-NIaPro$_8$ and the green filled circle that of the ancestral wild-type TEV. The black open circles represent the evolved 3-week (right) and 9-week (left) TEV-NIaPro$_2$-NIaPro$_8$ lineages. The evolved 9-week lineages, that contain genomic deletions, have a significant higher within-host competitive fitness (Mann-Whitney U test) than the evolved 3-week lineages without deletions. A linear regression has been drawn to emphasize the trend in the data.

The observation that gene duplication events result in decreases in fitness is not unexpected. Although all the duplications considered here could conceivably have advantages for viral replication or encapsidation, our results suggest that the any such advantages are far outweighed by the costs associated with a larger genome, increased protein expression or the effects on polyprotein processing. However, the duplication of *NIa-Pro* does not affect the viral accumulation rate. This could be explained by the fact that the *NIa-Pro* gene is much smaller than the other duplicated genes. Consequently, in conditions where selection has the least time to act between bottleneck events associated with infection of a new host (3-week passages), no deletions were observed. In the long-passage experiment selection has more time to act and increase the frequency of beneficial *de novo* variants, allowing them to be sampled during the bottleneck at the start of the next round of infection. In addition, the size of the gene duplication also seems to play a role. But what about the position and the nature of the duplicated gene? When duplicating the same gene, *NIb*, to either the first or second position in the TEV genome, we see clear differences in the deletion dynamics and fitness measurements (**Figs C2.4**, **C2.5**, **C2.6** and Chapter 1). Comparing the duplication and subsequent deletion of *HC-Pro*, *NIa-Pro* and *NIb* at the same second position, we observe that both accumulation and within-host competitive fitness cannot be completely restored by the virus that originally had two copies of the NIb replicase, whilst viruses that originally had two copies of the multi-functional protein HC-Pro or the main viral protease NIa-Pro do restore their fitness after deletion. However, these observations are true for the evolutionary time given in our experimental setup, there is a possibility that with more time, all viruses restore their fitness to similar levels.

2.4. Genome sequences of the evolved lineages

All evolved and ancestral lineages have been fully sequenced using the Illumina technology. The sequences of the ancestral lineages were used as an initial reference for the evolved lineages. After an initial mapping step, mutations were detected in the evolved lineages as compared to their corresponding ancestor (**Fig C2.8**).

Beside the large genomic deletions, different patterns of adaptive evolution were observed for each viral genotype (**Table C2.1** and **Fig C2.8**). For the evolved TEV-HCPro$_2$-HCPro$_3$ virus a convergent nonsynonymous mutation was found in 3/7 9-week lineages in the P1 serine protease (A304G), however, this mutation was also present in 1/5 9-week lineages of TEV. Another convergent nonsynonymous mutation was found in 3/7 9-week lineages in the P3 protein (U4444C), known to be implicated in virus amplification and host adaptation (Revers and García 2015). For the evolved TEV-NIaPro$_2$-NIaPro$_8$ virus fixed convergent nonsynonymous mutations were found in the duplicated *NIa-Pro* (C1466U) copy in 4/5 3-week lineages, and in *6K1* (A4357G) in 3/5 3-week lineages. The latter mutation was also found fixed in TEV in 1/5 3-week lineages. For the evolved TEV-NIb$_1$-NIb$_9$ virus a fixed convergent nonsynonymous mutation was found in the pseudogenized *NIb* copy (A1643U) in 2/5 3-week lineages. For the evolved TEV-NIb$_2$-NIb$_9$ virus one fixed synonymous mutation was found in the multifunctional CI protein (C6531U) in 2/5 9-week lineages. Other convergent mutations in all virus genotypes were found in *VPg*, *NIa-Pro* and *NIb*, however these mutations were also found in 2 or more lineages in the wild-type virus (**Fig C2.8**). Therefore, we do not consider these genotype specific adaptive mutations.

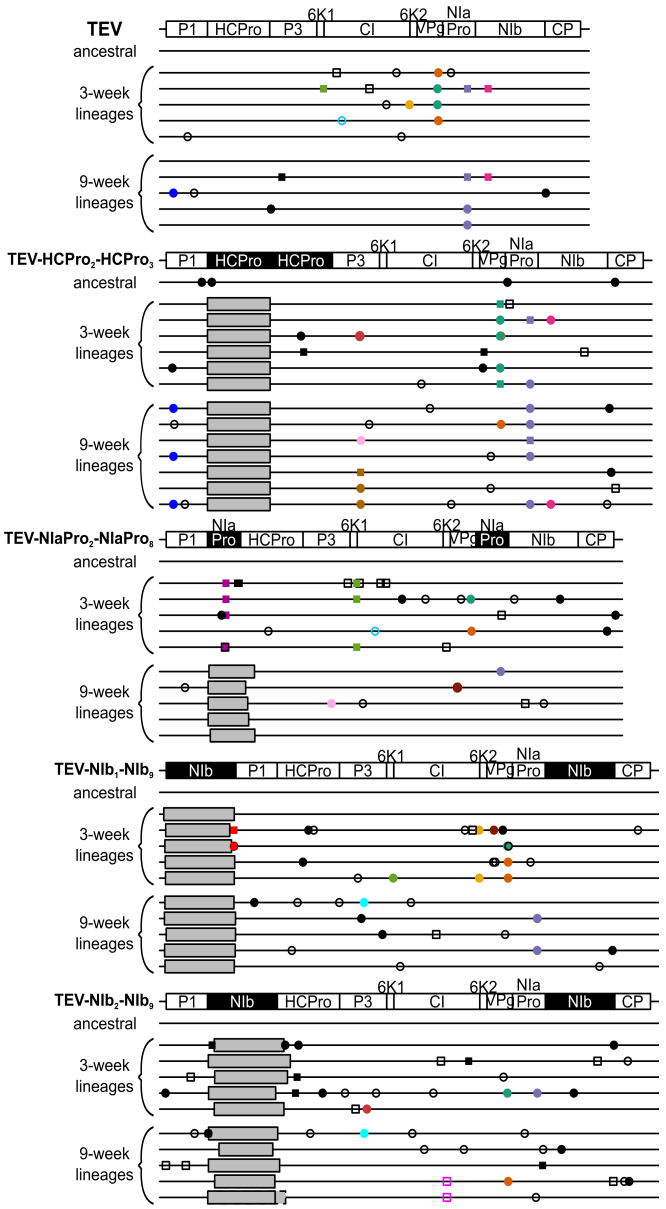

Figure 2.8. Genomes of the ancestral and evolved lineages. Mutations were detected using NGS data of the evolved virus lineages as compared to their ancestral lineages. The square symbols represent mutations that are fixed (> 50%) and the circle symbols represent mutations that are not fixed (< 50%). Filled symbols represent nonsynonymous substitutions and open symbols represent synonymous substitutions. Black substitutions occur only in one lineage, whereas color-coded substitutions are repeated in two or more lineages, or in a lineage from another virus genotype. Note that the mutations are present at different frequencies as reported by SAMtools. Grey boxes indicate genomic deletions in the majority variant.

Table C2.1. Adaptive convergent mutations within each virus genotype

Virus genotype	nt change at ancestral position	aa change	gene	nt position in gene
TEV-HCPro$_2$-HCPro$_3$	A304G	I→V	P1	160
	U4444C	S→P	P3	625
TEV-NIaPro$_2$-NIaPro$_8$	C1466U	S→L	NIa-Pro	410
	A4357G	I→V	6K1	148
TEV-NIb$_1$-NIb$_9$	A1643U	Y→F	NIb$_1$	1499
TEV-NIb$_2$-NIb$_9$	C6351U	Y→Y	CI	1173

nt: nucleotide; aa: amino acid

Although there are some convergent single nucleotide mutations, in most cases these occur only in a small fraction of lineages. The transient presence of the duplicated *NIa-Pro* copy in the 3-week lineages does seem to be linked to an adaptive mutation. However, the fitness cost of an increasing genome size cannot be overcome by this single nucleotide mutation. The main change in the evolved lineages is the deletion of the duplicated gene copy, which makes the different viral genotypes similar to the wild-type virus.

2.5. Genomic stability of TEV with duplications of homologous genes

To better understand the evolutionary dynamics of viruses with gene duplications, we developed a simple model of virus evolution that considers the maintenance of gene duplications. Based on amplicon sizes, the genome size for all evolved lineages was estimated for every passage (**Fig C2.4**). The genome size over time is a good indicator of genomic stability of homologous gene duplications within TEV. We developed a model to fit with the experimental

data for the 3-week and 9-week passages. The model consists of two coupled ordinary differential equations:

(1) $$\frac{dA}{dt} = rA\left(1 - \frac{A + \beta B}{\kappa_t}\right) - \Delta A$$

(2) $$\frac{dB}{dt} = sB\left(1 - \frac{B + \alpha A}{\kappa_t}\right) + \Delta A$$

where A is the number of virions containing a gene duplication, B is the number of virions with a reversion to a single copy, r is the initial growth rate of A, s is the initial growth rate of B, $\beta = s/r$ is a constant for determining the effect of the presence of B on replication of A, $\alpha = 1/\beta$ is a constant for determining the opposing effect of A on B, κ is the time-dependent carrying capacity of the host plant, and Δ is the recombination rate. We assume that κ increases linearly over time, being proportional to the estimated weight of collected plant tissue (2 g for the whole plant at inoculation, 200 g for the collected leaves at 9 weeks). At the start of each round of infection, there is a fixed bottleneck size of λ. The number of infecting virions of A is determined by a random draw from a binomial distribution with a probability of success $\lambda A/(A+B)$ and a size λ, and the number of infecting virions of B is then λ minus this realization from the binomial distribution.

Estimates of most model parameters could be obtained from previous studies. An estimate of s is provided by (Zwart *et al.* 2012), whilst r can be determined knowing the competitive fitness of the virus with duplication relative to the wildtype virus. The value of s used is 1.344, and r values are 1.175, 1.234, 1.185 and 1.134 for TEV-HCPro$_2$-HCPro$_3$, TEV-NIaPro$_2$-NIaPro$_8$, TEV-NIb$_1$-NIb$_9$, and TEV-NIb$_2$-NIb$_9$, respectively. The only model parameter that needed

to be estimated from the data is the recombination rate Δ. This parameter was therefore estimated individually for each data set simply by running the model for a wide range of recombination rates (10^{-20} to $10^{-0.1}$). One-thousand simulations were run for each parameter value.

The model was fitted to the experimental data by minimizing the multinomial likelihood of the observed categories: (*i*) populations with only *A*, (*ii*) populations with only *B*, and (*iii*) mixed populations of *A* and *B*. In all large populations recombinants wills arise almost instantaneously, but they are not necessarily carried over to the next round of passaging, due to the bottleneck at the start of infection. Nor would the RT-PCR assay necessarily detect all variants (Zwart *et al.* 2011; Majer *et al.* 2013). Given the detection constraints imposed by the assay, the threshold frequency was set to 0.1 for the detection of both virus variants (*A* and *B*).

The main question we can ask is whether knowing the fitness of duplicated viruses (*i.e., s*) is sufficient information to predict the stability of the inserted gene. Or does the data support a context-dependent recombination rate, with the context being (*i*) identity and position of the duplication, (*ii*) passage length, or (*iii*) both. Therefore, we considered four different situations that are represented in the following models:

Model 1: one recombination rate for all conditions (1 parameter);

Model 2: virus-genotype-dependent recombination rate (4 parameters);

Model 3: passage-duration-dependent recombination rates (2 parameters);

Model 4: virus-genotype- and passage-duration-dependent recombination rates (full model, 8 parameters).

The model estimates of Δ are given in **Table C2.2**. Note that the parameter is often a minimum (when the virus is very unstable) or a maximum value (when the virus is very stable). If the optimum is represented by more than one parameter value, the mean of these values is given.

Table C2.2. Model parameter estimates for deterministic recombination rate

Model	Estimates of $Log_{10}[\Delta]$ (Lower 95% fiducial limit, upper 95% fiducial limit)			
1	−6.2 (*)			
2	2HCPro ≥ −3.0 (*)	2NIaPro = −9.65 (−10.1, −9.0)	2NIb1 ≥ −2.9 (*)	2NIb2 = −4.45 (−5.2, −3.7)
3	3W = −6.2 (*)	9W = −9.65 (−10.1, −7.5)		
4	2HCPro 3W ≥ −3.0 (*)	2NIaPro 3W ≤ −9.0 (*)	2NIb1 3W ≥ −2.9 (*)	2NIb2 3W = −4.45 (−5.5, −3.7)
	2HCPro 9W ≥ −10.5 (*)	2NIaPro 9W = −9.6 (−10.1, −7.5)	2NIb1 9W ≥ −10.1	2NIb2 9W ≥ −11.5 (*)

* indicates the fiducial limit is identical to the parameter estimate, also when the parameter estimate is a range, 2HCPro: TEV-$HCPro_2$-$HCPro_3$; 2NIaPro: TEV-$NIaPro_2$-$NIaPro_8$; 2NIb1: TEV-NIb_1-NIb_9; 2NIb2: TEV-NIb_2-NIb_9; 3W: 3-week passages; 9W: 9-week passages

When comparing these models, we found that Model 2 is the best-supported model. Thus, only a genotype-dependent recombination rate is required to account for the data.

The fact that a passage-duration-dependent recombination rate is not required to account for the data, strongly suggests that the differences in results of the 3-week and 9-week passages can be explained entirely by the selection coefficients and the genetic bottleneck. For the model, the 3-week passages deletion variants are generated during infection, but the relatively short time

between bottleneck events does not allow them to reach a high enough frequency to have an appreciable chance of being sampled.

Table C2.3. Model selection for models with deterministic recombination

Model	Parameters	NLL	AIC	ΔAIC	Akaike Weight
1	1	254.284	510.569	443.470	0.000
2	4	29.549	67.099	−0.000	0.982
3	2	226.669	457.339	390.240	0.000
4	8	29.549	75.099	8.000	0.018

2.6. Concluding remarks

Here we test the stability of duplicated sequences that might contribute to the enhancement of a virus function or even exploration of new functions. None of the duplication events explored appeared to be beneficial for TEV overall. Either duplication of a gene resulted in an unviable virus or a significant reduction in viral fitness. In all cases the duplicated gene copy was deleted using the longer-duration-passages. The reduction in genome size resulted in increases in within-host competitive fitness. Nevertheless, knowing only the fitness levels is not enough to predict the stability of the duplicated genes. By fitting a model of virus evolution to the data, we show there is a context-dependent recombination rate, and specifically, where the identity and position of the duplication play a role. Given that the supply rate of variants with large deletions will be driven largely by homologous recombination, we expected such a context dependence. It would be highly interesting to know which other biological features constrain the likelihood of maintenance of duplicated genes.

Genetic redundancy is evolutionary unstable as it requires the maintenance of both gene copies for a long period after a duplication event. In this study we showed that viruses do not restore their fitness after gene duplication. This fitness cost can be related to three processes: (*i*) the increase in genome size, (*ii*) the extra cost of more proteins being expressed in the context of using more cellular resources, and (*iii*) a disturbance in correct polyprotein processing.

In addition to gene duplications, the model developed in this study can be applied to predict the stability of other types of gene insertions, like horizontal gene transfer. Understanding the stability of gene insertions in genomes has important implications within the context of biotechnology, expression systems and genome-architecture evolution.

Chapter 3: Introduction of functional exogenous sequences

1. Introduction

Viruses play an important evolutionary role as vectors for horizontal gene transfer (HGT) in the genomes of theirs hosts. Integrated virus genomes are commonly found in prokaryotes. For example, prophages can constitute 10-20% of bacterial genomes (Canchaya *et al.* 2003; Casjens 2003), which serve as vectors for HGT between bacteria and contribute to the genetic variability in several bacterial species. Among eukaryotes HGT between double stranded RNA viruses is widespread, and may play significant roles in their evolution (Liu *et al.* 2010). In addition to the transfer from viruses to host cells, HGT also occurs commonly from host cells to viruses and from viruses to viruses. For example, phylogenetic analysis suggest that many Mimivirus genes were acquired by HGT from the host organism or bacteria that parasitize the same host (Moreira and Brochier-Armanet 2008). Another example of HGT from host to virus is illustrated by plant closteroviruses, that encode homologs of Hsp70 (heat shock protein, 70 kDa). Viruses do generally not encode Hsp70s, however they are often recruited to aid virion assembly or genome replication (Cripe *et al.* 1995; Kelley 1998). In the case of closteroviruses, it seems that the *hsp70* gene was acquired by a common ancestor through recombination with a host mRNA coding for Hsp70 (Dolja *et al.* 1994).

As illustrated by the examples above, HGT is a widespread key mechanism in virus evolution. However, strong selection against increasing genome sizes in

viruses is an impediment to HGT (Chapter 2, Zwart *et al.* 2014; Holmes 2003). An exogenous sequence that is transferred to the recipient virus, must be accommodated into the virus genome. We postulate that accommodation of gene sequences could entail the regulation of expression levels, fine-tuning of interactions with other viral or host proteins, and the optimization of codon usage. However, the incorporation of exogenous sequences in a viral genome will generally have an appreciable fitness cost, typically resulting in the rapid loss of exogenous sequences. When a new sequence which may potentially be beneficial for the virus is incorporated in the genome, a race against time begins: will the new sequence be accommodated or will it be purged? The high instability of viral genomes may therefore hinder evolutionary innovation, by shortsightedly purging elements that might have been beneficial in the longer term (Chapter 1 and 2).

As a first exploration of this conundrum, Zwart *et al*. (2014) considered the stability of a fluorescent marker (eGFP) incorporated in the TEV genome. As the fluorescent marker itself is not expected to be able to acquire any functionality, this allows consideration of stability when the exogenous sequence has no potential to be beneficial, and the only accommodation that is possible is attenuation of fitness costs associated with the sequence. The eGFP sequence was quickly lost in long duration passages (Zwart *et al.* 2014), in agreement with work on the stability of other marker genes in the TEV genome (Dolja *et al.* 1993; Majer *et al.* 2013). Although these studies contribute to a better understanding of the stability of exogenous sequences, for understanding HGT it is important to consider experimentally the evolutionary fate of sequences that could potentially lead to improved or even new functions in virus replication or transmission.

In this study we experimentally explore the evolutionary fate of functional exogenous sequences introduced into the TEV genome, considering functions that could potentially be beneficial to the virus. Plant RNA viruses counteract the host plant's immune responses, and in the case of TEV the multi-functional HC-Pro protein is the suppressor of RNA silencing, a key immune response in plants. HC-Pro can be divided into three regions: (*i*) an N-terminal region associated with aphid transmission, (*ii*) a C-terminal region associated with proteinase and RNA silencing suppressor activities (Varrelmann *et al.* 2007), and (*iii*) a central region implicated in many functions, including RNA silencing suppression activity (Plisson *et al.* 2003). The silencing suppressor activity of HC-Pro is not always strong enough to nullify silencing effects on virus replication (Gammelgard *et al.* 2007). Moreover, the silencing suppression activity of this multifunctional protein may be subject to pleiotropic constraints, which arise due to its other functions.

We have simulated two possible HGT events that are related to RNA silencing suppression and might be beneficial for TEV. The first event represents the acquisition of a heterologous gene that duplicates a function that was already present in TEV, generating functional redundancy. We have chosen to use the *2b* gene from *Cucumber mosaic virus* (CMV; genus *Cucumovirus*, family *Bromoviridae*). The multi-functional 2b protein is implicated in polyprotein cleavage, virulence, viral movement, aphid-borne transmission and suppression of RNA silencing (Ding *et al.* 1995, 1996; Li *et al.* 1999; Guo and Ding 2002; Shi *et al.* 2003). However, the multi-functional HC-Pro protein in TEV also acts as a RNA silencing suppressor. The addition of a second silencing suppressor to the TEV genome could have both immediate benefits for replication, as well as second-order evolutionary benefits. Immediate benefits could arise simply due

to stronger silencing-suppression activity. Second-order benefits could occur during evolution, because pleiotropic constraints on HC-Pro's other functions have been ameliorated by the introduction of the *2b* gene and these functions can therefore be improved.

The second genome reorganization event represents the acquisition of a function that was not originally present in TEV. We have chosen to add an AlkB domain, which TEV in principle could acquire from one of its natural hosts. The AlkB protein is responsible for the repair of alkylation damage in DNA and RNA bases. Alkylating agents that produce this damage can be present in the environment and inside cells (Sedgwick *et al.* 2007). Homologues of AlkB are widespread in eukaryotes and bacteria (Aravind and Koonin 2001; Kurowski *et al.* 2003), as well as in plant RNA viruses (Aravind and Koonin 2001). Although the overall sequence similarity among the AlkB family is low, a conserved domain exists characterized as the 2OG-Fe(II) oxygenase superfamily (Aravind and Koonin 2001; Finn *et al.* 2014). The 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenases are a class of enzymes that commonly catalyze the oxidation of an organic substrate using a dioxygen molecule. The AlkB protein in plant RNA viruses is involved in the repair of methylation damage (van den Born *et al.* 2008). The host plant may use methylation as a defense mechanism to inactivate viral RNAs. Therefore, it has been suggested that AlkB homologues are involved in counteracting this defense mechanism (Bratlie and Drabløs 2005). However, selection for maintenance of the conserved AlkB domain could also involve environmental factors, such as pesticide use. AlkB domains are mainly found in plant viruses belonging to the *Flexiviridae* family, which are, like TEV, positive-sense single-stranded RNA viruses. Within the *Potyviridae* family, only one virus with an AlkB domain has

been reported: *Blackberry virus Y* (BVY) (Susaimuthu *et al.* 2008). In this particular virus AlkB is located within the P1 serine protease. Interestingly, BVY lacks the N-terminus of HC-Pro that is present in the potyvirus orthologs. This missing region is involved in silencing suppression and vector transmission (Revers *et al.* 1999; Urcuqui-Inchima *et al.* 2001; Stenger *et al.* 2006; Young *et al.* 2007). These are important properties for a potyvirus. It may be that AlkB has evolved to take over these functions in BVY. To test this idea in a model system, we have introduced AlkB within the P1 protein in TEV.

In this study we experimentally explore two different events of HGT in TEV; the acquisition of a gene that duplicates an existing function, and the acquisition of a new function. We found that the newly introduced *AlkB* gene is unstable in TEV in our experimental setup, and therefore a HGT event to acquire this new function from the host plant is unlikely. On the contrary, the functional redundant *2b* gene from CMV is maintained after more than half a year of evolution. By generating knockout mutants in two different regions of HC-Pro, we show that TEV carrying *2b* can induce wild-type symptoms while the wild-type TEV cannot. Our results suggest that *2b* can actually perform a function within the TEV genome.

2. Results and Discussion

2.1. Introducing an existing function

To simulate HGT between two virus species, we have introduced the 2b RNA silencing suppressor from CMV within the TEV genome (**Fig C3.1A**), between the *NIb* replicase gene and the *CP* (**Fig C3.1B**). Both ends of *2b* were modified to provide similar proteolytic cleavage sites for NIa-Pro as the ones that

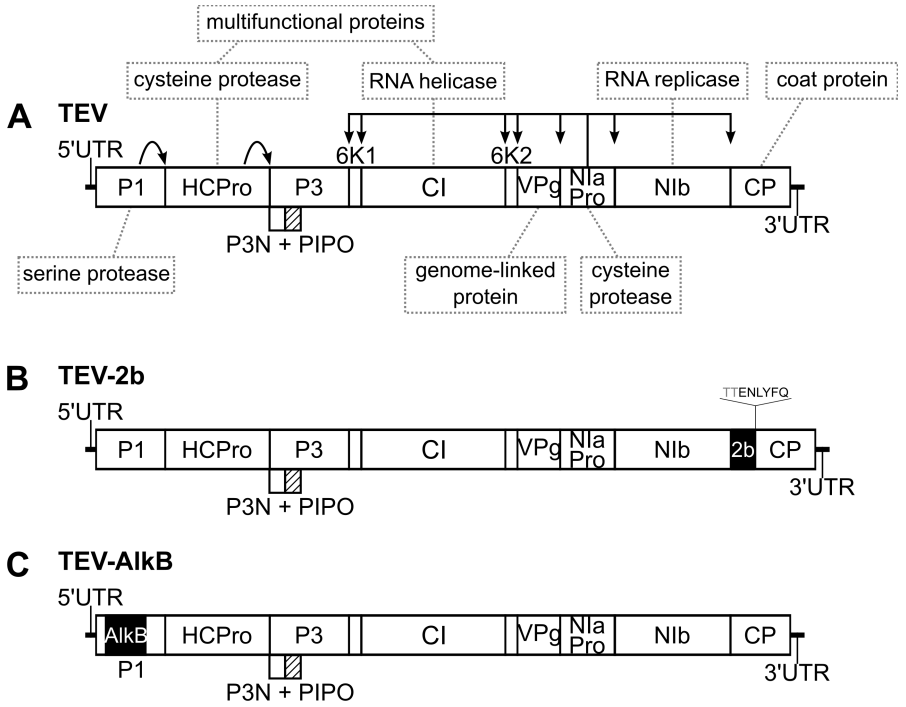

Figure C3.1. Schematic representation of the Tobacco etch virus genotypes with exogenous gene insertions. The wild-type TEV (**A**) codes for 11 mature peptides, including P3N-PIPO embedded within the 3 protein at a +2 frameshift. Relevant names of the protein products are indicated in the grey dotted boxes. The *2b* gene from *Cucumber mosaic virus* was introduced between *NIb* and *CP* (**B**). Both ends of 2b were modified to provide proteolytic cleavage sites. An AlkB domain from *Nicotiana silvestris* was introduced within the *P1* gene (**C**).

originally existed between *NIb* and *CP*. As the multifunctional HC-Pro protein also acts as a silencing suppressor within TEV, the introduction of *2b* generates functional redundancy within TEV. By evolving the functionally redundant TEV-2b virus we imagined the virus could go down three different trajectories: (*i*) either the 2b gene is genuinely redundant and will therefore be purged from the TEV genome over evolutionary time; or (*ii*) the fitness costs are minimal and the transgene is stable over evolutionary time; or (*iii*) the *2b* is selected for and will be maintained throughout evolutionary time. In the latter scenario, it is possible that the *2b* gene takes over the silencing suppressor function of HC-Pro, allowing HC-Pro to specialize one of its other functions. Nonetheless, *2b* could also be maintained because it increases the robustness of TEV, buffering the effects of deleterious mutations in the silencing suppressor region (and possibly other regions) of HC-Pro.

When we evolved the TEV-2b virus in *N tabacum* plants, using 3-week and 9-week passages, we never observed any deletions when RT-PCR amplifying the region encompassing *2b*. This results contrast all previous results with longer duration passages, where deletions were observed after genetic redundancy by means of gene duplications (Chapter 1 and 2), and after the introduction of a non-functional sequence (Dolja *et al.* 1993; Zwart *et al.* 2014). Therefore, we speculated that the TEV-2b virus might have a fitness advantage over the wild-type TEV. To test this idea, we measured within-host competitive fitness and viral accumulation of both the ancestral and evolved lineages. Additionally, we measured plant height as a measure of virulence; healthy plants grow higher compared to plants infected with TEV. We found that the ancestral TEV and TEV-2b have a very similar within-host competitive fitness (**Fig C3.2A**; *t*-test: $t_4 = 2.008$, $P = 0.115$), viral accumulation (**Fig C3.2B**; *t*-test: $t_4 = 1.389$, $P =$

0.237), and plants grow to similar heights when infected with either of these viruses (**Fig C3.2C**; *t*-test: t_4 = 0.447, P = 0.678). Comparing the evolved lineages, within-host competitive fitness of the TEV-2b lineages similar to those of TEV (**Fig C3.2A**; Mann-Whitney U = 4, P = 0.095). In addition, no significant difference in accumulation (**Fig C3.2B**; Mann-Whitney U = 18, P = 0.310) or plant height (**Fig C3.2C**; Mann-Whitney U = 8.5, P = 0.462) was found between the evolved lineages. All plants infected with TEV or TEV-2b were significantly lower in height compared to the healthy control plants (**Fig C3.2C**; asterisks indicate significance of a *t*-test with Holm-Bonferroni correction). These results therefore do not provide any evidence that the inserted *2b* gene is providing any advantage for TEV, whilst they also show that there do not appear to be any net fitness costs for having *2b* inserted.

All evolved and ancestral lineages were fully sequenced by Illumina technology. The sequences of the ancestral lineages were used as an initial reference for mapping of the evolved lineages. When looking at the genome sequences of the evolved TEV-2b lineages, we find no evidence of convergent adaptive evolution other than the evolved wild-type TEV lineages (**Fig C3.3**). Even so, one mutation that appears in the evolved wild-type lineages in *NIb* (U7262C), does not appear in the evolved TEV-2b lineages. The presence of *2b* could prevent the occurrence of this mutation. However, this is not supported by solid evidence as this nonsynonymous mutation only appears in 2/10 evolved wild-type TEV lineages.

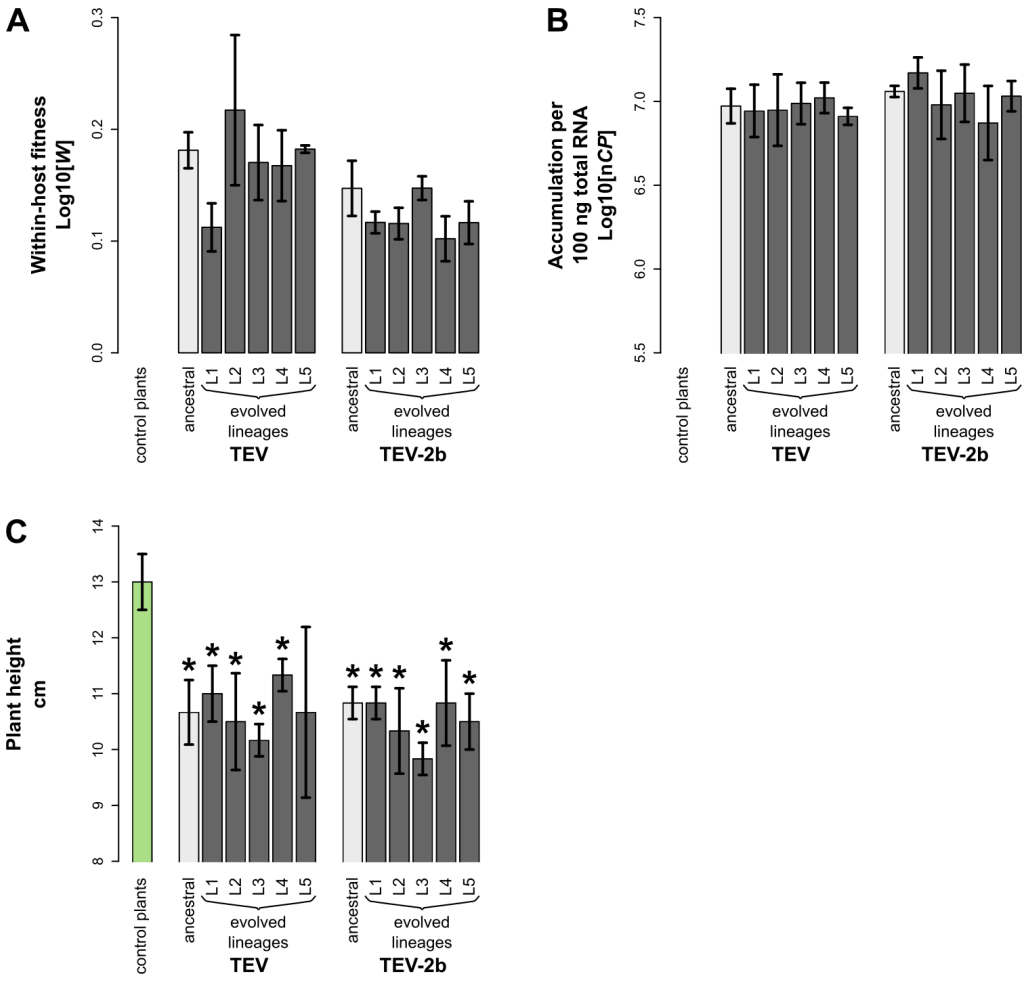

Figure C3.2. Fitness assays of the ancestral and evolved TEV-2b lineages. (**A**) Within-host competitive fitness (*W*), as determined by competition experiments and RT-qPCR, of TEV and TEV-2b with respect tot a common competitor; TEV-eGFP. (**B**) Virus accumulation (*nCP*), as measured by RT-qPCR, of TEV and TEV-2b. (**C**) Plant height, measured in cm, of TEV and TEV-2b. All three measurements were done at 7 dpi. The ancestral lineages are indicated with light-gray bars and the evolved lineages with dark-gray bars. The healthy control plants are indicated with a green bar. Both TEV and TEV-2b were evolved using 5 replicate lineages each (L1-L5), for a total of 27 weeks using three 9-week passages. Lineages that tested significantly different in plant height, compared to the healthy control plants, are indicated with an asterisk (*t*-test with Holm-Bonferroni correction for multiple tests).

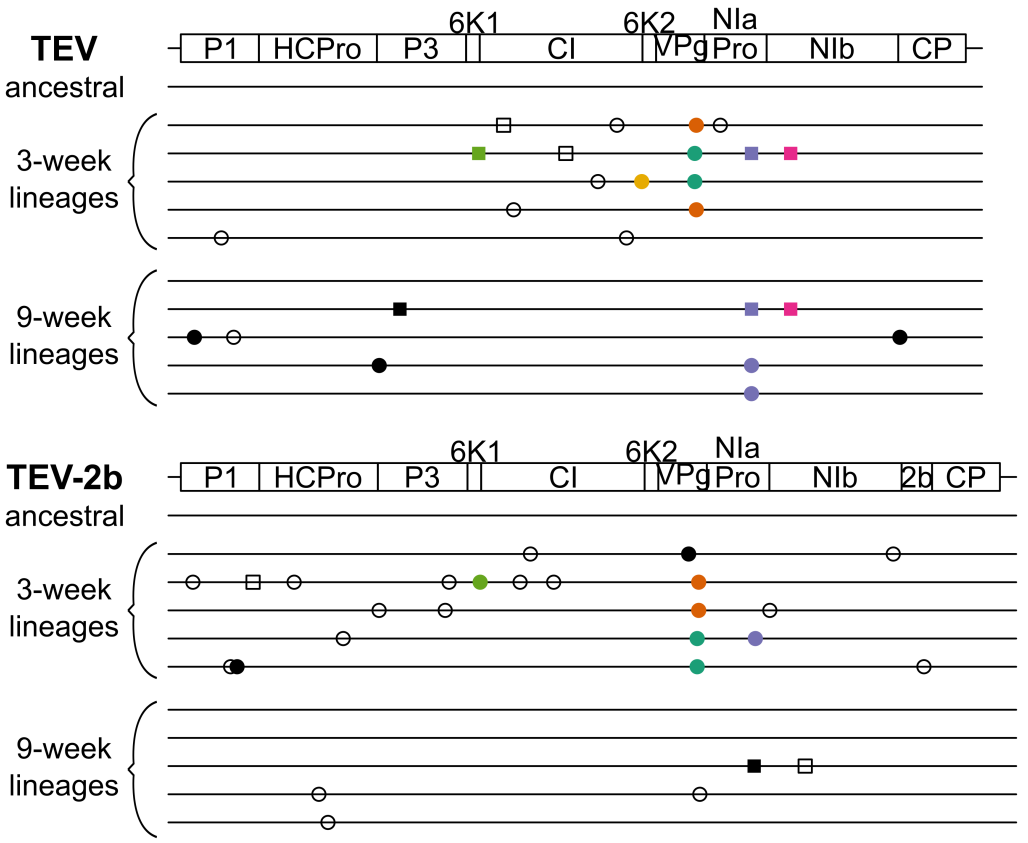

Figure C3.3. Genomes of the ancestral and evolved TEV-2b lineages. Mutations were detected using NGS data of the evolved virus lineages as compared to their ancestral lineages. The wild-type TEV is given for comparative purposes. The names in the left identify the virus genotypes and the lineages with their corresponding passage length. The square symbols represent mutations that are fixed (> 50%) and the circle symbols represent mutations that are not fixed (< 50%). Filled symbols represent nonsynonymous substitutions and open symbols represent synonymous substitutions. Black substitutions occur only in one lineage, whereas color-coded substitutions are repeated in two or more lineages, or in a lineage from another virus genotype. Note that the mutations are present at different frequencies as reported by SAMtools.

After remapping the cleaned reads against a new defined consensus sequence for each lineage, we looked at the variation within each lineage. Single nucleotide polymorphisms (SNPs) were detected from a frequency as low as 1%. In the evolved TEV-2b lineages a total of 507 (390 unique) SNPs were detected, with a median of 30.5 (17-149) per lineage. In the evolved TEV lineages a total of 379 (326 unique) SNPs were detected, with a median of 35 (17-63) per lineage. The increase in SNP count for TEV-2b is apparent in the number of nonsynonymous mutations, which is 277 SNPs (192 unique) in TEV-2b versus 141 SNPs (108 unique) in TEV. More SNPs accumulate within the *P1*, *HC-Pro*, *CI* and *NIb* genes in the evolved TEV-2b lineages compared to those of the wild-type. However, most of these SNPs are present at a low frequency. **Fig C3.4** illustrates the differences in the distribution of the SNP frequencies at each nucleotide position in the evolved TEV and TEV-2b genomes. The differences between TEV and TEV-2b in both the distribution of the SNP frequencies (Kolmogorov–Smirnov test; $D = 0.240$, $P < 0.001$) and the SNP positions (Kolmogorov–Smirnov test; $D = 0.094$, $P = 0.043$) are significant.

The accumulation of a higher number of SNPs suggests that the *2b* could make the virus more robust against deleterious mutations. Another possibility is that the fitness cost of *2b*, which is small in size, is so low that there is little selection for viruses with genomic deletions, and the insert would only be eventually removed by genetic drift. However, the introduction of two non-functional genes, the colored markers Rosea1 (Ros1) and eGFP, at the same position, did lead to instability and eventual deletion of these markers over evolutionary time (Majer *et al.* 2013).

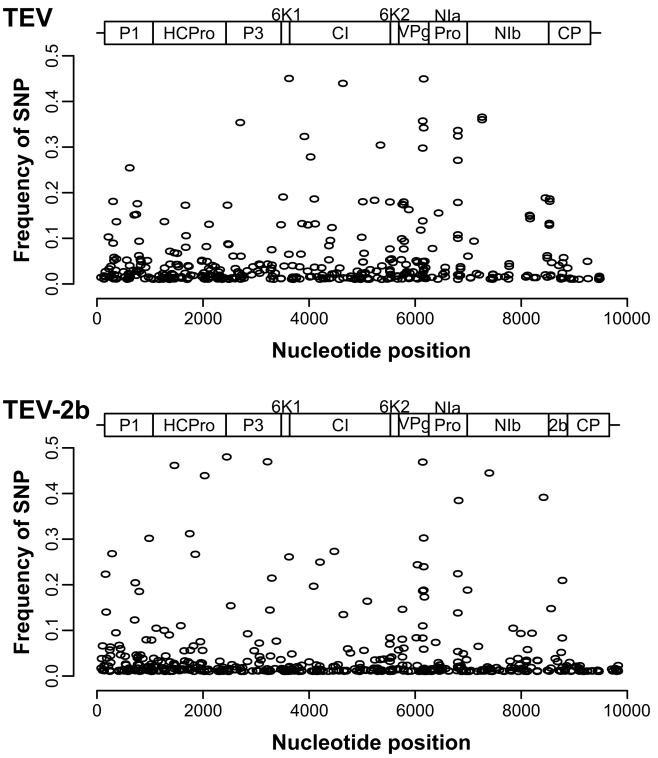

Figure C3.4. SNP frequencies within the evolved TEV-2b lineages. SNPs were detected within each population of the 10 independently evolved virus lineages. SNP detection was done from a frequency as low as 1%. The wild-type TEV is given for comparative purposes. The SNP frequencies are plotted (ordinates) at their corresponding nucleotide positions (abscissas). The genomes of TEV and TEV-2b, with the genes at the corresponding nucleotide positions are schematically represented above the plots.

Neither the genome sequences of the evolved populations, nor the fitness and virulence assays, suggest that *2b* is actually performing a function or adaptively accommodated in TEV. To test whether *2b* can actually perform a function when inserted in the TEV genome, we therefore generated knockout mutations in the silencing suppressor region of HC-Pro.

Two different mutant genotypes were considered. One mutant genotype consists of an amino acid change in the FRNK box, located in the central region of HC-Pro, to the FINK amino acid sequence. The FRNK box is a highly conserved motif within potyviruses which is required for small RNA binding and this activity is correlated with symptom severity (Shiboleth *et al.* 2007; Wu *et al.*

2010). The R183I mutation affects viral pathogenicity and has led to a significant reduction in symptom severity in other potyviruses (Gal-On 2000; Lin *et al.* 2007; Kung *et al.* 2014). The other mutant genotype has the E299A amino acid substitution, in the functional domain that is a protease and an RNA silencing suppressor, as well as being involved in viral movement. This mutant genotype in TEV, hereafter referred to as AS13, has shown reduced suppressor activity, low accumulation levels, and does not induce viral symptoms in *N. benthamiana* plants (Torres-Barceló *et al.* 2008, 2010).

After generating the FINK and AS13 variants in the TEV and TEV-2b background sequences, we performed infectivity assays in *N. tabacum* plants. Based on the scoring of the symptoms, we deduced infection of plants with the TEV-2b^{FINK} and TEV-2b^{AS13} viruses, while no infection was observed for the TEV^{FINK} and TEV^{AS13} viruses (**Fig C3.5**). The symptoms for TEV-2b^{FINK} and TEV-2b^{AS13} appear later compared to the wild-type TEV and TEV-2b viruses, however at 7 dpi all plants are infected. Event though it seems that TEV^{FINK} and TEV^{AS13} do not infect *N. tabacum* plants, the possibility exists that plants are infected and that these viruses do not induce symptoms and accumulate at low levels.

These results strongly suggest that the *2b* gene from CMV can compensate for what would otherwise be deleterious mutations in the HC-Pro region. Therefore, it seems reasonable to conclude that *2b* might be functional as a silencing suppressor in TEV and compensate for the hypo-suppression in the AS13 mutant. Even though initial infection is slower, pathogenicity appears to be restored 10 dpi in both TEV-2b^{FINK} and TEV-2b^{AS13} genotypes.

Here we have considered the acquisition of a gene conferring an existing

function, in the context of HGT from another plant virus to TEV. We therefore introduced *2b* from CMV into the TEV genome. The data suggest that the hyposuppressor knockout mutants of TEV are transformed to hypersuppressors of RNA silencing, when TEV carries the *2b* gene.

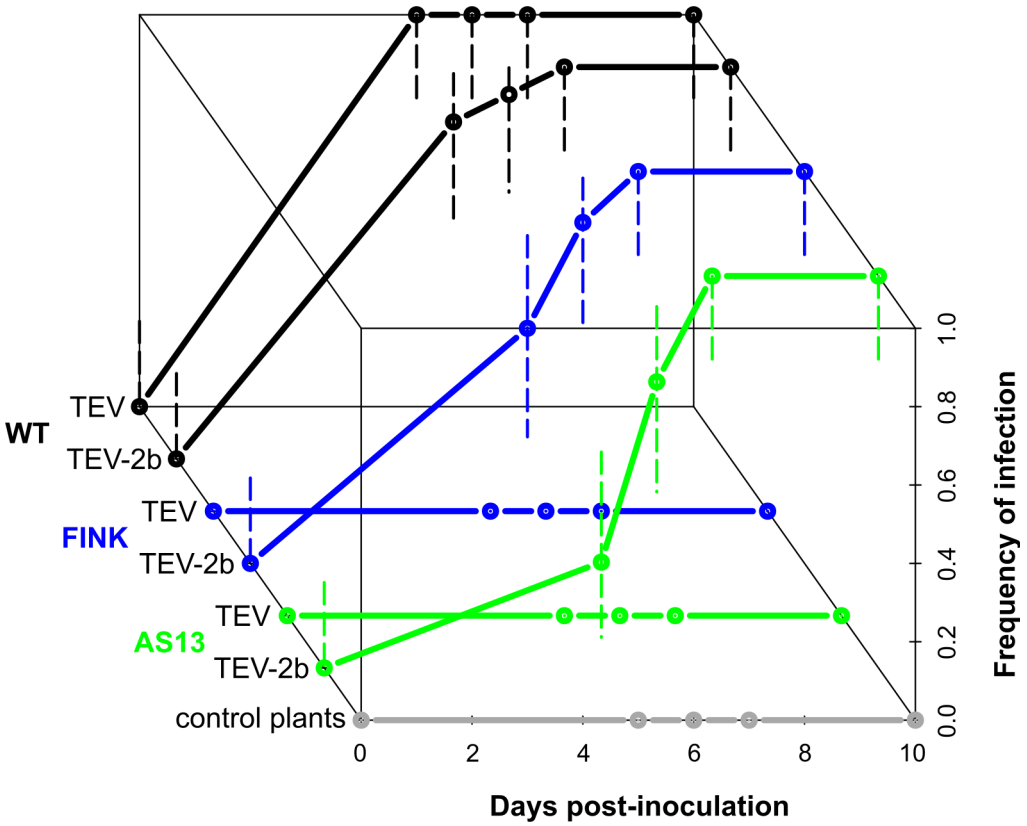

Figure C3.5. Infectivity of HC-Pro mutants. The cumulative frequency of infection in *N. tabacum* plants of the FINK and AS13 variants of TEV and TEV-2b versus days post-inoculation (dpi). Symptoms were screened at 5, 6, 7 and 10 dpi. Error bars represent the 95% confidence intervals of the estimated frequencies (binomial test).

2.2. Introducing a new function

Based on the position of *AlkB* in BVY, we have introduced the conserved 2OG-Fe(II) oxygenase domain of AlkB in the TEV genome (**Fig C3.1A** and **C**). To simulate HGT from host to virus, the introduced domain originated from *N. silvestris*. Avoiding to interrupt secondary RNA structures, the conserved AlkB domain was introduced within the *P1* gene, and hence there are no proteolytic cleavage sites. The TEV-AlkB virus (**Fig C3.1C**) was viable in *N. tabacum* plants, however, the symptoms induced in the host plant were very mild (**Fig C3.6**). As AlkB is also present within the host plant, the expression of AlkB in *N. tabacum,* hereafter referred to as Nt-AlkB, was measured by RT-qPCR relative to two stable reference genes (Schmidt and Delaney 2010). To avoid amplification of the conserved "viral" AlkB domain in the TEV-AlkB infected plants, primers outside the conserved domain were used to solely measure AlkB expression of the host plant. The expression of Nt-AlkB was higher in uninfected plants, compared to plants infected with TEV (**Fig C3.7**; *t*-test reference *L25*: $t_4 = 9.520$, $P < 0.001$; reference *EF-1α*: $t_4 = 27.592$, $P < 0.001$) or TEV-AlkB (**Fig C3.7**; *t*-test

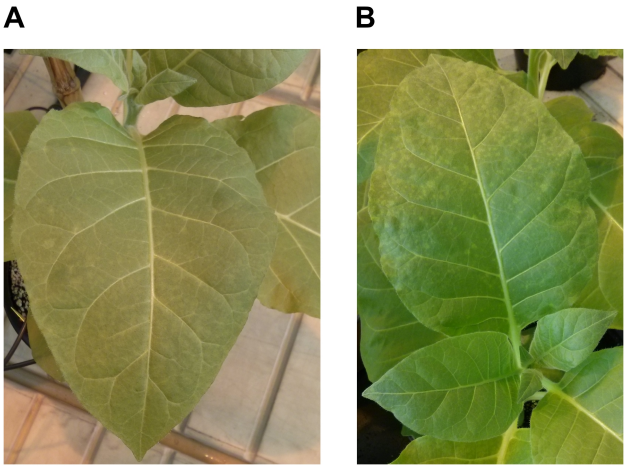

Figure C3.6. Symptoms TEV-AlkB in *N. tabacum*. Two examples of leaves displaying the mild symptoms induced by TEV-Alkb, several small chlorotic spots were observed (**A**), that in some lineages were observed at a higher density (**B**).

reference *L25*: $t_4 = 4.767$, $P = 0.009$; reference *EF-1α*: $t_4 = 13.863$, $P < 0.001$). The plants infected with TEV-AlkB, however, expressed Nt-AlkB at a significantly higher level than plants infected with TEV (*t*-test reference *L25*: $t_4 = 3.305$, $P = 0.030$; reference *EF-1α*: $t_4 = 8.775$, $P < 0.001$).

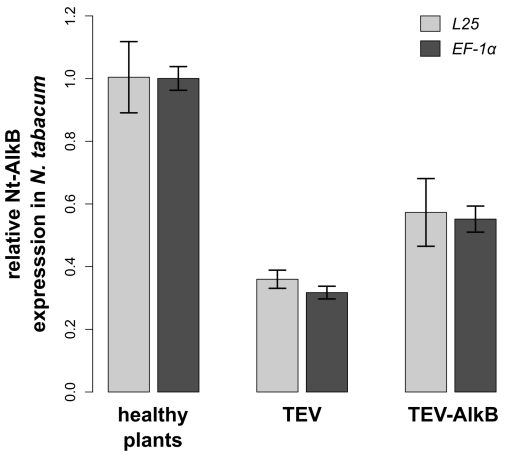

Figure C3.7. Relative Nt-AlkB expression of plants infected with TEV-AlkB. The expression of Nt-AlkB in *N. tabacum* plants was measured by RT-qPCR, relative to the two stable reference genes *L25* (ribosomal protein) and *EF-1α* (elongation factor 1α). Relative expression levels are given in healthy plants, plants infected with the wild-type TEV, and plants infected with TEV-AlkB. Light-gray bars indicate that L25 was used as a reference, and dark-gray bars indicate that *EF-1α* was used.

We then serially passaged the TEV-AlkB virus for 27 weeks, with each passage having a duration of either 3 or 9 weeks. We observed a variety of deletions already in the first passage, when RT-PCR amplifying the region encompassing the AlkB domain (**Fig C3.8**). A variety of small deletion variants were observed in the first 3-week passage (**Fig C3.8A**), while after a single 9-week passage there were apparently larger deletions (**Fig C3.8**B). At the end of the evolution experiment all TEV-AlkB lineages contained large deletions within AlkB, making this domain inactive.

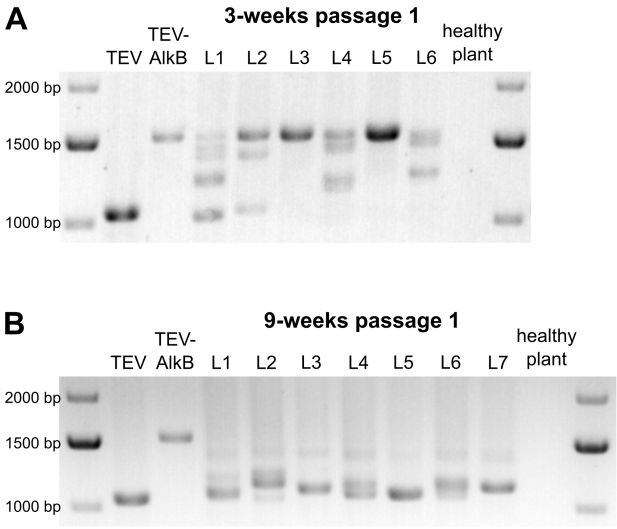

Figure C3.8. Deletion detection in the first passages of TEV-AlkB. Agarose gels with RT-PCR products of the region encompassing the AlkB domain. Deletions were detected in lineages (L1-L6) of the first 3-week passage (**A**) and in lineages (L1-L7) of the first 9-week passage (**B**). In both gels the product sizes of the TEV and TEV-AlkB are shown as a reference, followed by the independent lineages, followed by a healthy control plant.

Next, we measured within-host competitive fitness in direct competitions with a fluorescently labeled TEV, and we also measured viral accumulation. The introduction of the conserved AlkB domain within the TEV genome results in a significant decrease in within-host competitive fitness (**Fig C3.9A**; *t*-test comparing ancestral viruses: $t_4 = 7.846$, $P = 0.001$) and viral accumulation (**Fig C3.9B**; *t*-test comparing ancestral viruses: $t_4 = 3.587$, $P = 0.023$), compared to the wild-type TEV. After 27 weeks of evolution and the occurrence of partial or complete deletions of AlkB, both within-host competitive fitness (**Fig C3.9A**; comparing evolved lineages: Mann-Whitney $U = 10$, $P = 0.268$) and viral accumulation (**Fig C3.9B**; comparing evolved lineages: Mann-Whitney $U = 18$, $P = 1$) were restored to levels similar to the evolved lineages of the wild-type virus.

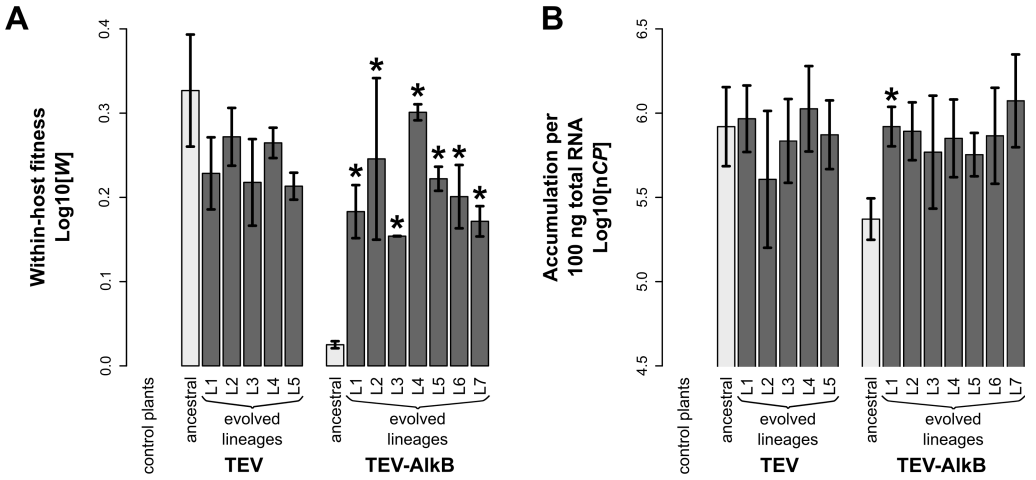

Figure C3.9. Fitness assays of the ancestral and evolved TEV-AlkB lineages. (**A**) Within-host competitive fitness (*W*), as determined by competition experiments and RT-qPCR, of TEV and TEV-AlkB with respect tot a common competitor; TEV-eGFP. (**B**) Virus accumulation (*nCP*), as measured by RT-qPCR, of TEV and TEV-AlkB. Both within-host competitive fitness and virus accumulation were measured at 7 dpi. The ancestral lineages are indicated with light-gray bars and the evolved lineages with dark-gray bars. Bot TEV and TEV-AlkB were evolved for a total of 27 weeks using three 9-week passages. For TEV 5 replicate lineages (L1-L5) were used, and for TEV-AlkB 7 replicate lineages (L1-L7) were used. Evolved lineages that tested significantly different compared to their ancestral virus are indicated with an asterisk (*t*-test with Holm-Bonferroni correction for multiple tests).

All evolved and ancestral lineages were fully sequenced by Illumina technology. After an initial mapping step using the ancestral sequence as a reference, the positions of the majority deletions were defined (**Fig C3.10**, Methods). Out of 13 evolved lineages, 8 lineages contain deletions within the inserted AlkB domain, of which 2 lineages contain an exact deletion of AlkB. For the remaining 5 lineages, deletions occurred either upstream (2/13 lineages) or downstream (3/13 lineages) AlkB, in the P1 serine protease. The P1 proteinase activity is not essential for viral infectivity (Verchot and Carrington

1995a), however the cleavage that separates P1 and HC-Pro is required (Verchot and Carrington 1995b). Based TEV-AlkB infectivity and the fitness data of the evolved lineages, it does not seem that the proteolytic activity of P1 is affected by the deletions in this gene.

When mutations were detected in the sequenced lineages, evidence for convergent evolution was found (**Fig C3.10**). A nonsynomous mutation was found within the AlkB domain (U439C) in 9/13 evolved lineages. However, in 8/9 lineages this mutation is present in the sequences of the minority deletion variants, in other words, this mutation falls within the deleted region of the majority deletion variant (**Fig C3.10**) and is supported by a low read coverage. Another convergent mutation was found in 12/13 evolved lineages in *VPg* (A6429C). This is a synonymous mutation and is fixed in all 12 lineages, however, this mutations was also found to be fixed within the ancestral population.

The reduction of Nt-AlkB expression in the host plant when infected with TEV, suggests that TEV might have a mechanism to suppress the expression of this gene in the host plant. Such activity would be beneficial for the host plant as a mechanism of immunity against pathogens like TEV. The introduction of the conserved AlkB domain within the TEV genome does not result in complete silencing of the plant *Nt-AlkB* gene. On the contrary, TEV-AlkB does not suppress the expression of Nt-AlkB to similar low levels as the wild-type virus does. The higher Nt-AlkB expression in the host plant when infected with TEV-AlkB is likely to be an effect slower replication of this virus. This is corroborated by the reduced symptomatology, within-host competitive fitness, and viral accumulation of TEV-AlkB. The pseudogenization of the AlkB

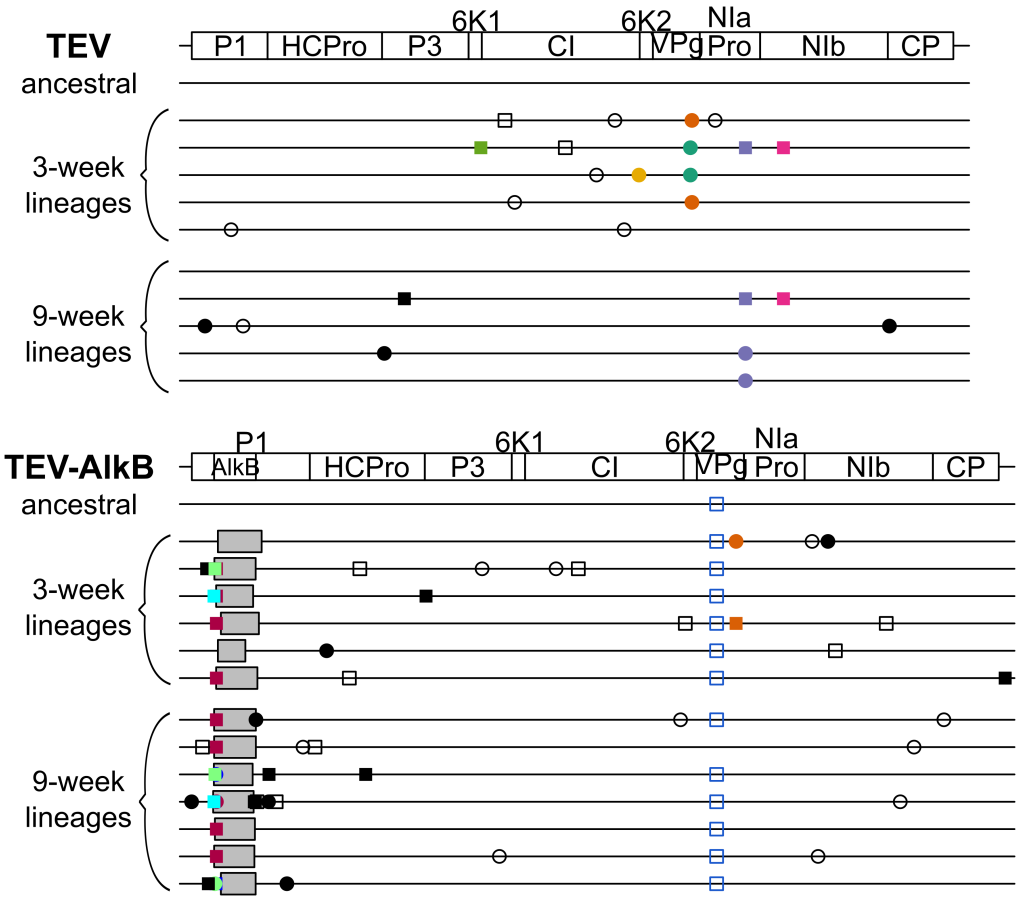

Figure C3.10. Genomes of the ancestral and evolved TEV-AlkB lineages. Mutations were detected using NGS data of the evolved virus lineages as compared to their ancestral lineages. The wild-type TEV is given for comparative purposes. The names in the left identify the virus genotypes and the lineages with their corresponding passage length. The square symbols represent mutations that are fixed (> 50%) and the circle symbols represent mutations that are not fixed (< 50%). Filled symbols represent nonsynonymous substitutions and open symbols represent synonymous substitutions. Black substitutions occur only in one lineage, whereas color-coded substitutions are repeated in two or more lineages. Note that the mutations are present at different frequencies as reported by SAMtools. Grey boxes indicate genomic deletions in the majority variant.

domain confirms that – for this experimental setup – this gene is not beneficial for TEV, or that the fitness effect of the conferred benefits is smaller than their cost on replication.

Here we have considered the acquisition of a gene conferring a new function, in the context of HGT from the host plant to TEV. We therefore introduced the he conserved 2OG-Fe(II) oxygenase domain of AlkB from *Nicotiana* into the TEV genome. This domain has very low similarity to the AlkB domain found in BVY, the only potyvirus with this domain. A similar experimental setup with the AlkB domain from BVY might therefore lead to a more readily accommodation in the TEV genome.

2.3. Concluding remarks

We have shown that HGT transfer of the conserved AlkB domain from a host plant to TEV is not very likely. The AlkB domain in BVY has orthologous domains of other viruses, and the similarity between these is higher than the similarity of BVY P1 orthologs in other potyviruses (Susaimuthu *et al.* 2008). Therefore, it is more likely that BVY has obtained this domain through a recombination event with another virus, instead of with its host.

On the contrary, we have shown that HGT of the *2b* gene from another virus family into the highly streamlined TEV genome, is a promising event. So far, from all insertions explored, this is the only stable gene insertion in TEV. In addition, this is the first gene insertion that has proven to be functional and beneficial for TEV when its own host plant defense mechanism fails.

Chapter 4: Fitness effects of exogenous sequences can be unpredictable in alternative hosts.

1. Introduction

From both applied and fundamental perspectives, virulence is a key phenotypic trait of microparasitic organisms. In medicine and agriculture, it is key to understand mechanistically how microparasites harm the host, in order to devise effective interventions. From a more fundamental perspective, evolutionary biologists have long been interested in understanding why many microparasites are highly virulent. It was originally suggested that virulence reduces between-host transmission, and selection would therefore act to maximize between-host transmission and reduce virulence (reviewed by Ewald 1983; Alizon et al. 2009). High virulence would signal maladaptation, for example following a host-species jump, and eventually be selected against. The ubiquity of microparasitic virulence and the fact that many virulent microparasites appeared to have high virulence led to a more sophisticated framework: the hypothesis that there are tradeoffs between virulence and transmission (de Roode *et al.* 2008; Alizon *et al.* 2009). This framework posits that high levels of replication could increase the probability of the microparasite being transferred to a new host, whilst also increasing the probability that the host would die quickly and the temporal window for transmission would be very brief. Under this more plausible framework, virulence evolves to the level that optimizes between-host transmission (Ebert and Weisser 1997; de Roode *et al.* 2008; Pagán *et al.* 2014).

The tradeoff hypothesis forms the cornerstone for theoretical frameworks

considering the evolution of virulence in many different pathosystems. Many important additions to the framework have been made, for example recognizing that within-host competition and opportunism can lead to increases in virulence (May and Nowak 1995; de Roode *et al.* 2005; Brown *et al.* 2012). Moreover, the importance of other factors at the between-host level have been given consideration, such as self-shading (Boots and Sasaki 2000). Self-shading occurs when the host population is structured, and a highly virulent microparasite kills all host organisms in a subpopulation before transmission to another subpopulation can be achieved. The effects of evolution on microparasitic virulence have therefore been given considerable attention, although the number of experimental studies that address this issue is still rather limited, especially for viruses (Bull and Lauring 2014).

The effects of evolution on microparasite virulence have been widely considered. However, virulence itself could also have profound effects on evolution, including its own evolutionary dynamics (Bull and Ebert 2008). This reversed causality is already apparent from the tradeoff model, under which microparasites with suboptimal virulence will undergo reduced between-hosts transmission. All other things equal, if a smaller number of hosts are infected effective population size will be decreased, increasing the strength of genetic drift and decreasing the mutation supply rate. However, besides these effects of virulence on evolution mediated by reduced between-host transmission, it is conceivable that a similar within-host effect could also occur, if virulence curtails host development and thereby limits host resources available to the microparasite. Interestingly, in these cases the general constraints on evolvability imposed by high virulence might limit the rate at which lower virulence evolves, meaning that high virulence might persist longer than

suggested by the tradeoff model.

There are many reasons why high virulence in host-pathogen interactions could emerge, but the most likely avenue is probably a change of host species. If a microparasite is confronted with a new host environment in which its level of virulence is altered, how does virulence affect is ability to adapt to the new host? Here we address these questions using *Tobacco etch virus,* a virus that infects a wide-range of host plants, and an experimental evolution approach. To consider the effect of virulence on virus adaptation, we looked for two natural host species in which (*i*) there was some evidence that TEV potential for adaptation would be roughly similar, and (*ii*) there was a large difference in virulence. The distribution of mutational fitness effects (DMFE) of TEV has been compared in eighth host species, and this study concluded that there were strong genotype-by-host interactions (Lalić *et al.* 2011). For many host species distantly related the TEV typical host, *Nicotiana tabacum*, the DMFE changed drastically; many mutations that were neutral or deleterious in *N. tabacum* became beneficial. For two alternative host species, *Nicotiana benthamiana* and *Datura stramonium*, all mutations tested remained neutral or deleterious (Lalić *et al.* 2011), implying that the fraction of beneficial mutations in both hosts is small and that the occurrence of adaptive evolution is not a given. Moreover, virus accumulation after one week of infection is also similar for both hosts (Bedhomme *et al.* 2012). On the other hand, TEV infection of *N. benthamiana* will typically result in heavy stunting and the death of the plant within a matter of weeks, whereas TEV infection of *D. stramonium* is virtually asymptomatic. Whilst there are many similarities between TEV infection in these two hosts, one key difference is therefore host-pathogen interactions and the levels of viral virulence brought about.

As a first exploration of the effects of virulence on microparasite evolution, we therefore decided to serially passage in TEV in *N. benthamiana* and *D. stramonium*. By serially passaging each independent lineage in a single plant, our study maximizes within-host selection. This setup allows us to exclusively focus on effects of within-host selection, although for our model system we expect to see large differences in the resulting population size and the scope for virus movement within the host. Moreover, to immediately gauge whether adaptive evolution might be occurring, we passaged a TEV variant expressing a marker protein (**Fig C4.1**), enhanced GFP (eGFP). Upon long-duration passages in *N. tabacum*, this exogenous sequence is quickly lost due to its high fitness cost, and its loss is reliably indicated by a loss of eGFP fluorescence (Zwart *et al.* 2014). We hypothesized that adaptive evolution would occur more quickly in the host species for which TEV has lower virulence, *D. stramonium*, than in the host species for which it has high virulence, *N. benthamiana*. Hence, we expected that in *D. stramonium* (*i*) the eGFP marker would be lost more rapidly, (*ii*) there would be more sequence-level convergent evolution, and (*iii*) there would be larger increases in within-host competitive fitness. However, the results clashed with our simple hypotheses, exemplifying the extent to which a host species jump can be a game changer for evolutionary dynamics.

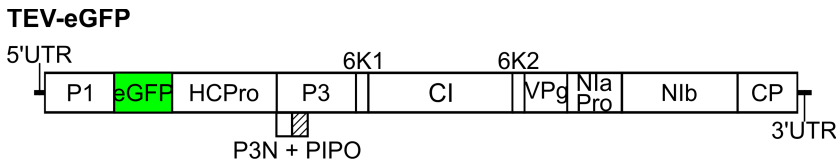

Figure C4.1. Schematic representation of TEV-eGFP. The *eGFP* gene is located between *P1* and *HC-Pro.* Proteolytic cleavage sites were provided at both ends of *eGFP*.

2. Results

2.1. Experimental setup and fluorescent marker stability upon passaging of TEV-eGFP

TEV-eGFP was passaged in *N. benthamiana* and *D. stramonium* by rub-inoculating finely ground tissue (Methods). Although 9-week passages could be performed in *D. stramonium*, for *N. benthamiana* this was not possible due to virus-induced host mortality. These plants died after 6 weeks of infection, and therefore we were forced to collect tissue at this time point. We chose to maximize infection duration, however, since we had previously noted that longer passages led to rapid and convergent evolution in *N. tabacum* (Zwart *et al.* 2014). We performed three 9-week passages in *D. stramonium* and – to keep the evolutionary time comparable – five 6-week passages in *N. benthamiana*. In *D. stramonium* all ten lineages initiated were completed, whereas in *N. benthamiana* only six out of ten lineages were completed. The remaining four *N. benthamiana* lineages failed to cause infection in subsequent rounds of passaging, and were therefore halted.

Based on previous results, we expected that the exogenous eGFP sequence would be rapidly purged (Dolja *et al.* 1993; Majer *et al.* 2013; Zwart *et al.* 2014), and as such would serve as a first indicator of the occurrence of adaptation. However, the usefulness of eGFP for determining the integrity of the eGFP marker was limited in both hosts, by (*i*) the high levels of autofluorescence in the highly symptomatic *N. benthamiana* leaves, and (*ii*) the "patchy" fluorescence in the *D. stramonium* tissue. Therefore, unlike for TEV-eGFP in *N. tabacum*, the fluorescent marker was of limited use here. Nevertheless, all *N. benthamiana* lineages appeared to have some fluorescence

until the end of the evolution experiment, and we observed a loss of fluorescence in only one out of ten *D. stramonium* lineages in the third 9-week passage.

After each passage, RNA was extracted from the collected leaf tissue, and RT-PCR with primers flanking the eGFP insert was performed. This RT-PCR assay can therefore detect deletions in the *eGFP* gene, even when deletions extend well into the downstream HC-Pro cistron (Zwart *et al.* 2014). In general, the RT-PCR results confirmed the fluorescence microscopy results: A large deletion was detected only in the one *D. stramonium* lineage with a loss of fluorescence (**Fig C4.2A**; 9-weeks passage 2 L8). This deletion variant went to a high

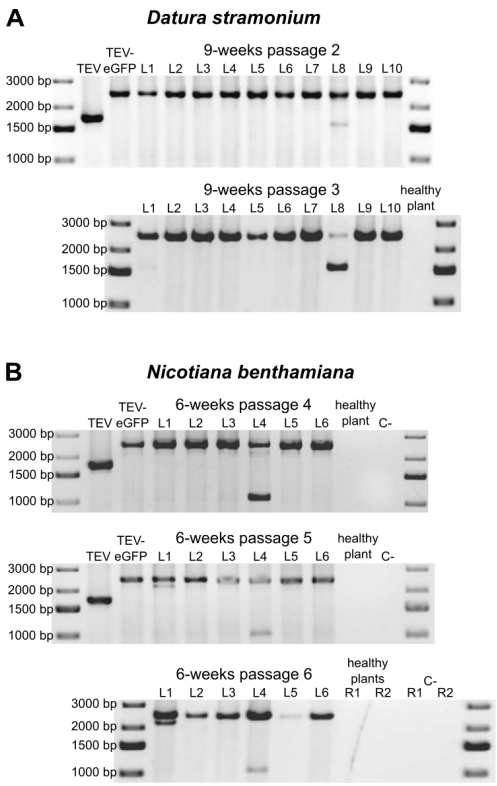

Figure C4.2. Deletion detection in the *eGFP* gene. Agarose gels with RT-PCR products of the region encompassing the *eGFP* gene. The TEV and TEV-eGFP are shown for comparative purposes. The negative controls are healthy plants and PCR controls (C-). (**A**) TEV-eGFP in *D. stramonium* has 10 independent lineages (L1-L10). A deletion encompassing the *eGFP* gene was detected in one lineage (L8) in the second 9-week passage. This deletion went to a high frequency in the subsequent passage. (**B**) TEV-eGFP in *N. benthamiana* has six independent lineages (L1-L6). A deletion bigger than the size of *eGFP* was detected in one lineage (L4) in the fourth 6-week passage. This deletion was not fixed in the two subsequent passages. A small deletion was detected in the fifth and sixth 6-week passage in L1.

frequency in the subsequent passage (**Fig C4.2A**; 9-weeks passage 3 L8). For *N. benthamiana* lineages, we did detect a low-frequency deletion in the eGFP cistron in one lineage (**Fig C4.2B**; 6-weeks passage 4 and 5 L4), but this deletion is so large that this variant is no probably longer capable of autonomous replication. We performed an extra round of passaging with all *N. benthamiana* lineages to check whether this variant would remain at a low frequency, and found exactly this result (**Fig C4.2B**; 6-week passage 6 L4). Furthermore, we detected a small deletion in one lineage (**Fig C4.2B**; 6-week passage 5 and 6 L1), that was maintained at a low frequency in subsequent passages of the virus population.

2.2. Whole-genome sequencing of the evolved lineages

All evolved and the ancestral TEV-eGFP lineages were fully sequenced by Illumina technology. The ancestral consensus sequence was used as a reference for mapping the evolved lineages. The deletion observed by RT-PCR (**Fig C4.2A**) in one of the *D. stramonium* lineages was confirmed by a low number of reads mapping inside the eGFP region (median coverage eGFP: 111.5), compared to a higher average coverage outside this region (median coverage *P1* gene: 19190, median overall genome coverage: 18460). The large deletion included the N-terminal region of *HC-Pro*, as observed for other deletions that occur after gene insertions before this gene (Chapter 1, 2 and Zwart et al. 2014). For all other lineages in *D. stramonium* and *N. benthamiana*, coverage over the genome was largely uniform and similar to the ancestral virus, indicating that there were indeed no genomic deletions present at appreciable frequencies.

Single nucleotide mutations were detected from a frequency as low as 1%,

comparing the evolved TEV-eGFP lineages in *N. benthamiana* and *D. stramonium* to the ancestral population. This detection was also performed for evolved TEV-eGFP lineages in *N. tabacum*, that were sequenced in a previous study (Zwart *et al.* 2014). In the evolved *N. benthamiana* lineages 165 unique mutations were found, with a median of 34.5 (27-47) mutations per lineage. In the evolved *D. stramonium* lineages 239 unique mutations were found, with a median of 31.5 (16-35) mutations per lineage. In the evolved *N. tabacum* lineages, 183 unique mutations were found, with a median of 21.5 (17-36) mutations per lineage. Note that the single nucleotide mutations detected here can be fixed (frequency > 50%) in the evolved lineages, as the detection was done over the ancestral population. Hence, it allows us to compare the mutations that arose by evolving TEV-eGFP in the different hosts.

We detected only one mutation (U6286C) that is shared between all three hosts. However, this mutation was present at a low frequency and not detected in all *D. stramonium* and *N. tabacum* lineages (**Table C4.1**). The *N. benthamiana* and *D. stramonium* lineages share more mutations (15) than either *N. benthamiana* or *D. stramonium* share with *N. tabacum* (4 and 9, respectively). However, most of these mutations are present in only a few lineages and at low frequency (**Table C4.1**).

In *D. stramonium* we found two mutations (A7479C and A8253C) that are present in all 10 lineages, and within each lineage they are detected at the same frequency (**Fig C4.3A**). The same two mutations were found when evolving TEV-eGFP in *N. tabacum* in 7/10 lineages (**Table C4.1**), also here the frequency of these two mutations is similar within every lineage (**Fig C4.3A**). These data suggest a strong linkage between the A7479C and A8253C

Table C4.1. TEV-eGFP mutations shared in the different hosts

			N. benthamiana		*D. stramonium*		*N. tabacum*	
nt change at postion	aa change	gene	#/total lineages	med freq	#/total lineages	med freq	#/total lineages	med freq
U6286C	Y→H	CI	6/6	0.028	2/10	0.022	4/10	0.027
A208G	M→V	P1	1/6	0.012	1/10	0.010	-	-
C1039U	H→Y	P1	1/6	0.013	1/10	0.035	-	-
G1332A	M→I	eGFP	1/6	0.176	1/10	0.139	-	-
U1556G*	V→G	eGFP	1/6	0.011	6/10	0.011	-	-
U1836G	S→S	HC-Pro	5/6	0.105	1/10	0.089	-	-
A1917G	V→V	HC-Pro	1/6	0.015	1/10	0.017	-	-
A6278G	E→G	CI	1/6	0.012	2/10	0.023	-	-
A6438G	E→E	6K1	1/6	0.012	1/10	0.024	-	-
C6547U	H→Y	VPg	1/6	0.013	1/10	0.014	-	-
U6747C	F→F	VPg	1/6	0.012	1/10	0.110	-	-
A6776G	D→G	VPg	2/6	0.279	1/10	0.023	-	-
G6803A	S→N	VPg	1/6	0.014	1/10	0.024	-	-
C8405G*	T→R	NIb	5/6	0.012	5/10	0.014	-	-
U9474C	N→N	CP	1/6	0.013	1/10	0.061	-	-
C9837U	N→N	CP	1/6	0.070	1/10	0.010	-	-
U3803C	I→T	P3	2/6	0.769	-	-	2/10	0.015
U3872C	V→A	P3	1/6	0.064	-	-	1/10	0.016
G4411A	V→I	CI	1/6	0.030	-	-	1/10	0.016
C4989U	V→V	CI	1/6	0.018	-	-	1/10	0.066
C548U	T→I	P1	-	-	1/10	0.011	1/10	0.024
G2928A*	R→R	HC-Pro	-	-	1/10	0.017	1/10	0.999
U7092C*	S→S	NIa-Pro	-	-	10/10	0.348	2/10	0.588
A7479C*	P→P	NIa-Pro	-	-	10/10	0.255	7/10	0.988
A7567G	K→E	NIa-Pro	-	-	4/10	0.018	10/10	0.143
G7710A	A→A	NIa-Pro	-	-	1/10	0.014	1/10	0.024
A8253C*	S→S	NIb	-	-	10/10	0.264	7/10	0.988
G9117A	A→A	NIb	-	-	1/10	0.321	2/10	0.023
U9249C	D→D	NIb	-	-	2/10	0.121	1/10	0.040

nt: nucleotide; aa: amino acid; med freq: median frequency
*mutation that is also detected in the ancestral population

mutations. There is a third mutation (U7092C) that never appears together with the former two mutations (**Fig C4.3**), suggesting that this mutation occurs in another haplotype and that there may be sign epistasis between these two combinations of mutations. This is well displayed in the *D. stramonium* lineages, as there is a lot of variation in the frequencies at which the A7479C and A8253C mutations occur (**Fig C4.3A**). In *N. tabacum* the third mutation (U7092C) only appears in 2/10 lineages, as the other two inverse correlated mutations are present at a frequency of ~ 100% in 6/10 lineages (**Fig C4.3B**). Despite of the U7092C, A7479C and A8253C mutations already being present in the ancestral population, at a frequency of ~40%, these mutations were not found when evolving TEV-eGFP in *N. benthamiana* (**Table C4.1**). A fourth

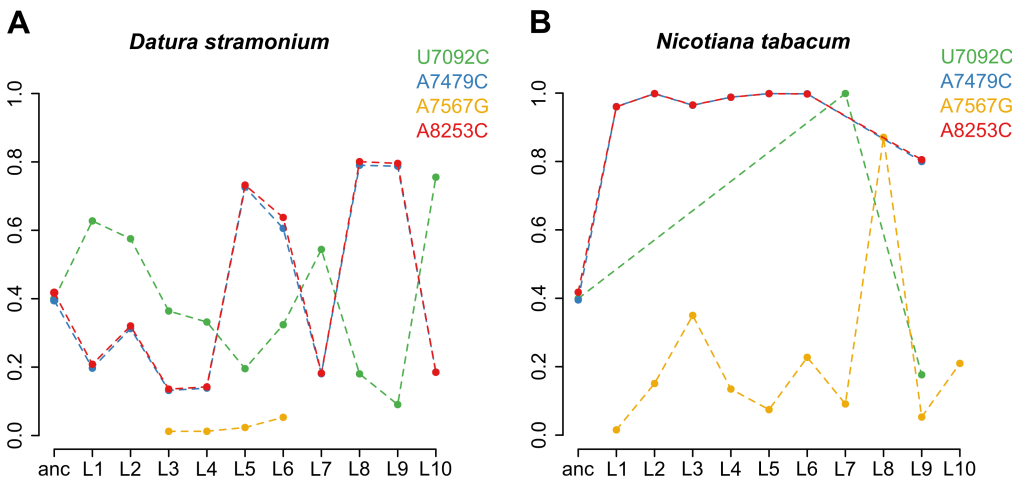

Figure C4.3. Frequency of mutations found in both *D. stramonium* and *N. tabacum*. Mutations detected in both *D. stramonium* and *N. tabacum* that were present in all the lineages of either one of these hosts. The frequency of these mutations in either the ancestral population or the different lineages (L1-L10) is given by the color-coded points. The points are connected by the broken lines to emphasize the trend in the data.

nonsynonymous mutation (A7567G) was detected at a low frequency in 4/10 *D. stramonium* lineages (**Fig C4.3A**), and at a varying frequency in 10/10 *N. tabacum* lineages (**Fig C4.3B**), however not in the *N. benthamiana* lineages. Unlike the other mutations described above, this mutation was not detected in the ancestral population. Moreover, there is no apparent linkage to the other mutations.

Host-specific mutations were mostly found in the evolved TEV-eGFP lineages of *N. benthamiana* (**Table C4.2**). In this host, a total number of 7 specific mutations were detected and all of them being nonsynonymous. In *D. stramonium* no host-specific mutations were detected. And in *N. tabacum* only 1 host-specific mutation was detected in the 3' UTR (**Table C4.2**). Note that host specific mutations were defined as mutations detected in at least half of the evolved lineages.

Table C4.2. Host specific mutations in the evolved TEV-eGFP lineages

	nt change at position	aa change	gene	#/total lineages	med freq
N. benthamiana	G3797A	G→E	P3	3/6	0.291
	G4380U	E→D	6K1	3/6	0.049
	U4387C	Y→H	6K1	4/6	0.013
	C4391U	T→M	6K1	6/6	0.114
	G4397A	S→N	CI	6/6	0.015
	A6771U	L→F	VPg	4/6	0.110
	G8909U*	W→L	NIb	5/6	0.034
D. stramonium	-	-	-	-	-
N. tabacum	G10253A	C→Y	3'UTR	10/10	0.037

nt: nucleotide; aa: amino acid; med freq: median frequency
*mutation that is also detected in the ancestral population

2.3. Viral accumulation and competitive fitness

We measured virus accumulation after one week of infection, by reverse-transcription quantitative polymerase chain reaction (RT-qPCR) for the coat protein gene (*CP*). In both host species, we found no statistically significant differences (*t*-test with Holm-Bonferroni correction) between TEV, TEV-eGFP and the lineages of TEV-eGFP evolved in that host (**Fig C4.4)**.

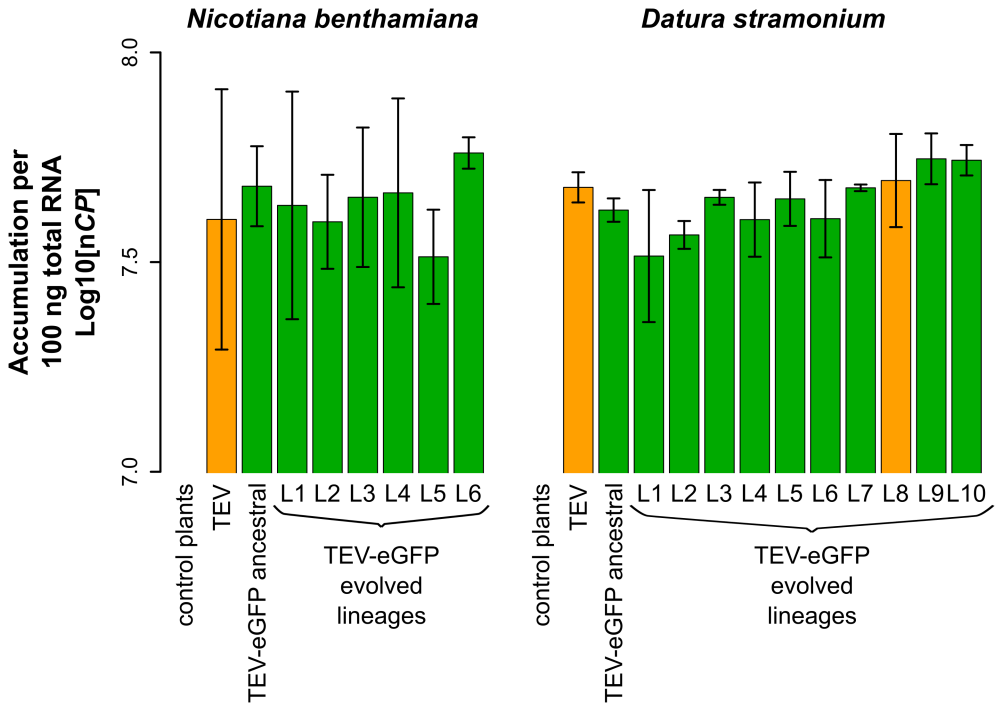

Figure C4.4. Virus accumulation of the evolved and ancestral lineages. Virus accumulation, as determined by accumulation experiments and RT-qPCR at 10 dpi, of the ancestral TEV and TEV-eGFP and the evolved TEV-eGFP lineages in the corresponding hosts. The ancestral TEV and the evolved lineage with a deletion in the *eGFP* gene are indicated with the orange bars. The ancestral TEV-eGFP and the evolved lineages with an intact *eGFP* gene are indicated with the green bars.

We then measured within-host competitive fitness by means of head-to-head competitions with TEV-mCherry, a virus with a different marker but similar fitness to TEV-eGFP (Zwart *et al.* 2014). Here we observed interesting differences between TEV and TEV-eGFP in the two different hosts. Whereas the TEV-eGFP had lower fitness than the wild-type virus in *D. stramonium*, there was no difference in *N. benthamiana* (**Fig C4.5**; compare ancestral TEV and TEV-eGFP). Our results therefore suggest that although there is a fitness cost associated with the eGFP cistron in *N. tabacum* (Zwart *et al.* 2014) and *D. stramonium*, there is none in *N. benthamiana*. Interestingly, *N. tabacum* and *N. benthamiana* are more closely related to each other than either species is to *D. stramonium*, and yet the host species has a strong effect on the costs of a heterologous gene.

For the lineages evolved in *D. stramonium*, only for one out of ten lineages was a significant increase in competitive fitness compared to the ancestral TEV-eGFP observed (**Fig C4.5**, L8; *t*-test with Holm-Bonferroni correction: t_4 = -6.890, P = 0.002). This lineage is the only one to have a deletion in the eGFP insert. In *N. benthamiana*, one out of six lineages had a significant increase in within-host fitness (**Fig C4.5**, L4; *t*-test with Holm-Bonferroni correction: t_4 = -5.349, P = 0.006). This increase in fitness can possibly be associated with recombination between variants with the intact *eGFP* gene and variants containing a big deletion in this lineage (**Fig C4.2B**; L4). However, it is unlikely that variants with this big deletions are viable, therefore single-nucleotide variation is probably the main driving force for the increase in fitness. These fitness measurements therefore show that most lineages failed to adapt to the new host species. However, in the two cases that there were significant fitness increases, the underlying genetic changes were consistent

with the expected route of adaptation. In *D. stramonium*, where the *eGFP* has a high fitness cost, this sequence was deleted; In *N. benthamiana*, where the *eGFP* apparently has not fitness cost, single-nucleotide variation was observed.

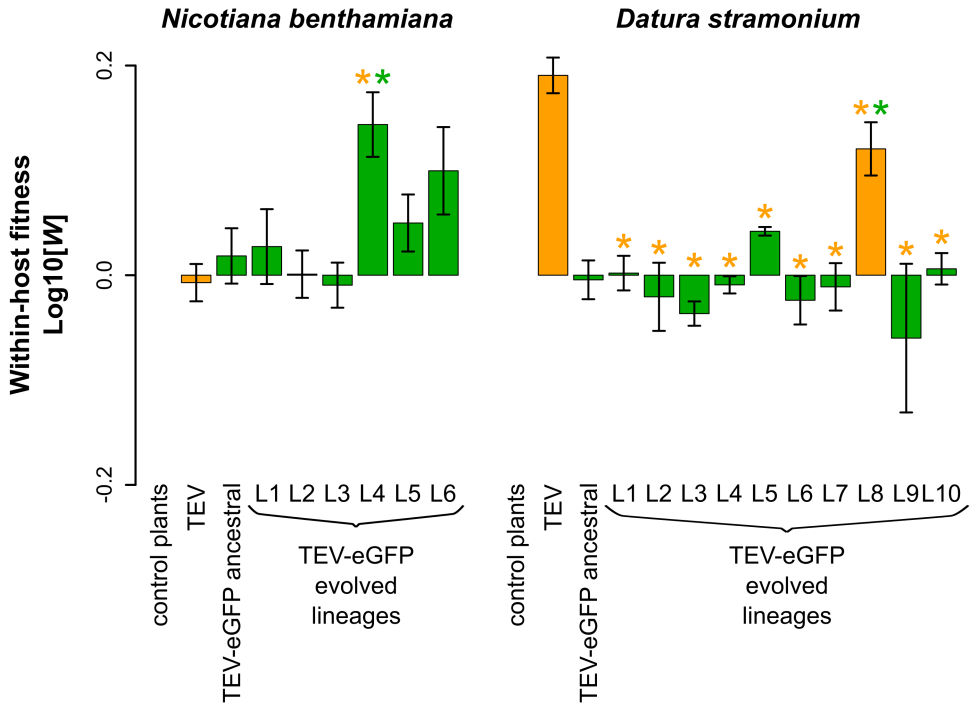

Figure C4.5. Within-host competitive fitness of the evolved and ancestral lineages. Fitness (*W*), as determined by competition experiments and RT-qPCR of the different viral genotypes with respect to a common competitor; TEV-mCherry. *W* was determined at 10 dpi, of the ancestral TEV and TEV-eGFP and the evolved TEV-eGFP lineages in the corresponding hosts. The ancestral TEV and the evolved lineage with a deletion in the *eGFP* gene are indicated with the orange bars. The ancestral TEV-eGFP and the evolved lineages with an intact *eGFP* gene are indicated with the green bars. The orange asterisks indicate statistical significant differences of the evolved lineages as compared to the ancestral TEV (*t*-test with Holm-Bonferroni correction). The green asterisks indicate statistical significant differences of the evolved lineages as compared to the ancestral TEV-eGFP (*t*-test with Holm-Bonferroni correction).

3. Discussion

We set out to explore the hypothesis that differences in virulence for different hosts could have an effect on the rate of virus adaptation in each host (Bull and Ebert 2008). Although we find this hypothesis simple and provocative, the observed patterns in our experiments suggest that even in a controlled laboratory environment, the reality will often be much more complex and harder to predict. We used a virus expressing an eGFP marker in the hope that the loss of this marker could serve as a real-time indicator of adaptation. However, there were complications with this method, and a loss of fluorescence was only observed in a single *D. stramonium* lineage. RT-PCR and Illumina sequencing confirmed the loss of the eGFP marker in this case, and its integrity in all other lineages. The data of our competitive fitness assay demonstrate why the maker sequence was probably rather stable in *N. benthamiana*; eGFP does not appear to have a cost in this host species. We expect that the marker will eventually be lost, but only due to genetic drift and therefore at a slow rate.

These results are at odds with our expectations, but they nevertheless have some interesting implications. First, host species changes can apparently ameliorate the costs of exogenous genes. Although strong virus genotype-by-host species interactions have been previously shown for TEV (Lalić *et al.* 2011), we did not anticipate that a such a simple difference (the presence of eGFP) could also be subject to such an interaction. These results suggest that when considering the evolution of genome architecture, host species might play a very important role, by allowing evolutionary intermediates to be competitive. For example, we have shown that for TEV with an altered *NIb* positioning, all intermediate steps lead to decreases in fitness, making the trajectory to alternative gene order

inaccessible (Chapter 1). If *NIb* duplication has a similar interaction with host species as the eGFP insert, then an alternative host species could act as stepping stone and hereby increase the accessibility of the evolutionary trajectory to alternative gene order. Similar effects of environmental change have been noted in other studies (de Vos *et al.* 2015). The generality of these results has not been addressed yet using other viruses with altered genome architecture, but the possibilities are tantalizing. Second, our results could also have implications for assessing the biosafety risks of the genetically modified organisms. Our results suggest that extrapolating fitness results from a permissive host to alternative hosts can be problematic, even when the scope for unexpected interactions appears to be limited, as would be the case for the addition of eGFP expression. In other model systems, unexpected interactions between heterologous genes and host species have also been reported (Hernández-Crespo *et al.* 2001).

Our results were not consistent with the hypothesis that high virulence could slow down the rate of adaptation, as in each host only a single lineage had evolved higher fitness. The low rate of adaptation observed was consistent with a previous report (Bedhomme *et al.* 2012), although we used passages of a longer duration here and had therefore expected more rapid adaptation (Zwart *et al.* 2014). Given the low rate at which lineages adapted in this experiment, however, we do not consider that our results provide strong evidence against the hypothesis. Nevertheless, our results do stress that differences in host biology can have a much stronger effect on evolutionary dynamics than differences in virus-induced virulence between host species. An alternative way to tackling the question of the effects of virulence on adaptation might be to use a biotechnological approach; hosts which have different levels of virulence can be engineered, to ensure the main difference between host treatment is

microparasite-induced virulence. For example, plant hosts could be engineered to express antiviral siRNAs at low levels. Such an approach would allow for a more controlled test of the hypothesis suggested here, whilst probably not being representative for natural host populations. On the other hand, such experiments could perhaps help shed light on the effects of virulence on adaptation in agroecosystems or vaccinated populations.

Final Discussion

RNA virus genomes are are highly streamlined. In this study, the potential of TEV to evolve alternative genome architectures has been explored. To asses whether these architectural changes have an overall effect, here we compare all TEV variants that were used in this study.

1. Genome complexity, fitness and stability

Following the Eigen paradox, RNA viruses are evolutionary constrained to have genomes of relatively low complexity and of a small size. Therefore, we expected that increases in genome size would lead to decreases in fitness. Indeed, when comparing the viruses of this study with an increase in genome size, there is a negative relationship between genome size and within-host competitive fitness (**Fig FD.1A**; black circles: Spearman's rho: $\rho = -0.810$, 8 d.f., $P = 0.022$). Yet, there is no significant relationship between viral accumulation and the genome size (**Fig FD.1B**; black circles: Spearman's rho: $\rho = -0.619$, 8 d.f., $P = 0.115$). The replication speed does not appear to be affected by the increase in genome size, and therefore the accumulation of deleterious mutations makes these viruses inferior competitors. Without mechanisms to get rid of the extra unwanted material, and increase in genome size could ultimately lead to virus extinction. Interestingly, the viruses with a change in gene order but no change in genome size, are in the lower range of within-host competitive fitness (**Fig FD.1A**; grey circles). The virus with the lowest competitive fitness is TEV-eGFP. This is surprising as we did not

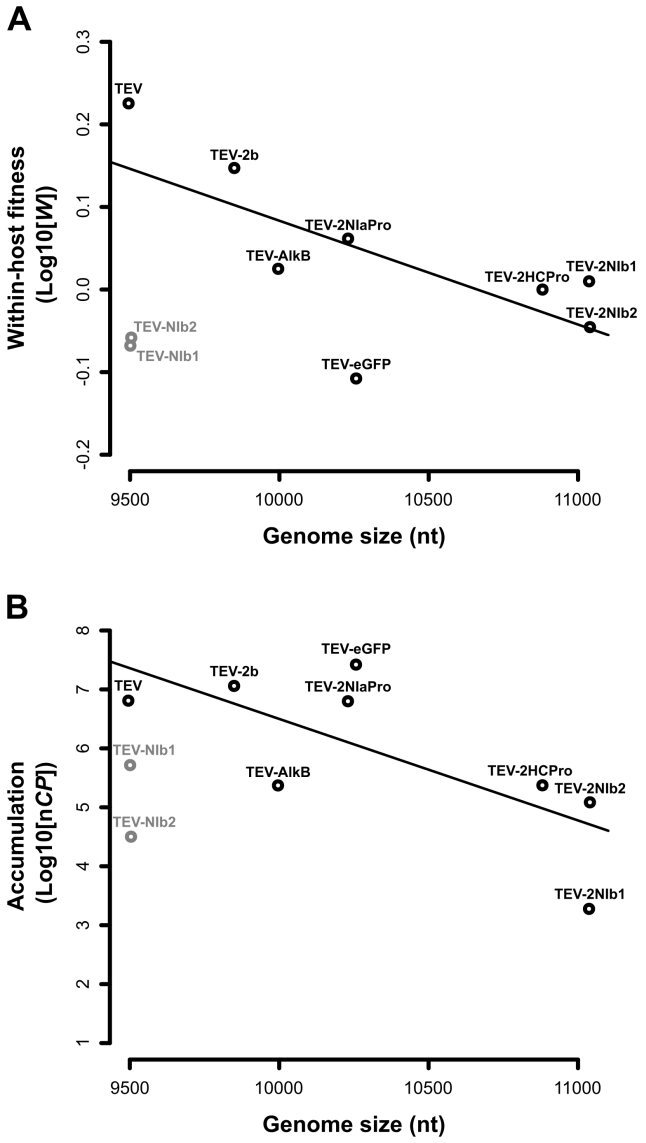

Figure FD.1. The relationship between genome size and viral fitness. Grey circles indicate the rearranged viruses with no change in genome size. Black circles indicate the viruses with an increase in genome size plus the wild-type TEV, for which a linear regression line is drawn to show the trend in the data. (**A**) Within-host competitive fitness. (**B**) Viral Accumulation. Some of the long viral names are shortened; TEV-NIb1: TEV-NIb$_1$-ΔNIb$_9$, TEV-NIb2: TEV-NIb$_2$-ΔNIb$_9$, TEV-2NIb1: TEV-NIb$_1$-NIb$_9$, TEV-2NIb2: TEV-NIb$_2$-NIb$_9$, TEV-2HCPro: TEV-HCPro$_2$-HCPro$_3$, TEV-2NIaPro: TEV-NIaPro$_2$-NIaPro$_8$.

expected that the insertion of a non-functional gene could have a higher impact on fitness than duplications of existing genes that in most cases lead to larger genome sizes. Similar to within-host fitness, the rearranged viruses are also in the lower range of accumulation (**Fig FD.1B**; grey circles). The low fitness of the rearranged viruses most likely reflects an alteration of the expression levels plus the incorrect timing of protein expression at the corresponding cellular location. This effect appears to be smaller after insertion or duplication of a gene.

The stability of the different viruses over time has been measured as the mean time to deletion of the different viral lineages of each genotype. The rearranged viruses are evolutionary stable (**Fig FD.2**: grey circles**)**, as essential virus genes

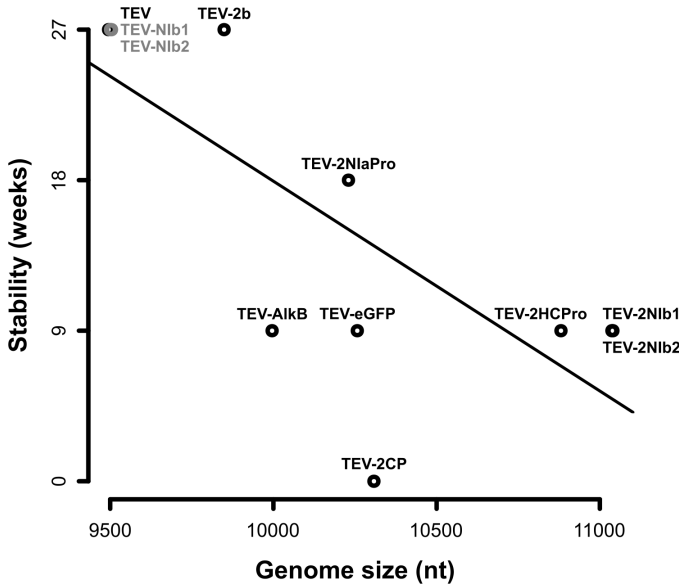

Figure FD.2. The relationship between genome size and genomic stability. The stability has been measured as the mean time to deletion of a inserted or duplicated gene, in the 9-week serial passage experiments. Grey circles indicate the rearranged viruses with no change in genome size. Black circles indicate the viruses with an increase in genome size plus the wild-type TEV. Note that the viruses with a genomic stability of 27 weeks are evolutionary stable in this time frame. The linear regression lines shows the trend in the data. Some of the long viral names are shortened; TEV-NIb1: TEV-NIb$_1$-ΔNIb$_9$, TEV-NIb2: TEV-NIb$_2$-ΔNIb$_9$, TEV-2NIb1: TEV-NIb$_1$-NIb$_9$, TEV-2NIb2: TEV-NIb$_2$-NIb$_9$, TEV-2HCPro: TEV-HCPro$_2$-HCPro$_3$, TEV-2NIaPro: TEV-NIaPro$_2$-NIaPro$_8$, TEV-2CP: TEV-CP$_{10}$-CP$_{11}$.

cannot be deleted without losing the ability to infect autonomously. However, such deletions can occur in viruses with gene duplications or insertions. When comparing viruses in which such deletions can occur without a loss of viability, there is an inverse relationship between stability and genome size (**Fig FD.2**: black circles: Spearman's rho: $\rho = -0.688$, 8 d.f., $P = 0.040$). This suggests that gene duplications or gene insertions of a larger size are likely to be deleted faster in our model system.

2. Genome complexity and mutational robustness

When exposing TEV and some of the variants to a mutagenic agent, the wild-type virus had the highest mutational robustness (**Fig FD.3**). The TEV-AlkB virus has the lowest mutational robustness. Interestingly, this is the only virus that has an insertion within a protein, instead of at an intergenic site. Comparing all five viruses (**Fig FD.3**) there is no significant relationship between mutational robustness and genome size (**Fig FD.3;** all circles: Spearman's rho: $\rho = -0.7$, 5 d.f., $P = 0.233$). However, there is a near significant relationship, when only considering the viruses with insertions or a duplication at an intergenic site (**Fig FD.3;** black circles: Spearman's rho: $\rho = -1$, 4 d.f., $P = 0.083$). To confirm whether a relationship between genome size and mutational robustness exists the effect of mutagenic treatment should be tested on a larger sample of viruses with different genome sizes.

More complex organisms are predicted to be more robust than simpler ones (Lenski *et al.* 1999). In particular, gene duplications are found to confer organisms with higher mutational robustness (Lynch and Conery 2000; Gu *et*

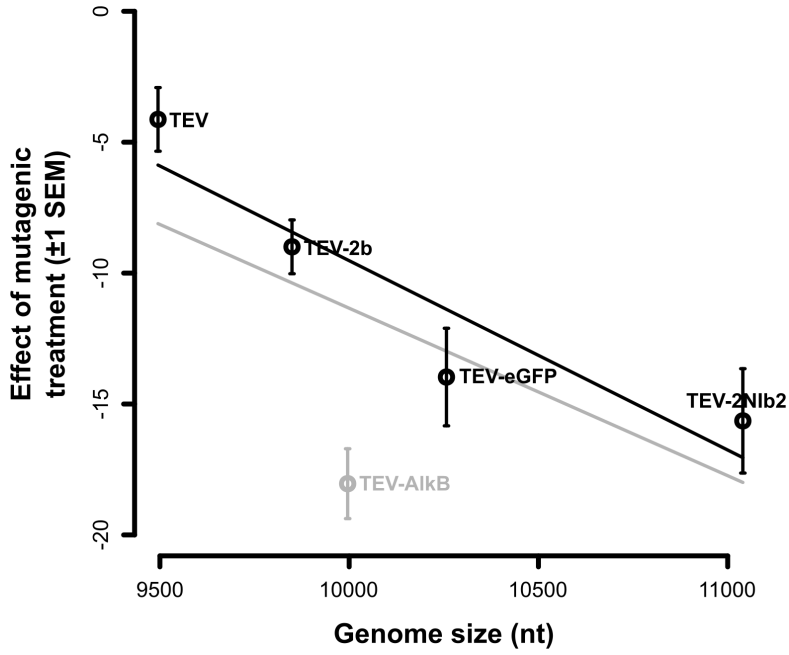

Figure FD.3. The relationship between genome size and mutational robustness. Linear regression lines are drawn to show the trend in the data. The grey line considers all viruses, whereas the black line considers only viruses with a gene insertion or duplication at an intergenic site, plus the wild-type TEV. TEV-2NIb2: TEV-NIb_2-NIb_9.

al. 2003; Conant and Wagner 2004; Hsiao and Vitkup 2008; Fares *et al.* 2013; Keane *et al.* 2014). In our model system, however, gene duplication does not confer higher mutational robustness (**Fig FD.3**; TEV-2NIb2). Gene duplications are evolutionary unstable in TEV, and several mechanisms can explain why duplications are not beneficial here. The high fitness costs of (*i*) maintaining additional genetic material, (*ii*) a disturbance in polyprotein processing, or (*iii*) yet unknown post-translational mechanisms could all play a role in the instability of genetic redundancy. The observation that the deletions occurred in

all cases of the duplicated gene copy, and not the original gene copy, provides additional support for the lack of accessible trajectories to the evolution of alternative gene orders. Contrarily to genetic redundancy, functional redundancy − generated by introducing an exogenous viral sequence − was found to be evolutionary stable in TEV. Even though functional redundancy could compensate for directed knock-out mutations, it did not provide for higher mutational robustness (**Fig FD.3**; TEV-2b).

We found that changes in the genome architecture of TEV go paired with high fitness costs and evolutionary barriers. The wild-type outcompetes the attenuated viruses resulting from genome rearrangements in many different areas. Gene duplications are evolutionary unstable, which limits the evolution of alternative gene orders, novel functions and mutational robustness. Therefore, the highly conserved gene order appears to be an optimal gene order for present-day potyviruses and it is highly unlikely that TEV will evolve any major rearrangements of its genome. Nevertheless, it cannot be discarded that important changes of TEV will occur in the future, as there is a potential for TEV to evolve novel functions or substitute existing functions through HGT. Finally, a host species jump can drastically change the evolutionary fate of genome rearrangements and can therefore open new roads to the the evolution of alternative genome architectures in TEV.

Final Conclusions

Multiple barriers to the evolution of alternative gene orders were identified, revealing important constraints on virus evolution. The intermediate steps for evolving an alternative gene order come with high fitness costs, which are not easily overcome. The lack of accessible trajectories to alternative gene orders are suggested to contribute to gene order conservation in potyviruses.

The gene duplications explored are evolutionary unstable and do not appear to be beneficial for TEV. Gene duplication results in fitness reductions which are restored by deleting the duplicated gene copy. In addition to viral fitness, the stability of the duplicated gene copy is affected by a context-dependent recombination rate, with the context being the identity and position of the duplicated gene.

Contrarily to gene duplications, the insertion of one functional exogenous sequence – the *2b* gene from CMV – has shown to be evolutionary stable. No fitness reductions were observed when TEV carried this gene. In addition, 2b has shown to be beneficial for TEV in the case of failure of its own RNA silencing suppressor mechanism. These results suggest that the HGT between plant viruses is a possible event and a promising result for future experimental evolution studies.

Host species jump can be a game changer for evolutionary dynamics. A non-functional exogenous sequence – eGFP – was more stable in alternative hosts, for which TEV has both lower and higher virulence. In addition, eGFP does not appear to have any fitness effects in the host for which TEV has high virulence.

This clashes with our hypothesis that high virulence slows down the rate of adaptation. The results suggest that when considering the evolution of genome architecture, host species jumps might play a very important role, by allowing evolutionary intermediates to be competitive.

Abbreviations

CMV: *Cucumber mosaic virus*

DMFE: distribution of mutational fitness effects

dpi: days post-inoculation

eGFP: enhanced Green Fluorescent Protein

ER: endoplasmatic reticulum

HGT: horizontal gene transfer

IAH: independent-action hypothesis

kb: kilobases

LUCA: last universal cellular ancestor

MOI: multiplicity of infection

NCLDVs: nucleocytoplasmic large DNA viruses

N_e: the effective population size

NGS: next-generation sequencing

NLL: negative log likelihood

ORF: open reading frame

PCR: polymerase chain reaction

Ros1: Rosea1

RT: reverse transcription

RT-PCR: reverse transcription polymerase chain reaction

RT-qPCR: reverse transcription quantitative polymerase chain reaction

TEV: *Tobacco etch virus*

VSV: *Vesicular stomatitis virus*

Resumen

Introducción

La evolución de la arquitectura del genoma – las dimensiones y la organización del material hereditario de un organismo – es poco conocida. Existe una variación asombrosa en la arquitectura genómica entre diferentes organismos. Los virus tienden a tener genomas pequeños, con un mínimo de secuencias intergénicas y típicamente con genes solapantes, los cuales se cree son una forma de compresión del genoma que permite al virus incrementar su número de proteínas sin aumentar su tamaño. Los procariotas tienen genomas compactos, pero con secuencias intergénicas más largas que los virus, y los genes con solapamientos largos son escasos. Los eucariotas presentan una amplia gama de tamaños de genoma, desde el rango de los procariotas a órdenes de magnitud más largos, y se distinguen por su organización génica en intrones y exones. Las secuencias intrónicas e intergénicas contribuyen al gran tamaño del los genomas eucariotas, pudiendo éstas constituir hasta un 60% del ADN. En contraste, los virus y procariotas consisten principalmente de ADN (o ARN) codificante, representando éste más del 85% del tamaño total del genoma.

Hay un aumento notable en la complejidad del genoma desde los virus, pasando por los procariotas, hasta los eucariotas, en términos de tamaño del genoma, el número de genes, el número de elementos móviles, el número y tamaño de los intrones, y la complejidad de las regiones reguladoras. Aun así, no existen límites claros en las diferencias de la arquitectura genómica entre organismos. El descubrimiento de los virus gigantes y de bacterias simbióticas con genomas

pequeños, ha eliminado la separación de genomas celulares y virales por tamaño. Por lo tanto, esto sugiere que los mecanismos principales promoviendo la divergencia de la arquitectura genómica son genético-poblacionales, y no los diferentes estilos de vida, estructuras celulares o fisiologías de los organismos.

Los mecanismos genético-poblacionales, como la tasa de mutación y el tamaño efectivo de la población, podrían desempeñar un papel en la conformación de arquitecturas genómicas divergentes. La Ley de Drake describe como la tasa de mutación en todo el genoma es más o menos constante entre los organismos, por lo tanto, existe una correlación negativa entre la tasa de mutación y el tamaño de genoma. Esto también es una consecuencia de la paradoja de Eigen sobre el umbral de error: la longitud de un genoma está limitada por la tasa de mutación. Cuando la tasa de mutación está por debajo del umbral de error, una población se mantendrá en el equilibrio entre la acción de la mutación y de la selección.

Los virus, que tienen genomas pequeños en comparación con las especies multicelulares, tienen una tasa de mutación más alta. En particular, las altas tasas de mutación observadas para los virus ARN son varios órdenes de magnitud superiores a las tasas de mutación observadas en la mayoría de los organismos de ADN. Una explicación para este contraste es que, a diferencia de los polimerasas de ADN, muchas polimerasas de ARN no tienen un mecanismo de corrección de errores, y por lo tanto son más propensas a introducir errores durante la replicación. Aún así, los límites de las tasas de mutación entre los virus con diferentes arquitecturas genómicas no están bien definidos, y las altas tasas de mutación de los virus de ARN son igualadas por algunos virus de ADN. Por lo tanto, se sugiere que otros aspectos como la arquitectura del genoma y la

velocidad de replicación, en lugar de la baja fidelidad de la polimerasa de ARN, podrían explicar las diferencias de las tasas de mutación en los virus.

Las transiciones desde organismos simples hasta organismos más complejos, parecen estar asociadas a grandes reducciones en el tamaño de la población. Cuando el tamaño poblacional es pequeño, el poder de la deriva genética aumenta, el cual crea oportunidades para la producción de nuevas y diversas características genómicas, que por el contrario serían eliminadas por la acción de la selección purificadora. Por lo tanto, se ha propuesto que un tamaño efectivo poblacional grande – como el de los virus de ARN – es un obstáculo importante para la evolución de los genomas complejos.

Para estudiar la evolución de la arquitectura genómica, se optó por trabajar con el virus del grabado del tabaco (TEV). Este patógeno de planta tiene un genoma de ARN de cadena sencilla y polaridad positiva. Los virus de ARN de polaridad positiva representan el grupo de virus conocido más grande hasta la fecha, y se clasifican en tres tribus: *picorna-* , *alfa-* y *flavi-like*. Estos virus se caracterizan por tener grupos de genes conservados, y en especial por la disposición helicasa-polimerasa, en donde el gen de la helicasa se encuentra normalmente corriente arriba del gen de la polimerasa. En particular, la tribu *picorna-like* se identifica por la conservación parcial de genes esenciales. Además, estos genes esenciales tienden a estar organizados en el mismo orden, mientras que los genes no esenciales son responsables de la reorganización genómica y la recombinación entre grupos distantes de virus de las tres tribus.

El genoma monopartito lineal de TEV es de aproximadamente 9.5 kb en longitud, y codifica para un único marco de lectura abierto, es decir una poliproteína. Después de la traducción, la poliproteína se procesa

proteolíticamente en diez péptidos maduros, mediante la acción de proteaseas que están codificadas en el propio virus. TEV está clasificado dentro del género *Potyvirus* de la familia *Potyviridae*. Esta familia se divide en ocho géneros y todos los virus en estos géneros tienen un genoma monpartito, a excepción de los virus del género *Bymovirus* que tienen un genoma bipartito. Curiosamente, el orden de los genes de los diferentes miembros de la familia *Potyviridae* está muy bien conservado, inclusive en los *Bymovirus*.

Objetivos, metodología y resultados

Esta tesis trata de entender tres conceptos básicos sobre la evolución de la arquitectura genómica de los virus de RNA: (*i*) el posible aumento de su complejidad mediante la adquisición de nuevos genes, (*ii*) la disminución de su complejidad por pérdida de material redundante o innecesario y (*iii*) la reorganización de elementos ya existentes. Utilizando como sistema modelo al TEV, hemos generado cambios importantes en el genoma viral (reorganizaciones, duplicaciones y la introducción de genes exógenos), seguidos por evolución experimental de estos genomas modificados para ver cómo acomodaban los cambios introducidos. Hemos comparado por secuenciación masiva, mediciones de acumulación, virulencia y experimentos de competencia los linajes evolucionados contra los ancestrales y entre si. Los cuatro objetivos principales de este estudio se corresponden con los cuatro capítulos en esta tesis.

Capítulo 1: Múltiples barreras a la evolución de órdenes alternativos de genes

El primer objetivo es comprender mejor la conservación del orden génico en los órdenes y familias de los virus. El orden en que los genes se organizan dentro de un genoma por lo general no se conserva entre especies poco relacionadas. Sin embargo, dentro de los órdenes y las familias virales, se observa una fuerte conservación en el orden génico. Los factores que limitan o promueven la diversidad en el orden génico son en su mayoría desconocidos, aunque se sabe que la regulación de la expresión génica es una limitación importante para los virus. Aquí investigamos por qué el orden de los genes se conserva en un virus de ARN de cadena positiva que codifica una única poliproteína, en el contexto de su huésped natural (multicelular). Inicialmente, hemos identificado la trayectoria más plausible por la cual los órdenes alternativos de genes podrían evolucionar. Posteriormente, se estudió la accesibilidad de los pasos clave a lo largo de esta trayectoria evolutiva mediante la construcción de dos pasos intermedios: (*i*) la duplicación de un gen seguido por (*ii*) la pérdida del gen ancestral. Se identificaron cinco barreras a la evolución de los órdenes alternativos de genes. En primer lugar, el número de posiciones viables para el reordenamiento es limitado. En segundo lugar, la eficacia viral dentro del huésped de los virus con duplicaciones de genes es baja en comparación con el virus silvestre. En tercer lugar, después de la duplicación, la copia del gen ancestral siempre se mantiene, y nunca la del gen duplicado. En cuarto lugar, los virus con un orden génico alternativo tienen una eficacia viral incluso más baja que los virus con duplicaciones de genes. En quinto lugar, después de más de medio año de evolución experimental, en aislamiento, los virus con un orden génico alternativo son aún mucho más inferiores que el virus silvestre. Nuestros

resultados muestran que todos los pasos a lo largo de las trayectorias evolutivas plausibles a los órdenes génicos alternativos, son muy poco probables. Estos resultados contribuyen a nuestro entendimiento acerca de los factores que limitan y/o promueven la conservación del orden génico.

Capítulo 2: Predicción de la estabilidad de duplicaciones de genes homólogos

El segundo objetivo es comprender mejor la estabilidad de la redundancia genética y cómo los virus evolucionan a tener genomas pequeños mediante la eliminación de ésta. Una de las características sorprendentes de muchos eucariotas, es la cantidad aparente de redundancia de los elementos codificantes y no codificantes de sus genomas. A pesar de sus ventajas evolutivas, hay menos ejemplos de secuencias redundantes en los genomas virales, particularmente en aquellos con genomas de ARN. La baja prevalencia de duplicación génica en los virus de ARN probablemente refleja las fuertes restricciones selectivas en contra del aumento en el tamaño de genoma. Mediante la evolución experimental de variantes de TEV hemos explorado como un genoma de ARN viral puede acomodar duplicaciones génicas potencialmente beneficiosas, y cómo la evolución adaptativa procede a eliminarlos. Hemos medido la estabilidad y los costes en la eficacia viral de los genomas del TEV con redundancia genética. Ninguno de los eventos de duplicación explorados parecen ser beneficiosos para TEV. La duplicación de un gen resultó en un virus no viable o en una reducción significativa en la eficacia viral. En todos los casos, se delecionó el gen duplicado, lo que condujo a la recuperación de la eficacia viral. Hemos observado diferencias en la dinámica de cómo se delecionan las secuencias, asociadas a la duración de los

pases evolutivos, y al origen, al tamaño y a la ubicación del gen duplicado. Basado en los datos experimentales, hemos desarrollado un modelo que permite predecir la estabilidad de la redundancia genética. Con este modelo, mostramos que hay una tasa de recombinación dependiente del contexto genómico, donde la identidad y la posición de la duplicación juegan un papel. Los resultados contribuyen a nuestra comprensión de que características biológicas limitan la probabilidad de la retención de genes duplicados.

Capítulo 3: Introducción de secuencias exógenas funcionales

El tercer objetivo es explorar el destino evolutivo de un aumento en el tamaño genómico, en el contexto de la transferencia horizontal de genes (THG). La THG es un mecanismo clave en la evolución de los virus y extendido en estos organismos. Sin embargo, la fuerte selección en contra de un aumento en el tamaño genómico de los virus es un impedimento para la THG. Además, la inestabilidad de los genomas virales puede impedir la innovación evolutiva. Aquí, consideramos dos posibles eventos de THG introduciendo secuencias exógenas funcionales en el genoma de TEV que están relacionadas con la supresión del silenciamiento de ARN. Mediante evolución experimental, hemos observado cómo se acomodan estas secuencias exógenas y hemos determinado su estabilidad. Uno de los eventos simula la adquisición de una nueva función, mediante THG del dominio AlkB, responsable de la reparación de daños de alquilación. Encontramos que AlkB es inestable en TEV, por lo menos en nuestra configuración experimental. El otro evento simula la adquisición de una función existente, mediante THG del supresor del silenciamiento 2b del virus del mosaico del pepino (CMV). Este gen foráneo es estable y no parece afectar

al TEV en términos de acumulación y eficacia darwiniana. Además, nuestros resultados sugieren que este gen puede hacerse cargo de la supresión de silenciamiento del virus, relajando la presión de selección sobre esta función en la proteína HC-Pro del TEV.

Capítulo 4: Los efectos sobre la eficacia viral de las secuencias exógenas pueden ser impredecibles en huéspedes alternativos

El cuarto objetivo es entender mejor los efectos de la virulencia y la transmisión sobre la evolución. Cuando predominan las presiones de selección entre huéspedes, la teoría sugiere que la alta virulencia podría dificultar la transmisión de microparásitos entre huéspedes, y por lo tanto, que la virulencia evolucionaría hacia niveles más bajos, lo cuál optimizaría la transmisión entre huéspedes. Los microparásitos con una alta virulencia también pueden restringir el desarrollo del huésped, lo cual limitaría los recursos del huésped disponibles tanto para ellos como para su propio tamaño poblacional efectivo. Por lo tanto, la alta virulencia puede reducir la tasa de mutación y aumentar la fuerza con la que la deriva genética actúa sobre las poblaciones de microparásitos, limitando así el potencial para adaptarse al huésped y, finalmente, quizás a la capacidad de evolucionar una virulencia baja. Como una primera exploración de esta hipótesis, hemos evolucionado experimentalmente un TEV protador del marcador eGFP en dos especies de huésped semipermisivas, *Nicotiana benthamiana* y *Datura stramonium*. Después de aproximadamente 30 semanas de evolución, hemos secuenciado los genomas de los linajes evolucionados y hemos medido su eficacia viral. Sorprendentemente, el marcador eGFP no tiene ningún costo de eficacia viral en *N. benthamiana*, mientras que sí tenia un coste

en *D. stramonium*, lo que sugiere que los efectos en la eficacia de las secuencias heterólogas pueden ser impredecibles en huéspedes alternativos. En *N. benthamiana*, el marcador eGFP es estable en todos los linajes. En uno de los seis linajes evolucionados donde observamos una eficacia mayor, sólo encontramos mutaciones puntuales. En *D. stramonium*, el marcador eGFP se perdió en sólo uno de los diez linajes siendo éste el único con un aumento significativo en la eficacia viral. Los patrones observados de la adaptación son consistentes con los costes de la inserción de eGFP en los diferentes huéspedes. Nuestros resultados no proporcionan soporte alguno a la hipótesis de que la alta virulencia impide la evolución. Más bien, los resultados sugieren que un salto entre huéspedes puede cambiar radicalmente la dinámica evolutiva.

Conclusiones

Se identificaron múltiples barreras a la evolución de órdenes génicos alternativos, revelando limitaciones importantes a la evolución viral. Los pasos intermedios para evolucionar un orden génico alternativo, están asociados a costes elevados en la eficacia viral, los cuales no son fáciles de superar. La falta de trayectorias accesibles a un orden génico alternativo sugiere una contribución a la conservación del orden génico en los potyvirus.

Las duplicaciones génicas que se exploraron en este estudio son evolutivamente inestables y no parecen ser beneficiosas para TEV. La duplicación de un gen resulta en una reducción en la eficacia viral, la cual se restaura mediante la eliminación de la copia duplicada del gen. Además de la eficacia viral, la estabilidad de la copia del gen duplicado se ve afectada por una tasa de

recombinación dependiente de la identidad y la posición de las duplicaciones génicas.

Contrariamente a la duplicación génica, la inserción de una secuencia exógena funcional – el gen *2b* del virus del CMV – ha demostrado ser evolutivamente estable, sin afectar a la eficacia viral. Además, *2b* ha demostrado ser beneficioso para TEV en el caso del fallo de su propio mecanismo de supresión de silenciamento de ARN. Estos resultados sugieren que la THG entre virus de plantas es un evento posible y prometedor para futuros estudios en evolución experimental.

Un salto de huéspedes puede cambiar radicalmente la dinámica evolutiva de un virus. Una secuencia exógena funcional – eGFP – era más estable en huéspedes alternativos, para los cuales TEV tenía una virulencia tanto menor como superior. Además, eGFP no parece afectar la eficacia viral en el huésped para el que TEV tiene una virulencia alta. Esto choca con nuestra hipótesis de que la alta virulencia ralentiza el ritmo de adaptación. Los resultados sugieren que cuando se considera la evolución de la arquitectura del genoma, los saltos de huéspedes podrían desempeñar un papel muy importante, al permitir que los pasos evolutivos intermedios sean competitivos.

References

Adams M. J., Antoniw J. F., Fauquet C. M., 2005a Molecular criteria for genus and species discrimination within the family *Potyviridae*. Arch. Virol. **150**: 459–479.

Adams M. J., Antoniw J. F., Beaudoin F., 2005b Overview and analysis of the polyprotein cleavage sites in the family *Potyviridae*. Mol. Plant Pathol. **6**: 471–487.

Alizon S., Hurford A., Mideo N., Van Baalen M., 2009 Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. J. Evol. Biol. **22**: 245–259.

Andersson D. I., Hughes D., 2009 Gene amplification and adaptive evolution in bacteria. Annu. Rev. Genet. **43**: 167–195.

Anindya R., Savithri H. S., 2004 Potyviral NIa proteinase, a proteinase with novel deoxyribonuclease activity. J. Biol. Chem. **279**: 32159–32169.

Aravind L., Koonin E. V, 2001 The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol. **2**: research0007.1–research0007.8.

Atreya C. D., Atreya P. L., Thornbury D. W., Pirone T. P., 1992 Site-directed mutations in the potyvirus HC-PRO gene affect helper component activity, virus accumulation, and symptom expression in infected tobacco plants. Virology **191**: 106–111.

Bald J. G., 1937 The use of numbers of infections for comparing the concentration of plant virus suspensions: dilution experiments with purified suspensions. Ann. Appl. Biol. **24**: 33–55.

Baltimore D., 1971 Expression of animal virus genomes. Bacteriol. Rev. **35**: 235–241.

Bamford D. H., Grimes J. M., Stuart D. I., 2005 What does structure tell us

about virus evolution? Curr. Opin. Struct. Biol. **15**: 655–663.

Barrell B. G., Air G. M., Hutchison C. A., 1976 Overlapping genes in bacteriophage φX174. Nature **264**: 34–41.

Beauchemin C., Boutet N., Laliberte J. F., 2007 Visualization of the interaction between the precursors of VPg, the viral protein linked to the genome of *Turnip mosaic virus*, and the translation eukaryotic initiation factor iso 4E in planta. J. Virol. **81**: 775–782.

Beauchemin C., Laliberté J. F., 2007 The poly(A) binding protein is internalized in virus-induced vesicles or redistributed to the nucleolus during turnip mosaic virus infection. J. Virol. **81**: 10905–10913.

Bedhomme S., Lafforgue G., Elena S. F., 2012 Multihost experimental evolution of a plant RNA virus reveals local adaptation and host-specific mutations. Mol. Biol. Evol. **29**: 1481–1492.

Bedoya L. C., Daròs J. A., 2010 Stability of *Tobacco etch virus* infectious clones in plasmid vectors. Virus Res. **149**: 234–240.

Belshaw R., Pybus O. G., Rambaut A., 2007 The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. **17**: 1496–1504.

Bennett G. M., Moran N. A., 2013 Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. Genome Biol. Evol. **5**: 1675–1688.

Blackman R. L., Eastop V. F., 2000 *Aphids on the world's crops. An identification and information guide.* John Wiley & Sons, Inc., New York.

Blasdell K. R., Voysey R., Bulach D., Joubert D. A., Tesh R. B., Boyle D. B., Walker P. J., 2012 Kotonkan and Obodhiang viruses: African ephemeroviruses with large and complex genomes. Virology **425**: 143–153.

Boetzer M., Pirovano W., 2012 Toward almost closed genomes with GapFiller.

Genome Biol. **13**: R56.

Boots M., Sasaki A., 2000 The evolutionary dynamics of local infection and global reproduction in host-parasite interactions. Ecol. Lett. **3**: 181–185.

Born E. van den, Omelchenko M. V, Bekkelund A., Leihne V., Koonin E. V, Dolja V. V, Falnes P. O., 2008 Viral AlkB proteins repair RNA damage by oxidative demethylation. Nucleic Acids Res. **36**: 5451–5461.

Boyko V. P., Karasev A. V., Agranovsky A. A., Koonin E. V., Dolja V. V., 1992 Coat protein gene duplication in a filamentous RNA virus of plants. Proc. Natl. Acad. Sci. USA **89**: 9156–9160.

Bratlie M. S., Drabløs F., 2005 Bioinformatic mapping of AlkB homology domains in viruses. BMC Genomics **6**: 1.

Brown S. P., Cornforth D. M., Mideo N., 2012 Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control. Trends Microbiol. **20**: 336–342.

Bull J. J., Ebert D., 2008 Invasion thresholds and the evolution of nonequilibrium virulence. Evol. Appl. **1**: 172–182.

Bull J. J., Lauring A. S., 2014 Theory and empiricism in virulence evolution. PLoS Pathog. **10**: 1–3.

Canchaya C., Proux C., Fournous G., Bruttin A., Brussow H., 2003 Prophage genomics. Microbiol. Mol. Biol. Rev. **67**: 238–276.

Carrasco P., Daròs J. A., Agudelo-Romero P., Elena S. F., 2007 A real-time RT-PCR assay for quantifying the fitness of *Tobacco etch virus* in competition experiments. J. Virol. Methods **139**: 181–188.

Carrington J. C., Freed D. D., Sanders T. C., 1989 Autocatalytic processing of the potyvirus helper component proteinase in *Escherichia coli* and *in vitro*. J. Virol. **63**: 4459–4463.

Carrington J. C., Haldeman R., Dolja V. V, Restrepo-Hartwig M. A., 1993 Internal cleavage and trans-proteolytic activities of the VPg-proteinase

(NIa) of *Tobacco etch potyvirus* in vivo. J. Virol. **67**: 6995–7000.

Casjens S., 2003 Prophages and bacterial genomics: what have we learned so far? Mol. Microbiol. **49**: 277–300.

Chung B. Y. W., Miller W. A., Atkins J. F., Firth A. E., 2008 An overlapping essential gene in the Potyviridae. Proc. Natl. Acad. Sci. USA. **105**: 5897–5902.

Clarke D. K., Duarte E. A., Moya A., Elena S. F., Domingo E., Holland J., 1993 Genetic bottlenecks and population passages cause profound fitness differences in RNA viruses. J. Virol. **67**: 222–228.

Codoñer F. M., Darós J. A., Solé R. V, Elena S. F., 2006 The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens. PLoS Pathog. **2**: e136.

Coffey L. L., Beeharry Y., Borderia A. V., Blanc H., Vignuzzi M., 2011 Arbovirus high fidelity variant loses fitness in mosquitoes and mice. Proc. Natl. Acad. Sci. USA. **108**: 16038–16043.

Colson P., Lamballerie X. de, Fournous G., Raoult D., 2012 Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. Intervirology **55**: 321–332.

Conant G. C., Wagner A., 2004 Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. Proc. R. Soc. B. **271**: 89–96.

Cripe T. P., Delos S. E., Estes P. A., Garcea R. L., 1995 *In vivo* and *in vitro* association of hsc70 with polyomavirus capsid proteins. J. Virol. **69**: 7807–7813.

Cronin S., Verchot J., Haldeman-Cahill R., Schaad M. C., Carrington J. C., 1995 Long-distance movement factor: a transport function of the potyvirus helper component proteinase. Plant Cell **7**: 549–559.

Cruz S., Roberts A., Prior D., Chapman S., Oparka K., 1998 Cell-to-cell and

phloem-mediated transport of *Potato virus X*. The role of virions. Plant Cell **10**: 495–510.

Dales S., Eggers H. J., Tamm I., Palade G. E., 1965 Electron microscopic study of the formation of poliovirus. Virology **26**: 379–389.

Dandekar T., Snel B., Huynen M., Bork P., 1998 Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. **23**: 324–328.

de Haan C. A. M., van Genne L., Stoop J. N., Volders H., Rottier P. J. M., 2003 Coronaviruses as vectors: position dependence of foreign gene expression. J. Virol. **77**: 11312–11323.

de la Iglesia F., Elena S. F., 2007 Fitness declines in *Tobacco etch virus* upon serial bottleneck transfers. J. Virol. **81**: 4941–4947.

Dietrich C., Maiss E., 2003 Fluorescent labelling reveals spatial separation of potyvirus populations in mixed infected *Nicotiana benthamiana* plants. J. Gen. Virol. **84**: 2871–2876.

Ding S. W., Li W. X., Symons R. H., 1995 A novel naturally occurring hybrid gene encoded by a plant RNA virus facilitates long distance virus movement. EMBO J. **14**: 5762–5772.

Ding S. W., Shi B. J., Li W. X., Symons R. H., 1996 An interspecies hybrid RNA virus is significantly more virulent than either parental virus. Proc. Natl. Acad. Sci. USA **93**: 7470–7474.

Dolja V. V., McBride H. J., Carrington J. C., 1992 Tagging of plant potyvirus replication and movement by insertion of beta-glucuronidase into the viral polyprotein. Proc. Natl. Acad. Sci. USA **89**: 10208–10212.

Dolja V. V, Herndon K. L., Pirone T. P., Carrington J. C., 1993 Spontaneous mutagenesis of a plant potyvirus genome after insertion of a foreign gene. J. Virol. **67**: 5968–5975.

Dolja V. V., Karasev A. V., Koonin E. V., 1994 Molecular biology and evolution

of closteroviruses: sophisticated build-up of large RNA genomes. Annu. Rev. Phytopathol. **32**: 261–285.

Domingo E., Sabo D., Taniguchi T., Weissmann C., 1978 Nucleotide sequence heterogeneity of an RNA phage population. Cell **13**: 735–744.

Domingo E., Holland J. J., 1997 RNA virus mutations and fitness for survival. Annu. Rev. Microbiol. **51**: 151–178.

Domingo E., 2000 Viruses at the edge of adaptation. Virology **270**: 251–253.

Domingo E., Sheldon J., Perales C., 2012 Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. **76**: 159–216.

Drake J. W., 1991 A constant rate of spontaneous mutation in DNA-based microbe. Proc. Natl. Acad. Sci. USA **88**: 7160–7164.

Drake J. W., Charlesworth B., Charlesworth D., Crow J. F., 1998 Rates of spontaneous mutation. Genetics **148**: 1667–1686.

Duarte E., Clarke D., Moya A., Domingo E., Holland J., 1992 Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet. Proc. Natl. Acad. Sci. USA **89**: 6015–6019.

Duffy S., Shackelton L. A., Holmes E. C., 2008 Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. **9**: 267–276.

Dufresne P. J., Thivierge K., Cotton S., Beauchemin C., Ide C., Ubalijoro E., Laliberté J. F., Fortin M. G., 2008 Heat shock 70 protein interaction with *Turnip mosaic virus* RNA-dependent RNA polymerase within virus-induced membrane vesicles. Virology **374**: 217–227.

Ebert D., Weisser W. W., 1997 Optimal killing for obligate killers: the evolution of life histories and virulence of semelparous parasites. Proc. R. Soc. B. **264**: 985–991.

Edwardson J. R., Christie R. G., 1996 *Cylindrical inclusions.* Agricultural Experiment Station, University of Florida, Gainesville, FL.

Eigen M., 1971 Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften **58**: 465–523.

Eigen M., Schuster P., 1977 A principle of natural self-organization. Naturwissenschaften **64**: 541–565.

Elena S. F., Carrasco P., Daròs J. A., Sanjuán R., 2006 Mechanisms of genetic robustness in RNA viruses. EMBO Rep. **7**: 168–173.

Elena S. F., Sanjuán R., 2007 Virus evolution: insights from an experimental approach. Annu. Rev. Ecol. Evol. Syst. **38**: 27–52.

Elena S. F., 2012 RNA virus genetic robustness: possible causes and some consequences. Curr. Opin. Virol. **2**: 525–530.

Endy D., You L., Yin J., Molineux I. J., 2000 Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. Proc. Natl. Acad. Sci. USA. **97**: 5375–5380.

Engler C., Gruetzner R., Kandzia R., Marillonnet S., 2009 Golden Gate Shuffling: A one-pot DNA shuffling method based on type IIs restriction enzymes. PLoS One **4**: e5553.

Escarmís C., Perales C., Domingo E., 2009 Biological effect of Muller's Ratchet: distant capsid site can affect picornavirus protein processing. J. Virol. **83**: 6748–6756.

Escarmís C., Dávila M., Domingo E., 1999 Multiple molecular pathways for fitness recovery of an RNA virus debilitated by operation of Muller's ratchet. J. Mol. Biol. **285**: 495–505.

Ewald P. W., 1983 Host-parasite relations, vectors, and the evolution of disease severity. Annu. Rev. Ecol. Syst. **14**: 465–485.

Fabre F., Moury B., Johansen E. I., Simon V., Jacquemond M., Senoussi R., 2014 Narrow bottlenecks affect *Pea seedborne mosaic virus* populations during vertical seed transmission but not during leaf colonization. PLoS Pathog. **10**: e1003833.

Fares M. A., Keane O. M., Toft C., Carretero-Paulet L., Jones G. W., 2013 The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. PLoS Genet. **9**: e1003176.

Fauquet C. M., Mayo M. A., Maniloff J., Desselberger U., Ball L. A., 2005 *Virus Taxonomy: Eighth Report of the International Committee on Taxonomy of Viruses*. Elsevier Academic Press, San Diego, CA.

Fazeli C. F., Rezaian M. A., 2000 Nucleotide sequence and organization of ten open reading frames in the genome of *Grapevine leafroll-associated virus 1* and identification of three subgenomic RNAs. J. Gen. Virol. **81**: 605–615.

Fernandez A., 1997 The motif V of plum pox potyvirus CI RNA helicase is involved in NTP hydrolysis and is essential for virus RNA replication. Nucleic Acids Res. **25**: 4474–4480.

Finn R. D., Bateman A., Clements J., Coggill P., Eberhardt R. Y., Eddy S. R., Heger A., Hetherington K., Holm L., Mistry J., Sonnhammer E. L. L., Tate J., Punta M., 2014 Pfam: the protein families database. Nucleic Acids Res. **42**: D222–D230.

Flint J. S., Enquist L. W., Racaniello V. R., Skalka A. M., 2009 *Principles of Virology, Third Edition*. ASM Press, Washington, DC.

Flint J., Racaniello V. R., Rall G. F., Skalka A. M., 2015 *Principles of Virology, Fourth Edition.* ASM Press, Washington, DC.

Forss S., Schaller H., 1982 A tandem repeat gene in a picornavirus. Nucleic Acids Res. **10**: 6441–6450.

Francki R. I. B., Fauquet C. M., Knudson D. L., Brown F., 1991 Classification and nomenclature of viruses. Fourth report of the International Committee on Taxonomy of Viruses. Fifth Rep. Int. Comm. Taxon. Viruses, Arch. Virol. **17**: 1.

Fukuhara T., Koga R., Aoki N., Yuki C., Yamamoto N., Oyama N., Udagawa T., Horiuchi H., Miyazaki S., Higashi Y., Takeshita M., Ikeda K., Arakawa M., Matsumoto N., Moriyama H., 2006 The wide distribution of endornaviruses, large double-stranded RNA replicons with plasmid-like properties. Arch. Virol. **151**: 995–1002.

Furumoto W. A., Mickey R., 1967 A mathematical model for the infectivity-dilution curve of *Tobacco mosaic virus*: Experimental tests. Virology **32**: 224–233.

Gabrenaite-Verkhovskaya R., Andreev I. A., Kalinina N. O., Torrance L., Taliansky M. E., Makinen K., 2008 Cylindrical inclusion protein of *Potato virus A* is associated with a subpopulation of particles isolated from infected plants. J. Gen. Virol. **89**: 829–838.

Gago S., Elena S. F., Flores R., Sanjuán R., 2009 Extremely high mutation rate of a hammerhead viroid. Science **323**: 1308.

Gal-On A., 2000 A point mutation in the FRNK motif of the potyvirus Helper Component-Protease gene alters symptom expression in cucurbits and elicits protection against the severe homologous virus. Phytopathology **90**: 467–473.

Gammelgard E., Mohan M., Valkonen J. P. T., 2007 Potyvirus-induced gene silencing: the dynamic process of systemic silencing and silencing suppression. J. Gen. Virol. **88**: 2337–2346.

Gorman O. T., Bean W. J., Webster R. G., 1992 Evolutionary processes in Influenza viruses: divergence, rapid evolution, and stasis. Curr. Top. Microbiol. Immunol. **176**: 75-97

Grangeon R., Jiang J., Laliberté J.-F., 2012 Host endomembrane recruitment for plant RNA virus replication. Curr. Opin. Virol. **2**: 683–690.

Gu Z., Steinmetz L. M., Gu X., Scharfe C., Davis R. W., Li W.-H., 2003 Role of duplicate genes in genetic robustness against null mutations. Nature **421**: 63–66.

Gubala A., Davis S., Weir R., Melville L., Cowled C., Walker P., Boyle D., 2010 Ngaingan virus, a macropod-associated rhabdovirus, contains a second glycoprotein gene and seven novel open reading frames. Virology **399**: 98–108.

Guo H. S., Ding S. W., 2002 A viral protein inhibits the long range signaling activity of the gene silencing signal. EMBO J. **21**: 398–407.

Gutiérrez S., Pirolles E., Yvon M., Baecker V., Michalakis Y., Blanc S., 2015 The multiplicity of cellular infection changes depending on the route of cell infection in a plant virus. J. Virol. **89**: 9665–9675.

Hawkin J. D., 1988 A survey on intron and exon lengths. Nucleic Acids Res. **16**: 9893–9908.

Hernández-Crespo P., Sait S. M., Hails R. S., Cory J. S., 2001 Behavior of a recombinant baculovirus in lepidopteran hosts with different susceptibilities. Appl. Environ. Microbiol. **67**: 1140–1146.

Himmelreich R., Plagens H., Hilbert H., Reiner B., Herrmann R., 1997 Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. Nucleic Acids Res. **25**: 701–12.

Holmes E. C., 2003 Error thresholds and the constraints to RNA virus evolution. Trends Microbiol. **11**: 543–546.

Holmes E. C., 2009 The evolutionary genetics of emerging viruses. Annu. Rev. Ecol. Evol. Syst. **40**: 353–372.

Hong Y., Hunt A. G., 1996 RNA polymerase activity catalyzed by a potyvirus-encoded RNA-dependent RNA polymerase. Virology **226**: 146–151.

Hsiao T. L., Vitkup D., 2008 Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet. **4**: e1000014.

Hughes J. F., Coffin J. M., 2001 Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat. Genet. **29**: 487–489.

Hurst G. D. D., Werren J. H., 2001 The role of selfish genetic elements in eukaryotic evolution. Nat. Rev. Genet. **2**: 597–606.

Iyer L. M., Aravind L., Koonin E. V., 2001 Common origin of four diverse families of large eukaryotic DNA viruses. J. Virol. **75**: 11720–11734.

Kambol R., Kabat P., Tristem M., 2003 Complete nucleotide sequence of an endogenous retrovirus from the amphibian, *Xenopus laevis*. Virology **311**: 1–6.

Karasev A. V, Boyko V. P., Gowda S., Nikolaeva O. V, Hilf M. E., Koonin E. V, Niblett C. L., Cline K., Gumpf D. J., Lee R. F., 1995 Complete sequence of the *Citrus tristeza virus* RNA genome. Virology **208**: 511–520.

Kasschau K. D., Carrington J. C., 2001 Long-distance movement and replication maintenance functions correlate with silencing suppression activity of potyviral HC-Pro. Virology **285**: 71–81.

Keane O. M., Toft C., Carretero-Paulet L., Jones G. W., Fares M. A., 2014 Preservation of genetic and regulatory robustness in ancient gene duplicates of *Saccharomyces cerevisiae*. Genome Res. **24**: 1830–1841.

Kelley W. L., 1998 The J-domain family and the recruitment of chaperone power. Trends Biochem. Sci. **23**: 222–227.

King A. M. Q., Lefkowitz E., Adams M. J., Carstens E. B., 2011 *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*. Elsevier Academic Press, San Diego, CA.

Klein P. G., Klein R. R., Rodriguez-Cerezo E., Hunt A. G., Shaw J. G., 1994 Mutational analysis of the *Tobacco vein mottling virus* genome. Virology **204**: 759–769.

Koboldt D. C., Zhang Q., Larson D. E., Shen D., McLellan M. D., Lin L., Miller C. A., Mardis E. R., Ding L., Wilson R. K., 2012 VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. **22**: 568–576.

Kolstø A. B., 1997 Dynamic bacterial genome organization. Mol. Microbiol. **24**: 241–248.

Koonin E. V, Dolja V. V, 1993 Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit. Rev. Biochem. Mol. Biol. **28**: 375–430.

Koonin E. V, Galperin M. Y., 1997 Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr. Opin. Genet. Dev. **7**: 757–763.

Koonin E. V, Wolf Y. I., Nagasaki K., Dolja V. V, 2008 The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. Nat. Rev. Microbiol. **6**: 925–939.

Koonin E. V, 2009 Evolution of genome architecture. Int. J. Biochem. Cell Biol. **41**: 298–306.

Kreuze J. F., Savenkov E. I., Valkonen J. P. T., 2002 Complete genome sequence and analyses of the subgenomic RNAs of *Sweet potato chlorotic stunt virus* reveal several new features for the genus *Crinivirus*. J. Virol. **76**: 9260–9270.

Kung Y. J., Lin P. C., Yeh S. D., Hong S. F., Chua N. H., Liu L. Y., Lin C. P., Huang Y. H., Wu H. W., Chen C. C., Lin S. S., 2014 Genetic analyses of the FRNK motif function of *Turnip mosaic virus* uncover multiple and potentially interactive pathways of cross-protection. Mol. Plant Microbe Interact. **27**: 944–955.

Kurowski M. A., Bhagwat A. S., Papaj G., Bujnicki J. M., 2003 Phylogenomic identification of five new human homologs of the DNA repair enzyme AlkB. BMC Genomics **4**: 48.

Lafforgue G., Tromas N., Elena S. F., Zwart M. P., 2012 Dynamics of the establishment of systemic Potyvirus infection: independent yet cumulative action of primary infection sites. J. Virol. **86**: 12912–12922.

Lalić J., Cuevas J. M., Elena S. F., 2011 Effect of host species on the distribution of mutational fitness effects for an RNA virus. PLoS Genet. **7**: e1002378.

Langmead B., Salzberg S. L., 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**: 357–359.

LaPierre L. A., Holzschu D. L., Bowser P. R., Casey J. W., 1999 Sequence and transcriptional analyses of the fish retroviruses walleye epidermal hyperplasia virus types 1 and 2: evidence for a gene duplication. J. Virol. **73**: 9393–9403.

Lauring A. S., Frydman J., Andino R., 2013 The role of mutational robustness in RNA virus evolution. Nat. Rev. Microbiol. **11**: 327–336.

Lenski R. E., Ofria C., Collier T. C., Adami C., 1999 Genome complexity, robustness and genetic interactions in digital organisms. Nature **400**: 661–664.

Léonard S., Viel C., Beauchemin C., Daigneault N., Fortin M. G., Laliberté J.-F., 2004 Interaction of VPg-Pro of *Turnip mosaic virus* with the translation initiation factor 4E and the poly(A)-binding protein in planta. J. Gen. Virol. **85**: 1055–1063.

Li X. H., Carrington J. C., 1995 Complementation of *Tobacco etch potyvirus* mutants by active RNA polymerase expressed in transgenic cells. Proc. Natl. Acad. Sci. USA. **92**: 457–461.

Li H. W., Lucy A. P., Guo H. S., Li W. X., Ji L. H., Wong S. M., Ding S. W., 1999 Strong host resistance targeted against a viral suppressor of the plant gene silencing defence mechanism. EMBO J. **18**: 2683–2691.

Li H., Roossinck M. J., 2004 Genetic bottlenecks reduce population variation in an experimental RNA virus population. J. Virol. **78**: 10582–10587.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 2009 The sequence alignment/map format and

SAMtools. Bioinformatics **25**: 2078–2079.

Lillo F., Krakauer D. C., 2007 A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. Biol. Direct **2**: 22.

Lim K., Yin J., 2009 Computational fitness landscape for all gene-order permutations of an RNA virus. PLoS Comput. Biol. **5**: e1000283.

Lin S. S., Wu H. W., Jan F. J., Hou R. F., Yeh S. D., 2007 Modifications of the Helper Component-Protease of *Zucchini yellow mosaic virus* for generation of attenuated mutants for cross protection against severe infection. Phytopathology **97**: 287–296.

Lin L., Luo Z., Yan F., Lu Y., Zheng H., Chen J., 2011 Interaction between potyvirus P3 and ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO) of host plants. Virus Genes **43**: 90–92.

Liu H., Fu Y., Jiang D., Li G., Xie J., Cheng J., Peng Y., Ghabrial S. A., Yi X., 2010 Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J. Virol. **84**: 11876–11887.

López-Moya J. J., García J. A., 2008 Potyviruses. p. 313–322. In: *Encyclopedia of Virology*, *Third Edition*, Elsevier, Oxford, United Kingdom.

Lorenzo-Redondo R., Borderia A. V., Lopez-Galindez C., 2011 Dynamics of *in vitro* fitness recovery of HIV-1. J. Virol. **85**: 1861–1870.

Lwoff A., Horne R., Tournier P., 1962 A system of viruses. Cold Spring Harb. Symp. Quant. Biol. **27**: 51–55.

Lynch M., Conery J. S., 2000 The evolutionary fate and consequences of duplicate genes. Science. **290**: 1151–1155.

Lynch M., Conery J. S., 2003 The Origins of Genome Complexity. Science. **302**: 1401–1404.

Lynch M., 2006 Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. **60**: 327–349.

Lynch M., 2010 Evolution of the mutation rate. Trends Genet. **26**: 345–352.

Majer E., Daròs J.-A., Zwart M., 2013 Stability and fitness impact of the visually discernible Rosea1 marker in the *Tobacco etch virus* genome. Viruses **5**: 2153–2168.

Majer E., Salvador Z., Zwart M. P., Willemsen A., Elena S. F., Daròs J. A., 2014 Relocation of the NIb gene in the *Tobacco etch potyvirus* genome. J. Virol. **88**: 4586–4590.

Mallory A. C., Reinhart B. J., Bartel D., Vance V. B., Bowman L. H., 2002 A viral suppressor of RNA silencing differentially regulates the accumulation of short interfering RNAs and micro-RNAs in tobacco. Proc. Natl. Acad. Sci. USA. **99**: 15228–15233.

Malpica J. M., Fraile A., Moreno I., Obies C. I., Drake J. W., García-Arenal F., 2002 The rate and character of spontaneous mutation in an RNA virus. Genetics **162**: 1505–1511.

Martínez F., Sardanyés J., Elena S. F., Daròs J. A., 2011 Dynamics of a plant RNA virus intracellular accumulation: stamping machine vs. geometric replication. Genetics **188**: 637–646.

May R. M., Nowak M. A., 1995 Coinfection and the evolution of parasite virulence. Proc. R. Soc. B. **261**: 209–215.

Moreira D., Brochier-Armanet C., 2008 Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. BMC Evol. Biol. **8**: 12.

Moreno A., Tjallingii W. F., Fernández-Mata G., Fereres A., 2012 Differences in the mechanism of inoculation between a semi-persistent and a non-persistent aphid-transmitted plant virus. J. Gen. Virol. **93**: 662–667.

Moya A., Elena S. F., Bracho A., Miralles R., Barrio E., 2000 The evolution of RNA viruses: A population genetics view. Proc. Natl. Acad. Sci. USA **97**: 6967–6973.

Muller H. J., 1964 The relation of recombination to mutational advance. Mutat. Res. Mol. Mech. Mutagen. **1**: 2–9.

Nasir A., Kim K., Caetano-Anolles G., 2012 Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. BMC Evol. Biol. **12**: 156.

Ng J. C. K., Perry K. L., 2004 Transmission of plant viruses by aphid vectors. Mol. Plant Pathol. **5**: 505–511.

Novella I. S., Duarte E. A., Elena S. F., Moya A., Domingo E., Holland J. J., 1995 Exponential increases of RNA virus fitness during large population transmissions. Proc. Natl. Acad. Sci. USA **92**: 5841–5844.

Novella I. S., 2004 Negative effect of genetic bottlenecks on the adaptability of *Vesicular stomatitis virus*. J. Mol. Biol. **336**: 61–67.

Novella I. S., Ball L. A., Wertz G. W., 2004 Fitness analyses of vesicular stomatitis strains with rearranged genomes reveal replicative disadvantages. J. Virol. **78**: 9837–9841.

Pagán I., Montes N., Milgroom M. G., García-Arenal F., 2014 Vertical transmission selects for reduced virulence in a plant virus and for increased resistance in the host. PLoS Pathog. **10**: e1004293.

Palade G. E., 1955 A small particulate component of the cytoplasm. J. Biophys. Biochem. Cytol. **1**: 59–68.

Peng C. W., Peremyslov V. V, Mushegian A. R., Dawson W. O., Dolja V. V, 2001 Functional specialization and evolution of leader proteinases in the family *Closteroviridae*. J. Virol. **75**: 12153–12160.

Pesko K., Voigt E. A., Swick A., Morley V. J., Timm C., Yin J., Turner P. E., 2015 Genome rearrangement affects RNA virus adaptability on prostate cancer cells. Front. Genet. **6**: 121.

Pfeiffer J. K., Kirkegaard K., 2005 Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. PLoS Pathog. **1**: e11.

Plisson C., Drucker M., Blanc S., German-Retana S., Gall O. Le, Thomas D., Bron P., 2003 Structural characterization of HC-Pro, a plant virus multifunctional protein. J. Biol. Chem. **278**: 23753–23761.

Pommier Y., Johnson A. A., Marchand C., 2005 Integrase inhibitors to treat HIV/Aids. Nat. Rev. Drug Discov. **4**: 236–248.

Puustinen P., Mäkinen K., 2004 Uridylylation of the potyvirus VPg by viral replicase NIb correlates with the nucleotide binding capacity of VPg. J. Biol. Chem. **279**: 38103–38110.

R Core Team, 2014 R: A language and environment for statistical computing.

Racaniello V. R., 2013 Picornaviridae: The Viruses and Their Replication, in *Fields Virology, Fourth Edition.* Lippincott Williams & Wilkins, Philadelphia, PA.

Raoult D., Audic S., Robert C., Abergel C., Renesto P., Ogata H., Scola B. La, Suzan M., Claverie J. M., 2004 The 1.2-megabase genome sequence of Mimivirus. Science. **306**: 1344–1350.

Revers F., Gall O. Le, Candresse T., Maule A. J., 1999 New advances in understanding the molecular biology of plant/potyvirus interactions. Mol. Plant Microbe Interact. **12**: 367–376.

Revers F., García J. A., 2015 Molecular Biology of Potyviruses. Adv. Virus Res. **92**: 101–199.

Riechmann J. L., Cervera M. T., García J. A., 1995 Processing of the plum pox virus polyprotein at the P3-6K1 junction is not required for virus viability. J. Gen. Virol. **76** (Pt 4): 951–956.

Rocha E. P. C., 2008 The organization of the bacterial genome. Annu. Rev. Genet. **42**: 211–233.

Rodrigo G., Zwart M. P., Elena S. F., 2014 Onset of virus systemic infection in plants is determined by speed of cell-to-cell movement and number of primary infection foci. J. R. Soc. Interface **11**: 20140555–20140555.

Roode J. C. de, Pansini R., Cheesman S. J., Helinski M. E. H., Huijben S., Wargo A. R., Bell A. S., Chan B. H. K., Walliker D., Read A. F., 2005 Virulence and competitive ability in genetically diverse malaria infections. Proc. Natl. Acad. Sci. USA **102**: 7624–7628.

Roode J. C. de, Yates A. J., Altizer S., 2008 Virulence-transmission trade-offs and population divergence in virulence in a naturally occurring butterfly parasite. Proc. Natl. Acad. Sci. USA **105**: 7489–7494.

Roossinck M. J., Sabanadzovic S., Okada R., Valverde R. A., 2011 The remarkable evolutionary history of endornaviruses. J. Gen. Virol. **92**: 2674–2678.

Roy S. W., Gilbert W., 2006 The evolution of spliceosomal introns: patterns, puzzles and progress. Nat. Rev. Genet. **7**: 211–221.

Salverda M. L. M., Dellus E., Gorter F. A., Debets A. J. M., Oost J. van der, Hoekstra R. F., Tawfik D. S., Visser J. A. G. M. de, 2011 Initial mutations direct alternative pathways of protein evolution. PLoS Genet. **7**: e1001321.

Sanjuán R., Agudelo-Romero P., Elena S. F., 2009 Upper-limit mutation rate estimation for a plant RNA virus. Biol. Lett. **5**: 394–396.

Sanjuán R., 2010 Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **365**: 1975–1982.

Sanjuán R., Nebot M. R., Chirico N., Mansky L. M., Belshaw R., 2010 Viral mutation rates. J. Virol. **84**: 9733–9748.

Sanjuán R., Cuevas J. M., Furió V., Holmes E. C., Moya A., 2007 Selection for robustness in mutagenized RNA viruses. PLoS Genet. **3**: e93.

Scherbakov D. V., Garber M. B., 2000 Overlapping genes in bacterial and phage genomes. Mol. Biol. **34**: 485–495.

Schmidt G. W., Delaney S. K., 2010 Stable internal reference genes for normalization of real-time RT-PCR in tobacco (*Nicotiana tabacum*) during

development and abiotic stress. Mol. Genet. Genomics **283**: 233–241.

Schmieder R., Edwards R., 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics **27**: 863–864.

Scola B. La, Audic S., Robert C., Jungang L., Lamballerie X. de, Drancourt M., Birtles R., Claverie J.-M., Raoult D., 2003 A giant virus in amoebae. Science. **299**: 2033.

Scola B. La, Desnues C., Pagnier I., Robert C., Barrassi L., Fournous G., Merchat M., Suzan-Monti M., Forterre P., Koonin E., Raoult D., 2008 The virophage as a unique parasite of the giant mimivirus. Nature **455**: 100–104.

Sedgwick B., Bates P., Paik J., Jacobs S., Lindahl T., 2007 Repair of alkylated DNA: recent advances. DNA Repair (Amst). **6**: 429–442.

Shah P., McCandlish D. M., Plotkin J. B., 2015 Contingency and entrenchment in protein evolution under purifying selection. Proc. Natl. Acad. Sci. USA. **112**: E3226–E3235.

Shi B. J., Miller J., Symons R. H., Palukaitis P., 2003 The 2b protein of cucumoviruses has a role in promoting the cell-to-cell movement of pseudorecombinant viruses. Mol. Plant Microbe Interact. **16**: 261–267.

Shi Y., Chen J., Hong X., Chen J., Adams M. J., 2007 A potyvirus P1 protein interacts with the Rieske Fe/S protein of its host. Mol. Plant Pathol. **8**: 785–790.

Shiboleth Y. M., Haronsky E., Leibman D., Arazi T., Wassenegger M., Whitham S. A., Gaba V., Gal-On A., 2007 The conserved FRNK box in HC-Pro, a plant viral suppressor of gene silencing, is required for small RNA binding and mediates symptom development. J. Virol. **81**: 13135–13148.

Sicard A., Yvon M., Timchenko T., Gronenborn B., Michalakis Y., Gutierrez S., Blanc S., 2013 Gene copy number is differentially regulated in a multipartite virus. Nat. Commun. **4**: 2248.

Siefert J. L., Martin K. A., Abdi F., Widger W. R., Fox G. E., 1997 Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. J. Mol. Evol. **45**: 467–472.

Simon-Loriere E., Holmes E. C., 2013 Gene duplication is infrequent in the recent evolutionary history of RNA viruses. Mol. Biol. Evol. **30**: 1263–1269.

Soitamo A. J., Jada B., Lehto K., 2011 HC-Pro silencing suppressor significantly alters the gene expression profile in tobacco leaves and flowers. BMC Plant Biol. **11**: 68.

Song D., Cho W. K., Park S. H., Jo Y., Kim K. H., 2013 Evolution of and horizontal gene transfer in the *Endornavirus* genus. PLoS ONE **8**: e64270.

Sorel M., Garcia J. A., German-Retana S., 2014 The *Potyviridae* cylindrical inclusion helicase: A key multipartner and multifunctional protein. Mol. Plant-Microbe Interact. **27**: 215–226.

Spencer D. H., Tyagi M., Vallania F., Bredemeyer A. J., Pfeifer J. D., Mitra R. D., Duncavage E. J., 2014 Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. J. Mol. Diagnostics **16**: 75–88.

Springman R., Badgett M. R., Molineux I. J., Bull J. J., 2005 Gene order constrains adaptation in bacteriophage T7. Virology **341**: 141–52.

Steinhauer D. A., la Torre J. C. de, Meier E., Holland J. J., 1989 Extreme heterogeneity in populations of *Vesicular stomatitis virus*. J. Virol. **63**: 2072–2080.

Stenger D. C., Hein G. L., French R., 2006 Nested deletion analysis of Wheat streak mosaic virus HC-Pro: Mapping of domains affecting polyprotein processing and eriophyid mite transmission. Virology **350**: 465–474.

Sullivan M. L., Ahlquist P., 1997 cis-Acting signals in bromovirus RNA replication and gene expression: networking with viral proteins and host

factors. Semin. Virol. **8**: 221–230.

Sung W., Ackerman M. S., Miller S. F., Doak T. G., Lynch M., 2012 Drift-barrier hypothesis and mutation-rate evolution. Proc. Natl. Acad. Sci. USA **109**: 18488–18492.

Susaimuthu J., Tzanetakis I. E., Gergerich R. C., Martin R. R., 2008 A member of a new genus in the *Potyviridae* infects *Rubus*. Virus Res. **131**: 145–151.

Thole V., Worland B., Snape J. W., Vain P., 2007 The pCLEAN dual binary vector system for agrobacterium-mediated plant transformation. Plant Physiol. **145**: 1211–1219.

Thornbury D. W., Patterson C. A., Dessens J. T., Pirone T. P., 1990 Comparative sequence of the helper component (HC) region of *Potato virus Y* and a HC-defective strain, *Potato virus C*. Virology **178**: 573–578.

Torres-Barceló C., Martín S., Daròs J. A., Elena S. F., 2008 From hypo- to hypersuppression: effect of amino acid substitutions on the RNA-silencing suppressor activity of the *Tobacco etch potyvirus* HC-Pro. Genetics **180**: 1039–1049.

Torres-Barceló C., Daròs J. A., Elena S. F., 2010 Compensatory molecular evolution of HC-Pro, an RNA-silencing suppressor from a plant RNA virus. Mol. Biol. Evol. **27**: 543–551.

Tristem M., Marshall C., Karpas A., Petrik J., Hill F., 1990 Origin of vpx in lentiviruses. Nature **347**: 341–342.

Tromas N., Elena S. F., 2010 The rate and spectrum of spontaneous mutations in a plant RNA virus. Genetics **185**: 983–989.

Tromas N., Zwart M. P., Lafforgue G., Elena S. F., 2014a Within-host spatiotemporal dynamics of plant virus infection at the cellular level. PLoS Genet. **10**: e1004186.

Tromas N., Zwart M. P., Forment J., Elena S. F., 2014b Shrinkage of genome size in a plant RNA virus upon transfer of an essential viral gene into the

host genome. Genome Biol. Evol. **6**: 538–550.

Turgeon R., 1989 The sink-source transition in leaves. Annu. Rev. Plant Physiol. Plant Mol. Biol. **40**: 119–138.

Turner P. E., Chao L., 1998 Sex and the evolution of intrahost competition in RNA virus φ6. Genetics **150**: 523–532.

Tzanetakis I. E., Postman J. D., Martin R. R., 2005 Characterization of a novel member of the family *Closteroviridae* from *Mentha* spp. Phytopathology **95**: 1043–1048.

Tzanetakis I. E., Martin R. R., 2007 Strawberry chlorotic fleck: Identification and characterization of a novel *Closterovirus* associated with the disease. Virus Res. **124**: 88–94.

Urcuqui-Inchima S., Haenni a L., Bernardi F., 2001 Potyvirus proteins: a wealth of functions. Virus Res. **74**: 157–175.

Valli A., López-Moya J. J., García J. A., 2007 Recombination and gene duplication in the evolutionary diversification of P1 proteins in the family *Potyviridae*. J. Gen. Virol. **88**: 1016–1028.

Valli A., Gallo A., Calvo M., Perez J. d. J., Garcia J. A., 2014 A novel role of the potyviral Helper Component proteinase contributes to enhance the yield of viral particles. J. Virol. **88**: 9808–9818.

Valverde R. A., Nameth S., Abdallha O., Al-Musa O., Desjardins P., Dodds A., 1990 Indigenous double-stranded RNA from pepper (*Capsicum annuum*). Plant Sci. **67**: 195–201.

Varrelmann M., Maiss E., Pilot R., Palkovics L., 2007 Use of pentapeptide-insertion scanning mutagenesis for functional mapping of the P*lum pox virus* helper component proteinase suppressor of gene silencing. J. Gen. Virol. **88**: 1005–1015.

Velasquez N., Hossain M. J., Murphy J. F., 2014 Differential disease symptoms and full-length genome sequence analysis for three strains of *Tobacco etch*

virus. Virus Genes. **50**: 442-449.

Verchot J., Koonin E. V, Carrington J. C., 1991 The 35-kDa protein from the N-terminus of the potyviral polyprotein functions as a third virus-encoded proteinase. Virology **185**: 527–535.

Verchot J., Carrington J. C., 1995a Evidence that the potyvirus P1 proteinase functions in trans as an accessory factor for genome amplification. J. Virol. **69**: 3668–3674.

Verchot J., Carrington J. C., 1995b Debilitation of plant potyvirus infectivity by P1 proteinase-inactivating mutations and restoration by second-site modifications. J. Virol. **69**: 1582–1590.

Vignuzzi M., Stone J. K., Arnold J. J., Cameron C. E., Andino R., 2006 Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature **439**: 344–348.

Vijayapalani P., Maeshima M., Nagasaki-Takekuchi N., Miller W. A., 2012 Interaction of the trans-frame potyvirus protein P3N-PIPO with host protein PCaP1 facilitates potyvirus movement. PLoS Pathog. **8**: e1002639.

Vos M. G. J. de, Dawid A., Sunderlikova V., Tans S. J., 2015 Breaking evolutionary constraint with a tradeoff ratchet. Proc. Natl. Acad. Sci. USA. **112**: 14906–14911.

Wakarchuk D. A., Hamilton R. I., 1990 Partial nucleotide sequence from enigmatic dsRNAs in Phaseolus vulgaris. Plant Mol. Biol. **14**: 637–639.

Walker P. J., Byrne K. A., Riding G. A., Cowley J. A., Wang Y., McWilliam S., 1992 The genome of bovine ephemeral fever rhabdovirus contains two related glycoprotein genes. Virology **191**: 49–61.

Wang Y., Walker P. J., 1993 Adelaide river rhabdovirus expresses consecutive glycoprotein genes as polycistronic mRNAs: new evidence of gene duplication as an evolutionary process. Virology **195**: 719–731.

Watanabe H., Mori H., Itoh T., Gojobori T., 1997 Genome plasticity as a

paradigm of eubacteria evolution. J. Mol. Evol. **44** Suppl 1: S57–64.

Wei T., Huang T. S., McNeil J., Laliberté J.-F., Hong J., Nelson R. S., Wang A., 2010 Sequential recruitment of the endoplasmic reticulum and chloroplasts for plant potyvirus replication. J. Virol. **84**: 799–809.

Wen R. H., Hajimorad M. R., 2010 Mutational analysis of the putative pipo of *Soybean mosaic virus* suggests disruption of PIPO protein impedes movement. Virology **400**: 1–7.

Wertz G. W., Perepelitsa V. P., Ball L. A., 1998 Gene rearrangement attenuates expression and lethality of a nonsegmented negative strand RNA virus. Proc. Natl. Acad. Sci. USA. **95**: 3501–3506.

Wu H. W., Lin S. S., Chen K. C., Yeh S. D., Chua N. H., 2010 Discriminating mutations of HC-Pro of *Zucchini yellow mosaic virus* with differential effects on small RNA pathways involved in viral pathogenicity and symptom development. Mol. Plant Microbe Interact. **23**: 17–28.

Xiao C., Chipman P. R., Battisti A. J., Bowman V. D., Renesto P., Raoult D., Rossmann M. G., 2005 Cryo-electron microscopy of the giant Mimivirus. J. Mol. Biol. **353**: 493–496.

Young B. A., Hein G. L., French R., Stenger D. C., 2007 Substitution of conserved cysteine residues in *Wheat streak mosaic virus* HC-Pro abolishes virus transmission by the wheat curl mite. Arch. Virol. **152**: 2107–2111.

Zhang J., 2003 Evolution by gene duplication: an update. Trends Ecol. Evol. **18**: 292–298.

Zwart M. P., Hemerik L., Cory J. S., de Visser J. A., Bianchi F. J, Van Oers M. M., Vlak J. M., Hoekstra R. F., Van der Werf W., 2009 An experimental test of the independent action hypothesis in virus-insect pathosystems. Proc. R. Soc. B. **276**: 2233–2242.

Zwart M. P., Daròs J. A., Elena S. F., 2011 One is enough: in vivo effective

population size is dose-dependent for a plant RNA virus. PLoS Pathog. **7**: e1002122.

Zwart M. P., Daròs J. A., Elena S. F., 2012 Effects of potyvirus effective population size in inoculated leaves on viral accumulation and the onset of symptoms. J. Virol. **86**: 9737–9747.

Zwart M. P., Tromas N., Elena S. F., 2013 Model-selection-based approach for calculating cellular multiplicity of infection during virus colonization of multi-cellular hosts. PLoS ONE **8**: e64657.

Zwart M. P., Willemsen A., Daròs J. A., Elena S. F., 2014 Experimental evolution of pseudogenization and gene loss in a plant RNA virus. Mol. Biol. Evol. **31**: 121–134.