PhD THESIS

Evolutionary responses of fast adapting populations to opposing selection pressures

Jaime Iranzo Sanz

UNIVERSIDAD CARLOS III DE MADRID

PhD THESIS

Evolutionary responses of fast adapting populations to opposing selection pressures

Author

Jaime Iranzo Sanz

Supervisor

Susanna Cuevas Manrubia

DEPARTMENT OF MATHEMATICS

Leganés, May 2013

a mis padres

Agradecimientos

No se sorprenderá quien me conozca al leer que escribir una tesis se asemeja a conquistar una montaña. Así como el éxito en una ascensión comienza a fraguarse mucho antes de acometer las primeras rampas, la lectura de la tesis culmina un proceso que se inicia muchos años atrás, quizá cuando todavía siendo un alumno de secundaria a uno le empieza a rondar el gusanillo de la Ciencia. Es por ello que el haber llegado hasta aquí se lo deba en gran medida a mis profesores, tanto en Zaragoza como más tarde en Pamplona y Madrid, puesto que de todos ellos he recibido enseñanzas útiles y una gran motivación.

Estoy muy agradecido a mis padres por despertar en mí el interés por el conocimiento y la naturaleza; por enseñarme a ser curioso, a plantear tantas veces la pregunta por qué...?, y por contestarla otras tantas. También por su comprensión y paciencia soportando tantas ausencias, a las que la ocupación de estudiante y aprendiz de científico me ha llevado.

En el ámbito académico, tengo una gran deuda con el Dr. Pablo Villoslada, mi mentor cuando todavía era estudiante de Biología en Pamplona, quien me brindó la oportunidad de conocer de cerca el trabajo de laboratorio y generosamente me ayudó a abrirme camino en el terreno de la Biología Teórica. También fueron claves en mi "reconversión" las estancias breves realizadas en 2008 con Ricard Solé, Mario Floría y Yamir Moreno, durante las que comencé a familiarizarme con las herramientas que hoy uso a diario. A los tres, gracias por la oportunidad brindada. Y ya durante el transcurso de la tesis, doy las gracias a Kim Sneppen y Eugene Koonin por acogerme en sus grupos a lo largo de sendas estancias; de las dos volví cargado de ideas nuevas y buenos recuerdos.

Una buena parte del contenido de esta tesis es fruto de diversas colaboraciones. Esteban Domingo y Celia Perales han hecho posible contrastar los modelos teóricos de evolución viral con experimentos de laboratorio; Francisco López de Saro y Manuel Gómez me llevaron a acometer el estudio de los elementos móviles en el genoma; Eugene Koonin, Alex Lobkovsky y Yuri Wolf despertaron mi interés por el sistema immunitario CRISPR-Cas en bacterias. A todos ellos agradezco su colaboración y disponibilidad para resolver dudas, compartir resultados y discutir ideas. Por los mis-

mos motivos estoy más que agradecido a José Cuesta y Anxo Sánchez, que también han desarrollado un papel importante a lo largo de esta tesis.

No puedo cerrar esta sección sin acordarme de todos aquéllos que han contribuido a que mi estancia en Madrid haya sido tan agradable como provechosa: Gracias a mi hermana, con quien una salida al campo se convierte en el remedio perfecto para un empacho de ecuaciones. A mis compañeros del grupo de Sistemas Evolutivos: Jacobo, Capi, Carlos, Michael, Jacob, Ester... y a tantos en el Centro de Astrobiología por echarme una mano (y a veces dos) siempre que lo he necesitado. Gracias a todos aquellos amigos con los que he compartido camino, tienda y cena bajo las estrellas, mientras conversábamos sobre por qué el mundo es cómo es, por qué cooperamos, cómo empezó la vida en la Tierra y llegó a ser lo que vemos hoy; lejos por unas horas de los libros, poyatas y ordenadores que, a veces, no nos dejan ver la realidad con los ojos desnudos.

Gracias finalmente a mi directora de tesis, Susanna C. Manrubia, por confiar en un biólogo que al comienzo de esta aventura ni siquiera sabía programar. A su buen consejo, continua disponibilidad y brillante intuición le debo el haber llegado hasta el final con el convencimiento de haber elegido bien.

Jaime Iranzo
Torrejón de Ardoz, Abril de 2013

This work has been possible thanks to finantial support from projects FIS2008-05273 (Ministerio de Ciencia e Innovación), FIS2011-27569 (Ministerio de Economía y Competitividad) and MODELICO-CM S2009/ESP-1691 (Comunidad de Madrid), and a contract from Comunidad de Madrid.

Contents

1

Introduction

1

Introduction

Nothing in biology makes sense except in the light of evolution.
—Theodosius Dobzhansky.

1.1 A brief history of evolutionary theory

The theory of evolution based on natural selection was originally proposed by Charles Darwin and Alfred Russel Wallace (Fig. 1.1, who independently presented their results to the Linnean Society of London in 1858. It was, however, in 1859 when Darwin's celebrated book *On the Origin of Species* became published. The theory of evolution, as conceived by Darwin, was based on two principles: (1) individuals in a population show a certain degree of variation—spontaneously originated during reproduction—that can be at least partially inherited by their progeny; and (2) due to resource limitations, organisms are engaged into a struggle for life, such that only a small fraction of seeds, eggs or other kind of descendants survives to adulthood and leaves offspring. Provided that some specific variations improve the ability of an individual to survive, it follows that individuals carrying favourable variations will succeed in the struggle for life and such variations will be preserved. It is the accumulation of small useful variations what results in adaptation and, eventually, evolution. In Darwin's words:

> "I have called this principle, by which each slight variation, if useful, is preserved, by the term of Natural Selection, in order to mark its relation to man's power of selection." (Darwin 1859)

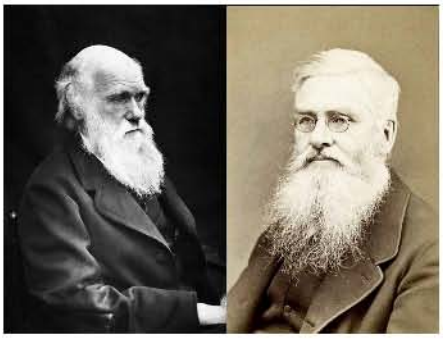


Figure 1.1: *Charles Darwin (left) and Alfred Russel Wallace (right).* Both proposed the theory of evolution through natural selection in 1858. Photographs were taken in 1869 and ca. 1880, respectively.

The multidisciplinary approaches Darwin took during the development of his theory are remarkable: not only did he merge empirical observations with formal reasoning but also designed and performed experiments that gave support to some of his hypotheses. For instance, when considering the possibility that plants be transported across the sea and settle down in new continents he writes:

> "Until I tried, with Mr. Berkeley's aid, a few experiments, it was not even known how far seeds could resist the injurious action of sea-water. To my surprise I found that out of 87 kinds, 64 germinated after an immersion of 28 days, and a few survived an immersion of 137 days."

Or, when discussing the high mortality in the struggle for existence:

> "on a piece of ground three feet long and two wide [...] I marked all the seedlings of our native weeds as they came up, and out of the 357 no less than 295 were destroyed, chiefly by slugs and insects."

A point that remained unsolved in Darwin's works was that of the mechanisms behind variation and inheritance. At his time, it was widely believed that the offspring characters were an average of those of their parents—the blending inheritance hypothesis—, which would make preservation of useful variants hardly feasible. One of the strongest advocates of this idea was Darwin's first cousin Francis Galton, for whom this averaging precluded the possibility of departing from the initial types, and thus of generating new species through Natural Selection. This difficulty was not overcome until the rediscovery of Mendel's laws by Hugo de Vries, Carl Correns and Erich von Tschermak at the turn of the century and the advent of population genetics in the decade of 1920.

A brief digression must be done here to discuss two pioneer works on mathematical biology that were carried out in the first quarter of the 20th century. On the one hand,

Figure 1.2: *Spirals in nature.* The ubiquity of some geometrical structures in nature was addressed by D'Arcy Thompson in his book *On Growth and Form*, hinting at the importance of Mathematics and Physics in understanding Biology.

there was the publication of the book *On Growth and Form* by D'Arcy W. Thompson in 1917. This singular book described multiple examples of parallelisms between biological forms and mechanical properties and explored the degree to which differences in the form of related animals could be described by means of relatively simple mathematical transformations (Fig. 1.2). D'Arcy argued that some morphological and structural traits might be the result of physical laws and mechanics acting on organisms rather than a product of natural selection. Somewhat, his work foresaw the influence that mathematical tools and physical concepts—such as dynamical systems, reaction-diffusion equations, and self-organization—would later have in Biology:

> "And while I have sought to shew the naturalist how a few mathematical concepts and dynamical principles may help and guide him, I have tried to shew the mathematician a field for his labour—a field which few have entered and no man has explored." (Thompson 1917)

On the other hand, it is also worth to mention the works of Alfred J. Lotka and Vito Volterra, who independently arrived in 1925 and 1926 to the Lotka-Volterra equations for describing predator-prey interactions. There is no need to insist on the influence that ecological models, many of them derived from those equations, have exerted on many branches of Biology—ecology, epidemiology and microbiology, among others—and also on further disciplines, such as economics.

Back to evolutionary theory, the decade of 1920 set the dawn of population genetics. Population genetics is the discipline that studies the genetic composition of a population and its change under the influence of evolutionary processes. It was founded on the works of Ronald A. Fisher, John B. S. Haldane, and Sewall Wright, who combined Mendelian genetics, statistical analysis, and mathematical models; and showed that inheritance according to Mendel's laws is consistent with natural selection and gradual evolution (Haldane 1924; Fisher 1930). In such a way, they took the first step in producing a unified theory of how evolution works. The modern evolutionary synthesis crystallised between 1936 and 1947 and still remains, to a large extent,

the current paradigm in evolutionary biology. Major figures in the development of the modern synthesis were zoologist Julian Huxley (who coined the term), geneticist Theodosius Dobzhansky, taxonomist Ernst Mayr, paleontologist George G. Simpson, zoologist Bernhard Rensch and botanist G. Ledyard Stebbins.

Modern synthesis explains evolution as the result of changes in the genetic composition of a population along successive generations. The course of evolution is ruled by the interplay of four main processes. First, mutation, which is the main source of genetic variation in a population. Mutation is usually the consequence of errors in genome replication and gives rise to new alleles (genetic variants). The particular set of alleles that one individual possesses (out of all possible allele combinations) is what defines its genotype. Second, natural selection, which consists of the differential survival and reproduction of individuals with different characteristics. Selection acts on the phenotype, which is the observable manifestation of the genotype—i.e., an abstraction of the characteristics displayed by an individual with a given genotype. The quantitative effect that a phenotype exerts on survival and reproduction defines the *fitness*[1] of such a phenotype. Third, genetic drift, which is the stochastic variation in the composition of a population due to random sampling along generations. Drift is specially relevant in small populations or when population bottlenecks occur. Fourth and finally, gene flow or transfer, which accounts for genetic exchange between different populations (for instance, due to migration).

Population genetics provides a collection of models that can be used to predict, under different circumstances, the probability that a new allele, produced through mutation in a single individual, spreads to the whole population and reaches *fixation*. In a converse manner, it also allows to calculate the probability that an old allele becomes extinct. In short, the fixation probability for a new allele is larger the higher the fitness associated to its phenotype, although random drift in small populations may counteract fitness and lead to fixation of slightly detrimental alleles. Classical population genetics mostly deals with homogeneous populations, for which the concept of fixation makes sense—each time a mutation occurs, the new allele either disappears or gets fixed before further mutations come about. In contrast, if the rate at which new mutations are produced is sufficiently high, populations become heterogeneous and multiple genetic variants coexist at any time. The study of evolution on such heterogeneous populations requires a wider framework—the quasispecies theory—that was introduced by Eigen and Schuster (1979). Due to its relevance to the study of viral evolution, an ampler discussion on quasispecies theory is carried out later in this chapter.

The fact that mutations take place at the genotype level while selection acts on the phenotype has deep implications for evolutionary dynamics. That is because there is no straightforward correspondence between genotype and phenotype—actually, the general case is that many genotypes produce the same phenotype. As a result, the fitness effects produced by mutations greatly vary. Indeed, as Motoo Kimura hypothesized after studying the aminoacidic sequences of proteins that became available in the decade of 1960, it may be the case that most mutations are almost neutral to natural selection

[1] Here, as in the rest of the text, words in italics denote key terms whose definitions appear in the Glossary.

(Kimura 1968). According to the Neutral Theory, the vast majority of evolutionary changes at the molecular level are due to random drift of selectively neutral mutants. Even though Kimura's neutral model only referred to molecular evolution, the idea that neutral, non-adaptive processes may play a more relevant role than previously thought was later extended to the evolution of organismal traits (Gould and Lewontin 1979) and, more recently, to the structure of ecosystems (Hubbell 2001), genomes (Lynch and Conery 2003), and protein interaction networks (Fernández and Lynch 2011). Nowadays, bridging the gap between genotype and phenotype and determining the degree at which selective and neutral processes contribute to evolutionary change remain central problems in evolutionary biology (Wagner 2008; Wagner 2011; Manrubia 2012).

1.2 Viruses, quasispecies and genome structure

1.2.1 Viral biology

Viruses are among the simplest biological entities in nature. They display all kinds of complex adaptive behavior and have found a large number of evolutionary solutions that turn them into a paradigmatic model to study evolution at the molecular and population levels.

From a structural point of view, viruses are generally composed of a genome—one or several molecules of either DNA or RNA—packed into a proteic or glicoproteic coat—the viral capsid. In some viruses there is also an envelope of lipids that surrounds the capsid, and a few enzymes that accompany the genetic material. Free viral particles are called virions and can survive in the environment for a limited period of time before degradation.

Although there are many kinds of viruses and the details differ among them, it is possible to extract some generalities of viral biology (Wagner and Hewlett 2004). Above all, viruses must infect cells in order to reproduce. The viral cycle, depicted in Fig. 1.3, starts when a virion adheres to a cell and gets into it[2]. The viral genome is then released from the capsid. The next step varies according to the virus, but it basically implies the replication of the viral genome and the expression of viral genes. In this phase, the virus "hijacks" the cell's machinery for its own benefit. The expression of viral genes implies the translation of genomic information into viral proteins, some of which will constitute the capsids. On the other hand, genome replication gives rise to a high number of genome copies. The viral cycle ends with the packing of genomes into capsids and the release of the new virions out of the cell.

From an epidemiological point of view, it is usual to characterize viruses according to two basic traits: infectivity and replicative ability. The former refers to the ability of the virus to enter new cells and carry out productive viral cycles; the latter is related to the number of genome copies (or alternatively, new virions) produced at each infection event.

[2]Note, however, that some viruses directly inject their genetic material into the cell, the capsid remaining outside.

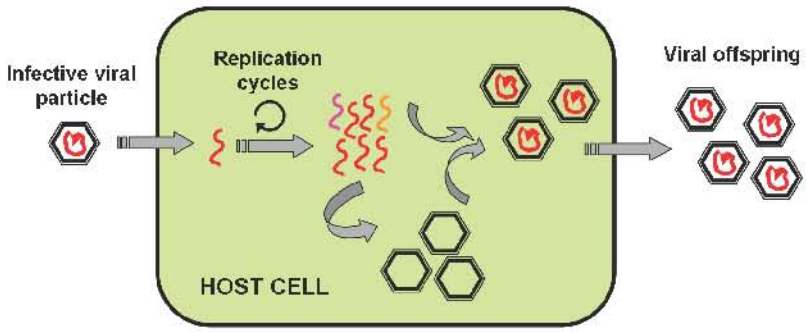


Figure 1.3: *Schematic of the viral cycle.* Hexagons represent viral capsids; red scrawls are viral genomes (packed into capsids or unpacked), while those coloured in orange and pink correspond to mutant genomes.

There are some molecular aspects of viral replication that are worth to delve into. First of all, when viral genomes are replicated some errors can be introduced, so that the resulting genomic sequences are not exact copies of the original. As already stated in the previous section, such errors are termed mutations and the rate at which they appear is the *mutation rate* (in this context the term *error rate* can be used with a similar meaning). Mutations occur from time to time in all living beings, but the mutation rate is especially high in viruses with RNA genomes. The reason is that proofreading and repair mechanisms—that account for fidelity in the copy of DNA genomes—do not work during replication of RNA (Steinhauer et al. 1992; Ferrer-Orta et al. 2006; Friedberg et al. 2006). The mutation rates in RNA viruses range from 10^{-3} to 10^{-5} mutations introduced per nucleotide copied (Batschelet et al. 1976; Drake and Holland 1999; Sanjuán et al. 2010). That roughly corresponds to one mutation per genome and cell infection cycle. In comparison, it decreases to circa 10^{-2} mutations per genome in the case of DNA viruses and prokaryotes, with the mutation rate per nucleotide ranging from 10^{-6} to 10^{-10}. As we will see later, evolutionary dynamics of RNA viruses is strongly influenced by their high mutation rates.

Viral genomes code for a number of proteins with different functions: forming the capsid, recognising and adhering to the host cell, replicating the viral genome, recruiting the cell machinery required for gene expression, and mediating the assembly of virions, among others. In standard conditions, all proteins are required for a successful infection cycle to be completed. Viable genomes are those able to produce the whole set of functional proteins. In contrast, mutant genomes that produce non-functional proteins and are unable to complete the viral cycle by themselves are termed *defective genomes*. A key aspect of viral biology is that some proteins can be shared by viral genomes inside the same cell—it is said that proteins act in *trans*. As a result of *trans*-acting interactions, some defective genomes can still survive if they are accompanied by viable ones. This phenomenon is called *complementation*.

1.2.2 Quasispecies theory

As already mentioned, classical population genetics deals with a scenario where mutations scarcely appear in the population. It assumes that once a genome mutates there is enough time for the mutation to become lost or fixed before new mutations occur. From a formal point of view, that requires the processes of mutation and fixation to take place at different time scales. On the one hand, the time scale for the occurrence of new mutations is equal to the mutation rate (μ). On the other hand, neutral mutations get fixed in a time scale equal to the inverse of the effective population size[3] (N_e, measured in number of generations). Thus, classical population genetics applies in situations where $\mu N_e \ll 1$. If the error rate is large ($\mu \gg 1/N_e$) new mutations appear before the preceding ones reach fixation. The result is a heterogeneous population comprising multiple, coexisting mutants.

The quasispecies theory of molecular evolution was proposed to explain self-organization and adaptability of primitive RNA or RNA-like genetic elements—also termed replicators or replicons—affected by error-prone template copying, and thus endowed with a huge population diversity (Eigen 1971; Eigen and Schuster 1979). The evolutionary dynamics of a quasispecies is determined by the interplay between mutation and selection. Let us identify replicons with RNA sequences, then mutations occur through errors made in the process of copying already existing sequences. Selection arises because different sequences replicate at different rates—slow-replicating sequences tend to disappear in favor of sequences that replicate faster. The key finding of quasispecies theory is that, in the regime of heterogeneous populations, selection does not lead to the final loss of all but the fastest replicating sequence—termed the *master sequence*. The reason is that mutations affecting the master sequence continuously replenish slower replicating ones. At equilibrium, mutation and selection balance and shape the composition of the population, which usually consists of an heterogeneous repertoire of fast and slow replicating sequences.

One of the main predictions of quasispecies theory is the existence of an *error threshold* above which the population cannot maintain its identity (Eigen 1971; Swetina and Schuster 1982; Schuster and Stadler 1994). The error threshold accounts for a condition where the selective advantage of the master sequence is unable to compensate for its continuous losses via mutations. As the mutation rate crosses the threshold, the master sequence is lost from the population, which wanders in the sequence space as an unstructured cloud of mutants (Fig. 1.4). This situation—termed *error catastrophe*—is equated with the extinction of the quasispecies, understood as the disintegration of its genetic information content.

The initial quasispecies theory made strong assumptions, that conditioned most quasispecies models and the design of many experments based on them. For instance, a one-to-one correspondence between genotype and phenotype was assumed, as well as the absence of compensatory or beneficial mutations. Such assumptions proved inadequate when contrasting quasispecies models with empirical observations on the

[3]The effective population size is a measure of the number of individuals that contribute with their offspring to the next generation

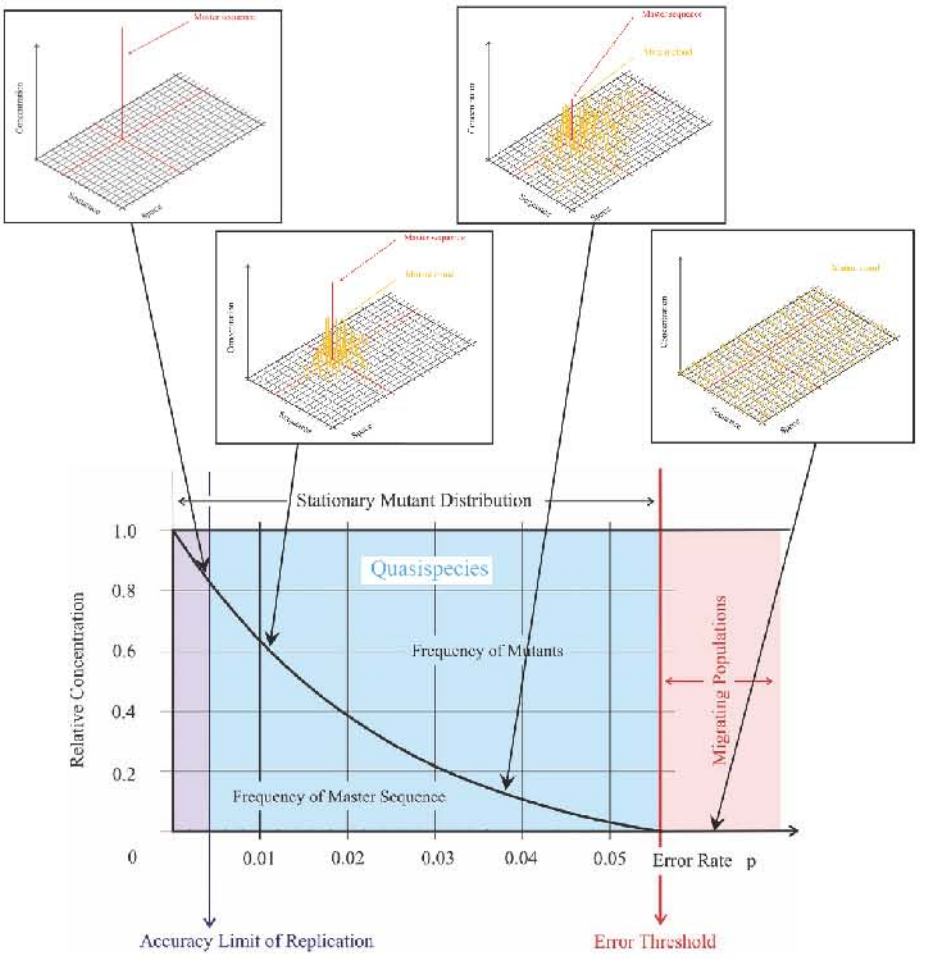


Figure 1.4: *Transition to error catastrophe in quasispecies theory.* The sequence space is represented as a bidimensional grid, with the master sequence corresponding to the position in red. All mutant sequences have the same fitness, which is smaller than the fitness of the master sequence. As the error rate increases, the relative abundance of the master sequence decreases as the population spreads on the sequence space and forms a mutant cloud. Above the error threshold, the master sequence disappears and the mutant cloud dissolves. Modified from Schuster and Stadler (1994) and http://www.tbi.univie.ac.at/ pks.

evolution of RNA viruses. The theory has been subsequently modified to include back mutations and a degenerate genotype-phenotype correspondence, so that models now account for the richer and often counterintuitive behavior of natural viruses (Manrubia et al. 2010). An ever increasing dialogue between experimental and theoretical studies has turned to be essential is fostering these advances.

1.2.3 Evolutionary dynamics of viral quasispecies

Quasispecies theory has exerted great influence in virology because of the observation that RNA viruses replicate in their hosts as complex ensembles of mutant genomes (Domingo et al. 1978). Formation of such mutant distributions—also termed *mutant spectra* or clouds—is fuelled by the high mutation rates of RNA genome replication. Furthermore, the population structure of RNA viruses resembles that of the primitive replicons as postulated by quasispecies theory (Domingo 2006; Lauring and Andino 2010; Domingo et al. 2012). The term *viral quasispecies* was therefore coined to describe the complex structure and organization of viral populations.

The initial quasispecies theory emphasized that the population as a whole, rather than individual genetic elements, was the true target of selection (Eigen and Schuster 1979). Within a quasispecies, genomes are mutationally coupled, such that the abundance of a given variant is not only a function of its replicative ability in isolation but also depends on how often it is produced through mutation of neighbouring genotypes. For many years, the implications of a quasispecies organization were not tested with viruses, fundamentally because of the lack of suitable experimental designs. However it is now well established experimentally and through theoretical models that *interference* and *complementation* occur within mutant spectra and influence viral performance (reviewed in Domingo, Sheldon, and Perales 2012). Interaction among genomes in the viral quasispecies is at the basis of phenomena such as *lethal defection* and genome segmentation, described in chapters 2 and 4, respectively.

RNA viruses count amongst the most plastic organisms on Earth: the huge genetic diversity whithin their quasispecies allows them to adapt to new environmental conditions in relatively short time spans. Viral evolutionary dynamics is thus characterized by fast responses to changes in the selective pressures, which are strongly dependent on the environmental conditions. For instance, production of defective (not necessarily interferring) mutants may either lead to stochastic extinction or be inocuous for the virus, depending on the external constraints (the former in persistent infections, the latter in lytic ones). Even more, if such defective mutants appear in a context of lytic infections at a high *multiplicity of infection*[4] and are capable of complementation, they may give rise to a new version of the virus with a segmented genome. All these cases will be thoroughly analysed in the following chapters. We will conclude that a careful assessment of the selective pressures and environmental constraints acting on the virus is essential when it comes to interpret the evolutionary oucomes observed in the laboratory or in the wild.

[4]The multiplicity of infection (MOI) is the average number of viral particles infecting the same cell, see Chapter 4

The enormous adaptability and fast evolution of RNA viruses is a major obstacle for the design of successful therapeutic strategies able to control their proliferation and propagation (Richman 1996; Domingo et al. 1997). Drugs exert constant selection pressures on viral populations and, as such, the question is not whether a resistant form of the virus will appear, but when it will occur. In the search for novel antiviral strategies, the idea of an error threshold derived from quasispecies theory has suggested that virus-specific mutagenic agents can drive viral populations into extinction. Such a phenomenon, that could be used to control viral infections, is known as *lethal mutagenesis* (Eigen 2002; Biebricher and Eigen 2005; Domingo et al. 2005). In this respect, experiments with a variety of RNA viruses—HIV and hepatitis C, among others—have confirmed that mutageninc drugs inflict an adverse effect on viral production (Loeb et al. 1999; Crotty et al. 2001; Graci and Cameron 2004; Anderson et al. 2004). It must be pointed out, however, that the mechanism by which an excess of mutations leads to the loss of viral infectivity is probably different from that postulated to occur during the transition to error catastrophe in simple replicons. Actually, viral populations include multiple viable variants with different fitness values, and the loss of the fittest—which can be regarded as the equivalent to the loss of superiority of the master sequence—need not imply elimination of other components of the mutant spectrum (Manrubia et al. 2010). Mutagen-induced viral extinction presumably involves diverse pathways that become manifest at different mutagen doses (e.g. competition and interference with defective mutants, observed at moderate mutation rates). A deeper discussion on the mechanisms by which mutagens cause viral extinction is developed in Chapter 2. The implications that derive from the viral quasispecies concept entail great clinical relevance, since they can be applied to viruses such as HIV (Hinkley et al. 2011; Woo and Reifman 2012), influenza (Koelle et al. 2006), hepatitis C (Jardim et al. 2013) and poliovirus (Lauring and Andino 2011)—the causal agent of poliomyelitis.

A remarkable aspect of viral quasispecies is that population and evolutionary dynamics take place together, within a time scale that can be captured in laboratory experiments. Thus, it is possible to see in "real time" how the viral population evolves and adapts to changing environments. A large fraction of this thesis deals, indeed, with the theoretical interpretation of phenomena that have been observed for the first time in experiments of viral evolution. It can be argued that evolutionary theory has been historically quite often decoupled from observation, be it due to a limitation in the data, to technical restrictions or to the interference with previously unknown processes. Our lack of direct experience with natural selection—which usually acts at time scales that we cannot probe—impairs our intuition for evolutionary processes. For this reason, the continuous dialogue between theory and experiment is a must in our way to developing a reliable evolutionary theory.

1.2.4 Evolution of genomes: neutrality or selection?

As already sketched when introducing viruses, genomes show a large structural diversity in different organisms. At the whole genome scale, the number of molecules—chromosomes—that constitute the genome, as well as their size and chemical nature,

vary: eukaryotes possess DNA genomes that are segmented in a set of lineal chromosomes, their total length ranging from 10^7 to 10^{11} base pairs; prokaryotes ususally have a single circular DNA chromosome—although there are exceptions—with a size of 10^6-10^7 base pairs, together with smaller pieces of circular DNA termed plasmids; viruses display the largest diversity in what concerns genomes, which can be made of either DNA of RNA, single- or double-stranded, in a single molecule or segmented into several fragments. Such a diversity may partly be the result of the multiple origins that viruses may have had (Forterre 2006; Koonin et al. 2006; Wessner 2010). In contrast, the various degrees of genome segmentation may reveal different evolutionary strategies adopted by organisms in connection with a necessity of exchanging genes. For instance, the evolutionary implications of a segmented genome in viruses have been largely discussed as an analogy to sexual reproduction, i.e. a mechanism that promotes genetic exchange through segment recombination (Chao 1991). The case of multipartite viruses, presented in Chapter 4, is other example where selective pressures may have played a role in shaping genome structure.

The role of natural selection in determining the size of the genome is a classical problem in the field of genome evolution. The traditional view stated that, provided that small genomes can be replicated at a smaller cost, selection would positively favour genomes as reduced as possible. Two works challenged this view around a decade ago: first, Mira et al. (2001) proposed that genome reduction could be the natural outcome of prokaryotic genomes evolving under a deletional bias, with no need of positive selection favouring it; next, Lynch and Conery (2003) compared prokaryotic and eukaryotic genomes with their characteristic population sizes and came to the conclusion that genome size in prokaryotes is controlled by *purifying selection*, while the larger size of eukaryotic genomes results from the lack of it—i.e. genome complexity in eukaryotes is a non-selected trait.

The structure of genomes also varies at a smaller scale. Genomes are composed of a series of genomic elements—genes, introns, transposons, and so on—and their number and distribution also differs notably among genomes. A current matter of research deals with determining whether genome composition at this level is shaped by natural selection or it is the result of neutral processes (Koonin 2011). A series of universal behaviours in the dependencies and distribution of several genomic features suggests the importance of mathematical and physical processes— not necessarily selection—behind global aspects of genome evolution. In contrast, selection can be detected in some particular cases, such as the frequency of specific genes (Lobkobsky et al. 2013) or the abundance of group II introns in bacteria (Leclercq and Cordaux 1997). We discuss in Chapter 5 the case of *transposable elements*, a class of genomic elements able to move within and between genomes, formerly considered a paradigm of selfish DNA and known to contribute to genome plasticity and evolvability.

Finally, an additional factor that can produce genomic changes is the interaction between parasites and hosts. Some viruses can insert themselves into their host's genome or transport pieces of DNA from one to another host. Furthermore, son e organisms can modify their genomes in response to parasites. This is the case of prokaryotes harbouring the CRISPR-Cas antiviral defense system(Koonin and Makarova 2009). Such

a system incorporates small pieces of viral DNA—spacers—to the host genome, and uses them in order to recognize, target and destroy the virus in future infections (van der Oost et al. 2009). We will explore the influence of parasite-host interactions on prokaryotic genomes via the CRISPR-Cas system in Chapter 6.

1.3 Mathematical modelling of evolutionary processes

There is a great variety of approaches to model biological processes: depending on the desired degree of detail, the objective pursued with the model, the available information and many other factors, different kinds of models can be built. With no aim of making a comprehensive introduction to this topic, we will introduce here some aspects of the models that we will use in this thesis. A key point about them is that we generally intend to understand a phenomenon rather than make accurate prediction. As a result, we will mostly deal with simple, easy to interpret models.

All the models presented in this work are phenomenological models, in the sense that they do not attempt to explain the observed phenomena from its very primary causes—genotypes, molecular interactions, etc—but from a higher-level starting point. Such a starting point depends on the desired degree of detail desired in each particular case. When studying evolution, phenomenological models are interesting because they work directly on the phenotype and do not require an explicit exploration of the genotype-phenotype relation, which is often unknown.

Many of the models discussed in this thesis are also toy models, i.e. very simple models aimed to reproduce an observed phenomenon with as few parameters and processes as possible. Toy models are not only useful as a first approximation to a problem, but also when the interest lies in capturing general features of a system while allowing for an intuitive interpretation of the phenomenon. Thus, these models often teach us that simple mechanisms may be behind apparently complex phenomenology. From a conceptual point of view, simple models are also appealing since they frequently can be analytically solved. Analytical solutions make it easier to explore the parameter space and detect alternative behaviors.

Capturing the subtleties of a phenomenon sometimes requires building more realistic—and complicated—models. For instance, we will see in Chapter 6 how the introduction of host-virus ecological dynamics in a model of antiviral immunity results in an easier explanation of some features observed in nature. More realistic models require a more detailed knowledge of the underlying biological processes, what limits the degree of detail that can be achieved without going into extensive hypothetical assumptions. At odds with toy models, realistic ones must be numerically simulated and their parameter space extensively explored (unless good estimates for the parameters are available *a priori*). It is worth to mention that making a model more complicated does not always result in a richer phenomenology; sometimes, the same qualitative behavior can be observed with simple models, with the advantage that they allow for an easier understanding of the phenomenon.

From a more technical point of view, simple models describing the evolution of a population in time can be expressed in terms of either differential or discrete time equations. Differential equations are appealing since they can be solved by simpler mathematical procedures. However, discrete time equations may be preferable when it comes to compare models with experimental data. Let us show that by taking the example of viral replication, which occurs through discrete replication rounds. Viral replication can be characterized by two key values: the number of genomes produced from each parent genome and the mutation probability per replication round. If replication rounds are taken as the time unit, it is straightforward to identify the parameters of a discrete time model with the aforementioned biological quantities.

We will show in Chapter 3 how a simple model can be used to make predictions on the efficacy of antiviral therapies. To a certain extent, it is surprising and unusual that models with so few parameters can make predictions, and the merit should be probably given to the relative simplicity of the viral system in the experimental conditions assayed. However, a note of caution should be stated about the limitations of predictive models. Mathematical models aimed at yielding specific predictions need to be formulated in conjunction with empirical results and should focus on a single or few observations to improve their predicting power (Manrubia 2012). Moreover, models tailored to a particular population and environment necessarily suffer from restricted applicability and should only be applied to other experimental systems (a different virus, for instance) once the specific features of the new system have been formally taken into account.

1.4 Outline of the thesis

This thesis deals with the mathematical modeling of evolutionary processes that take place in heterogeneous populations. Its leitmotif is the response of complex ensembles of replicating entities to multiple (and often opposite) selection pressures. Even though the specific problems addressed in different chapters belong to different organizational levels—genome, population, and community—all of them can be conceptualized as the evolution of a heterogeneous population—let it be a population of genomic elements, viruses, or prokaryotic hosts and phages—facing a complex environment. As a result, the mathematical tools required for their study are quite similar. In contrast, the strategies that each population has discovered to perpetuate vary according to the different evolutionary challenges and environmental constraints that the population experiences.

Along this thesis, there has been a special interest on connecting theoretical models with experimental results. To that end, most of the work presented here has been motivated either by laboratory findings or by the bioinformatic analysis of sequenced genomes. We strongly believe that such a multidisciplinary approach is necessary in order to improve our knowledge on how evolution works. Moreover, experiments are a must when it comes to propose antiviral strategies based on theoretical predictions, as we do in Chapter 3.

This thesis is structured in two main blocks. The first one focuses on studying instances of viral evolution under the action of mutagenic drugs, paying particular attention to their possible application to the development of novel antiviral therapies. This block comprises chapters 2 and 3; the former dicussing the phenomenon of lethal defection and stochastic viral extinction; the latter dealing with the optimal way to combine mutagens and inhibitors in multidrug antiviral treatments. The second block is devoted to the study of the evolutionary forces underlying genome structure. In chapter 4, we propose a mechanism through which multipartite viruses could have originated. Interestingly, the pathway leading to genome segmentation shares some steps with lethal defection, but each outcome is reached at specific environmental conditions. Chapter 5 analyses the abundance distributions of transposable elements in prokaryotic genomes, with the aim of determining the key processes involved in their spreading. We explicitly explore the hypothesis that transposable elements follow a neutral dynamics, with a negligible fitness cost for their host genomes. A higher level of organization is studied in Chapter 6, where an agent based coevolutionary model based on Lotka-Volterra interactions is used to investigate the evolutionary dynamics of the prokaryotic antiviral immunity system CRISPR-Cas. This chapter also examines the environmental factors that are responsible for its maintenance or loss. Finally, Chapter 7 summarizes the main results obtained along the thesis and sketches possible lines of work based on them.

II

Viral evolution and therapeutical applications

2

Stochastic extinction of viral populations mediated by defectors

2.1 Mutagen induced viral extinction: from lethal mutagenesis to lethal defection

As we saw in the Introduction, populations of RNA viruses evolve at high mutation rates and form heterogeneous groups—quasispecies—that can adapt to environmental changes with relative easiness. The theory of quasispecies (Eigen 1971), a first attempt to formalize the collective behavior of such populations, predicts the existence of an error threshold—a critical value of the mutation rate—above which genetic information can no more be conserved. With the help of mutagenic drugs it would be possible to increase the viral mutation rate and push the virus beyond the error threshold, thus provoking its extinction. This idea is the foundation of lethal mutagenesis, an antiviral strategy that aims to kill the virus by inducing an excess of lethal mutations.

2.1.1 Life at the edge of the error catastrophe?

The feasability of lethal mutagenesis as an antiviral strategy relies on the premise that natural mutation rates are not too far from the critical threshold, so that small doses of a mutagen are capable of producing viral extinction. Quasispecies theory estimates the value of the critical mutation rate as the inverse of the genome length. Interestingly, studies with RNA viruses seem to confirm that natural mutation rates are close to that

value (Drake and Holland 1999). The evolutionary explanation of this fact argues that near the error threshold information is still preserved, while diversity (assumed to determine adaptability) is maximal. Hence, RNA viruses living at the edge of the error catastrophe, it can be expected that little increases in the mutation rate be lethal for the virus.

At odds with those theoretical expectations, experimental observations in viruses (Crotty et al. 2001) and ribozymes (Kun et al. 2005) reveal that they can withstand mutation rates 3 to 8 times above their natural ones and still maintain their viability. Such a disagreement may arise from the fact that the original quasispecies theory focuses on the genotype—the genetic sequence—while the survival of the virus depends on the phenotype—the true object of selection. There is a very large number of potencial genotypes expressing the same phenotype, therefore not every mutation affects the phenotype. In consequence, the critical mutation rate takes different values at the genotypic and phenotypic levels. An example of the large genotype-phenotype degeneracy is provided by the average number M_n of RNA sequences—genotypes—of length n whose folded state is compatible with a given secondary structure—a proxy for their phenotype: $M_n = 1.402n^{3/2}1.748^n$ (Stein and Waterman 1978). More realistic models of quasispecies distinguishing between genotype and phenotype predict an error threshold at mutation rates several-fold higher than those expected only from considerations on the genotype (Takeuchi et al. 2005; Saakian and Hu 2006). This hints at the possibility that natural quasispecies are not that close to the error threshold. It has been suggested, instead, that the natural mutation rate might result from a process that minimizes adaptability time, the latter emerging from a compromise between minimizing search time in the genome space –this occurs at high mutation rates—and obtaining a rapid fixation of advantageous mutants—this takes place at low mutation rates (Stich et al. 2007). As viruses do not live at the edge of the error catastrophe, lethal mutagenesis requires high doses of mutagenic drugs.

2.1.2 Lethal mutagenesis: Error catastrophe or mutational meltdown?

Though increased mutagenesis is a robust experimental way to produce the extinction of a viral population, a current matter of concern is whether extinction truly occurs through crossing an error threshold (Bull et al. 2005; Wilke 2005; Bull et al. 2007; Takeuchi and Hogeweg 2007; Manrubia et al. 2010), as postulated by classical quasispecies theory. A different form of extinction is *mutational meltdown*, where all genotypes in the quasispecies disappear simultaneously. At present, this seems to be the mechanism that better describes experimental observations of viral extinction under a strong increase in the mutation rate.

Let us illustrate those two extinction mechanisms with a simple example. Consider a quasispecies formed by two phenotypes characterized by replicating at rates $\sigma_1 = \sigma > 1$ and $\sigma_2 = 1$. At time t, each type is represented by $\nu_1(t)$ and $\nu_2(t)$ individuals,

whose abundances evolve according to

$$\begin{aligned} \nu_1(t+1) &= \sigma(1-\mu)\nu_1(t) + \mu'\nu_2(t) \\ \nu_2(t+1) &= (1-\mu')\nu_2(t) + \sigma\mu\nu_1(t) - \pi\nu_2(t)\,. \end{aligned} \tag{2.1}$$

Faster replicators mutate to lower replicators at a rate μ, while backward or compensatory mutations occur at a rate $\mu' < \mu$. Slower replicators can be hit by lethal mutations at a rate π. The error threshold is by definition the point where the high-fitness class is lost from the population while low-fitness classes are maintained. According to the model in Eq. (2.1), this would happen when $\nu_1(t \to \infty) = \nu_1^*$ becomes zero, while $\nu_2^* \neq 0$. It can be easily shown that both populations maintain positive values for any $\mu' \neq 0$, irrespectively of the initial condition: There is no extinction threshold if the class of fast replicators can be regenerated by the slower class. For $\mu' = 0$, the extinction threshold occurs at $\sigma(1-\mu) \leq (1-\pi)$. Note that it always exists in the absence of lethal mutations while, for $\pi \geq \mu$, σ cannot simultaneously fulfill the previous inequality and be larger than one. The most relevant result is that the presence of backward mutations, unavoidable in any model describing the phenotype (Manrubia et al. 2003; Takeuchi et al. 2005), precludes the existence of an error threshold.

The mutational meltdown takes place when the population cannot replicate fast enough to sustain itself, and both classes disappear simultaneously. Mathematically, this happens when the largest eigenvalue of the matrix describing the dynamics of the system becomes smaller than one, meaning that the average number of offspring is less than one per parent individual. In the previous example, its value can be exactly calculated. To first order in μ', mutational meltdown occurs when

$$\sigma(1-\mu) + \frac{\mu\sigma}{\sigma+\pi-1-\mu\sigma}\mu' < 1\,. \tag{2.2}$$

At odds with extinction through an error threshold, mutational meltdown seems to be a generic mechanism through which viral populations can undergo extinction.

2.1.3 From lethal mutagenesis to lethal defection

There might be still other mechanisms behind the loss of viability of viral populations. For instance, it has been reported that mild increases in the mutation rate—smaller than those required for lethal mutagenesis—can cause a loss of infectivity in viruses under certain circumstantces. Grande-Pérez et al. (2005)performed experiments during which persistent infections of lymphocytic choriomeningitis virus (LCMV) were treated with a small amount of mutagen. Though the replicative ability of the virus inside cells was not affected, the virus eventually lost its ability to produce infective particles. As a result, the infection could no more propagate to new cells and came to an end (see Fig. 2.1). The fact that this phenomenon was observed in persistent infections is not accesory, but it provides some clues about the underlying mechanisms. In a persistent infection, intracellular selection for higher replication rates is acting all the time. However, viral particles are slowly released to the external medium and the

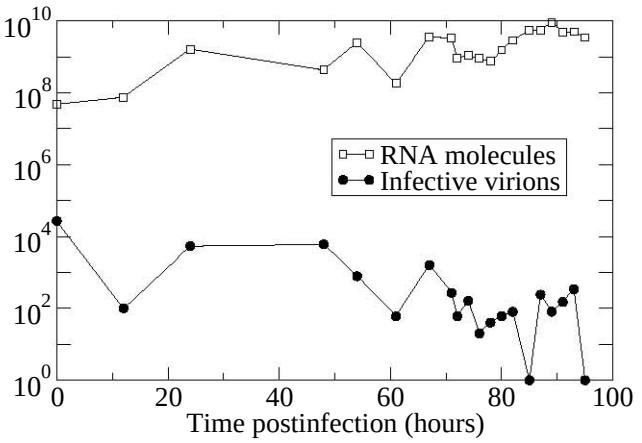


Figure 2.1: *Experimental observations support lethal defection.* Absolute number of RNA molecules and infective virions inside cells persistently infected with LCMV and treated with 100μg/ml of the mutagen 5-fluorouracil. After about 90 hours of infection, the cells cease to produce infective virions, though RNA replication is not impaired (Grande-Pérez et al. 2005).

ability of the virus to infect new cells is not selected for. This situation is remarkably different from lytic infections, where release of viral particles to the medium is a fast step that implies breaking the cell. In the latter case, selection for replicative ability is naturally coupled to selection for infectivity.

The extinction of infectivity in persistent infections of LCMV cannot be understood in the framework of present quasispecies theory. It requires, instead, to explicitly consider infectivity and replicative ability as two separate (decoupled) traits contributing to phenotype. The molecular mechanisms behind the decoupling of traits can be summarized as follows. The replicative ability of a genome is related to its capacity to bind to and be copied by replication enzymes (polymerases). This depends strictly on the sequence of the genome. Unavoidable mutations in the copy process may affect the binding and copying of a sequence. On the other hand, infectivity depends on the performance of proteins codified by that same genome. Hence, changes in infective ability are conditional on the genome experiencing a mutation, though not all mutations have an effect in proteins, and thus only a fraction of those will affect infectivity. The key issue is that, in a persistent infection, replicative ability and infectivity evolve under different selective pressures: genomes able to replicate compete inside each cell, while infectivity behaves as a neutral trait. A neutral trait, by definition, is not useful in the current environment and thus can accumulate random mutations. Those mutations may result in a loss of viability in the long run. It was conjectured (Grande-Pérez et al. 2005) that the role of the mutagen is to enhance the appearance of a class of defective mutants, able to replicate but unable to infect susceptible cells. This parasitic

subclass eventually invades the cell and induces the extinction of the whole population. Hence, this new pathway to viral extinction, that takes place at low mutagen doses and is mediated by defective mutants, was termed *lethal defection*.

2.1.4 Towards a more realistic model of persistent viral infections

A step forward towards modelling real systems is to consider that phenotype is a multi-trait feature that can be only rarely reduced to a single variable. Actually, there are abundant examples in the literature where two phenotypic traits need to be considered in order to appropriately describe the evolution and adaptation of heterogeneous populations. Among them, growth rate and yield (Novak et al. 2006), robustness and evolvability (Lenski et al. 2006; Aguirre et al. 2009), or virulence and replicative ability in competition assays (Herrera et al. 2007) have been pondered as characteristics that simultaneously affect the survival ability of a viral population. Motivated by the experiment of extinction of infectivity in LCMV, we here introduce a model for the evolution of a population whose individuals are characterized by two traits subject to positive and neutral selective pressures, respectively.

2.2 Model of a quasispecies with a two-traits phenotype

We consider a quasispecies formed by four different classes. Fast replicators have an average of R offspring per replication cycle; slow replicators have r. Either type can take a viable or a defective form. We assume that viable forms maintain the integrity of their genomes and correctly code for the proteins that permit replication and infection. Thus, replication of either type is only possible if individuals of the viable type are present. Defective forms can only replicate in the presence of an individual of the viable type. The four types and the corresponding transition rates are depicted in Fig. 2.2. The replicative ability decreases (increases) with probability p (q). Changes in this trait fix new mutations that can affect viability. With probability w, an individual mutating to the class of slow replicators can simultaneously lose its viability; with the same probability, viability is recovered conditional on experiencing a mutation increasing the replicative ability. The model includes lethal mutations with rate p affecting slowly replicating individuals. The rates p, q, and w, actually stem from a microscopic mutation rate characteristic of each virus. They can be treated as constant on the average for a given genome (i.e. population, species, or organism). Dynamics proceeds through discrete generations and the population size N is constant. The matrix M characterizing the mean-field dynamics of the system reads

$$\mathsf{M} = \begin{pmatrix} R(1-p) & q & 0 & qw \\ Rp(1-w) & 1-p-q & 0 & 0 \\ 0 & 0 & R(1-p) & q(1-w) \\ Rpw & 0 & Rp & 1-p-q \end{pmatrix} \tag{2.3}$$

We set $r = 1$ without loss of generality, thus fixing the time scale. Let $\mathbf{n}(g) = \{n^V(g), n^v(g), n^D(g), n^d(g)\}$ be the vector describing the evolution of the fraction of

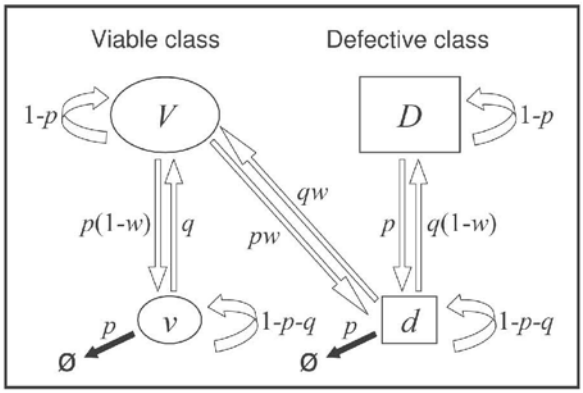


Figure 2.2: *Schematic representation of the four types forming the quasispecies.* Permitted transitions between types are indicated as arrows, with their corresponding rates.

individuals in each type. The composition of the population evolves according to the following equation:

$$\mathbf{n}(g+1) = \alpha(g)\mathsf{M}\mathbf{n}(g)/[\alpha(g)\lambda]\,, \tag{2.4}$$

where $\alpha(g) = n^V(g)+n^v(g)$ is the fraction of viable individuals at generation g, and λ is the largest eigenvalue of M. As initial condition we assume $\mathbf{n}(0) = \{1,0,0,0\}$. Note that though $\alpha(g)$ actually does not affect the average composition of the population, it is the cause of extinction, since disappearance of the viable types means disappearance of the whole population.

Assuming that viability is a neutral trait implies that cell-to-cell transmission is not represented in the model. Hence, Eq. (2.4) describes intracellular dynamics, with a typical time scale shorter than that of transmission of the infection. The size of the system thus corresponds to the number of viral genomes inside a single cell.

This model has an explicit solution $\mathbf{n}^* = \mathbf{n}(g\to\infty) \equiv \{n^V, n^v, n^D, n^d\}$,

$$\mathbf{n}^* = \mathcal{N}^{-1}\left\{2q, (1-w)(c-a_+), 2q(1-w), c-a_+\right\} \tag{2.5}$$

with $\mathcal{N} = (2-w)(c-a_-)$, $a_\pm = (R-1)(1-p) \pm q$, $c = [(1-p)^2(R-1)^2 + 2(R(1+p)-(1-p))q+q^2]^{1/2}$, and $\lambda = 1/2[(R+1)(1-p)-q+c]$. As with the example discussed in the introduction, no extinction threshold is found for $q \neq 0$, that is, when backward mutations exist. Mutational meltdown is possible and holds for an asymptotic growth rate at the mutation-selection equilibrium below one, that is $\lambda < 1$.

The solution given in Eq. (2.5) represents well the dynamics of the quasispecies only for sufficiently large populations. For small population sizes the dynamics are qualitatively different and dominated by the intermittent appearance of class D. In this regime, stochastic extinction is a common event.

There are different limits of the model worth mentioning. The case $w = 0$ corresponds to a quasispecies described only by its replicative ability where back and lethal mutations are considered (only classes V and v sustain finite populations). The case $w = 1$ is formally identical, with classes V and d surviving. This model has been analyzed for instance in (Bull et al. 2005). The case $q = 0$ is particularly interesting. Since class D can only be generated through (rare) beneficial mutations appearing in class d, class D can remain empty for extended periods of time when the population size is small enough. Thus, in the biologically relevant limit of small N and $q \to 0$, the case $q = 0$ should approximate accurately the intervals where $n^D(g) = 0$. The stationary populations $\mathbf{n}_0^* = \{n_0^V, n_0^v, n_0^D, n_0^d\}$ in this limit are

$$\mathbf{n}_0^* = \mathcal{N}_0^{-1} \left\{ \mathcal{N}_0 - Rp, Rp(1-w), 0, Rpw \right\} , \tag{2.6}$$

with $\mathcal{N}_0 = R + p - 1$ and $\lambda_0 = R(1-p)$.

In order to check the accuracy of our analytical results, we have performed numerical simulations of the dynamical model. As initial condition, we take N individuals in class V. At each generation g, the population replicates deterministically (with rates R and 1) to generate the individuals at generation $g+1$, which then mutate according to the probabilities described. This step introduces stochasticity in the system. If the population $n(g+1) > N$, a random subset of N individuals is selected. This keeps the population size bounded. When, as a result of fluctuations, the number of viable individuals reaches zero, the population is considered extinct and the simulation halts.

2.3 Results

The different dynamical regimes of the population are illustrated in Fig. 2.3. As the size of the population N increases, the behaviour changes from a stochastic regime dominated by the intermittent appearance of class D and with average values well described by Eq. (2.6) to a mean-field regime with average values following Eq. (2.5). Though the transition is smooth, it will be shown that there exists a characteristic system size N_m where the stochastic regime crosses-over to the mean-field regime.

2.3.1 Stochastic regime

For small q and finite system size, the population of defective, fast replicating individuals appears in bursts that either are terminated after a finite number of generations or (also in finite time) invade the population, thus causing its extinction. In this limit, the probability p_0 that class d does not produce any individual of class D in one generation is $p_0 \simeq (1 - q(1-w)/\lambda)^{N n_0^d}$. Hence, the probability $P_0(g) = p_0^g$ of having an interval of g generations without individuals of class D reads

$$P_0(g) \simeq \exp\left\{ -gN \frac{Rpw}{R+p-1} \ln\left(1 - \frac{q(1-w)}{R(1-p)}\right) \right\} . \tag{2.7}$$

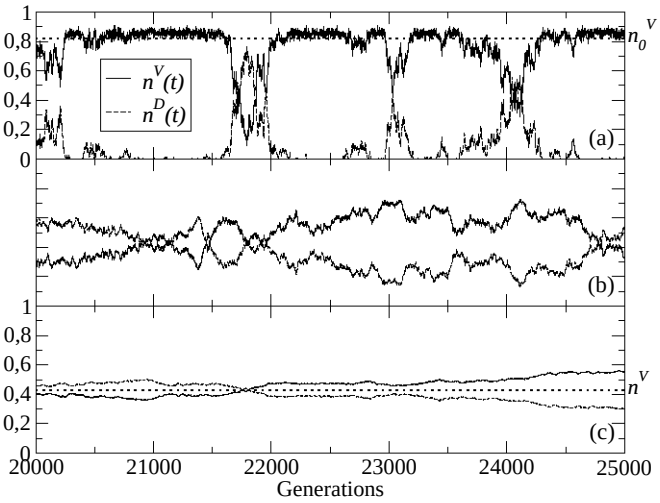


Figure 2.3: *Dynamical regimes of the model.* (a) Stochastic regime, $N = 200$. For small system sizes, the dynamics are dominated by the intermittent appearance of class D. The dotted line corresponds to the value obtained in the approximation $q = 0$, $n_0^V \simeq 0.8182$. (b) Transition regime, $N = 2500$. For system sizes $N \simeq N_m$, the population of D individuals is always above zero, though fluctuations are still large. (c) Mean-field regime, $N = 10^5$. For $N \to \infty$, the population in each class approaches the mean-field value. The dotted line corresponds to the asymptotic solution, $n^V \simeq 0.4317$. Parameters for all simulations are $p = 0.1$, $q = 0.01$, $w = 0.1$, $R = 2$, which yield $N_m \simeq 2116$ (see below).

The exponent of this distribution, $\ln p_0$, is represented in Fig. 2.4(a) together with the results of numerical simulations. The outbreaks of the D class for sufficiently small p and q start with a single individual and follow the dynamics of a branching process with branching ratio m. To a first approximation, the value of m is the average number of offspring of class D per individual in that class ($\approx R(1-p)$) divided by the asymptotic growth rate of the population λ_0. Hence, in this limit, where the contribution from class d is neglected, $m = 1$ and the dynamics follows a critical branching process (Harris 1963). The corresponding generating function, $f_1(s) = \mathrm{e}^{s-1}$, allows to obtain a number of exact results. The probability of termination of the outbreak at any time in the future is the solution of $f_1(s^*) = s^*$, which has the known result $s^* = 1$. The probability of termination after g^b generations is $P(g^b) = f_{g^b}(0) - f_{g^b-1}(0)$, where $f_k(s) = f[f_{k-1}(s)]$. It can be iteratively obtained and, asymptotically, $P(g^b) \propto (g^b)^{-2}$. This function is compared with numerical simulations in Fig. 2.4(b). The existence of a neutral trait and the critical branching dynamics of the defective class are two sides of the same coin: Any coupling between

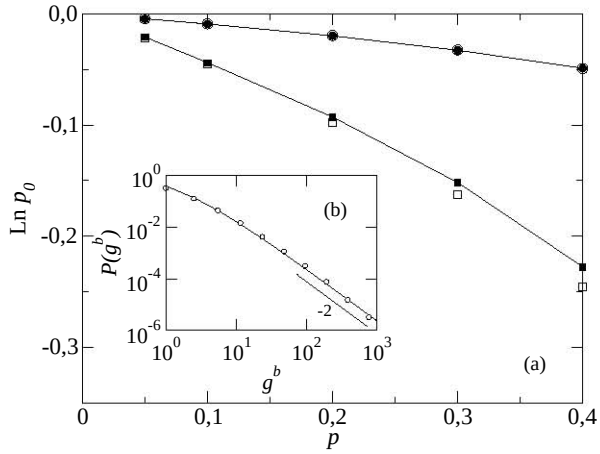


Figure 2.4: *Numerical and analytical results for the model in the stochastic regime.* (a) Exponent of the distribution of interval lengths (in number of generations) without individuals of the D class. Two values of $q = 0.01$ (circles) and $q = 0.05$ (squares) are shown for $N = 100$, $R = 4$ and $w = 0.5$. Solid symbols are results from simulations; open symbols are analytical results as in Eq. (2.7). (b) Probability $P(g^b)$ of a $D-$outbreak of length g^b generations. The solid line is the prediction of critical branching processes; circles correspond to numerical simulations for $p = 0.1$, $q = 0.01$, $w = 0.2$, $R = 2$, and $N = 100$.

traits would imply deviations from neutral behaviour and values of the branching ratio different from one.

2.3.2 Mean-field regime

As the size of the system increases, so does the duration of the outbreaks. At some system size N_m, the previously isolated bursts merge, and the approximation of the dynamics of D as a critical branching process is no longer valid. For $N > N_m$ all types are continuously represented in the quasispecies, albeit fluctuations in population sizes might still be large. The system size N_m can be estimated as the value of N where class d contributes on average one individual per time step to class D, $N_m n^d q(1-w)/\lambda \simeq 1$, which yields

$$N_m \simeq \frac{(R+p-1)}{q}\frac{(2-w)}{(1-w)}\frac{(R(1+p)+q-1-R-c)}{(R(1+p)+q-1+R-c)}, \tag{2.8}$$

using the value of n^d in Eq. (2.5) and the corresponding λ. Series expansion of N_m in powers of q yields

$$N_m = \frac{2-w}{1-w}\left[\frac{1-p}{q} + \frac{R}{(1-p)(R-1)} + O(q)\right], \tag{2.9}$$

so N_m diverges as $q \to 0$. Hence, for finite p, w, and R, in situations where the probability of hitting beneficial mutations is small enough, the dynamics is systematically dominated by population fluctuations. The system size N_m separates the two relevant dynamical regimes. Below N_m, the dynamics is mostly determined by stochastic effects and well described by the solution $q = 0$ plus the probabilistic appearance of critical $D-$bursts: extinction is common. Above N_m, the dynamics is well described by the mean field asymptotic solution. As N grows, extinction becomes increasingly unlikely.

The transition between the stochastic and the mean-field regimes can be further characterized through the distribution of probability densities for each of the four subpopulations. In the stochastic regime, the abundances of viable and defective types proceed in anti-phase, such that when the population of $V + v$ is high that of $D + d$ is low (as in Fig. 2.3(a)). In this case the average population values agree with Eq. (2.6): the distributions of V, v and d present a maximum near those values and the abundance of D is close to zero. When outbreaks of D appear, the populations of V and v decrease strongly while population D becomes abundant. Extinction supervenes if the number of $V + v$ attains zero, $\alpha(g) = 0$ in Eq. (2.4). In the mean-field regime, the size of the system is large enough to sustain finite populations of all four classes. The maxima of the population size distributions move towards the average values predicted by Eq. (2.5). In Fig. 2.5 we plot the main quantities characterizing the transition.

The average time to extinction T_{ext} grows exponentially with the system size, $T_{ext} \propto \exp\{kN\}$, with k depending on the model parameters. For the case shown in Fig. 2.5, $k = 0.0054(1)$, so T_{ext} increases more than a thousand-fold between $N = 100$ and $N = 10^3$. We do not have evidence that $T_{ext} \to \infty$ at finite N, though its rapidly increasing value asserts that, in practice, extinction will be rarely observed once in the mean-field regime.

2.4 Discussion

Fitness is a multitrait feature with different expression in different environments. In lytic infections, where cells are killed after a number of replication cycles, the requirement to maintain the ability of infecting susceptible cells acts as a purifying selection pressure that regularly removes non-infective particles from the population. When infections are persistent, selection pressure over infectivity is released. Since the number of viral particles inside cells is relatively small (about 10^{2-3}), population fluctuations are large, and, in the presence of one trait not subject to selection, a defective subpopulation able to induce the extinction of the whole might appear. Stochastic extinction through lethal defection (Grande-Pérez et al. 2005) becomes possible. From another viewpoint, stochastic extinction occurs only if the characteristic time between infections of susceptible cells is larger than the time to extinction T_{ext}. Every new infection

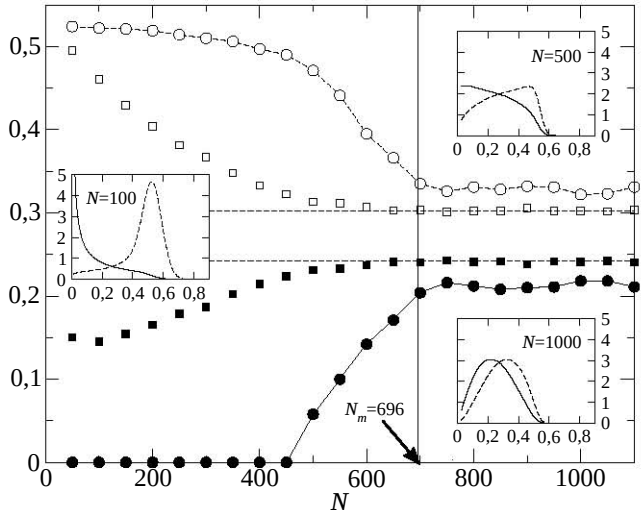


Figure 2.5: *Transition from stochastic to mean-field regimes.* We show the position of the maxima for the distribution of populations D (solid circles) and V (open circles). Solid squares correspond to the average value of $n^D(g)$, open squares to that of $n^V(g)$. From Eq. (2.6), $n_0^V = 0.53$, $n_0^D = 0$ (in agreement with the maximal values at the stochastic regime), while dashed lines signal asymptotic mean-field values obtained from Eq. (2.5). The maxima of the distributions eventually converge to those values as fluctuations disappear in the limit $N \to \infty$. The insets show three representative probability distributions for $n^D(g)$ (solid line) and $n^V(g)$ (dashed line), below, during, and above the transition. The estimated system size separating stochastic from deterministic behavior is $N_m \simeq 696$ according to Eq. (2.8). Parameters are $p = 0.3$, $q = 0.01$, $w = 0.2$, $R = 2$.

event acts as a filter cleaning the population from defectors, unable to infect, and thus resetting the dynamics to the initial condition. This mechanism can be generalized to situations where a previously essential trait is temporarily unneeded (not selected for) and then becomes essential again. This could be the case of genes that respond to uncommon environmental conditions or get rarely switched on: the absence of activity could lead to the loss of viability. The molecular processes that turn a gene into nonfunctional include deleterious mutations and insertion of transposable elements (TE). In the latter case and according to our results in Chapter 5, gene interruption is reversible; as a result the population can be rescued if any reversion occurs before the gene becomes essential. Interestingly, TE increase their mobility under stress conditions, which could be a strategy to recover the functionality of interrupted genes and avoid lethal defection-like extinction (McGraw and Brookfield 2006).

The model here presented shows how simple evolutionary mechanisms can cause the extinction of populations of fast mutating pathogens under environmental changes,

and strongly suggests that one could devise strategies to take advantage of those mechanisms in fighting viral infections. In this context, tuning the balance among intracellular replication, frequency of infection of new cells and multiplicity of infection, or applying mild increases in viral mutation rate, appear as therapies alternative to the massive use of drugs. In a broader framework, a better understanding of the complex population dynamics typical of these organisms should make possible to identify and manage selection pressures over target traits, resulting in the development of new control strategies at the host level for infectious diseases.

3

Tempo and mode of multidrug antiviral therapies

3.1 Inhibitors, mutagens, and multidrug antiviral therapies

RNA viruses are an iconic example of populations persistently escaping the action of antiviral drugs (Richman 1996; Domingo et al. 1997). As explained in the previous chapter, alternative therapies are continuously sought with the aim of impeding the appearance and fixation of resistance mutants (Endy and Yin 2000; Domingo et al. 2008; Vignuzzi et al. 2008). In order to succeed, such therapies must take into account the evolutionary properties of the viral population, for instance, by inducing opposite selective pressures that complicate viral adaptation. One example of how selective pressures can be managed in order to facilitate viral extinction has already been discussed in the context of lethal defection. In this chapter we explore an alternative strategy, namely the use of multiple drugs—possibly with different action mechanisms—in the same treatment.

Multidrug treatments have been established as an efficient way of delaying the appearance of resistant forms. Such treatments take advantage of the combination of two or more antiviral drugs, that are administered in a concrete manner—tempo and mode—in order to maximize their efficiency. However, in order to understand the mode of action of combined treatments it is urgent to clarify the degree of interaction between the different drugs used and the quantitative impact they have on the virus they are affecting. The joint action of two drugs can rarely be reduced to the simple addition of their independent effects (Torella et al. 2010), and the same drug might elicit dif-

ferent responses in different viral systems, including the appearance of compensatory mutations able to induce resistance (Handel et al. 2006). Thus, the complete characterization of a multidrug treatment applied to a particular viral system might require in the last term a systematic assay with a large number of drug doses delivered under different administration protocols. A full characterization of the response of the viral system *in vitro* seems prior to any *in vivo* assay of the treatment. This procedure demands a costly amount of time and resources that can be significantly reduced through the guide offered by formal approaches to therapies.

Knowledge derived from *in vitro* studies of the response of pathogens to the action of one or a number of drugs might be combined with the information yielded by well-designed mathematical models involving the relevant mechanisms of action of and interaction among the drugs (Fitzgerald et al. 2006; Yeh et al. 2006), virus-host interactions (Bonhoeffer et al. 1997), or the role of the immune response (Komarova et al. 2003). An important feature, rarely taken into account, is the intrinsic replicative ability of the pathogen. Many different ways to encode genomic information (double- and single-stranded RNA and different polarities) and a repertoire of replicating strategies have been selected by different RNA viruses. Hence, the knowledge gained for a particular virus may not be extrapolable to other systems, though suitable modifications of dynamical models can likely account for these different replication modes (Sardanyés et al. 2009; Loverdo et al. 2012).

The dynamics of the picornavirus foot-and-mouth disease virus (FMDV) have been explored under different experimental regimes with the aim of disclosing protocols able to cause its extinction during replication in cell culture. It has been demonstrated that a combination of mutagenic agents and antiviral inhibitors is an efficient way to drive FMDV to extinction (Pariente et al. 2003; Pariente et al. 2005). The mutagen succeeds in raising the fraction of defective and lethal mutants in the population, thus decreasing its overall fitness (see section 2.1). Defectors may also interfere with infectivity as shown with specific mutants (Perales et al. 2007) and preextinction viral populations (González-López et al. 2004), and thus accelerate extinction. For its part, the inhibitor contributes to extinction by reducing the viral load[1] (Fig. 3.1).

However, the joint effect of a mutagen and an inhibitor cannot be reduced to the addition of their individual effects (Fitzgerald et al. 2006). Since the dawn of quasispecies theory (Eigen and Schuster 1979), the use of mutagens has been proposed as a plausible strategy to induce viral extinction (Eigen 2002; Domingo et al. 2005). A significant increase in the mutation rate has been, indeed, successful to cause the extinction of infectivity in many different viral systems (Lee et al. 1997; Crotty et al. 2001), though the mechanisms through which extinction supervenes are diverse and related to a variety of molecular and population responses (Manrubia, Domingo, and Lázaro 2010, see also section 2.1). In particular, increased mutagenesis can also bear beneficial effects for viral populations through an enhancement of their diversity, which promotes adaptation of low-fitness viruses (Cases-González et al. 2008) and facilitates

[1] Viral inhibitors are drugs that interfere with the viral cycle and impair the replicative ability of the virus.

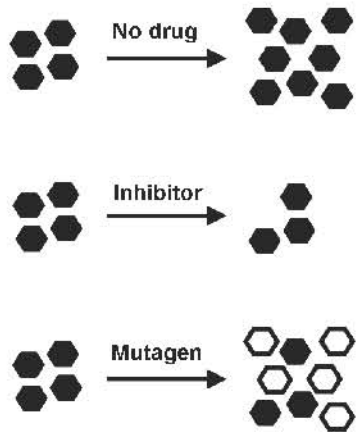


Figure 3.1: *Schematic of the effect of inhibitors and mutagens on viral growth.* In the absence of drugs (top), viral growth is faster than degradation, so that the number of viral particles increases exponentially. Inhibitors (middle) interfere with viral proliferation, which results in a reduction in the viral population. In turn, mutagens (bottom) increase the fraction of viral offspring that is non-viable (represented in white), thus reducing in practice the infective population.

the appearance of resistance mutants when an inhibitor of viral replication is present (Perales et al. 2009).

There is a well stablished result in clinical practice, according to which the optimal way of combining multiple inhibitors –with no mutagens involved– is dispensing them simultaneously (Bonhoeffer et al. 1997; Ribeiro and Bonhoeffer 2000). In contrast, the increased likelihood of developing resistance to the inhibitor when dispensed simultaneously with a mutagen advocates the sequential administration of the two drugs, with the inhibitor followed by the mutagen. In a first study with FMDV subjected to the action of the mutagen ribavirin (R) and the inhibitor of viral replication guanidinium hydrochloride (GU), it was shown that under the experimental conditions assayed the sequential therapy performs better than the simultaneous administration of both drugs (Perales et al. 2009). This is an interesting result that immediately queries the range of applicability and relative success of combined or sequential administration of those two dissimilar drugs. In this chapter, we face the characterization of the response of a general viral system to the action of an inhibitor of its replication and a mutagenic agent through a model for the dynamics of two different viral classes in the viral population. We obtain several exact results that predict the preferred therapy as a function of intrinsic viral characteristics and of the administered drug doses. Our predictions are tested and confirmed by new experiments with FMDV at different doses of inhibitor.

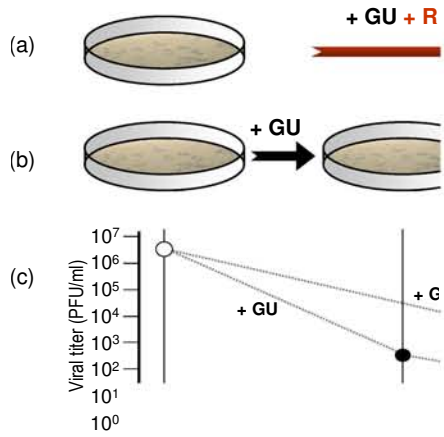


Figure 3.2: *Schematic outline of the experimental protocol in cell culture.* (a) Combined therapy. A dose d_{GU} of guanidine and a dose d_R of ribavirin are simultaneously added to the initial population. The viral titer $Y_T^{\mathcal{C}}$ is calculated after a single passage. (b) Sequential therapy. A dose d_{GU} is added to the initial population. After 24 hours, the inhibitor is removed and a dose d_R of ribavirin is added. The viral titer $Y_T^{\mathcal{S}}$ is obtained after this second passage. (c) Experimental results in a case example. The initial population has a titer of 3.7×10^6 PFU/ml (PFU—plaque forming units—is a measure of the number of viral particles able to infect cells). The applied doses are $d_{GU} = 18$ mM and $d_R = 5$ mM. With these drug doses, the viral titers after applying the protocols described in (a) and (b) are $Y_T^{\mathcal{C}} = 2 \times 10^2$ PFU/ml and $Y_T^{\mathcal{S}} = 10$ PFU/ml, respectively. Passages are performed and quantified as described in Appendix A.

3.2 Mathematical model

The present model is intended to reproduce the dynamics of a viral population after the addition of mutagens and/or replication inhibitors following the protocol represented in Fig. 3.2. Mutagens mostly promote the appearance of deleterious variants, though some particular mutations may however confer resistance to the inhibitor, thus favouring its survival. In order to capture this double role of the mutagen, the model considers that the dynamics is dominated by two types of individuals: viable susceptible to the inhibitor (v) and viable resistant to the inhibitor (V). Viable individuals are able to infect cells and replicate by themselves.

Let w_0 be the rate at which viable individuals produce, under replication, a non-viable class including lethal variants, defective interfering forms, or any other mutant unable to complete an infection cycle on its own [$(1 - w_0)$ is the intrinsic copying fidelity of the virus]. Under low multiplicity of infection, only viable individuals can infect new cells. Independently of the size of the population of non-viable mutants, they are unable to produce infection in the next passage due to the low MOI (from 10^{-5}

to 10^{-1} PFU/cell, see Appendix A). This is the reason why the dynamic equations for non-viable types are not explicitly considered. Resistant forms appear at a rate $\mu_0 = kw_0$. Since mutations providing resistance to the inhibitor are rare compared to deleterious ones, $k \ll 1$. Addition of a mutagen is implemented by increasing the mutation rate from its natural value w_0 to a higher $w > w_0$. As a consequence, the rate of appearance of resistant mutants, $\mu = kw$, increases when the mutagen is added. Each time a viral genome replicates inside the cell m copies are produced, so m is the replicative ability per genome and replication cycle. The effect of an inhibitor is to slow down replication of the susceptible type by multiplying parameter m by a factor $0 < i \leq 1$.

For the sake of simplicity, back mutations are not considered in the model: they would represent a small contribution to the population numbers of type v in the form of an additive term of order $m\mu$ and do not significantly affect population dynamics. This is so because, initially, only susceptible individuals are present. Note that a term of the same order is however required to trigger the growth of population V. Once resistant individuals have appeared, their dynamics are dominated by the term of order $(1-w)m$, in front of which the first term could be then discarded.

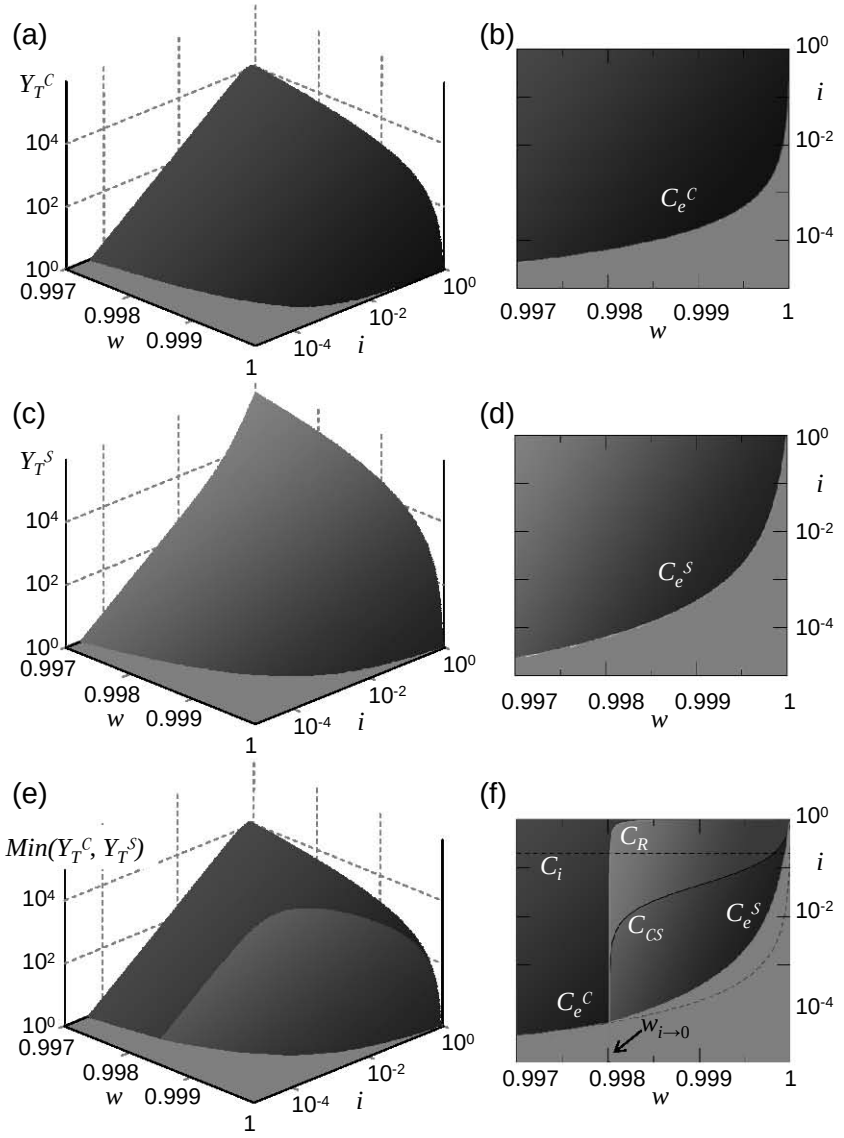
Let $v(g)$ and $V(g)$ denote the number of individuals of each type at replication cycle g. Populations after one replication cycle in the presence of a mutagen and an inhibitor will be

$$\begin{aligned} v(g+1) &= i(1-\mu-w)mv(g) \\ V(g+1) &= i\mu mv(g) + (1-w)mV(g) \end{aligned} \tag{3.1}$$

To study the dynamics of the model and to compare the sequential versus the combined administration of the drugs, we will mimic the protocol described in Fig. 3.2. At cycle 0 (initial condition) we will assume that the population is formed by S_0 individuals of type v, which due to mutations will populate the class or resistant mutants in successive replication cycles. A full solution of the model is given in Appendix A. The dynamics of the combined therapy correspond to applying equation (3.1) iteratively for $g = 1, \ldots, G$ cycles with parameters $i < 1$ and $w > w_0$; the dynamics of the sequential therapy imply replication for G cycles in absence of the mutagen ($i < 1$ and $w = w_0$) and use of the so generated population as initial condition to replicate for G additional cycles in absence of the inhibitor and presence of the mutagen ($i = 1$, $w > w_0$).

3.3 Model analysis

Viral particles are released from the cell after G replication cycles. This finishes the process for the combined therapy, while there is a second passage to undergo in the sequential therapy. The size of the infective population –let us call it $Y_T(G, w, i)$ to make explicit the experimental parameters involved– is obtained by adding the popu-


(a)
$Y_T{}^C$
10^4
10^2
10^0
0.997
0.998
w
0.999
1
10^0
10^{-2}
10^{-4}
i
(b)
10^0
i
10^{-2}
10^{-4}
$C_e{}^C$
0.997
0.998
0.999
1
w
(c)
$Y_T{}^S$
10^4
10^2
10^0
0.997
0.998
w
0.999
1
10^0
10^{-2}
10^{-4}
i
(d)
10^0
i
10^{-2}
10^{-4}
$C_e{}^S$
0.997
0.998
0.999
1
w
(e)
$Min(Y_T{}^C, Y_T{}^S)$
10^4
10^2
10^0
0.997
0.998
w
0.999
1
10^0
10^{-2}
10^{-4}
i
(f)
10^0
i
10^{-2}
10^{-4}
C_R
C_i
C_{CS}
$C_e{}^S$
$C_e{}^C$
$w_{i\to 0}$
0.997
0.998
0.999
1
w

lation of individuals in the two relevant classes after the first passage, $Y_T(G, w, i) = v(G) + V(G)$ (Appendix A), and therefore

$$Y_T(G, w, i) = S_0 m^G \left(\frac{i^G (1-\mu-w)^G (1-i) + i\mu(1-w)^{G-1}}{1 - i\frac{1-\mu-w}{1-w}} \right) \tag{3.2}$$

which for $i = 1$ (absence of inhibition) yields $Y_T(G, w, 1) = S_0 m^G (1-w)^G$. The size of a viable viral population after one passage with a mutagen and/or a replication inhibitor is, thus, given by equation (3.2) with the corresponding parameters w and i. In the absence of drugs, the total population obtained is $S_0 m^G (1-w_0)^G$, such that the initial population is simply multiplied by $m(1-w_0)$ at each replication cycle, which is the basic reproductive ratio of the population.

3.3.1 Comparison between treatments. Viral titre

Let us call $Y_T^{\mathcal{C}}$ the viral titer after the application for one passage of the combined treatment. According to the experimental protocol, the mutagen and the inhibitor have been simultaneously administered, so the viral yield is immediately given by Eq. (3.2): $Y_T^{\mathcal{C}} = Y_T(G, w, i)$. Analogously, $Y_T^{\mathcal{S}}$ is defined as the viral yield after the sequential

Figure 3.3 *(facing page)*: *Theoretical phase space representation of the model.* (a,b) Viral titres after application of the combination (a) or sequential (b) treatment. (d,e) Doses that fall on the grey area cause viral extinction under application of the combination (d) or sequential (e) treatment. The frontiers with the blue and red regions corresponds to values $\{i, w\}$ yielding $Y_T^{\mathcal{C}} = 1$ and $Y_T^{\mathcal{S}} = 1$, respectively, which define curves $C_e^{\mathcal{C}}$ and $C_e^{\mathcal{S}}$. (c) Comparison of the titres produced by combination and sequential therapies. In the $z-$axis we represent the minimum value of the viral yield for each pair of $\{i, w\}$ values. (f) In the grey area extinction occurs if the appropriate therapy is used. Blue and red lines correspond to extinction with the combination or sequential treatment, respectively . Black curve is $C^{\mathcal{CS}}$. It signals the pairs of $\{i, w\}$ values where both therapies produce the same titre. Curve C_R indicates when both treatments produce the same amount of resistants. In the violet region between C_R and $C_{\mathcal{CS}}$ curves the combined treatment yields the lower titre, but the sequential treatment is less prone to produce resistants. In the limit of large amounts of mutagen ($i \to 0$), both therapies perform equally well in regard to titre and resistant production. The expression for the limit value of mutagen $w_{i\to 0}$ (Appendix A) is indicated with an arrow in the plot. Finally, the dotted line stands for curve C_i, which indicates which values of i (below the curve) cause an efficient decrease of the population. Curve C_w, representing the values of the mutagen above which the population of viable individuals decreases from passage to passage is out of the range shown in the plots. Parameters for all plots are $m = 50$, $G = 2.5$, $w_0 = 0.9$, $k = 0.01$.

administration of the drugs, first the inhibitor for one passage and then the mutagen for a second passage

$$Y_T^{\mathcal{S}} = S_0^{-1} Y_T(G, w_0, i) Y_T(G, w, 1)\,, \tag{3.3}$$

where the initial condition for the second passage requires S_0 to be substituted by the yield obtained, $Y_T(G, w_0, i)$. The comparison of the yields obtained following either treatment determine which of them is more efficient for each dose of mutagen and inhibitor. The curve $C_{\mathcal{CS}}$ defined by $Y_T^{\mathcal{C}} = Y_T^{\mathcal{S}}$ separates two regions in the space of parameters $\{i, w\}$ where one therapy or the other are better suited to obtain a low viral titre (Appendix A has an approximate expression of curve $C_{\mathcal{CS}}$ and its asymptotic value at large values of the inhibitor). Figures 3.3(a), (b), and (c) show the viral titre corresponding to the numerical solution of the equations above for the combined and the sequential treatments.

3.3.2 Comparison between treatments. Appearance of resistant mutants

In addition to the viral titre produced after application of one or another therapy, it is important to know what is the fraction of resistant mutants that each treatment produces on average. This quantity corresponds to the final populations of type $V(G)$ after one passage with both the mutagen and the inhibitor (combination therapy) or one passage with the inhibitor plus a second passage with the mutagen (sequential therapy). Let us call $R^{\mathcal{C}}$ and $R^{\mathcal{S}}$ the two final populations of resistants. The exact expressions for the two quantities are obtained as $R^{\mathcal{C}} = V(G, w, i)$ and $R^{\mathcal{S}} = S_0^{-1} v(G, w_0, i) V(G, w, 1) + m^G (1-w)^G V(G, w_0, i)$, analogous to the way in which the titres were obtained. The amount of resistants produced with either treatment will be equal at values of w and i that fulfill $R^{\mathcal{C}} = R^{\mathcal{S}}$. This defines curve C_R. A convenient way of calculating some of the properties of that curve is to express the difference in the number of resistants generated for a pair of values $\{w, i\}$ as $\Delta R(w, i) = \Delta Y(w, i) - \Delta S(w, i)$ where $\Delta R(w, i) = R^{\mathcal{C}} - R^{\mathcal{S}}$, $\Delta Y(w, i) = Y_T^{\mathcal{C}} - Y_T^{\mathcal{S}}$, and $\Delta S(w, i) = S_0 (im(1-\mu-w))^G (1 - m^G (1-\mu_0-w_0)^G)$ is defined as the difference in the amount of susceptible virus produced by the combined and the sequential treatments, with $v(G, w, i)$ as calculated in Appendix A.

The values of $\Delta R(w, i)$ when the yields of the two treatments are equal (that is on curve $C_{\mathcal{CS}}$, where $\Delta Y(w, i) = 0$) are always positive, since $\Delta S(w, i) < 0$ due to the biological fact that the replication rate of the susceptible type is larger than one in the absence of drugs: $m(1 - \mu_0 - w_0) > 1$, and $1 \geq \mu + w$. The production of resistant mutants is thus lower through the sequential treatment when both therapies are equally eficient in terms of total titre. This also implies that curve C_R is always above curve $C_{\mathcal{CS}}$ [see Fig. 3.3(f)]. Hence, there is a region of doses of mutagen and inhibitor, between curves C_R and $C_{\mathcal{CS}}$, where the combination treatment leads to lower titres, but where at the same time the sequential treatment causes a lower population of resistants.

In the limit of high doses of the inhibitor, $i \to 0$, and since $\lim_{i \to 0} \Delta v(w, i) \to 0$ (the population of susceptibles is completely suppressed) the total population coincides

with the population of resistants and $\Delta R(w,i) \rightarrow \Delta Y(w,i)$. As a result, the limit value of w on curve C_R coincides with the limit value of w on curve $C_{\mathcal{CS}}$ (Eq. (A.5) in Appendix A).

3.3.3 Effect of treatment and virus parameters

Not all possible drug doses are experimentally meaningful. First, there is a minimum amount of mutagen required for the therapy to be effective, and it corresponds to those values of w large enough to cause a decrease of viral yield in the absence of the inhibitor. This condition takes the formal expression $m(1-w) < 1$. We thus define curve C_w as those points where $m(1-w) = 1$, separating the regions of increase and decrease of the population size under the action of the mutagen. A similar reasoning leads to curve C_i, which marks the values of i below which the inhibitor is able to cause a decrease of the viable population before resistance appears, $(im)^G(1-w_0)^G = 1$.

There is a limit under high enough values of the drug dose where either treatment may cause the extinction of the virus with the applied protocol. Curve $C_e^{\mathcal{C}}$ is determined by those combinations of $\{i,w\}$ where the population size of the combined treatment falls below 1. The exact solution for curve $C_e^{\mathcal{C}}$ can only be given in implicit form as those values solution of $Y_T(G,w,i) = 1$. To first order in μ, the curve fullfils the approximate expression $m^G S_0(1-w)^{G-1}(i^G(1-w) + kw\gamma) \simeq 1$. Analogously, curve $C_e^{\mathcal{S}}$ is defined by those pairs $\{i,w\}$ at which $S_0^{-1}Y_T(G,w_0,i)Y_T(G,w,1) = 1$. Figure 3.3 illustrates in a representative case the response of a viral population to the action of a combined or a sequential treatment. All curves defined above (except for C_w, which occurs at values of w quite far from the domains shown) are represented in Fig. 3.3(b), (d), and (f).

Parameters m and G depend significantly on the mode of replication of the virus inside an infected cell. Single-stranded RNA viruses are characterized by large values of m and probably quite low values of G. In double-stranded genomes with semi-conservative replication one would expect a value $m \approx 2$ and a larger value of G. In either case, the accumulation of mutations proceeds differently, and so also w_0 would vary accordingly. The model presented here cannot be directly applied to viruses that use other replication modes, such as retroviruses (with a DNA provirus phase) or DNA viruses such as the herpesviruses that include a latency step. Extension to such systems would require significant modifications of the dynamical rules.

3.4 Experimental validation of the model

The parameters of the model can be grouped into two categories, those related to the virus and those depending on the therapeutical treatment. Most of them can be directly estimated by means of simple experiments that are here described. As a case example we use data obtained with FMDV under sequential and combined therapies involving the inhibitor guanidine and the mutagen ribavirin.

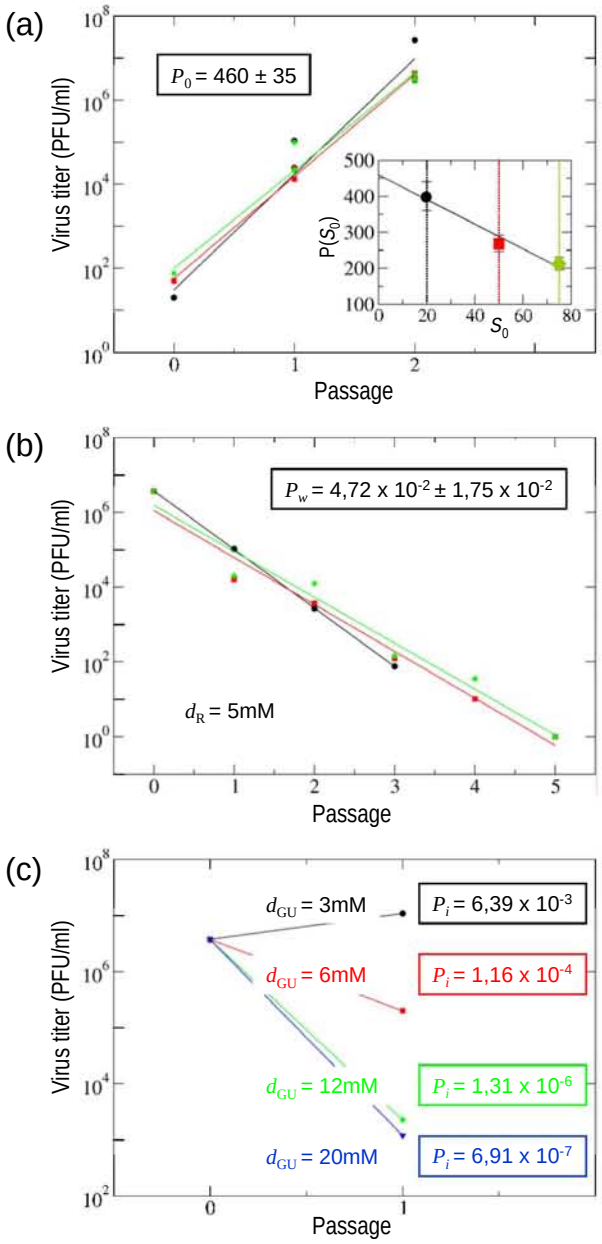
3.4.1 Viral parameters

Several parameters of the model describe the intrinsic characteristics of the virus in the cells it is infecting. These are the number of replication cycles inside the cell (G), the number of copies per viral genome and replication cycle (m), the rate of production of non-viable mutants (w_0), and the ratio (k) between resistance conferring mutations and deleterious ones.

The experimental measure of the viral productivity $P_0 = m^G(1-w_0)^G$ in the absence of drugs is easy to perform. If there is no limitation on resources, the population grows exponentially at a rate equal to the viral productivity. As a result, if the population size along succesive passages is plotted in a semilogarithmic graph we obtain a straight line whose slope is equal to the logarithm of the productivity. Viral productivity is thus obtained by quantifying population size in several succesive passages (Fig. 3.4). It is measured in units of PFU/ml and per passage.

When modelling the dynamics of single-stranded RNA viruses it should be noticed that each replication cycle requires the synthesis of a complementary antisense RNA, that is used as a template for the synthesis of new positive-sense genomes. For this reason, the number of copies per replication cycle is equal to the mean number of genomes produced from a single antisense RNA. Thus, parameter m could be estimated as the ratio between sense and antisense genomes inside the cell. The few estimations of that ratio available indicate a large excess of positive strands, from 50 to 1000 per negative strand (Herrera et al. 2008).

Figure 3.4 *(facing page): A case example of determination of model parameters from experimental data.* (a) Natural growth of the population. Three different sets of experiments using triplicates for each condition have been used. The titre of the virus (main plot) has been determined starting with different initial conditions, that is initial population sizes $S_0 = 20$ (black circles), 50 (red squares), and 75 (green diamonds) at two consecutive passages. Least-squares regression with exponential functions yield the productivity $P(S_0)$, which depends on S_0. Representation of the so obtained productivities as a function of S_0 (inset) allows extrapolation to $S_0 \to 1$, which finally yields the basal productivity $P_0 = 460 \pm 35$. (b) Decrease of the population in presence of the mutagen, without the inhibitor. We show the exponential decay in the viral titre for three independent realizations of the experiment with a dose of mutagen $d_R = 5$ mM. The slopes of the curves have been averaged to obtain an estimation of the productivity P_w in the presence of the mutagen, yielding $P_w = m^G(1-w)^G = 4.72 \times 10^{-2} \pm 1.75 \times 10^{-2}$. (c) Population growth in the presence of the inhibitor and absence of the mutagen. Productivity in the presence of the inhibitor has been determined using a single passage to avoid confounding effects due to the appearance of resistant forms. The slope of each curve corresponds to $P_i = im^G(1-w_0)^G = i^G P_0$. All productivities are measured in units of PFU/ml and per passage.


(a)
P_0 = 460 ± 35
Virus titer (PFU/ml)
Passage
P(S_0)
S_0
(b)
P_w = 4,72 x 10^{-2} ± 1,75 x 10^{-2}
d_R = 5mM
Virus titer (PFU/ml)
Passage
(c)
d_{GU} = 3mM
P_i = 6,39 x 10^{-3}
d_{GU} = 6mM
P_i = 1,16 x 10^{-4}
d_{GU} = 12mM
P_i = 1,31 x 10^{-6}
d_{GU} = 20mM
P_i = 6,91 x 10^{-7}
Virus titer (PFU/ml)
Passage

Our parameter w_0 comprises those genomes that carry lethal mutations and all defective genomes unable to replicate by themselves. This is so because the former do not play any further role in the dynamics and the latter are cleared up from the population at each passage when MOI is sufficiently low, as in the case studied. The explicit consideration of the defective type does not modify the quantitative predictions reached. This natural mutation rate can be estimated from a number of experiments which have determined the effect of mutations on viral fitness and the ratio of defective forms to wild type genomes. Lethal mutations affect from 20% to 55% of total genomes produced under replication (Parera et al. 2007; Sanjuán 2010), so a reasonable though still quite rough estimate for w_0 would be between 0.4 and 0.9, assuming half of the mutations are lethal and half cause defects that prevent completion of a viral infectious cycle.

Regarding the mutation ratio k, biological knowledge about the virus is also necessary to estimate its value. In any case, since it can be hypothesized that only one (or a few) mutations at specific sites of the genome generate resistance to the inhibitor, the ratio k should be of the order of the inverse of the genome size.

3.4.2 Parameters related to experimental conditions

The treatment is described by two experimental parameters: the inhibition factor i and the increased mutation rate $w \geq w_0$.

In a particular realization of the therapy, both parameters can be experimentally obtained as the decrease in viral productivity in the presence of the inhibitor or the mutagen separately. In the presence of a mutagen, viral productivity is $P_w = m^G(1 - w)^G$. This value can be experimentally measured in the same way as before, by plotting the evolution of the population size in semilogarithmic graph and taking the slope of the resulting line. If the productivity with mutagen is lower than one, that slope will be negative, which means that the population is decreasing. Once the productivity has been determined the parameter w can be calculated by using the previously estimated values for m and G. The case for the inhibitor is analogous. Now the expression of the productivity becomes $P_i = (im(1-w_0))^G$ and parameter i can be determined once the productivity has been experimentally measured. In this case some care must be taken when calculating productivity, since the slope changes as resistant mutants appear after a few passages.

3.4.3 Predicted parameter values for FMDV

Comparison of the titres experimentally obtained with those predicted by the model allows to fix the values of all model parameters. We use now the mathematical expressions obtained for the combination ($Y_T^{\mathcal{C}}$) or sequential ($Y_T^{\mathcal{S}}$) treatment and do as follows. First, we select a pair of values m and w_0 within the estimated interval and, by using the basal productivity $P_0 = 460 \pm 35$, obtain that G is bounded between 1 and 3.8. Second, use of P_w immediately yields the range of possible values of w, which is bounded between 0.992 and 0.999. Similarly, estimation of the productivity

in the presence of different doses of the inhibitor permits estimating parameter i as $i = (P_i/P_0)^{1/G}$. More accurate measurements of the viral parameters by means of specific experiments directed to quantify m, G, and/or w_0 could significantly narrow the intervals compatible with the experimental results. Third, we represent the experimentally obtained titre as a function of the calculated i. Since all parameters are now fixed, the titre predicted by the model with varying i is obtained by direct substitution into the expressions for $Y_T^{\mathcal{C}}$ and $Y_T^{\mathcal{S}}$. Finally, we evaluate the error produced by this set of parameters by calculating the sum of the squared distance between data and titre estimated through the model. The steps above and the evaluation of the error are repeated for all compatible pairs m and w_0. The combination yielding the smaller error is accepted as optimal given the experimental data.

3.4.4 The case of FMDV with ribavirin and guanidine

Results with FMDV subjected to the therapeutic protocols described under fixed doses of R and GU revealed that the sequential treatment could be better suited to cause viral extinction (Perales et al. 2009). However, in the light of the model described, the adequacy of a sequential versus a combination treatment with a mutagen and an inhibitor of the viral replication depends (i) on the administered doses and (ii) on the natural productivity of the virus when infecting a given host (i.e. parameters m, w_0, and G). In particular, the model predicts that, in most cases, at a sufficiently low amount of inhibitor and for a fixed value of the amount of mutagen, the combination treatment should yield a lower titre, the same change holding for a fixed amount of inhibitor and a sufficiently low amount of mutagen. Increase of the doses d_{GU} or d_R above a threshold value changes the treatment that produces the lower amount of infectious virions. In order to test this prediction, we carried out several experiments at increasing doses of inhibitor and compared the viral yield after both therapeutic protocols (Fig. 3.5(a)). At low doses of inhibitor, the combined therapy causes lower viral yields than the sequential therapy. However, for a dose of the inhibitor between 6 and 12 mM, it is the sequential treatment that begins to yield lower viral titers. The effect of the therapies is thus exchanged, as theoretically predicted (Fig. 3.5(b)).

3.5 Discussion

Ever since the quasispecies dynamics was revealed as general for pathogenic RNA viruses and retroviruses, the problem of treatment failure due to selection of drug-escape viral mutants was recognized (Domingo and Holland 1992). Several approaches have been used in medical practice to minimize selection of viral mutants resistant to antiviral agents. The most successful strategy was the implementation of combination therapy involving the simultaneous administration of two or more drugs directed to different viral targets. The advantages of combination therapy over monotherapy stem from basic statistical considerations on the frequency of generation of multidrug-resistant mutants (Domingo and Holland 1992). The advantage of combination versus

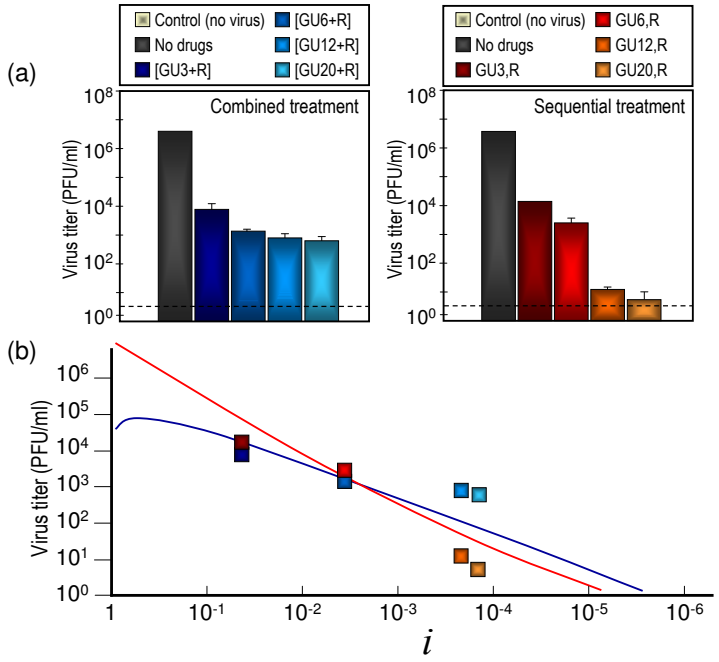


Figure 3.5: *Experimental results employing FMDV infection of cell cultures.* Comparison with theoretical predictions. (a) Experimental results. Passages were carried out by infecting 2×10^6 BHK-21 cells with pMT28 strain of FMDV (0.2 ml of supernatant from the previous passage), and infectivity levels were determined as detailed in Appendix A. The combination or sequential protocol therapies were applied in the presence of a dose of mutagen $d_R = 5$ mM and increasingly high doses of viral inhibitor, d_{GU} =3, 6, 12, or 20 mM, as indicated. The combination treatment is indicated in brackets (left panels), while for the sequential treatment the order of administration of GU and R is separated by a comma (right panels). A control measure is shown for reference (first left bar in each panel), other columns represent the mean $\pm$ SD (error bars) from triplicate determinations. The discontinuous lines in the virus titers indicate the limit of detection of infectivity. (b) Comparison of experimental results with the predictions of the model. Viral titers expected after appliying the sequential or the combination protocol in the mathematical model are shown as continuous lines (blue: combined treatment; red: sequential treatment; color code for data as in (a)). Curves obtained with parameters that yield the least-squares deviation to the logarithm of experimental data are shown as blue (combination) and red (sequential) curves. The obtained parameters are $w_0 = 0.76$, $w = 0.998$, $k = 0.005$, $m = 195$. The values of the inhibitor $i = 4.25 \times 10^{-2}, 3.47 \times 10^{-3}, 2.10 \times 10^{-4}$, and 1.41×10^{-4}, corresponding to the four doses assayed, are shown in the $x-$axis.

various regimens of sequential or treatment interruption (structured treatment interruptions or drug holidays) regimens has been supported by results of multiple clinical trials as well as by theoretical models of viral dynamics (Domingo 1989; Ho 1995; Bonhoeffer et al. 1997; Ribeiro and Bonhoeffer 2000; Domingo et al. 2008; Müller and Bonhoeffer 2008; Nijhuis et al. 2009). However, whenever residual viral replication is allowed, viral rebound and treatment failure often occur. For this reason, new proposals for the administration of drugs have been made. A recent one, yet to be tested in clinical trials, is the so called pro-active treatment in two steps: an induction regime aimed at reducing the background of viral mutants, followed by a maintenance regime to control the viral load for extensive time periods (von Kleist et al. 2011). An alternative strategy is to target cellular factors needed for viral replication, either alone or in combination with drugs directed to viral functions (Geller et al. 2007). Provided no toxic effects on the cells intervene, resistance to inhibitors of cellular functions can also develop when the host factor interacts with viral nucleic acids or proteins.

Lethal mutagenesis introduces an important new element in antiviral therapy in that a mutagenic agent is involved in treatment. When administered together with a non-mutagenic inhibitor, the mutagen can play a dual and opposite role: to deteriorate viral functions due to the excess mutations it provokes, and to increase the frequency of mutations that confer resistance to the partner, co-administered inhibitor. An advantage of a sequential over a combination treatment during lethal mutagenesis protocols was suggested by a study of lethal mutagenesis of FMDV (Perales et al. 2009), and this prompted the theoretical generalization reported in the present study and that has been further supported by additional experiments also reported herein.

In this chapter we have developed a theory that quantitatively describes the response of a viral population under two different protocols which involve the action of an inhibitor of viral replication and a mutagenic drug. A limited number of simple experiments allows to estimate the parameters that describe the dominant processes, and knowledge of those few relevant parameters permits to predict the behaviour of the model for other combinations of drugs. Two important outcomes of the treatments have to be considered when choosing between a combined or a sequential administration of the drugs: in a region of mutagen and inhibitor doses that can be calculated, the viral titre produced by one or another therapy is minimized. Further, there is a domain of drug doses where combination treatment yields the lower titre, but the probability of appearance of mutants resistant to the inhibitor is lower with a sequential treatment. The use of analogous models can significantly reduce the number of *in vitro* assays to be performed in other viral systems as well, where different replication strategies should translate into dynamical equations similar to our Eq. (3.1). The simulation of an *in vivo* situation entails additional difficulties, like the development of a large viral population from an infecting seed, the interaction with the immune system, or environmental and individual characteristics that may not lead to deterministic equations, but should be included as noisy, fluctuating dynamical variables. The predictions of the model, once tested *in vitro*, could be taken only as a rough guide to apply one or another administration protocol and to infer minimum drug doses in *in vivo* assays.

Our predictions acquire more relevance in view of the evidence that lethal mutagenesis can indeed be effective *in vivo* (Ruiz-Jarabo et al. 2003), and that some clinical trials for AIDS patients that involve administration of a mutagenic nucleoside analogue have been implemented (Mullins et al. 2011). Our results will be particularly relevant when considering a lethal mutagenesis approach to combat viruses for which a repertoire of non-mutagenic inhibitory agents is already available. Detailed predictions of the viral response tailored to the particular system under study are now possible.

III

Evolution of genomes

4

Genome segmentation in multipartite viruses

4.1 Multipartite viruses: An evolutionary puzzle

The origin and evolutionary history of viral genomes is a classical problem that has inspired a long series of questions and hypotheses in evolutionary biology (Eigen 1993; Roosinck 1997; Manrubia and Lázaro 2006). One of those questions is the adaptive meaning of genome segmentation, since it appears to be a common trait in a very broad variety of viruses (Szathmáry 1992; Turner and Chao 1998; Ojosnegros et al. 2011). The case of multipartite viruses is particularly striking, since their genome segments achieve complete independence at the apparent cost of reducing its infectivity (Palukaitis and García-Arenal 2003; Betancourt et al. 2008; González-Jara et al. 2009).

Multipartite viruses have their genomes fragmented into two or more (up to eight) segments, each packed into a separate capsid and containing one or more genes that are essential for the virus to complete an infection cycle. These viruses require complementation, since each genomic segment must be completed (i.e. complemented) with the rest of segments in order to produce viral offspring. The complementation requirements for multipartite viruses have a strong impact in the way they are transmitted: many viral particles have to enter each cell in order to assure that at least one representative of each segment will be present. The *multiplicity of infection* (MOI) is thus a key quantity in the biology of multipartite viruses.

A noticeable fact about multipartite viruses is their asymmetry in host distribution: while they are common among plant viruses, no multipartite virus has been described

infecting animals (Lazarowitz 2007). It has often been claimed that a larger characteristic MOI in plant infections may be behind this phenomenon (Nee 1987), but a quantitative analysis of this claim has not been carried out up to now. From an evolutionary perspective, this asymmetry could be understood by taking into account a trade-off between opposite selective pressures: while the complementation requirement acts as a limiting factor –viral extinction supervenes if transmission bottlenecks occur– there should be evolutionary forces that promote the fragmentation of viral genomes. Actually, the first step towards fragmentation might be the generation of incomplete genomes. The latter are known to arise spontaneously under replication of wild type (wt) viruses (Holland 1990), which often produce the so-called defective interfering particles (DIPs) –that is, incomplete genomes able to infect cells but unable to complete the infection cycle in the absence of the wt (Bangham and Kirkwood 1990; Roux et al. 1991). The effect of DIPs in the dynamics and evolution of viruses has been studied by means of mathematical models (Kirkwood and Bangham 1994; Frank 2000), and particular attention has been paid to the mechanisms allowing for the coexistence of wt and defective forms (Szathmáry 1992; Szathmáry 1993; Wilke and Novella 2003). Nonetheless, the advantage of fragmentation and especially the individual encapsidation of the fragments still remain open questions.

Faster replication of shorter genomes and higher replication fidelity have been classically presented as factors favouring genome segmentation (Nee 1987; Chao 1991), though there is no conclusive empirical evidence of their evolutionary advantage in viruses. Recent experimental work, on the other hand, has compared the performance of the wt, complete form of foot-and-mouth disease virus, with a fragmented (bipartite) counterpart obtained by evolution in cell culture (García-Arriaza et al. 2004). Competition experiments have shown that the latter, shorter genomes, may possess a larger average lifetime between infective events (Ojosnegros et al. 2011). These results point at the stability of viral particles as a relevant feature that could counterbalance the disadvantage of high MOIs required to produce infection.

While the possibility of obtaining complementary, segmented variants of an originally non-segmented virus had been confirmed through isolation of variants with genomes separated into two molecules (O'Neill et al. 1982) and by means of genetic engineering techniques (Geigenmüller-Gnirke et al. 1991; Kim et al. 1997), the experimental evolution of a bipartite virus from a complete, non-segmented wt provides an insight into how multipartite viruses could have originated in nature. In this chapter, we explore the hypothesis that multipartite viruses are the evolutionary outcome of the competition among genomic segments of different lengths. These segments would be naturally produced through deletions in the replication process of the original, wt virus (García-Arriaza et al. 2006). Provided that shorter genomes enjoy a certain evolutionary advantage, a set of segments may be able to outcompete the wt virus if the MOI is high enough to guarantee complementation among the segments. We will focus on reduced degradation as the selective pressure favouring segmentation, although alternatives will also be considered, and their formal equivalence investigated. We discuss the likeliness that multipartite viruses with a large number of fragments could have originated in that scenario.

4.2 Model of multipartite virus dynamics

4.2.1 Formal scenario

A schematic of the basic mechanisms included in the model is depicted in Figure 4.1. Let us consider a mixed population that contains wt viruses (those with a complete genome, denoted as wt) as well as two smaller components whose genomes are complementary segments (denoted as $\Delta 1$ and $\Delta 2$). Both components $\Delta 1$ and $\Delta 2$ constitute together a bipartite variant of the wt virus. Each component requires complementation for replication, either with the other component or with the wt. At each generation, viruses in the population infect a set of cells at a given MOI m. Although, in general, m is a quantity that depends on the viral load, we will study only the case of constant m for the sake of simplicity. Replication takes place inside the cells depending on complementation requirements. Thus, the amount of viruses of each class produced by a single cell depends on the initial composition of the (small) infecting population –it is an instance of frequency-dependent fitness (Smouse 1976). The sum of all viruses produced by all cells constitutes the primary offspring. It is exposed then to differential degradation, which preferentially affects the wt class. Survival of the wt relative to the segmented classes is given by a parameter $\sigma < 1$. The viral population that results after degradation is considered to be the infecting population for the next generation. In such a way, the composition of the population can be traced for several generations in an iterative way.

4.2.2 Evolution equation with differential degradation

The composition of the population can be expressed by a vector

$$\mathbf{p} = (p_{\Delta 1}, p_{\Delta 2}, p_{wt})^T \tag{4.1}$$

where $p_{\Delta 1}$, $p_{\Delta 2}$ and p_{wt} are the fractions of classes $\Delta 1$, $\Delta 2$ and wt in the population, such that $p_{\Delta 1} + p_{\Delta 2} + p_{wt} = 1$. At generation n the composition of the population will be denoted as $\mathbf{p}_n$.

Let us begin by considering a single cell that has been infected by a, b and c viral particles of classes $\Delta 1$, $\Delta 2$ and wt, respectively. The MOI in this case would be $m = a + b + c$. Vector $(a, b, c)^T$ will be referred to as the *infection configuration*. The viral offspring produced by the infected cell can be expressed as a product of a fitness matrix $\mathsf{M}_{a,b,c}$ and the infection configuration. The fitness matrix allows the introduction of frequency-dependent fitness, provided that production of each viral class is affected by the abundances of other classes in a linear way.

The complementation requirement can be implemented as follows. For a given genome to reproduce there are two limiting factors: first, the availability of the own genome; second, the availability of essential proteins (that depends itself on the amount of genomes coding for them). As a result, a simple way to consider complementation consists of assuming that the number of genomes of a given class produced inside a cell is equal to the minimum between the number of genomes of that class that infected

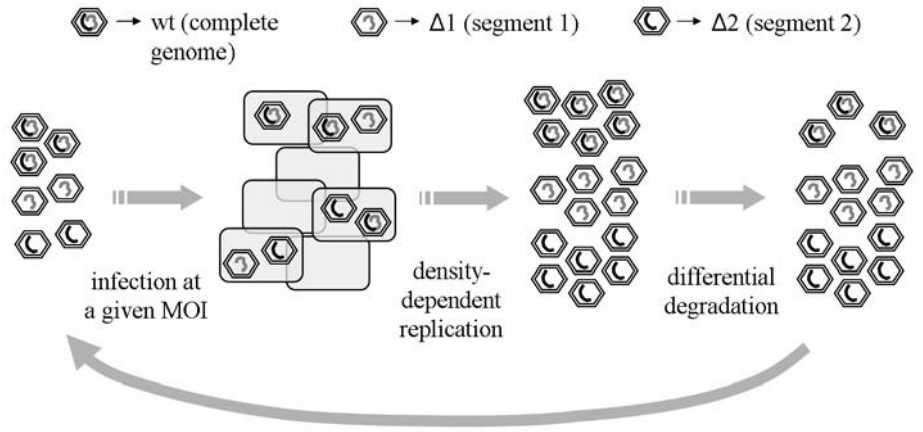


Figure 4.1: *Schematic of the infective process.* The MOI is the (average) number of viral particles infecting a cell. Incomplete genomes ($\Delta 1$ and $\Delta 2$) have an average lifetime larger than that of the wt between infective events, as shown graphically by the decrease in number of the wt relative to genomes with deletions. The process is iterated until the equilibrium state is attained.

the cell and the number of the corresponding complementary genomes. For instance, segment $\Delta 1$ requires $\Delta 2$ or wt for complementation; so the number of viral particles of class $\Delta 1$ produced will be $\min\{a, b+c\}$. On the other hand, wt genomes do not need complementation, and their final number will depend only on their initial abundance c. Without loss of generality, we will fix the replication rate for the wt class equal to one. Taking all this into account, the offspring produced by a single cell can be written as

$$\mathsf{M}_{a,b,c}\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \min\{a, b+c\} \\ \min\{b, a+c\} \\ c \end{pmatrix} \equiv \begin{pmatrix} f_{\Delta 1|a,b,c} \\ f_{\Delta 2|a,b,c} \\ \sigma^{-1} f_{wt|a,b,c} \end{pmatrix} \tag{4.2}$$

The previous expression defines conditional (frequency-dependent) fitness $f_{i|a,b,c}$, $i \in \{\Delta 1, \Delta 2, wt\}$.

The next step is to obtain the global offspring produced by the whole set of cells. Let us assume that the number of cells is large enough such that the global offspring can be calculated by averaging the offspring of a single cell (eqn. 4.2) over the probability of a given infection configuration. That probability depends on the MOI m as well as on the population composition $\mathbf{p}$. As an example, we consider here the case where the MOI is Poisson distributed,

$$\Pr(a, b, c|\mathbf{p}) = e^{-m}\, \frac{m^{(a+b+c)}}{a!\, b!\, c!}\, p_{\Delta 1}^{a}\, p_{\Delta 2}^{b}\, p_{wt}^{c}\,. \tag{4.3}$$

Note that this joint distribution is equivalent to the product of three independent Poisson distributions with averages mp_j, with $j \in \{\Delta 1, \Delta 2, wt\}$. The case of an MOI following a multinomial distribution is dealt with in Appendix B.

Differential degradation is applied by multiplying the global offspring by a degradation matrix D, that takes the form of a diagonal matrix with value one in the first two positions (corresponding to $\Delta 1$ and $\Delta 2$) and $\sigma < 1$ in the third position (reduced survival for the wt),

$$\mathsf{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma \end{pmatrix} \tag{4.4}$$

All steps can be written in a single equation that provides the composition of the population in succesive generations.

$$\mathbf{p}_{n+1} = Z^{-1}\,\mathsf{D} \sum_{a,b,c} \Pr(a,b,c|\mathbf{p}_n)\,\mathsf{M}_{a,b,c} \begin{pmatrix} a \\ b \\ c \end{pmatrix} \tag{4.5}$$

where Z is a normalization factor (see Appendix B)

Equation 4.5 can be iterated in order to obtain the evolution of a mixed population of single-particle wt virus and bipartite mutants derived from it, provided that the selective advantage of the bipartite mutants is due to reduced degradation. Note that according to the definition of conditional fitness (eqn. 4.2), it can be written in the form of a replicator equation. Indeed, let $\langle f_i \rangle$ be the average value of the conditional fitness for a generic class i,

$$\langle f_i \rangle = \sum_{a,b,c} \Pr(a,b,c|\mathbf{p}) f_{i|a,b,c} \tag{4.6}$$

If p_i is the fraction of class i in the population and we denote with a superscript the generation in which a given quantity is measured, the evolution equation is equivalent to the replicator equation

$$p_i^{(n+1)} = \frac{\langle f_i \rangle^{(n)}}{Z^{(n)}} \qquad i \in \{\Delta 1, \Delta 2, wt\} \tag{4.7}$$

where

$$Z^{(n)} = \sum_i \langle f_i \rangle^{(n)} \tag{4.8}$$

The evolutionary dynamics reaches an equilibrium state for such compositions that are fixed points of the replicator equation. They must fulfil the equilibrium condition

$$p_i^* = \frac{\langle f_i(\mathbf{p}*) \rangle}{Z}\,, \qquad \forall i \in \{\Delta 1, \Delta 2, wt\} \tag{4.9}$$

In particular, we are interested in equilibrium points that are attractors of the evolutionary dynamics. A detailed study of these points and their stability can be found in Appendix B.

4.2.3 Generalization of the evolution equation

The evolution equation in the previous section can be modified to include additional selective pressures. The set of fitness matrices will contain new parameters accounting, for instance, for replication and mutation rates, and in some cases, the number of replication cycles inside the cell has to be explicitly considered. A detailed derivation of generalized equations can be found in Appendix B.

Different replication rates

Owing to their smaller size, it has been argued that bipartite genomes may replicate faster than wt ones (Nee 1987; Chao 1991). Without loss of generality, let $R > 1$ be the replication rate of the segments $\Delta 1$ and $\Delta 2$ relative to that of the wt. In a discrete-time model, R^{-1} is the average number of genomes of wt class produced after one cycle of intracellular replication. Several replication cycles can take place before the viral offspring is released out of the cell, let G be that number of cycles. Under these circumstances, the conditional fitness that allow the evolutionary process to be written as a replicator equation are the following:

$$\begin{aligned} f_{\Delta 1|a,b,c} &= \min\{a, b + cR^{1-G}\} \\ f_{\Delta 2|a,b,c} &= \min\{b, a + cR^{1-G}\} \\ f_{wt|a,b,c} &= \sigma c R^{-G} \end{aligned} \tag{4.10}$$

Loss of segments through mutation and replication fidelity

Let us consider the possibility that a genomic segment is lost during replication with probability ρ. As a result, the probability that a wt virus replicates its whole genome without losing any segment is $(1-\rho)^2$, where the square means that a genome with two putative segments is considered. On the other hand, the probability that a single-segment virus $\Delta 1$ or $\Delta 2$ replicates without errors is $1-\rho$. That is the reason why bipartite mutants have, in principle, a higher probability of error-free copy than wt ones. In addition, single-segment viruses can be produced from the wt with probability $\rho(1-\rho)$. In this setting, the evolution of the population follows a replicator equation where conditional fitness can be expressed as

$$\begin{aligned} f_{\Delta 1|a,b,c} &= (1-\rho)^G \left[\min\{a, b + c(1-\rho)^{1-G}\} + c\big(1-(1-\rho)^G\big)\right] \\ f_{\Delta 2|a,b,c} &= (1-\rho)^G \left[\min\{b, a + c(1-\rho)^{1-G}\} + c\big(1-(1-\rho)^G\big)\right] \\ f_{wt|a,b,c} &= \sigma c (1-\rho)^{2G} \end{aligned} \tag{4.11}$$

Constant per-cell viral productivity

The expressions in Eq. (4.2) implicitly assume that there are no restrictions to the maximal number of viral particles that an infected cell can produce. However, because cellular resources are limited, viral production may be bound. This situation can be tackled by normalizing the total number of viral particles produced. Without loss of generality, we suppose that this value equals one and study the dynamics under the conditional fitness

$$\begin{aligned} f_{\Delta 1|a,b,c} &= z_{a,b,c}^{-1} \min\{a, b+c\} \\ f_{\Delta 2|a,b,c} &= z_{a,b,c}^{-1} \min\{b, a+c\} \\ f_{wt|a,b,c} &= z_{a,b,c}^{-1} \sigma c \,, \end{aligned} \tag{4.12}$$

with $z_{a,b,c} = \min\{a+b, 2a+c, 2b+c\} + c$.

4.2.4 Multiple segments

The model can be easily expanded to describe the dynamics of genomes that are susceptible to being partitioned into more than two segments. In a multiple-segment model, viral classes are defined by the genomic segments they conserve. The extreme cases are still the wt virus, that contains the complete genome, and the single-segment classes, that constitute the genuine multipartite version of the virus. In addition, there will also be classes with an intermediate number of segments. Provided that a genome is composed of n putative segments, the total number of different viral classes is $2^n - 1$ (the class containing no segments has been discounted), and the number of classes containing s segments is $\binom{n}{s} = n!/(s!(n-s)!)$.

Complementation implies that replication is limited by the less abundant genomic segment inside the cell. In the simplest scenario, we assume that the selective advantage favouring shorter genome lengths is proportional to the number of segments that a given class contains. In the case of degradative advantage, degradation is assumed to be zero for the single-segment classes and to increase linearly with the number of segments up to the value $1 - \sigma$ for the wt virus. By taking these considerations into account, the extension of the bipartite model to the multipartite case is straightforward, the main difficulties arising from the high dimensionality of the classes space. The case with three segments is developed in detail in Appendix B. Other relationships between genome length and degradative advantage are possible: the case where the selective advantage is proportional to the volume of the packed genome—emphasizing the role played by the interaction with the capsid—is studied in Appendix B.

4.3 Results

4.3.1 Evolutionary shift from wt to a bipartite form

Analysis of the evolution equation (4.5) or the equivalent replicator equation (4.7) reveals two possible outcomes for the evolutionary process, depending on the values of parameters σ and m. If degradation of the wt virus is high enough when compared to that of the segments, then the wt gets extinct and single-segment variants $\Delta 1$ and $\Delta 2$ take over the population (Fig. 4.2(a)). Therefore, in this regime bipartite variants of a single-particle virus are able to outcompete the latter and reach fixation in the population. On the other hand, coexistence is the expected outcome if degradation of the wt is low (Fig. 4.2(b)). Both regimes are separated by a critical value of the survival parameter σ_{crit}, so that $\sigma < \sigma_{crit}$ leads to extinction of the wt while $\sigma > \sigma_{crit}$ allows for coexistence.

An analytic expression for the critical value σ_{crit} can be obtained by means of simple invasibility arguments. Provided that the point $(1/2, 1/2, 0)^T$ (corresponding to a pure equilibrated population of the bipartite form) is a stable equilibrium point in the absence of the wt class, the key point is to study its stability when an infinitesimal amount of wt is introduced. The equilibrium point becomes unstable at a critical value $\sigma = \sigma_{crit}$, where

$$\sigma_{crit} = \frac{2}{m} \sum_{a,b} \Pr(a, b|\tfrac{1}{2}, \tfrac{1}{2}, 0) \min\{a, b\} \,. \tag{4.13}$$

Using Eq. (4.3) and after some algebra (see Appendix B), the critical value separating coexistence of all types from extinction of the wt can be written in terms of modified Bessel functions of the first kind $I_\alpha(m)$:

$$\sigma_{crit} = 1 - e^{-m} I_0(m) - e^{-m} I_1(m) \,. \tag{4.14}$$

A simple expression is obtained in the limit of large MOI, $m \gg 1$,

$$\sigma_{crit} \sim 1 - \sqrt{\frac{2}{\pi}}\, m^{-1/2} \,, \tag{4.15}$$

and this same asymptotic result holds for a multinomial distribution of MOIs.

Figure 4.3 compares the numerical and asymptotic values of σ_{crit} as a function of m, and reveals that the asymptotic approximation actually recovers well the behaviour of the system also at relatively small values of m. As intuitively expected, increasing the relative degradation of wt virus implies increasing the selective pressure favourable to the bipartite virus, what permits its fixation. Alternatively, an increase of the multiplicity of infection makes complementation easier, as there are more genomes inside the cell providing complementation. As a consequence, greater MOI also favours fixation of the bipartite virus. Finally, note that there is no parameter region for which the wt virus outcompetes the bipartite one.

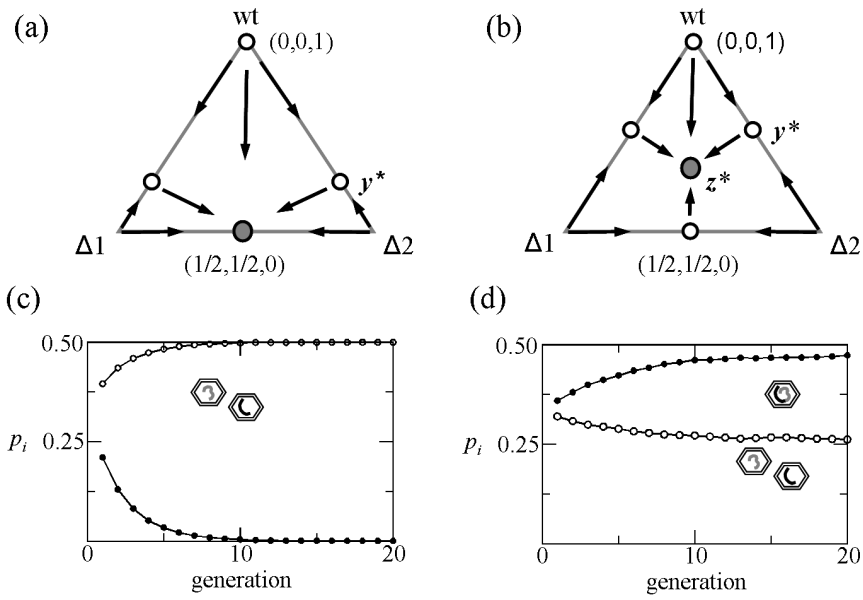


Figure 4.2: *Evolutionary outcomes for the competition between a single particle virus (wt) and its bipartite variant (*$\Delta 1$ *and* $\Delta 2$*).* (a,b) Equilibrium points in the population simplex (open circles: unstable, full circles: stable), arrows indicating the evolutionary trajectories. (c,d) Temporal evolution of population composition for a characteristic realization (solid symbols: wt, open symbols: $\Delta 1$ and $\Delta 2$, legend for each viral form as in the previous figure). (a,c) Fixation of the bipartite variant takes place if survival of the wt is below a critical value $\sigma < \sigma_{crit}$. (b,d) Coexistence of all viral classes occurs if $\sigma > \sigma_{crit}$. An analytical expression for σ_{crit} is given in the main text.

The results obtained earlier remain qualitatively unchanged if the selective advantage of the bipartite virus relies on a faster replication or if mutations leading to the loss of segments are considered. In the former case, the conditional fitness defined by eqn. (4.10) can be used to derive an analogous critical condition such that σ_{crit} is replaced by R^{-G}. In the latter, σ_{crit} is substituted by $(1-\rho)^G$. When resources are limited by the cell, the critical condition is derived from a slightly more involved mathematical relationship. Nonetheless, there are no qualitative differences with the situation reported, but only minor quantitative differences (see Appendix B).

4.3.2 Viruses with multiple segments

As in the bipartite case, evolutionary outcomes include coexistence of all possible viral classes and fixation of the single-segment classes. The latter would result in a net evo-

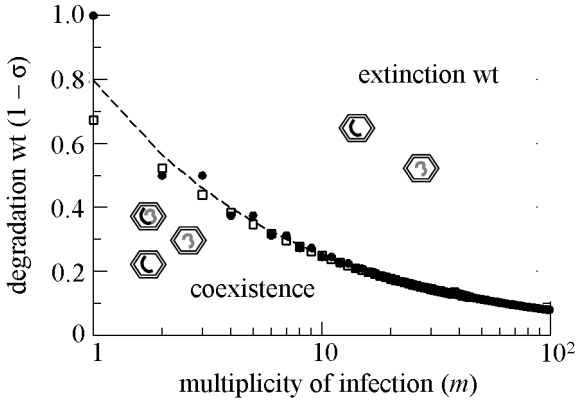


Figure 4.3: *Evolutionary regions depending on selective pressures (MOI and degradation of wt class).* Regions corresponding to coexistence and extinction of wt (fixation of bipartite virus) are separated by a series of critical points, whose values vary slightly according to the probability distribution that governs the infection process (solid circles, multinomial; open squares, Poisson; dashed line, asymptotic behaviour).

lution from a single-particle virus to a multipartite virus with as many particles as genomic segments. As a novelty, there appears a whole range of intermediate equilibrium states that successively lack the wt, the second longest classes and so on. Figure 4.4(a) shows a map of the evolutionary regions for a genome with 3 possible segments. The intermediate region corresponds to an equilibrium state that contains the three possible two-segment classes as well as the three single-segment classes. This region is limited by two series of critical points that separate it from the total coexistence region below and the multipartite fixation region above. A comparison with the two-segment case (Fig. 4.3) reveals that fixation of a multipartite virus with three segments requires a much higher MOI. This result is expected because, in this case, complementation of a single segment requires the presence of two different complementary segments. For low values of the MOI, it may be impossible for the single segments to get fixation, even with the maximum degradative advantage. This is because of the linear relation between the degradative advantage and the number of segments: the minimum survival for the wt virus, 0% related to that of a single segment, translates into a relative survival of 50% for two-segment classes. That can be enough for the two-segment classes to avoid extinction at not very high MOIs.

A relevant question is how high the MOI must be so that a multipartite virus with a given number of segments can reach fixation. To address that point, let us take a fixed value for the survival parameter and observe the critical MOI values that separate one evolutionary region from another. Results for $\sigma = 0.5$ are shown in Figure 4.4(b).

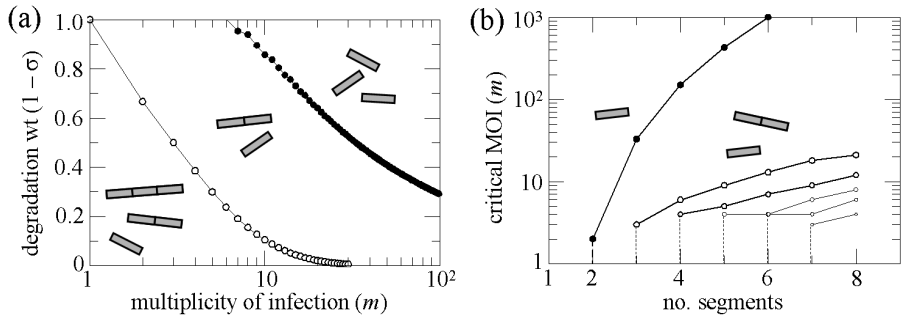


Figure 4.4: *Evolutionary outcomes for genomes with multiple segments.* (a) Coexistence and multipartite fixation regions for three segments. Rectangles schematically indicate the number of segments that viral classes contain in a certain region (all classes with such a number of segments will be present). Compare these curves with σ_{crit} for two segments, shown in Figure 4.3. (b) Critical value of the MOI required for fixation of single-segment classes (thick line on the left), for $\sigma = 0.5$. Other lines indicate further transitions: from double-single coexistence to triple-double-single, and so on.

The thick black curve on the left-hand side of the figure indicates the minimum MOI required so that the single-segment classes are able to outcompete other classes with longer genomes and, in consequence, a fully multipartite population is established. That critical MOI rapidly increases with the number of segments in the genome, being equal to two for two segments, around 30 for three segments and higher 100 for four and more segments. If the selection pressure favouring shorter genomes is stronger –let us take for instance $\sigma = 0.1$ for the survival of the wt– then critical MOIs decrease (two for two segments, nine for three segments), but still remain above 100 for five or more segments. The possibility that such a high MOI can be reached in nature is clearly unrealistic (not so many viral particles are expected to enter one cell in biological conditions); therefore, if the evolutionary origin of multipartite viruses is due to genome segmentation and competition among genomes of different lengths, no multipartite virus with more than three or four segments should be expected. This result remains unchanged if the advantage of segmented forms is proportional to the volume they occupy, as shown in the Appendix B.

4.4 Discussion

We have presented a model of how viral genomes may become segmented and give rise to a multipartite virus. Provided that mutants with shorter genomes can be produced through deletion events and that these mutants are able to replicate if they receive complementation, evolution of a multipartite variant of the original virus may be the result

of the competition among genomes with different lengths. Two opposite selective pressures determine if the multipartite virus will be able to reach fixation. On the one hand, complementation among segmented genomes requires co-infection of cells by at least one segment of each class. That is only possible if the MOI is high enough. On the other hand, shorter genomes benefit from a reduced degradation rate, which favours division of the genome into smaller segments (Ojosnegros et al. 2011). Other hypothetical benefits of segmentation, such as faster replication or higher fidelity in the replication process have also been considered, with no qualitative changes in the overall results. The main evolutionary regions –coexistence of wt and fragmented forms and extinction of the wt– are essentially unchanged when the production of viral particles is limited by the amount of cellular resources. This leads us to conclude that it is how the complementation rules are implemented what eventually determines the equilibrium state, and not the absolute number of viral particles produced. At the biochemical level, the expression used to evaluate the fitness of the different infection configurations implicitly assumes that gene products affected by segmentation are partially shared, though the genome coding them can use them preferentially. That is, gene products act in *trans* and partly in *cis*. Other models for complementation (Novella et al. 2004) have analysed the case of gene products acting only in *trans*. Compared to our scenario, this latter prescription confers a larger advantage to segmented forms. Hence, segmented genomes would be fixed at lower values of MOI, other parameters being equal.

The case with two genomic segments giving rise to a bipartite virus has been solved in a comprehensive, analytical way. It shows that an evolutionary shift from a single segment virus to a bipartite one is the expected outcome when the MOI during evolution exceedes a critical value. This result explains, from a theoretical point of view, the experimental observations by García-Arriaza et al. (2004), where a bipartite virus was obtained after culture of foot-and-mouth disease virus (a single-particle virus) at a high MOI.

Our results for the bipartite case can be compared to those obtained in previous theoretical scenarios. In a complementation model characterized by hyperbolic growth (Szathmáry 1992), both coexistence and fixation of the bipartite form were found, together with a third regime where the wt virus cannot be invaded by segmented mutants. The assumption of a hyperbolic viral growth is essential for this third regime to exist. However, hyperbolic viral growth requires a tight spatio-temporal coupling between synthesis of replication proteins and genome replication, a situation that is not representative for many viruses (Ball 2007; Belsham 2005). A related work (Chao 1991) studied differences in replicative ability and replication fidelity as the selective pressures driving genome segmentation. Interestingly, the model also considered an additional class of parasitic-like, defective mutants that need complementation by single-segment genomes to be replicated and provide no complementation to the rest of the population. For high enough MOI, it was found that defective mutants were able to invade a population of bipartite virus, driving it to extinction. This phenomenon is conceptually similar to that of *lethal defection* (Grande-Pérez et al. 2005; Iranzo and Manrubia 2009) that was presented in Chapter 2 and has not been considered here for simplicity.

Finally, we have extended our study to the case of multiple (more than two) segments. The main result is that a multipartite virus with a small number of segments can outcompete the single-particle one and get fixed in the population at realistic values of the MOI. According to experimental assays, the MOI in plant infections oscillates between 2 and 13 (González-Jara et al. 2009; Gutiérrez et al. 2010), which would allow selection of multipartite viruses with two and three segments. For a greater number of segments, total segmentation at realistic MOIs should not be expected in the framework of our model.

4.4.1 Multipartite viruses are found only in plants

It is known from experimental assays that the need for complementation reduces the infectivity of multipartite viruses to efficiencies below 10% (González-Jara et al. 2009). One can calculate the probability that a multipartite virus with n segments achieves complementation when infecting cells at MOI m. If we require that this probability reaches a level of 0.1 (one out of 10 cells would effectively get infected), we find that an MOI between once and twice the number of genome segments yields that efficiency, even for multipartite viruses with many segments. This prediction coincides with the MOI values that are indeed found experimentally in plants (Betancourt et al. 2008). Contrary to that, infections in animals are frequently subject to bottlenecks, events for which the MOI becomes severely reduced. The onset of viral infections in animals, as well as intra-host dispersal, are processes that involve a very small number of viral particles (Frost et al. 2001; Kuss et al. 2008), too small for a multipartite virus to achieve efficient infection. Hence, the asymmetry in the host distribution of multipartite viruses may derive from differences in the characteristic MOIs for plants and animals, which in fact are a consequence of the physiological constraints governing viral transmission and dispersal in different organisms.

4.4.2 On the origin of multipartite viruses

Should the evolution of multipartite viruses in nature proceed through competition among segments with different lengths, then a relatively small number of segments is expected in the light of our model. This may have been the case for multipartite viruses belonging to the families *Geminiviridae*, *Secoviridae* and *Bromoviridae* (the former two composed of two segments, and the latter of three segments). Among them, the family *Geminiviridae* is interesting, as it contains bipartite genera as well as non-segmented ones (Lazarowitz 2007).

However, some of the multipartite viruses found in nature present a much larger number of segments. In particular, members of the family *Nanoviridae* are composed of six or eight segments (Gronenborn 2004), so that the MOI that is required to get them fixed becomes of the order of 100 (or even higher). This value is very unlikely to be attained in nature. Therefore, other conceptual frameworks are needed in order to explain the origin of these highly multipartite viruses. Two alternative hypotheses can be proposed at this respect. The first is that viral capsids and genome size have

co-evolved, in such a way that as the genome becomes segmented a smaller capsid is recruited. As the stability of the viral particle depends on the chemical interaction between genome and capsid, it can be expected that if the new capsid fits the size of the genome segments, the relative fitness advantage of a further segmentation will increase. An argument supporting this idea is the fact that viral capsids in nanovirus indeed fit the segment size; so no multiple-segment hypothetical progenitor could be packed into them. A second hypothesis consists of accepting that there has been only one (or maybe two) true segmentation events, favourably selected by a moderate MOI, and the rest of segments have been recruited as genes captured from other viruses. In this respect, we recall that interspecific recombination and gene transfer events are widespread in multipartite viruses (Chare and Holmes 2006; Lefeuvre et al. 2009) and are thought to have played a role in the particular evolution of nanoviruses (Gibbs and Weiller 1999; Hu et al. 2007).

5

Neutral punctuations of mobile elements in prokaryotic genomes

5.1 Introducing transposable elements

Transposable elements (TE) are pieces of DNA that can move within the genome that hosts them, through a process termed transposition. They are widely distributed in prokaryotes and eukaryotes, and in some cases they constitute substantial fractions of the genome. Due to their relative autonomy, proliferative ability, and apparent lack of a useful function, they were considered in the past a paradigm of selfish DNA, i.e. a molecular parasite that proliferates at the cost of the genome it "infects" (Doolittle and Sapienza 1980; Orgel and Crick 1980). Nowadays, the relationship between TE and host genomes is known to be much more complex. Particular TE insertions may be benefitial for the host, for instance by inactivating genes whose expression is no more required (Schneider and Lenski 2004). Furthermore, some TE constitute a vehicle for the exchange of useful genes (e.g. antibiotic resistance genes). TE also facilitate recombination, which promotes genomic plasticity and can accelerate adaptation to fast environmental changes (Kazazian 2004; Treangen et al. 2009; Pál and Papp 2013). Finally, even if TE did not play any benefitial role, genomes often possess regulatory mechanisms that keep TE under control and minimize the risk of possibly deleterious insertions (Kleckner 1990; Chandler and Mahillon 2002; Slotkin and Martienssen 2007).

Insertion sequences (IS) are the simplest form of TE, coding only for the information required for their mobility and frequently found in prokaryotic genomes (Mahillon and Chandler 1998; Chandler and Mahillon 2002). IS are classified in several families, according to their sequence homology and other molecular features. The incorporation of new IS families to a genome takes place through *horizontal gene transfer* (HGT); afterwards, they are vertically inherited. Once an IS has settled into the genome, it can transpose—change its location—through mechanisms whose details depend on the particular IS family. It is a quite general fact, however, that IS elements often increase its number during transposition. That can occur either as a part of the transposition mechanism or if the transposition takes place during genome replication and the element moves from a location that has already been replicated. IS elements can be lost if they are unable to reinsert during transposition—a phenomenon termed excision—or through non-homologous recombination and large deletions. In addition to that, small deletions and other deletereous mutations result in IS inactivation and subsequent erosion and loss.

The study of the abundance of ISs in different genomes has revealed several cases of relative recent IS expansions, which have been related to episodes of host restriction and/or environmental changes (Moran and Plague 2004; Mira et al. 2006). It is a matter of debate wether such IS expansions ultimately lead to the extinction of the host (Wagner 2009) or if they represent transitory punctuations that may even play a role on host evolution (Zeh et al. 2009). A key question that remains unclear concerns the causes and nature of IS expansions: are they a natural outcome of the IS transposition dynamics or the consequence of environmental perturbations? Are reductions in the host population size the main cause of IS expansions? Is it possible that IS stably coexist with their hosts or does it constitute a doomed relationship? In order to solve these questions, a better understanding of IS dynamics is required.

The first works modelling TE dynamics date back to the decade of 1980 (Langley et al. 1983; Kaplan et al. 1985; Moody 1988; Basten and Moody 1991). Inspired by the idea that IS are selfish elements, they depicted a scenario where TE spontaneously tend to proliferate and either host regulatory mechanisms or purifying selection keeps them under control (Charlesworth and Charlesworth 1983; Moody 1988). Due to the limited data on TE abundance and distribuion available at that time, those works either remained mostly theoretical or were mainly addressed to the study of eukaryotic TE (Montgomery and Langley 1983). Two lines of progress have influenced more recent approaches to the modelling of TE. The first one is the discovery of a complex repertoire of interactions between TE and the host genome, and among TE themselves, which has lead to propose an analogy between genomes and ecosystems (Brookfield 2005; Venner et al. 2009). Such an analogy has crystallized in complex models that consider competition and complementation among TE, as well as different degrees of activity and fitness cost to the host genome, reproducing some features of the long-term TE dynamics in eukaryotes (Le Rouzic et al. 2007). The second line of progress is the ever increasing number of sequenced genomes, which has provided us with an unprecedented amount of data on the abundance and distribution of prokaryotic IS. This has made possible the evaluation of a series of hypotheses concerning IS dynamics (Mira

et al. 2006; Wagner 2006; Touchon and Rocha 2007; Cerveau et al. 2011; Bichsel et al. 2012). In particular, Wagner (2006) reported a high homology of IS copies within genomes, which was interpreted on the basis of a fast proliferation dynamics following the arrival of an IS element and ultimatelly leading to the extinction of the host. This view has been challenged by Cerveau et al. (2011), who found a large proportion of IS remnants in *Wolbachia* genomes, revealing that IS proliferation does not necessarily imply extinction. Touchon and Rocha (2007) used an statistical approach to investigate the causes behind IS abundance, finding that it correlates with genome size but not with HGT rate, host pathogenicity or life style. The fitness cost of IS elements was estimated by Bichsel et al. (2012), by comparing a simple model with the genomic data available for IS5. They found that the fitness cost is small enough to consider that, in practice, IS may be neutral or almost neutral for the host genome.

One of the main problems when modelling IS dynamics is the lack of reliable data on the transposition rate. Even though some studies have overcome that by using estimates obtained for complex transposons—such as Tn10—it is unclear if simple IS behave in the same way (Kleckner 1990). Moreover, transposition rate is known to vary with environmental conditions such as stresses (Levy et al. 1993), what makes experimental measures in laboratory conditions difficult to extrapolate to what is happening in the wild. On the other hand, it is usually accepted in the models that the main mechanism for IS loss is excision, with an estimated rate as small as 10^{-10} for Tn10 (Kleckner 1989). Alternatively, reversion of mutants with an inserted IS3-like element suggests an excision rate of the order of 10^{-6} (Christie-Oleza et al. 2008). In comparison to such a wide range of values, it is conceivable that regular deletions and erosion through accumulation of deleterious mutations play also a role in IS loss. The time scale of IS dynamics probably falls in between those of laboratory experiments and nucleotide substitutions, making it difficult to study it in the lab or with conventional phylogenetic approaches (Wagner 2009).

In this chapter, we take advantage of the huge amount of genomic data currently available and study the abundance distributions of 36 IS families in 1811 bacterial chromosomes belonging to 1685 different strains. This allows us to evaluate simple models of IS spreading, obtaining parameter estimations from the maximum likelihood fits. We evaluate the roles of transposition, HGT, deletions, and selection in determining IS dynamics. Specifically, we find that the observed abundances are compatible with a neutral model of IS spreading, where IS proliferation is controlled by deletions instead of purifying selection. Our approach also allows for a detection of recent IS expansions, that can be interpreted as transient events—punctuations—during which the equilibrium coexistence state of IS elements and host breaks down.

5.2 Models of IS spreading and loss

We aim at capturing the main mechanisms that are responsible for the spreading and loss of ISs within and among genomes. To that end, we start by proposing a neutral model that takes into account the following key processes: (a) IS duplication through

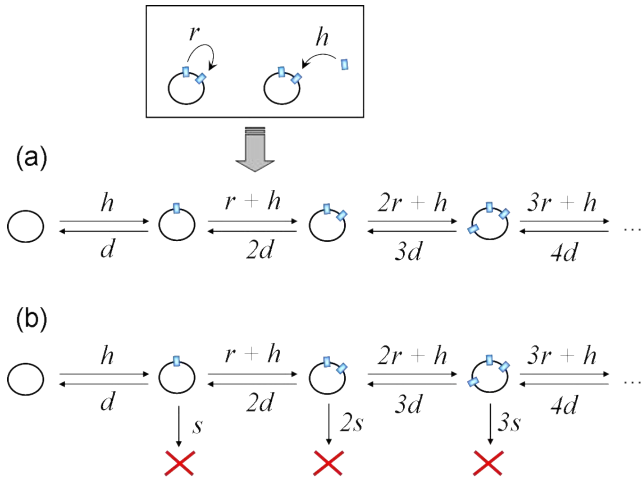


Figure 5.1: *Schematic of the neutral (a) and selection (b) models.* A genome containing k copies of an IS may increase its copy number through duplication of the extant elements, at a rate kr, or through horizontal transfer, at a rate h. Copies are lost through deletion at a rate kd. In addition, if the IS has a fitness cost s, genomes that contain it will die at a rate ks. Model parameters are defined as $\alpha = r/d$ (duplicacion-deletion ratio), $\beta = h/d$ (HGT-deletion ratio), and $\sigma = s/d$ (cost-deletion ratio).

replicative transposition, (b) IS loss through excision, deletion or accumulative deleterious mutations, and (c) IS incorporation through horizontal gene transfer (HGT) from an external source. Notice that the latter is the only mechanism able to introduce new ISs into a genome with no former copies. As an alternative to this neutral model, we also consider the case of IS copies entailing a fitness cost. The processes of duplication, deletion and HGT, complemented with a fitness cost that depends linearly on the IS copy number, define a model of IS spreading with selection.

A schematic of the models is shown in Fig. 5.1. The key processes in the neutral model can be summarized into two parameters: the duplication-deletion ratio (α) and the HGT-deletion ratio (β). The model with selection includes an extra parameter, the cost-deletion ratio (σ). The advantage of working with relative ratios becomes clear, provided the difficulty of obtaining reliable estimates of the actual duplication, deletion and HGT rates. From a formal perspective, working with relative ratios simply amounts to rescaling time, in such a way that the deletion rate sets the time unit. Furthermore, the duplication-deletion ratio can be easily interpreted in terms of the deletion bias, a subject that will be further discussed in this chapter.

Both models can be solved to obtain the expected abundance distribution of an IS family in the long-term stationary state. The models provide, for each IS family, the probability of finding a genome with a given number of copies. By comparing that

with the observed IS abundances one can estimate the value of the model parameters and test whether the neutral model or the model with selection are valid to explain the genomic abundances of ISs.

5.2.1 Neutral model

We study the neutral evolution of the number of copies in the genome as a generalized birth and death process (Fig. 5.1(a)). A complete analysis of this kind of processes applied to the study of proteomes has been carried out by Karev et al. (2002).

The neutral model focuses on a particular IS family in a single genome. Elements belonging to the family can be duplicated through replicative transposition at a rate r and lost through excision or deleterious mutations at a rate d. In addition, new copies can be inserted through horizontal transfer at a rate h. We define the state of the genome as the number of copies that it carries, with no upper limit for such copy number.

A genome with k copies will turn into a state with $k+1$ copies after duplication or HGT. Under the assumption that copies behave independently and HGT rate is a constant, the transition rate $k \rightarrow k+1$ is equal to $kr + h$. On the other side, the transition rate $k \rightarrow k-1$ due to copy deletion is equal to kd. As described in Fig. 5.1, the relevant parameters in this case are α (duplication-deletion ratio) and β (HGT-deletion ratio).

The duplication, deletion and transfer processes reach a stationary state where the probability p_k of finding a genome with k copies is equal to the following expression (Karev et al. 2002):

$$\begin{aligned} p_0 &= (1-\alpha)^{\beta/\alpha} \\ p_k &= (1-\alpha)^{\beta/\alpha}\, \beta\, \frac{\alpha^{k-1}}{k!}\, \frac{\Gamma(k+\beta/\alpha)}{\Gamma(1+\beta/\alpha)} \end{aligned} \tag{5.1}$$

The duplication-deletion ratio, α, plays a central role in the dynamics. If $\alpha > 1$ the number of copies inside the genome increases further and further until it invades the genome. This scenario, termed supercritical, is unrealistic in the absence of natural selection. In contrast, if $\alpha < 1$ duplications are slower than deletions and the copies inside the genome tend to dissapear. In this subcritical scenario the extinction of the IS is prevented by the external contribution of horizontal transfer.

5.2.2 Model with selection

Adding natural selection to the model requires considering a whole population of genomes instead of a single genome. Inside each genome the dynamics of duplication, deletion and horizontal transfer remains the same as in the neutral model. In addition, the IS copy number k determines a fitness cost s_k on the host genome. A schematic of the resulting process is depicted in Fig. 5.1(b). For simplicity we assume that the fitness cost is linear in the number of copies, $s_k = ks$, and define the cost-deletion

ratio $\sigma = s/d$. From a mathematical point of view, the model with natural selection can be seen as a multitype branching process whose stationary behavior is described by its generating matrix A (Moody 1988; Bichsel et al. 2012).

$$\mathsf{A} = \begin{pmatrix} -\beta_0 & 1 & 0 & 0 & \cdots \\ \beta_0 & -\phi_1 & 2 & 0 & \cdots \\ 0 & \alpha+\beta & -2\,\phi_2 & 3 & \cdots \\ 0 & 0 & 2\,\alpha+\beta & -3\,\phi_3 & \cdots \\ 0 & 0 & 0 & 3\,\alpha+\beta & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{5.2}$$

where $\phi_k = 1 + \sigma + \alpha + \beta/k$.

The population evolves according to the following dynamical equation:

$$\dot{\mathbf{p}} = \mathsf{A}\mathbf{p} + \sigma \left(\sum_k k p_k \right) \mathbf{p} \tag{5.3}$$

The stationary composition of the population is described by the eigenvector $\mathbf{p}^*$ associated with the greatest real eigenvalue of A. The stationary abundance distribution p_k^* is equal to the $(k+1)$-th component of $\mathbf{p}^*$. (Note that p_k takes values from $k = 0$, which corresponds to the first component of $\mathbf{p}$). It is worth to mention that the neutral model can be derived from the selection model in the limit $\sigma \to 0$.

5.3 Results

5.3.1 Neutral evolution explains abundance and distribution of ISs

Starting from a dataset of 1811 bacterial chromosomes, we selected 1079 chromosomes belonging to 1014 different bacterial species—multiple strains of the same species were discarded in order to avoid redundancies. IS elements within those chromosomes were detected and classified, and their abundance distributions fitted to both models by means of a maximum likelihood approach (see Appendix C for further details).

The majority of the 36 IS families studied show abundance distributions that fit well to the neutral model (Fig. 5.2 shows a representative example). This assertion is supported by the goodness of fit tests, that render non-significant p-values even if no correction for multiple comparisons is applied. The only exception is IS21 ($p = 0.016$), but it becomes non-significant once corrected for the 36 comparisons. The detailed results of the fits are provided in Appendix C. It is remarkable how a simple, neutral model can explain the data with only two free parameters. We checked if the addition of an extra parameter, namely different HGT rates to empty and infected genomes, can improve the fits. That is not the case for 34 of the 36 families, once corrected for multiple comparisons, thus suggesting that the HGT rates are similar regardless of the genome copy number.

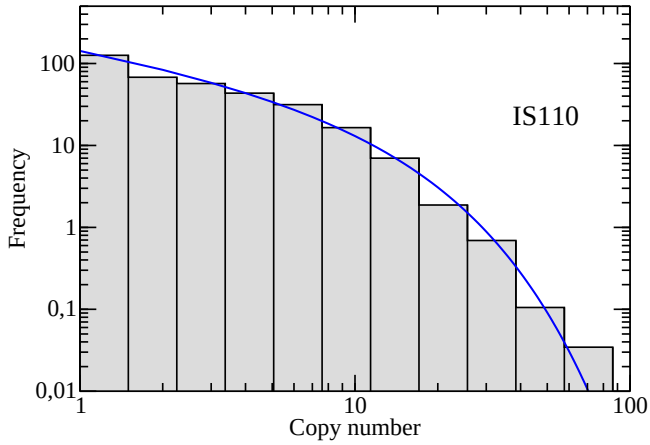


Figure 5.2: *Model fit to the IS110 abundance distribution.* The histogram is the real distribution obtained from the genomic data; the blue line is the fit to the neutral model. For this IS family, the model with selection provides a fit as good as that of the neutral model. The estimated parameters are $\alpha = 0.91$, $\beta = 0.27$ (goodness of fit $p = 0.636$).

Next, we took the values of α estimated in the neutral setting and tried to refine the fits by adding fitness cost and selection. We found that the optimal values of the selection parameter σ lay close to zero. In concordance, selection does not significantly improve the fit for any of the IS families (detailed results in Appendix C). That remains true even if small changes in α are considered. As an alternative, we also explored the selection model by adopting a completely different range of values of α, between 10^2 and 10^3, as suggested by Bichsel et al. (2012). In that scenario, duplications are overwhemingly more frequent than deletions and negative selection is the only factor able to prevent an explosive proliferation of the IS. As in the previous case, no improvement in the fits is observed compared to the neutral model. It is worth to mention that the estimated selection parameter σ is typically tenfold smaller than the duplication-deletion ratio. Taken together, our results show that selection need not be invoked to explain the abundance and distribution of ISs.

5.3.2 Relevance of duplications and HGT for IS spreading

A global analysis of the estimated parameters for the whole set of IS families reveals that most families behave in a strikingly similar way, with α close to 0.9 (Fig. 5.3(a)). Noticeable exceptions are Tn3 and Tn7, for which much smaller values of the duplication-deletion ratio are found. Interestingly, complex regulatory mechanisms have been described for those IS families.

In order to evaluate the relevance of HGT in determining the abundance distributions of ISs we studied the correlation between the HGT rate for different IS fami-

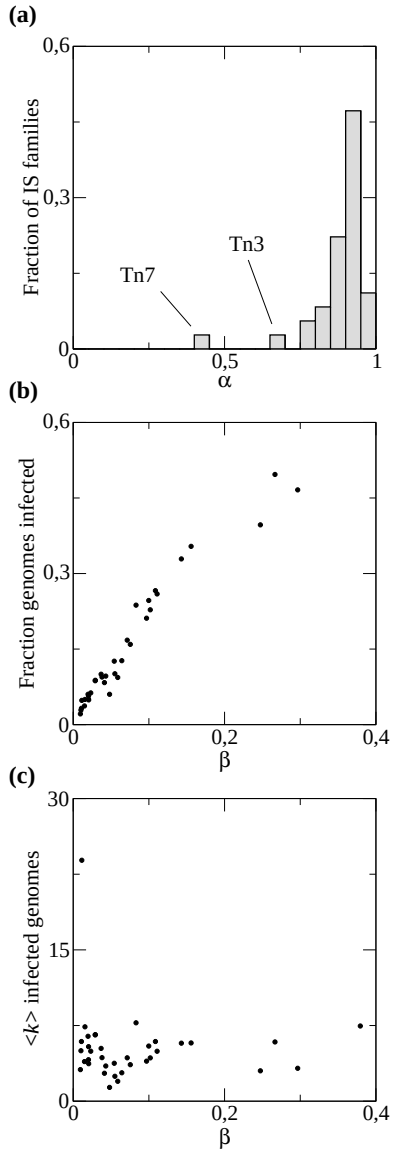


Figure 5.3: *Duplication and HGT play distinct roles in IS dynamics.* (a) Histogram of the estimated duplication-deletion ratios (α) for the whole set of IS families. (b) Correlation between HGT-deletion ratios (β) and the fraction of genomes that contain the IS family ($R = 0.979$, each point corresponds to an IS family). (c) Lack of correlation between HGT-deletion ratios and the mean copy number within genomes with at least one copy ($R = -0.061$).

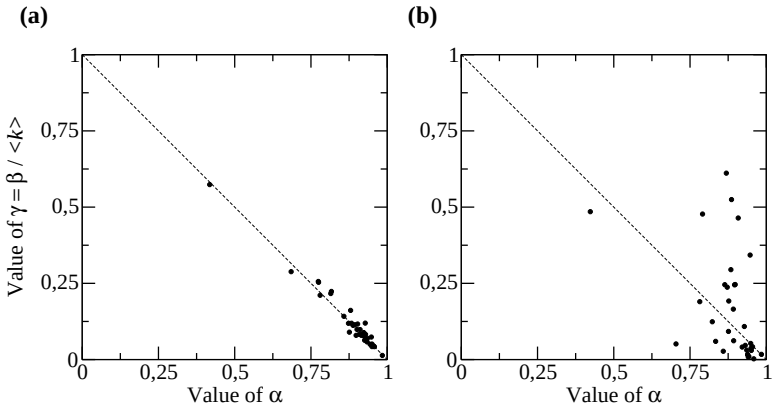


Figure 5.4: *Criticality in IS dynamics.* Each point corresponds to an IS family. The dashed line indicates the critical relation between duplication and gross HGT-deletion ratios: $\alpha + \gamma = 1$. (a) Genomic data obey the critical relation ($R = -0.983$). (b) Simulated data resembling non-equilibrium states do not follow the critical relation ($R = -0.230$).

lies (measured as parameter β) and the fraction of genomes containing such a family (Fig. 5.3(b)). A strong correlation is found ($R = 0.979$), which reflects the fact that the entry of new IS families into the genome totally relies on HGT. In contrast, no correlation is found between the HGT rate and the mean number of copies within genomes with at least one copy (Fig. 5.3(c), $R = -0.061$). This is in agreement with the idea that duplication-deletion processes, rather than HGT, is what determines the copy number once the genome has become "infected" (Touchon and Rocha 2007).

5.3.3 Criticality in IS dynamics

So far in this chapter, we have been taking the HGT rate as an independent parameter. It makes sense to hypothesize, however, that whenever a piece of DNA is incorporated via HGT, the probability that it carries a given IS is proportional to the abundance of such an IS in the host population. In other words, our parameter β, referred to the incorporation of an IS via HGT, could be interpreted in terms of the product of a "gross" HGT rate (γ) times the mean abundance of the IS family in the population $\langle k \rangle$. Conversely, we define the gross HGT-deletion ratio $\gamma = \beta / \langle k \rangle$. Notice that this definition does not affect the process of parameter estimation from genomic data, but only the interpretation of the results.

Figure 5.4(a) shows the relation between the estimated parameters α and γ for all the IS families studied. It reveals a trend of the data to be located close to the line $\alpha + \gamma = 1$ (correlation coefficient $R = -0.983$). Parameters α and γ were estimated independently in order to ensure that the observed trend is not a product of the

fitting algorithm (see Appendix C). If parameters are estimated jointly, the correlation coefficient rises to $R = -0.999$.

It can be proven (see Appendix C) that there exists a critical relation $\alpha + \gamma = 1$ so that the IS dynamics remains stable. If $\alpha + \gamma > 1$, the IS proliferates "explosively", whereas if $\alpha + \gamma < 1$, the IS gets quickly extinct. In this way, if ISs and genomes are to coexist for long periods of time, duplication, deletion and HGT rates must balance according to the critical relation, as it indeed happens.

Interestingly, the critical relation allows for discrimination between equilibrium and IS expansion or regression states. To check for that, we generated datasets by mimicking situations where the HGT rate remains stable while the duplication rate increases (IS expansion) or decreases (IS regression). We found strong deviations from the critical relation, even if the simulated values of α and β were kept inside the previously observed range (Fig. 5.4(b)).

5.3.4 Recent IS expansions are detected as outliers

The models in this work account for the dynamics of ISs in an equilibrium state. The fact that real abundance distributions fit the theoretical curves means that ISs are in equilibrium in most genomes. We can also take advantage of the model distributions to detect outliers: genomes that show an abnormally large copy number for a given IS family (see Appendix C for further details on the detection procedure). From the perspective of the model, outliers can be interpreted as the result of transient imbalances in duplication, deletion and HGT rates, that break down the critical relation.

The search for outliers gave as a result a set of 35 strains (of a total of 1685), that span over a small number of species. For instance, all 12 strains of *Yersinia pestis* are outliers with respect to IS200, and three of them also with respect to IS21. Genomes belonging to the genus *Shigella* (*S. boydii*, *S. dysenteriae*, *S. flexneri* and *S. sonnei*) are overcrowded with IS1, IS3 and IS4a. Other examples are four strains of *Xanthomonas oryzae* (outliers for IS1595, IS5a, IS5b and IS701) and three strains of *Salmonella enterica* subsp. *enterica* (outliers for IS200). The full list can be found in Table 5.1.

5.4 Discussion

Sequencing techniques have experienced a revolution in recent years, providing researchers with an ever growing amount of data on whole prokaryotic genomes. Nowadays, it is becoming possible to exploit all that information in order to address fundamental questions on genome evolution. In this chapter, we combined bioinformatics, statistical analysis and mathematical modelling of genome dynamics in order to obtain a better understanding of the processes that govern the spreading and extinction of transposable elements within genomes. Specifically, we focused on studying the abundance distribution of ISs in prokaryotic genomes, and found that it can be explained as the result of a random process that involves duplications, deletions and horizontal transfer. Remarkably, only two parameters –the duplication-deletion ratio and the

Table 5.1: **List of outlier genomes.**

Strain name	IS families
Acinetobacter baumanii SDF	IS5c (138), IS982 (141)
Bordetella pertussis CS	IS481 (199)
Bordetella pertussis Tohama I	IS481 (215)
Clavibacter michiganensis subsp. *sepedonicus*	IS481 (89)
Microcystis aeruginosa NIES-843	IS630 (116)
Mycobacterium ulcerans Agy99	IS256 (71), ISAs1 (185)
Salmonella enterica serovar Typhi str. CT18	IS200 (27)
Salmonella enterica serovar Typhi str. P-stx-12	IS200 (27)
Salmonella enterica serovar Typhi str. Ty2	IS200 (27)
Shigella boydii CDC 3083-94	IS1 (194)
Shigella boydii Sb227	IS1 (171), IS3 (132)
Shigella dysenteriae Sd197	IS1 (477)
Shigella flexneri 2002017	IS1 (107), IS3 (96)
Shigella flexneri 2a str. 2457T	IS1 (111)
Shigella flexneri 2a str. 301	IS1 (116), IS3 (108)
Shigella flexneri 5 str. 8401	IS1 (110)
Shigella sonnei 53G	IS1 (172), IS3 (94), IS4a (35)
Shigella sonnei Ss046	IS1 (172), IS3 (104), IS4a (32)
Streptococcus suis ST1	IS200 (41)
Xanthomonas oryzae KACC10331	IS1595 (68), IS5a (70), IS5b (80), IS701 (65)
Xanthomonas oryzae MAFF 311018	IS1595 (73), IS5a (83), IS5b (73), IS701 (61)
Xanthomonas oryzae PXO99A	IS1595 (76), IS5a (89), IS701 (95)
Xanthomonas oryzae BLS256	IS1595 (27)
Yersinia pestis A1122	IS200 (66), IS21 (44)
Yersinia pestis Angola	IS200 (99)
Yersinia pestis Antiqua	IS200 (68), IS21 (69)
Yersinia pestis biovar Medievalis str. Harbin 35	IS200 (60)
Yersinia pestis biovar Microtus str. 91001	IS200 (47)
Yersinia pestis CO92	IS200 (64), IS21 (43)
Yersinia pestis D106004	IS200 (58)
Yersinia pestis D182038	IS200 (63)
Yersinia pestis KIM 10	IS200 (52)
Yersinia pestis Nepal516	IS200 (63)
Yersinia pestis Pestoides F	IS200 (54)
Yersinia pestis Z176003	IS200 (62)

Genomes that contain an abnormally high copy number for any IS family, which reveals a non-equilibrium state deriving from recent IS expansions. The number in parentheses is the copy number.

HGT-deletion ratio– are required to recover the real distributions of all the 36 IS families considered. The simplicity of this result is surprising, provided that transposable elements are supposed to be engaged in a broad repertoire of intragenomic "ecological" interactions, that include, among others, competition and complementation (Brookfield 2005; Le Rouzic et al. 2007; Venner et al. 2009). Our analysis suggests, though, that such complex interactions do not play a leading role in determining the dynamics of ISs in bacteria.

By fitting the genomic data to a neutral duplication-deletion-HGT model, we were able to observe two general trends: first, the estimated duplication rates are typically one order of magnitude greater than the estimated HGT rates; second, the HGT rate correlates with the number of genomes that host a given IS family, but does not correlate with the IS genomic abundance. These findings together let us conclude, in agreement with Touchon and Rocha (2007), that transposition and HGT play different roles in the dynamics of ISs. Whereas HGT determines the spreading of ISs across genomes, it only plays a minor role once a genome already contains a given IS. Inside such infected genomes, the abundance of IS copies is mainly driven by stochastic duplications and deletions. When looking at the duplication-deletion ratio, we found that it takes a value slightly smaller than one, which can be easily interpreted in terms of a genomic deletion bias at the level of ISs (Mira et al. 2001). Such a deletion bias makes HGT essential for the long term persistence of ISs: in the absence of an external income via HGT, IS copies tend to be deleted faster than they duplicate and, eventually, they disappear. This mechanism offers a possible explanation to the loss of ISs in organisms whose life conditions limit their HGT rates, e.g. in anciently host-restricted endosymbionts (Moran and Plague 2004).

The duplication rate in our model is restricted to those insertion events that are not lethal for the host genome. As revealed by a recent work (Plague 2010), IS copies tend to be located in places where their interference with gene expression is least, which suggests that not every IS insertion is equally acceptable. From the perspective of a neutral scenario, new insertions are assumed either neutral or lethal (if they interfere with gene expression). In the latter case, the host genome dies shortly after the insertion and contributes no more to population dynamics. Hence, only non-lethal insertions add to the effective duplication rate.

Our results show that purifying selection at the host level needs not be invoked to explain the abundance and distribution of ISs, because the genomic data are fully compatible with a neutral scenario. In fact, the small differences in the distributions derived from neutral and selection models may be not enough for discriminating between both scenarios. There are, however, some clues, that challenge the prevailing role tradicionally given to selection. First, provided that there is a deletion bias, purifying selection is no more essential to control ISs. Second, even if there were no deletion bias and duplications greatly overwhelmed deletions (let us term that *duplication bias*), the values we found for the selection-deletion ratio –typically ten-fold smaller than the duplication-deletion ratio– bring along the possibility that IS control takes place in a weak selection scenario. This same idea had been pointed out by Bichsel et al. (2012),

who studied the abundance distribution of IS5 under the assumption of a strong duplication bias.

In a context of weak selection, the composition of the host population experiences random variations that may allow fixation of slightly deleterious genotypes (Kimura 1968). Hence, when the host population dynamics is taken into account, opposite predictions are derived from deletion and duplication biased scenarios. In the former case, the IS copy number is controlled by deletions, and selection may be neglected, which results in an effectively neutral dynamics. In the latter case, an explosive IS proliferation would be the expected outcome, provided that weak purifying selection is unable to compensate for IS duplications (see Appendix C for technical calculations). Therefore, finding weak selection rates in a duplication biased scenario necessarily implies that host genomes are out-of-equilibrium systems, in their way to becoming fully invaded by ISs (Wagner 2006; Wagner 2009). Since the efficiency of selection increases with the effective population size, it is expected that large host populations are able to fight IS invasion and prevent explosive IS proliferation. Thus, the actuality of a duplication bias could be ideally tested by comparing experiments where the population sizes ranged from small –weak selection scenario– to large –potentially efficient purifying selection. Alternatively, one could compare the genomes of bacteria with different estimated population sizes and look for differences in the abundance of ISs.

At odds with the aforementioned scenario of non-equilibrium proliferative dynamics, our results point towards a stable coexistence of ISs and hosts. Despite the fact that molecular mechanisms of transposition vary (Chandler and Mahillon 2002), all the 36 IS families considered show strikingly similar values of the dynamical parameters. Even more, duplication, deletion, and HGT rates balance according to a critical relation, that allows for evolutionary persistence without explosive proliferation. Such a narrow range of parameter values suggests an implicit role of stabilizing selection in promoting ISs that somewhat behave like mild, persistent parasites (Nuzhdin 2000). In fact, IS mutants that fall below the critical relation are doomed to dissapear; those that surpass it proliferate quickly and –even if they entail a minimal fitness cost– eventually kill their local host populations, thus causing their own extinction (Rankin et al. 2010).

Nonfunctional IS copies constitute a hallmark of the neutral dynamics based on deletion bias: ISs are controlled via deletions, which turn functional IS copies into nonfunctional. In contrast, if ISs are to be controlled via purifying selection, full genomes rich in ISs tend to disappear, without generation of any IS remnants. At this respect, it is worth to discuss the case of *Wolbachia*, a genera of anciently host-restricted endosymbiotic bacteria. *Wolbachia* endosymbionts have reduced genomes ($\sim$ 1Mb) and their effective population sizes are though to be very small. The strains of *Wolbachia* can be divided into two gropus: those associated to arthropods (e.g. *Drosophila melanogaster* and *Culex quinquefasciatus*), and those associated to filarial nematodes (*Brugia malayi* and *Onchocerca ochengi*). Importantly, the arthropod-associated group is known to coinfect hosts and undergo HGT (Werren and Bartos 2001); while the nematode-associated group seems to be transmitted in a strictly vertical way, which greatly limits HGT (Bandi et al. 1998). In agreement with the idea that HGT is required for the maintenance of ISs, only the former group hosts functional IS copies.

The key finding, however, arises when looking for nonfunctional copies. In a comparative analysis, Cerveau et al. (2011) observed that more than 70% of IS copies in arthropod-associated *Wolbachia* are nonfunctional. Those nonfunctional copies span over several IS families, which are also represented in nematode-associated *Wolbachia* with no functional copies. This fact suggests that nonfunctional, fragmentary IS copies may be prevalent in bacterial genomes, even if they have experienced strong reductions in size; and that deletions are an important force leadind to the loss of ISs. In contrast, group II introns—another kind of TE in prokaryotes—display a smaller fraction of fragmentary copies and, possibly, its dynamics is driven by selection (Leclercq and Cordaux 1997).

The neutral dynamics that we present here can give rise to punctuated events of IS proliferation. They occur whenever the HGT, duplication and deletion rates become imbalanced and the critical relation breaks down. We have identified some of those events by applying an outlier detection algorithm on the abundance distributions. According to our analysis, the fraction of such outliers is small, hence confirming that non-equilibrium states are the exception rather than the norm. It is not rare that multiple IS families show expansions within the same genome, which suggests that the mechanisms behind IS punctuations do not lie at the IS but at the bacterial genome level. Indeed, some of the IS expansions that we detected have been associated to episodes where bacteria underwent host restriction (Moran and Plague 2004; Mira et al. 2006). Tradicionally, the reduced efficiency of purifying selection in smaller populations has been invoked to explain such expansion events. There are other mechanisms, though, that may account for IS punctuations in the absence of selection. Transitory alterations in the deletion and HGT rates may play the same role, as well as stress induced down-regulation of host regulatory mechanims that limit IS transposition (Levy et al. 1993; Zeh et al. 2009). In an indirect way, ecological changes –such as host restriction– may imply reductions in the fraction of essential genes (Lan and Reeves 2002; Holden et al. 2009), which leads to a higher probability of IS insertions being non-lethal, and increases the effective duplication rate (Touchon and Rocha 2007).

In sum, our results indicate that the persistence of ISs in bacterial genomes may result from a neutral process, with little role for purifying selection. Most genomes contain IS abundances compatible with an equilibrium state; albeit punctual imbalances in the HGT, duplication and deletion rates –but not necessarily in the host population size– may produce transient IS expansions. Provided the important role of transposable elements in adaptation and genome evolution (Kazazian 2004; Schneider and Lenski 2004; Oliver and Greene 2009; Zeh et al. 2009), understanding the actual causes behind IS expansions becomes an appealing challenge. From an "ecological" perspective, the majority of IS families share closely similar values of the relevant dynamical parameters, which suggests that ISs and host genomes have coevolved towards an state of stable coexistence. The apparent equivalence of different IS families brings to mind the concept of a neutral ecosystem (Volvok et al. 2003). Hence, it would be interesting to further explore the parallelisms between IS dynamics and neutral ecology, which could provide us with novel insights into the processes that rule the architecture of genomes.

6

Coevolution of phages and prokaryotic immunity

The contents of this Chapter are the result of a collaboration with Alexander Lobkovsky, Yuri Wolf and Eugene Koonin, who share the authorship of the ideas and results presented here.

6.1 The CRISPR-Cas immunity system

The ubiquitous arms race between viruses and their hosts to a large extent shapes the evolution of both (Forterre and Prangishvili 2009; Stern and Sorek 2011; Koonin and Wolf 2012). All cellular life forms have evolved numerous, extremely diverse and elaborate antiviral defense systems that occupy a substantial part of the genome, at least in free-living organisms (Haaber et al. 2010; Makarova et al. 2011; Makarova et al. 2013). Although some widespread defense mechanisms of bacteria and archaea, in particular the restriction-modification systems, have been known for many years and thoroughly characterized, recent advances in comparative genomics and experimental study of virus-host interaction have revealed new antiviral defense mechanisms some of which function on novel, unexpected principles (Blower et al. 2011; Leplae et al. 2011; Blower et al. 2012; Makarova et al. 2013; Vasu and Nagaraja 2013). Arguably, the foremost of these new advances is the discovery of the adaptive immunity system that became known as CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR-associated genes) (Deveau et al. 2010; Marraffini and

Sontheimer 2010; Bhaya et al. 2011; Wiedenheft et al. 2012; Koonin and Makarova 2013).

The CRISPR-Cas system employs a unique defense mechanism that involves incorporation of virus DNA fragments into CRISPR repeat arrays and subsequent utilization of transcripts of these inserts (spacers) as guide RNAs to cleave the cognate virus genome (Jansen et al. 2002; Haft et al. 2005; Mojica et al. 2005; Makarova et al. 2006; Koonin and Makarova 2009). Thus, CRISPR-Cas represents bona fide adaptive immunity that until the discovery of this system has not been known to exist in prokaryotes (Goren et al. 2012). However, an important distinction between CRISPR-Cas and animal immune systems is that CRISPR-Cas modifies the host organism's genome in response to infection and hence provides heritable immunity. Thus, CRISPR-Cas is the most compelling known case of Lamarckian inheritance whereby an organism responds to an environmental cue by generating a heritable modification of the genome that provides an adaptive response to that specific cue (Koonin and Wolf 2009). The role of CRISPR-Cas in antivirus defense was initially predicted on the basis of the detection of spacers identical to short sequence segments from virus and plasmid genomes and through comparative analysis of Cas protein sequences (Makarova et al. 2006). At present, such a role has been successfully demonstrated experimentally (Barrangou et al. 2007). In the few years that elapsed since this key breakthrough, the CRISPR research evolved into a highly dynamic field of microbiology with major potential for applications in epidemiology, biotechnology and genome engineering (Bikard et al. 2012; Carroll 2012; Fabre et al. 2012).

The mechanism of CRISPR-Cas is usually divided into three stages: 1) adaptation, when new 30-84 base pair long, unique spacers homologous to proto-spacer sequences in viral genomes or other alien DNA molecules are integrated into the CRISPR repeat cassettes; 2) expression and processing of pre-crRNA into short guide crRNAs; and 3) interference, when the alien DNA or RNA is targeted by a complex of Cas proteins containing a crRNA guide and cleaved within the unique target site (van der Oost et al. 2009; Makarova et al. 2011; Wiedenheft et al. 2012).

Viruses can evade CRISPR-Cas through minimal mutational or recombinational changes in proto-spacer regions. In several experiments, single proto-spacer mutations have rendered CRISPR-Cas ineffectual (Barrangou et al. 2007; Andersson and Banfield 2008; Deveau et al. 2008) although other CRISPR-Cas systems showed less rigid specificity (Semenova et al. 0011). Conversely, hosts can regain antiviral immunity through new spacer additions (Andersson and Banfield 2008; Deveau et al. 2008; Horvath et al. 2008; Heidelberg et al. 2009), thus driving coevolutionary arms races between the mutating virus and the spacer-incorporating host. This arms race apparently can go multiple rounds and takes unexpected turns as demonstrated by the recent finding that certain bacteriophages encode their own CRISPR-Cas system which targets host innate immunity loci, thus turning a defense mechanism into an assault weapon (Seed et al. 2013).

The CRISPR-Cas systems show a remarkably non-uniform distribution among prokaryotes, with nearly all sequenced hyperthermophiles (mostly archaea) but less than 50% of the mesophiles (largely bacteria) encompassing CRISPR-Cas loci (Makarova

et al. 2006; Makarova et al. 2011; Weinberger et al. 2012). In bacteria, the CRISPR-Cas loci demonstrate notable evolutionary volatility, with many cases reported when some of several closely related bacterial strains possessed CRISPR-Cas but the others lacked it (Jorth and Whiteley 2012; Pleckaityte et al. 2012). Numerous cases of apparent horizontal transfer (HGT) of CRISPR-Cas loci also have been reported (Horvath et al. 2009; Chakraborty et al. 2010). Furthermore, the CRISPR-Cas loci have been shown to abrogate acquisition of foreign DNA via HGT (Bikard et al. 2012; Weinberger and Gilmore 2012) and consequently are rapidly lost under selective pressure for horizontal gene transfer as demonstrated by the propagation of antibiotic-resistant CRISPR- strains of Enterococcus faecalis derived from a CRISPR+ progenitor in a hospital environment (Palmer and Gilmore 2010). Rapid acquisition and loss of CRISPR spacers leading to intra-population heterogeneity also has been observed in experiments on both archaeal (Erdmann and Garrett 2012) and bacterial (López-Sánchez et al. 2012) models. Findings like these introduce the more general subject of the fitness cost incurred by the maintenance of the CRISPR-Cas loci (Weinberger et al. 2012; Westra et al. 2012) that in addition to the curtailment of HGT, is likely to involve the strong deleterious effect of autoimmunity caused by an occasional incorporation of proto-spacers from the self DNA (Stern et al. 2010; Paez-Espino et al. 2013).

6.2 Precedent models of CRISPR-Cas dynamics

The arms race between the immune system and viruses, the common events of loss and horizontal transfer of CRISPR-Cas loci and the fitness cost apparently incurred by CRISPR-Cas combine to yield complex evolutionary dynamics. These types of dynamics provide fertile ground for mathematical modeling with a potential to elucidate the interactions between different evolutionary processes and possibly discover unexpected evolutionary regimes. Thus, recently, several mathematical models of CRISPR-Cas-virus coevolution have been developed and studied, using different assumptions and approaches. Essentially, these modeling efforts focused on explaining the striking features of the CRISPR-Cas systems that became apparent through comparative genomic analyses (Han et al. 2013), namely their fast evolution, enormous diversity and old end uniformity.

Kupczok and Bollback used maximum likelihood estimates to analyze purely mechanistic models of CRISPR evolution in which spacers are added and removed stochastically (Kupczok and Bollback 2013). The fits of the model to the observed CRISPR-Cas loci content in collections of closely related bacteria yielded estimates of the spacer addition and deletion rates, and indicated that single spacer deletions were more likely than deletions of groups of spacers.

He and Deem were the first to introduce a stochastic population model with explicit CRISPR dynamics (He and Deem 2010) to analyze the dependence of the spacer diversity on the relative position in the CRISPR array. Their approach has substantial limitations in that the CRISPR cassettes in the model have an unrealistically short fixed CRISPR length (as few as two CRISPR units in the mean field approach) and the im-

munity is decoupled from the virus growth rate. In a follow-up study, the model was extended to include the effect of viral recombination (Han et al. 2013).

A similar approach, but with an explicit coupling between immunity and virus grows rate, also resulted in the observation that leading spacers were more likely to confer immunity and that a small probability of CRISPR failure was irrelevant to the dynamics (Childs et al. 2012).

Haerter et al investigated the conditions under which the diversity of CRISPR loci can be maintained in a spatially inhomogeneous, agent-based model with a small finite number of viral strains (Haerter et al. 2011; Haerter and Sneppen 2012). These studies have concluded that spatial structure was required to explain the observed diversity of the CRISPR loci.

Levin explored the conditions for the maintenance of a costly CRISPR-Cas locus using a parameter-rich mean field model in which CRISPR immunity was parameterized rather than derived from an explicit coevolution dynamics of the spacer and proto-spacer populations (Levin 2010). This study led to the conclusion that there were narrow parameter regimes under which CRISPR-Cas provided bacteria with an advantage over CRISPR-lacking counterparts with a higher Malthusian fitness and that selection for maintaining CRISPR-Cas was weak, suggesting that antivirus defense might not be the principal function of CRISPR-Cas. When the model was compared to experiments that measured the phage/host population dynamics, several apparent disagreements prompted the authors to conclude that the basic assumptions of the coevolutionary arms race models of CRISPR had to be reevaluated (Levin et al. 2013).

Weinberger et al. (2012) aimed to explain the old end uniformity of the CRISPR loci by examining the dynamics of the diversities of the host and viral populations while keeping the total population size fixed in a model that derived immunity directly from the CRISPR locus dynamics. A variant of this model with the additional dynamics of acquisition and loss of the entire CRISPR-Cas locus yielded the prediction of a viral diversity threshold above which CRISPR-Cas became ineffective and was therefore lost due to the fitness cost associated with its maintenance (Weinberger et al. 2012). This study further tested the hypothesis that CRISPR-Cas is nearly ubiquitous in hyperthermophiles but much less common in mesophiles due to the decreased rate of mutation fixation in viruses infecting hyperthermophiles. Simulations that included competition between CRISPR+ and CRISPR- hosts as well as loss and HGT of CRISPR-Cas loci showed that the immunological benefits provided by CRISPR-Cas outweigh the costs under moderate virus diversity that appears to be characteristic of hyperthermophilic environments. These results offered a possible explanation for the higher prevalence of CRISPR-Cas in hyperthermophiles compared to mesophiles and more generally identified the conditions for the evolutionary stability of sensor-type defense mechanisms.

We sought to investigate CRISPR-virus coevolution with as few simplifying assumptions as possible and to this end, incorporated explicit population dynamics, allowing virus and host extinction, unlike the previous model (Weinberger et al. 2012) that formally assumed constant population sizes. This model setting allowed us to exploit the thoroughly characterized stochastic agent based predator-prey framework, which reduces to the Lotka-Volterra (LV) formalism in the limit of the infinite popula-

tion size when fluctuations can be neglected (Hofbauer and Sigmund 1998; May 2001; Grimm and Railsback 2004), for understanding virus-host coevolution in the presence of CRISPR-Cas immunity. The results of the model analysis indicate that CRISPR-Cas stabilizes the stochastic LV system in the intermediate range of viral mutation rate, i.e. leads to extended coexistence of viruses and their microbial hosts. The model further reveals the dependence of CRISPR-Cas efficacy (and accordingly, evolutionary stability) on the population size, spacer incorporation efficiency, number of proto-spacers per virus, and viral mutation rate.

6.3 A CRISPR-Cas model with explicit ecological dynamics

The model developed here aims to reproduce the coevolutionary process shaped by immune interactions between viruses and bacteria or archaea carrying the CRISPR-Cas system, with an underlying ecological dynamics that controls the population of hosts and viruses. The model takes into account the fitness cost of the CRISPR-Cas loci as well as the possibility of their loss and gain via horizontal transfer. Thus, the model is suitable to study the evolutionary and ecological conditions that determine the efficacy of CRISPR-Cas and its long-term fate in the host population.

The analyzed system consists of variable numbers of CAS-positive hosts (N_{b+}), CAS-negative hosts (N_{b-}) and viruses (N_v). Hosts (bacteria or archaea) are abstracted as variable size sets of (possibly non-unique) spacers. Viruses are represented as sets of N_s distinct proto-spacers. The host-virus system evolves according to a stochastic dynamics that is simulated using the Gillespie algorithm (Gillespie 1977). We consider the following events: (i) growth of a CAS- host population with rate N_{b-} (this sets the scale of time); (ii) growth of a CAS+ host population with rate $N_{b+}/(1+c)$, where c is the fitness cost of the CRISPR system; (iii) encounter of viruses and hosts with rates $bN_{b+}N_v$ and $bN_{b-}N_v$ for CAS+ and CAS- respectively; (iv) viral degradation with rate dN_v; and (v) CRISPR-Cas loci horizontal transfer from CAS+ to CAS- hosts with rate $\sigma N_{b+}N_{b-}/(N_{b+}+N_{b-})$. This implementation of horizontal gene transfer as frequency-independent assumes that the DNA exchange mechanisms are saturated. The effect of a non-saturated scenario will be briefly described later in this chapter, in the context of a mean field model.

An encounter between a virus and a host may be immune or productive. An immune encounter occurs if a CAS+ host contains at least one spacer that matches any of the viral proto-spacers. Alternatively, both CAS+ and CAS- hosts can experience "innate" immune encounters (whereby the immunity is provided by defence systems other than CRISPR-Cas that do not depend on spacer acquisition), with a small probability s. Otherwise, the encounter is productive and results in the death of the host and a viral burst of size M.

The model further incorporates genome-level dynamics of the host spacers and viral proto-spacers. Every time a CAS+ host divides, the daughter cell may lose its CRISPR-Cas locus and all the spacers with probability λ. Moreover, single spacers are deleted with probability ℓ (per spacer). New spacers can be incorporated every time an

immune encounter takes place: each proto-spacer of the infecting virus is added to the spacer list of the host with probability a. Finally, the new viruses produced at every viral burst mutate their proto-spacers with probability μ (per proto-spacer). We assume an infinite allele scenario where mutations always give rise to novel proto-spacers.

6.3.1 Parameter setting

As we are dealing with a stochastic, agent-based model, the choice of parameter values is limited by the computational cost. We set the model parameters in such a way that simulation times become affordable while the key properties of the population remain as realistic as possible. We varied the viral burst size M between 2 and 90 while the degradation rate was fixed at $d = 0.5$. This parameter choice provides an equilibrium composition with viruses being 10 to 100-fold more abundant than hosts, which is close to the actual virus to cell ratios observed in various habitats (Breitbart and Rohwer 2005; Suttle 2007; Rohwer and Truber 2009). The mutation rate per proto-spacer μ was varied between 0 and 0.1; the encounter rate b (whose inverse controls the population size) was varied between 10^{-3} and 10^{-4}. The CRISPR-related parameters were set as follows: virus size $N_s = 10$ to 50 proto-spacers per virus, spacer loss probability $\ell = 0.05$ or 0.1 and the incorporation probability a between 0 and 0.1. The probability of innate immunity was fixed to $s = 0.1$. These parameters translate into spacer deletion bias so that in the absence of adaptive immunity, the host loses its spacers. In the region of the parameter space that was chosen to explore, the parameter values combine to yield a realistic range (10 to 100) of the steady state size of the CRISPR cassette (Bhaya et al. 2011).

Simulations start with viral and host population sizes equal to their Lotka-Volterra (LV) equilibrium values (see next section). Viruses are allowed to mutate once prior to the beginning of the simulation, whereas hosts start as CAS+ with no spacers. Simulation results were averaged over 100 independent realizations.

6.4 Results

6.4.1 Effect of CRISPR-Cas on the host-virus system dynamics

We first study the effect of the CRISPR-induced immunity on the dynamics of the host-virus system. As a starting point, we simplified the model by assuming that the CRISPR-Cas loci are constitutively maintained in the host population, i.e. there is neither loss of CRISPR-Cas loci nor HGT and the fitness cost incurred by the CRISPR-Cas system is negligible. In terms of the model parameters, these assumptions translate into $\lambda = \sigma = c = 0$.

To evaluate the effect of CRISPR-induced immunity, it is first necessary to characterize the behaviour of the virus-host system in the absence of CRISPR-Cas. When the population size is large and fluctuations can be neglected, the hosts and the viruses comprise a Lotka-Volterra (LV) system that oscillates around an equilibrium state with N_b^* hosts and N_v^* viruses. As shown in the Appendix D, the equilibrium sizes of the

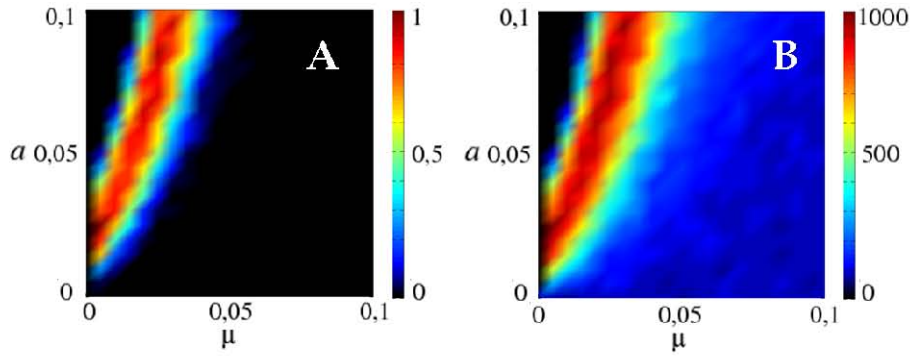


Figure 6.1: *Effect of CRISPR in host and virus survival.* A) Fraction of simulations that survive extinction for at least 10^3 generations (color bar), for varying values of the viral mutation rate μ and the spacer incorporation probability a. The small black region at the upper left corner corresponds to viral extinction driven by adaptive immunity whereas the main black region corresponds to the stochastic extinction of hosts. B) Mean survival time of the population (color bar). The maximum survival time correspond to populations with coexistence at the end of the simulation. Note that the survival time has a peak as a function of μ or a when all other parameters are fixed.

host and virus populations are $N_b^* = d/(b(M - Ms - s))$ and $N_v^* = 1/(b(1 - s))$ respectively, and the period of the LV oscillations is $2\pi/\sqrt{d}$. Because the host and virus populations are finite, either viruses or hosts can become extinct. Simulations of the model in the absence of adaptive immunity ($a = 0$) with $b = 10^{-3}$ show that the mean survival time for the hosts under the chosen parameters setting is approximately 10^2 generations whereas survival probability at $T = 10^3$ generations is negligible. Thus, stochastic extinction (Donalson and Nisbet 1999) of the hosts within a timespan of $T = 10^3$ generations is the expected outcome whenever the CRISPR system is unable to provide antiviral immunity.

We assessed the effect of CRISPR-induced immunity on the host-virus system at varying values of the viral mutation rate μ and the spacer incorporation rate a, after a simulation time of $T = 10^3$ generations (Fig. 6.1A and 6.1B). According to the final fate of the system, three regimes become apparent: (i) viral extinction at low μ and moderate a, (ii) long-term coexistence at an intermediate range of the parameters, and (iii) stochastic extinction of hosts at greater μ values. Viral extinction is fast and occurs if the hosts achieve an average fraction of immune encounters greater than $1 - M^{-1}$ which makes the mean viral yield drop below one per encounter (this regime corresponds to the black, upper left region in Fig. 6.1A and 6.1B). In contrast, host extinction is the result of stochastic fluctuations in the discrete LV model and requires much longer times to occur (main, right region in Fig. 6.1A and 6.1B). Within a rather narrow range of both parameters lies the regime of stable virus-host coexistence (the coloured area in Fig. 6.1A and 6.1B).

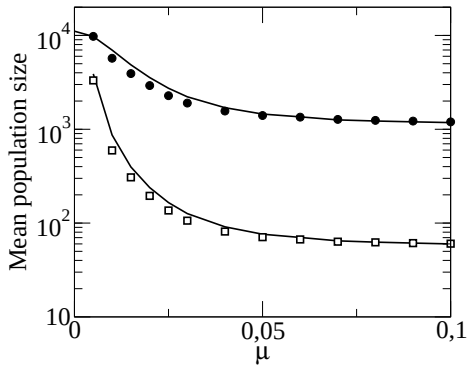
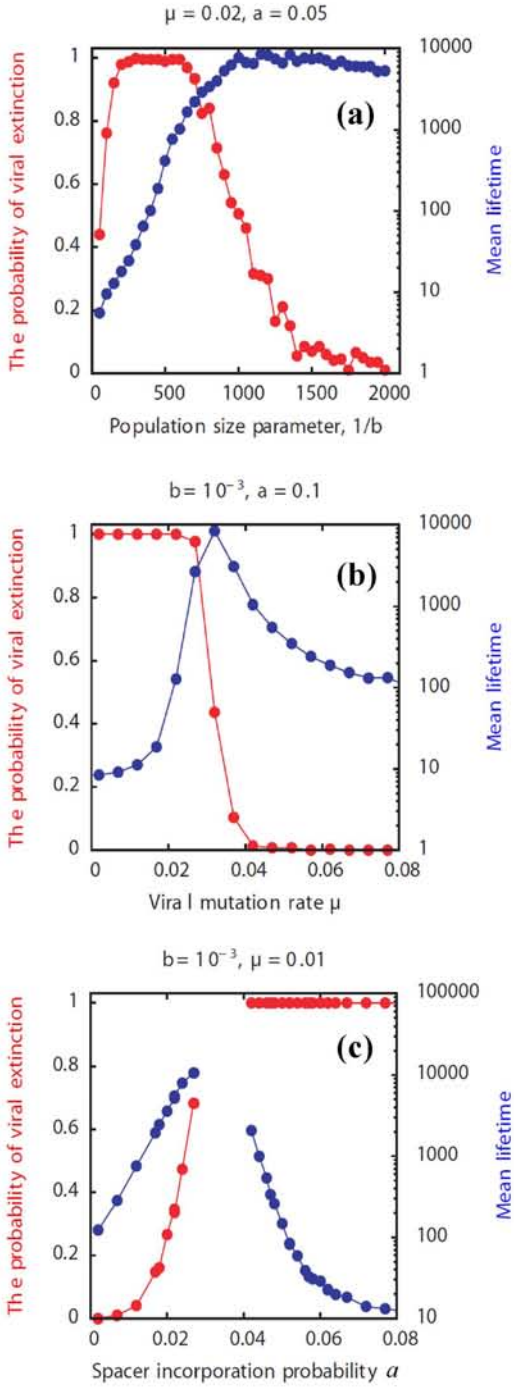


Figure 6.2: *CRISPR immunity affects population sizes.* Average number of hosts (open squares) and viruses (black circles) in the population at different values of μ. The incorporation probability is fixed at $a = 0.05$. The solid curves correspond to the analytical estimates based on the average degree of immunity in each simulation.

An important by-product of the CRISPR-induced immunity is that it increases the host and the viral population sizes (Fig. 6.2). This effect might appear paradoxical at the first glance but it is a direct consequence of the degree of CRISPR-mediated adaptive immunity p_c achieved by the hosts and can be captured by using p_c instead of the innate immunity s in the above expressions for N_b^* and N_v^* (solid lines in Fig. 6.2). A higher level of immunity leads to increased survival of the hosts, and the larger the host population, the more viruses it can sustain. There exists an optimal virus mutation rate that maximizes the mean lifetime of the system before extinction and therefore the total amount of viral particles produced during the infection (Fig. 6.3(b)). This optimal mutation rate is associated with a relatively high level of CRISPR-induced immunity, but not as high as to quickly extinguish the virus. Immunity allows for large populations and long-term coexistence, which translates into sustained production of viral particles. Conversely, virus mutation rates that render CRISPR-Cas ineffective result in a decreased size of the host populations and consequently lead to a low total production of viral particles. In a qualitatively similar manner, both the mean lifetime

Figure 6.3 *(facing page)*: *Effective immunity stabilizes the virus-host system.* (a) Exponential growth of the mean lifetime before extinction indicates that CRISPR stabilizes the stochastic virus-host system when hosts have the upper hand. (b) The constant a transect shows that stabilization is most effective in the intermediate viral mutation regime when viruses and hosts coexist. (c) A transect of Fig. 6.1B for a constant μ also shows that virus-host coexistence and LV stabilization are most pronounced in the intermediate range of spacer acquisition rates a. The gap in the graph corresponds to extremely long lifetimes.


μ = 0.02, a = 0.05
(a)
The probability of viral extinction
Mean lifetime
Population size parameter, 1/b
b= 10⁻³, a = 0.1
(b)
The probability of viral extinction
Mean lifetime
Viral mutation rate μ
b= 10⁻³, μ = 0.01
(c)
The probability of viral extinction
Mean lifetime
Spacer incorporation probability a

of the system and the probability of virus extinction show sharp dependencies on the spacer acquisition rate a, with the longest lifetime at intermediate a values and a steep drop in the lifetime associated with the deterministic virus extinction at higher a values (Fig. 6.3(c)).

Furthermore, the exponential increase of the mean lifetime before extinction with the population size (Fig. 6.3(a)) indicates that effective CRISPR-Cas stabilizes the stochastic LV virus-host system (a linear increase in the life time is expected without the stabilization effect). When viruses coexist with the hosts (see Fig. 6.1), the populations are in a quasi-steady state in which the length L of the CRISPR array, the number N_t of distinct proto-spacers in the viral population and the probability p_c that CRISPR provides immunity fluctuate around their time-average values. We show in Appendix D that the probability p_c that there is a match between a spacer and a proto-spacer in a encounter between a random virus and a random host can be expressed as a function of the ratio L/N_t and the magnitude of the correlation between the relative abundances of proto-spacers and the matching spacers. Fig. 6.4 illustrates that the steady state value of the CRISPR associated immunity p_c is well approximated by

$$p_c = 1 - \left(1 - \alpha \frac{L}{N_t}\right)^{N_s} \tag{6.1}$$

where α is a constant which depends on the strength of the correlation between the relative abundances of matching spacers and proto-spacers. The spacer-proto-spacer correlation reflected by α increases with the burst size M but does not seem to depend on virus size N_s. However, because the ratio L/N_t does not seem to depend on N_s, and N_s appears in the exponent, adaptive immunity p_c grows rapidly with N_s.

To get a handle on the dependence of the CRISPR immunity p_c on the model parameters, we examine L and N_t in Eq. (6.1) separately. In steady state the decay of the CRISPR array due to spacer loss is balanced by the growth due to immune encounters with viruses

$$\begin{aligned} \frac{1}{2}L &= bN_v p N_s a \\ L &= \frac{2N_s a p}{1-p} \end{aligned} \tag{6.2}$$

where $p = s + p_c(1-s)$ is the total immunity, and we used the expression $N_b = 1/(b(1-p))$ for the average viral population size. Equation (6.2) is obtained under the assumption that fluctuations in L and p are small and uncorrelated with each other across a particular population. The empirically computed steady state value of L is consistently above the prediction indicating that, not surprisingly, L and p are positively correlated leading to a larger average L for the same p (see Fig. 6.5).

6.4.2 Effect of CRISPR-Cas on viral diversity

The selective pressure exerted by the CRISPR-Cas system on frequent viral proto-spacers suggests that CRISPR-Cas might directly promote viral diversity. On the other

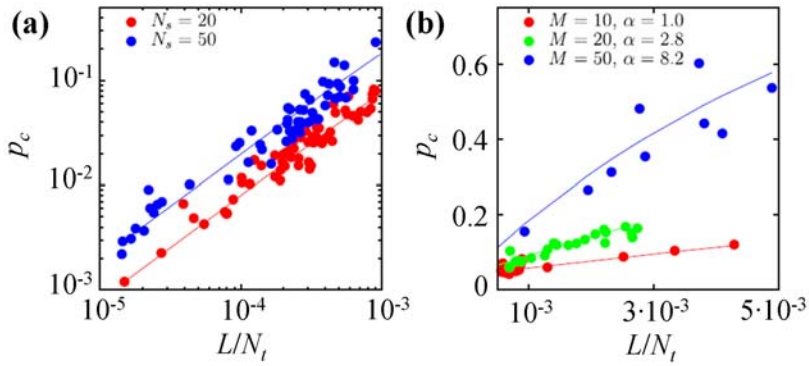


Figure 6.4: *Factors influencing CRISPR immunity.* (a) CRISPR immunity p_c is a function of only the ratio of the length L of the CRISPR array and the total number N_t of distinct viral proto-spacers. The graph is obtained by varying the spacer incorporation probability a and the viral mutation rate μ across the range of virus-host coexistence while keeping the rest of the parameters fixed at $M = 10, b = 10^{-4}, d = 0.5, \ell = 0.05$, $s = 0.1$. Solid lines are the predictions of Eq. (6.1) with parameter $\alpha = 4$ which reflects the strength of correlation between spacers and proto-spacers. This correlation seems to be independent of the number N_s of protospacers per virus. (b) Varying the viral burst size M at a fixed $N_s = 20$ shows that the correlation between the spacers and proto-spacers reflected by the parameter α grows roughly linearly with the burst size M.

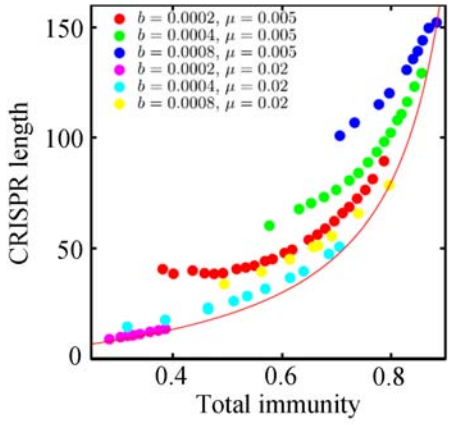


Figure 6.5: .The steady state length L of the CRISPR array computed for a range of parameters. Spacer incorporation probability a is varied in each group of like-colored points. The computed value of L is above the prediction (solid curve) of Eq. (6.2).

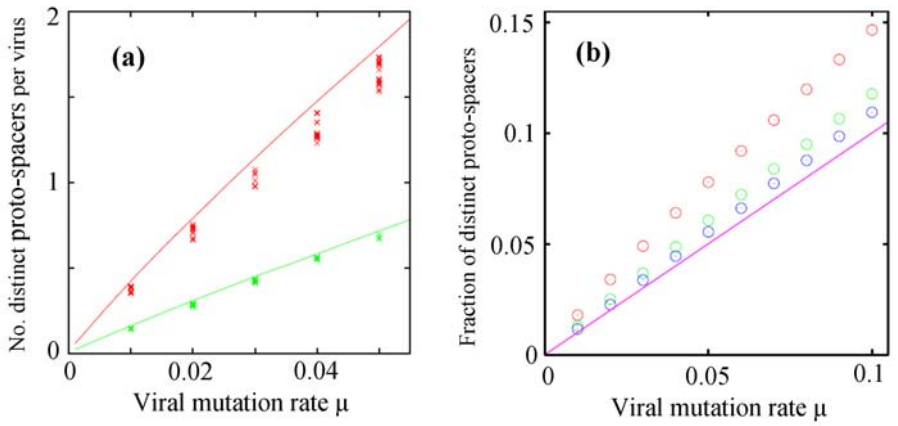


Figure 6.6: (a) *CRISPR does not promote viral diversity*. The number of distinct viral proto-spacers per virus N_t/N_v (symbols, y-axis) computed for a wide range of a and μ and $b = 10^{-4}$ is slightly below what it would be in a freely evolving viral population of the same size (solid curves). Two different combinations of virus and burst sizes are represented (red: $N_s = 20$, $M = 10$; green: $N_s = 10$, $M = 20$). (b) *Mutation rate and burst size determine viral diversity*. The number of distinct proto-spacers N_t in a freely evolving viral population divided by the product of the viral population size N_v and the virus size N_s is proportional to μ (solid line has unit slope) in the limit of the large burst size M. Red, green and blue circles correspond to $M = 15$, 45 and 95, respectively.

hand, new viruses that manage to escape adaptive immunity tend to rapidly proliferate thereby reducing viral diversity. We found that neither mechanism operates and that the steady state number of viral proto-spacers N_t is closely approximated by the diversity in a viral population of the same size evolving in the absence of CRISPR-mediated immunity ($a = 0$) (see Fig. 6.6(a)). Thus, the proto-spacer diversity increases in the presence of CRISPR-Cas only inasmuch as the virus population grows. This counter-intuitive finding is likely the result of the high number of proto-spacers per viral genome which means that the beneficial effect of a mutation in a single proto-spacer is small and accordingly positive selection driving the evolution of new proto-spacers is weak if not negligible.

The diversity of a freely evolving virus population (Fig. 6.6(a)) is described by a remarkably simple expression. In the limit of the large burst size we obtain

$$N_t \approx \mu N_s N_v \tag{6.3}$$

Perhaps counter-intuitively, N_t is a declining function of the burst size M for a fixed viral population size N_v (see Fig. 6.6(b)). This behavior can be explained by noting that when μ is small and the total number of proto-spacers in a viral population is fixed, each

burst of a virus carrying a particular set of proto-spacers produces a relatively greater fraction of these spacers in the whole population and thus results in the reduction of the number of distinct proto-spacer types.

6.4.3 Conditions for the maintenance of the CRISPR system

Motivated by the patchy distribution of the CRISPR-Cas system in prokaryotic genomes, we explored the conditions that govern the maintenance or loss of *cas* genes in the host population. We first asked how the virus mutation rate affects the efficacy of CRISPR-Cas. Eq. (6.1) can be used to derive a characteristic viral mutation rate μ_c at which CRISPR associated immunity p_c is equal to innate immunity s. When the viral mutation rate is much smaller than μ_c, CRISPR immunity dominates over the innate immunity and *vice versa*. When $s \ll 1$ we obtain

$$\mu_c \approx \frac{\alpha L}{s N_v} \approx \frac{4\alpha N_s a}{N_v} \tag{6.4}$$

where we used Eq. (6.2) to obtain the second expression. Eq. (6.4) predicts that the threshold mutation rate of the virus, below which CRISPR-Cas is effective, is proportional to the virus size N_s and the spacer acquisition probability a and inversely proportional to the viral population size N_v. If viruses present a larger target for CRISPR or if hosts are more efficient at incorporating viral genetic material, the viruses have to mutate faster to escape immunity. Conversely, as the viral population grows, the concomitant growth of the proto-spacer diversity renders CRISPR ineffective. In other words, if the viral mutation rate is fixed, there exists a critical viral population size below which CRISPR provides immunity and above which it is useless.

To further investigate the evolutionary dynamics of CRISPR-Cas, we explore the "three species system" that consists of CRISPR+ and CRISPR- hosts and viruses. Here we drop the simplifying assumptions of the preceding sections and assume that the CRISPR-Cas system entails some fitness cost c and that the CRISPR-Cas loci can be lost or horizontally transferred at rates λ and σ, respectively. As a first approach to the problem, let us introduce the mean field approximation that is valid when fluctuations can be ignored and the fraction of immune encounters in CAS+ hosts is assumed to be a constant parameter p. The population of CAS+ hosts (N_{b+}), CAS- hosts (N_{b-}) and virus (N_v) follows the equations:

$$\begin{aligned}
\dot{N}_{b+} &= N_{b+}\left(\frac{1-\lambda}{1+c} - b(1-p)N_v + \frac{\sigma N_{b-}}{N_{b+}+N_{b-}}\right) \\
\dot{N}_{b-} &= N_{b-}\left(1 - b(1-s)N_v - \frac{\sigma N_{b+}}{N_{b+}+N_{b-}}\right) + \frac{\lambda N_{b+}}{1+c} \\
\dot{N}_v &= N_v\big(b(M-Mp-p)N_{b+} + b(M-Ms-s)N_{b-} - d\big)
\end{aligned} \tag{6.5}$$

The analysis of the system of equations (6.5) shows that a minimum degree of CRISPR-induced immunity is required if CAS+ hosts are to survive. That efficacy

threshold, denoted as p_{min}, is equal to:

$$p_{min} = 1 - \left(\frac{1-\lambda}{1+c} + \sigma \right) (1-s) \tag{6.6}$$

If the degree of immunity provided by CRISPR-Cas is smaller than p_{min}, the CRISPR-Cas loci are lost. Horizontal transfer and deletion rates are involved in the expression for p_{min} together with the fitness cost. Thus, deletion bias plays a role equivalent to that of the fitness cost with respect to the maintenance of *cas* genes. Conversely, an enhanced rate of horizontal transfer might compensate for fitness cost. Here we analyse a scenario with equal rates of deletion and horizontal transfer, $\lambda = \sigma = 0.1$, and focus on the consequences of fitness cost. In modelling horizontal transfer, a scenario with saturated DNA exchange was chosen. It is easy to generalize the model to include non-saturated scenarios where the horizontal transfer rate is proportional to the number of hosts (see Appendix D). In such a case, the value of p_{min} that allows for CAS maintenance depends on the population size, with greater p_{min} required in smaller populations.

The results of simulations addressing the maintenance of a costly CRISPR system are plotted in Fig. 6.7. With the fitness cost set to $c = 1$, the fate of *cas* genes in the host population has been studied for varying values of the mutation and spacer incorporation rates Fig. 6.7A). Three regimes can be distinguished: (i) viral extinction, (ii) coexistence of virus and host, with CRISPR-Cas maintained with and (iii) CRISPR-Cas loss. Not surprisingly, these regimes roughly correspond to those obtained in Fig. 6.1A for the case without cost. When CRISPR-Cas is ineffective it is rapidly lost (Fig. 6.7B), whereas the stochastic extinction of hosts takes much longer, especially in large populations. When CRISPR-Cas drives viruses to extinction, the absence of new infections renders CRISPR-Cas useless and eventually causes its loss. That would not be the case if new viruses were introduced stochastically before Cas loss occurred.

The fate of the *cas* genes is determined by the effectiveness of the CRISPR-Cas immune system. With the parameters used in Fig. 6.7, Eq. (6.6) predicts that a minimum fraction of immune encounters $p_{min} = 0.505$ is required in order to retain CRISPR-Cas. Such a degree of immune encounters is achieved if the viral mutation rate is $\mu < 0.03$ (at $a = 0.05$) or $\mu < 0.04$(at $a = 0.1$) (Fig. 6.7C). The coincidence between these values and the boundary of the CRISPR-Cas maintenance region (Fig. 6.7A) supports the idea of an efficacy threshold given by Eq. (6.6).

6.4.4 Population size, fitness cost, burst size and the number of proto-spacers

To study the effect of the population size on the efficacy and evolutionary fate of CRISPR-Cas, we focused on the encounter parameter b that is inversely proportional to the time average population size in the model. By varying parameter b across one order of magnitude, we find that CRISPR-Cas fails to provide immunity in large populations, and as a result, large populations lose CRISPR-Cas loci (Fig. 6.8). This is a direct consequence of the increase in the viral diversity reflected in Eq. (6.3) and the resulting decrease in CRISPR-induced adaptive immunity predicted by Eq. (6.1).

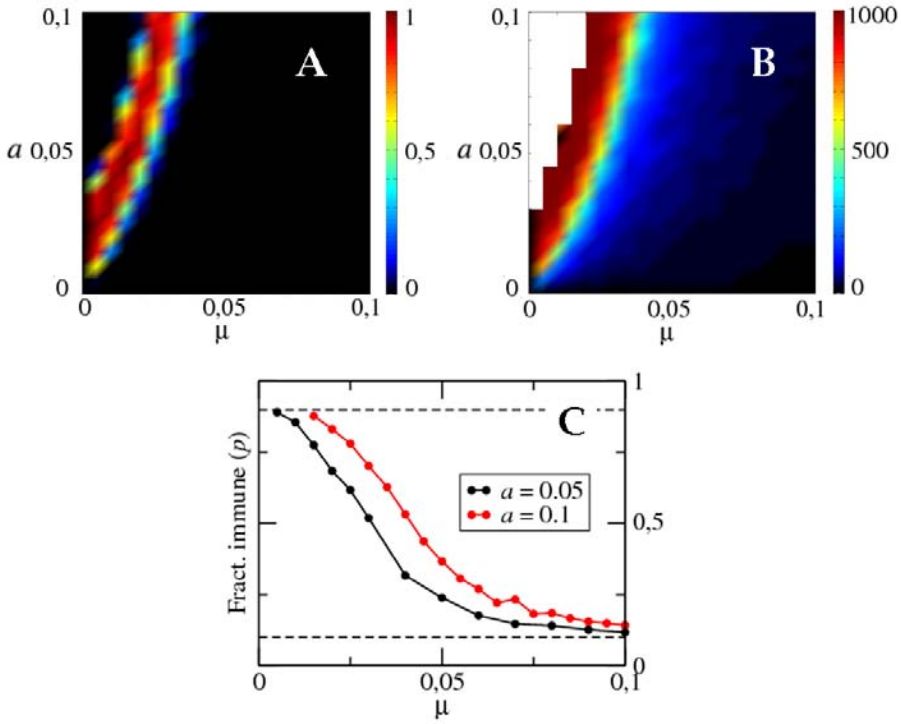


Figure 6.7: *The maintenance of a costly CRISPR-Cas system depends on its effectiveness.* A) Fraction of simulations where host and virus coexist and *cas* genes are maintained after 10^3 generations (color bar), as a function of a and μ. The small black region at the upper left corner corresponds to viral extinction. The main black region corresponds to the loss of *cas* genes. B) The average duration of CAS maintenance in the population. In the white region, where the virus gets extinct, CRISPR becomes useless and *cas* genes are lost unless periodic reinfections occur. C) Fraction of encounters between virus and CAS+ hosts with an immune outcome. The dashed line at the bottom corresponds to the innate immunity $s = 0.1$. The dashed line at the top corresponds to the 90% probability of an immune encounter at which the viral productivity is below one for $M = 10$ and causes viral extinction.

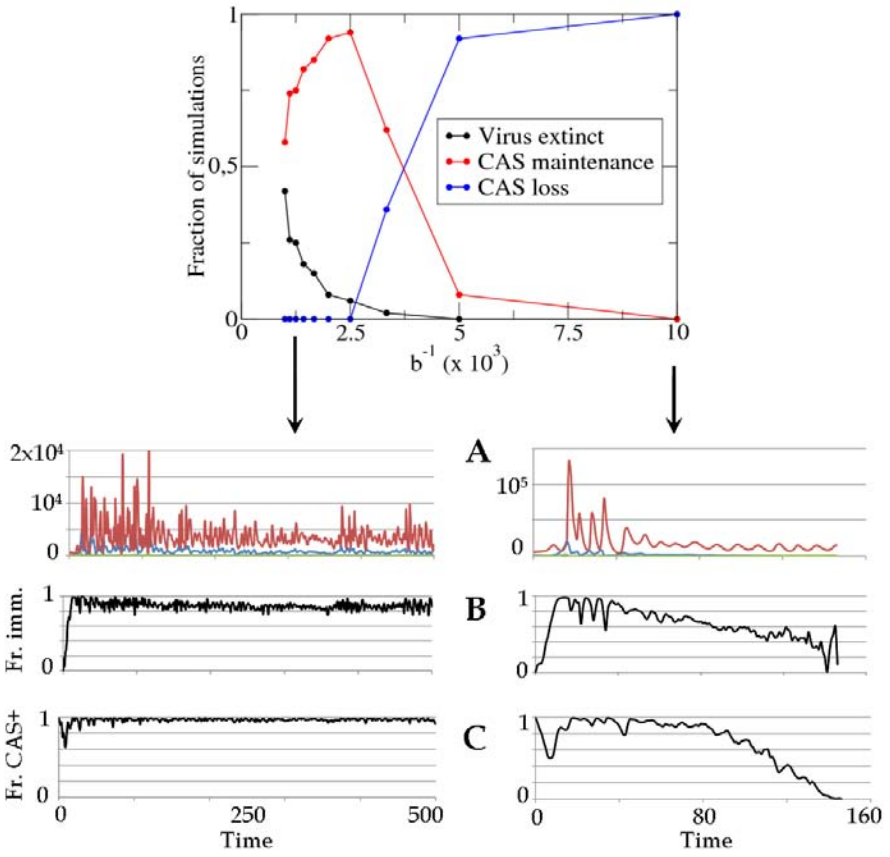


Figure 6.8: *Large population size leads to CAS loss.* Upper panel: fraction of simulations where *cas* genes are retained (red) and lost (blue) as a function of the population size parameter (inverse encounter rate b^{-1}). Lower panels: typical evolution of single realizations in small ($b^{-1} = 10^2$, left) and large ($b^{-1} = 10^3$, right) populations. A) Number of viruses (red), CAS+ hosts (blue) and CAS− hosts (green). B) Fraction of immune encounters. C) Fraction of hosts that conserve CAS.

A closer examination of the explicit dynamics of the model yields further insight into the mechanism of CRISPR-Cas loss (Fig. 6.8). There is an initial phase where the degree of immunity is high in both small and large populations. However, the increasing viral diversity reached in large populations leads to a gradual decrease in the efficacy of CRISPR, resulting in the eventual loss of CRISPR-Cas. Again, it is the increasing diversity of viruses, reflected by the total number N_t of proto-spacers in the viral population, which gradually makes the immunological memory ineffectual.

The effect of other biological parameters on the maintenance of *cas* genes is summarized in Fig. 6.9. The evolutionary outcome does not depend on the value of the CRISPR-Cas fitness cost as long as it is moderately high. Even for a small fitness costs, an ineffective CRISPR system becomes an almost neutral trait that may be lost through neutral drift and population bottlenecks. This feature seems to explain the loss of CRISPR-Cas at small fitness costs and moderate mutation rates (Fig. 6.9A) even when Eq. (6.6) predicts its retention. To summarize, the magnitude of the fitness cost for CRISPR-Cas does not qualitatively affect the outcome of virus-host interaction. The viral burst size does not seem to perceptibly affect the results either (Fig. 6.9B). In contrast, changes in the number N_s of proto-spacers per virus dramatically change the evolutionary fate of CRISPR-Cas. It is easy to see that the greater the number of proto-spacers per virus, the more difficult it is for the virus to escape the immune memory. This dependence translates into a sharp transition in the long-term maintenance of CRISPR when immunity becomes greater than the threshold value p_{min} as the number of proto-spacers increases slightly.

6.5 Discussion

The model described here seems to be more realistic than the previous models of CRISPR-Cas evolution because it includes changing population sizes of both the hosts and the viruses coupled to the explicit dynamics of the CRISPR array and thus provides for explicit analysis of the population dynamics within a stochastic agent based predator-prey framework which reduces to an LV system in the large population limit. This general modelling framework provides for an explicit analysis of the virus and host population dynamics and leads to several non-trivial results.

Although it might seem counter-intuitive, CRISPR-Cas immunity stabilizes the virus-host system for intermediate values of the viral mutation rate, i.e. promotes the long-term virus-host coexistence, rather than leading to the extinction of the virus. Alternatively, if the mutation rate is fixed, this stabilization only occurs for intermediate values of the host and virus population size. In large populations CRISPR-Cas is lost, due to the increase in viral diversity, whereas when populations are small, stochastic extinction reduces the mean lifetime of the system. This observation links the present results to those of our earlier modelling study of CRISPR-Cas in which we have shown that the immunological memory provided by CRISPR-Cas can be effective only at moderate virus diversities (Weinberger et al. 2012). This result has been suggested to explain the ubiquity of CRISPR-Cas in hypethermophiles as opposed to the substan-

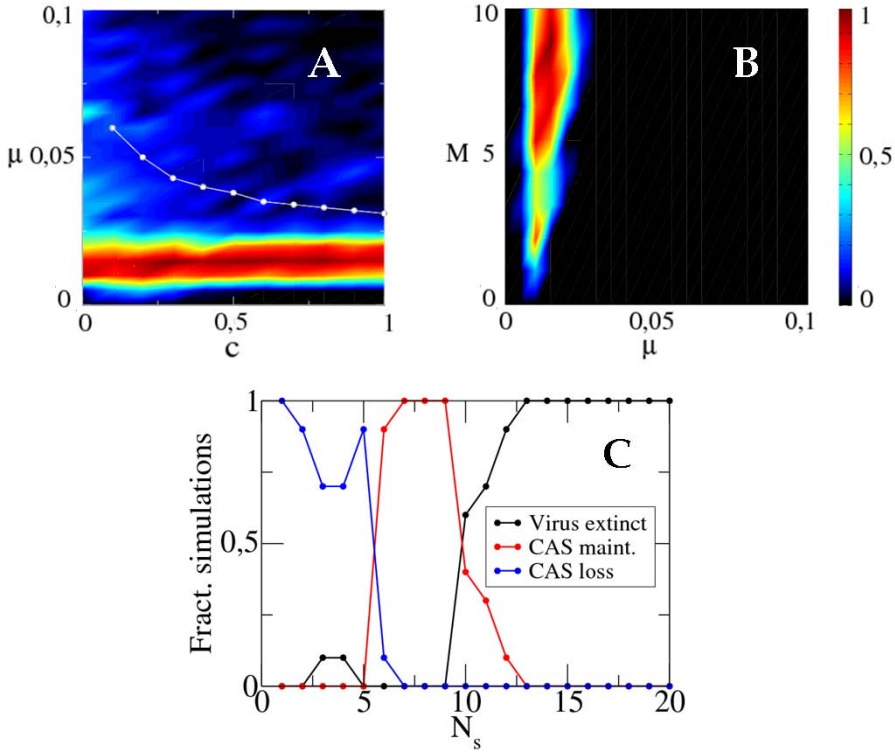


Figure 6.9: *The effect of the fitness cost (A), burst size (B) and number of proto-spacers per virus (C) on the evolutionary fate of CRISPR-Cas.* A) Fraction of simulations where CRISPR-Cas is retained after 10^3 generations, as a function of the mutation rate μ and fitness cost c. Cases of stochastic host extinction before CRISPR-Cas is lost are also included. The white dots correspond to the threshold values predicted by Eq. (6.6). B) Fraction of simulations with CRISPR-Cas conservation at different values of the mutation rate and the burst size (M). The black band to the left is due to viral extinction. The main black area corresponds to CRISPR-Cas loss. C) Fraction of simulations with viral extinction (black), CRISPR-Cas conservation (red) and CRISPR-Cas loss (blue) at increasing values of the number of proto-spacers per viral genome. Parameter values: $a = 0.05$, ($\mu = 0.01$ in C).

tially lower prevalence in mesophiles given that the rates of mutation fixation are much lower in hypthermophiles (both hosts and viruses) than in mesophiles (Zeldovich et al. 2007; Drake 2009). The present findings offer a complementary and simpler perspective on the difference in the prevalence of CRISPR-Cas between hypethermophiles and mesophiles: extremely large populations in which CRISPR-Cas apparently becomes useless are known only for mesophiles whereas hyperthermophiles typically exist as smaller populations (Fuhrman 2002; Schrenl et al. 2003; Whitaker and Banfield 2006; Wilmes et al. 2009) in which CRISPR-Cas immunity is predicted to be efficient.

Because the mean lifetime of the system (LV-stabilization) has a sharp peak at intermediate values of the viral mutation rate, the maximum total virus yield before either viruses or hosts become extinct is reached at a certain intermediate value of the viral mutation rate. Such a value is neither as high as to make CRISPR-Cas ineffective nor as low as to allow CRISPR to extinguish the virus.

Counter-intuitively, in the present model, CRISPR-Cas immunity does not specifically promote viral diversity in the sense of driving positive selection for emergence of new spacers, presumably because the selection pressure on any single spacer is too weak. In fact, when the viral mutation rate is sufficiently low, the clonal bloom dynamics results in slightly reduced viral diversity compared to a freely evolving (without pressure to escape adaptive immunity) population. Virus diversity is directly proportional to the mutation rate and the population size, and accordingly, CRISPR-Cas promotes virus diversity only inasmuch as the immunity leads to an increase in the population size.

The model results indicate that the efficacy of CRISPR increases with the number of proto-spacers per viral genome. Because maintenance of CRISPR-Cas is a threshold phenomenon, a small decrease in can lead to CRISPR-Cas loss. This finding might explain, at least in part, why the Protospacer-Associated motif (PAM) the presence of which in a viral genome is essential for a proto-spacer acquisition has a low information content (i.e. consists of only two or three nucleotides) (Mojica et al. 2009; Fischer et al. 2012): it is critical for the host to be able to use multiple proto-spacers. The fact that the specificity in the selection of proto-spacers exists at all, might reflect the trade-off between the benefits of the utilization of multiple proto-spacers for efficient immunity and the avoidance of autoimmunity. Clearly, the mechanisms of self-nonself discrimination in CRISPR-Cas require further, detailed exploration.

IV

Conclusions

7

Conclusions and open questions

Natural History [...] is either the beginning or the end of physical science.
—Sir John Herschel.

7.1 Lethal defection in persistent infections

The concept of fitness is central to evolutionary theory. Individuals in a population survive and reproduce according to their fitness, so that fitness determines –together with chance– the evolutionary fate of different genetic variants. However, an operative definition of fitness is not always straightforward: several traits may contribute to fitness, in a way that vary with environmental conditions and evolutionary constraints. As soon as the simplest evolutionary toy models are left behind, multiple selection pressures differentially affect traits contributing to fitness. The interplay between selective pressures and fitness traits gives rise to novel, sometimes non-intuitive phenomenology, as it is the case described in Chapter 2.

A successful viral cycle comprises multiple steps, which depending on the enviromental conditions may result in different selection pressures. In its simplest form, viral survival requires cell infection, protein expression, genome replication and virion assembly. In persistent infections, the virus stays for long periods of time inside the same cell, thus the selective pressure on infectivity (the ability to infect new cells) is relaxed. Moreover, if viral proteins are shared inside the cell, defective genomes that are unable to produce viable proteins can survive provided they are accompanied by

the wt. In these conditions, wt and defective viruses compete inside the cell in equal conditions, which may result in the stochastic extinction of the former and fixation of a defective virus. Eventually, this leads to the extinction of the viral population, because even in persistent infections the virus must infect new cells from time to time. This phenomenon is termed lethal defection.

Lethal defection is just an example of the complex outcomes arising from evolution under multiple selective pressures. It requires infections to be persistent and defective genomes to be unable to provide complementation or establish productive infections. In contrast, the process of genome segmentation described in Chapter 4 can take place if infections are lythic and defective genomes remain infective and capable of mutual complementation. Thus, two apparently unrelated phenomena –lethal defection and genome segmentation– may share a common origin in the basal production of defective genomes, with physiological details and environmental conditions making the difference.

The model in Chapter 2 is a toy model, and as such, far too simple to account for all the potential subtleties involved in lethal defection. However, its simplicity allows for an easy understanding of the stochastic extinction process and emphasizes its conceptual significance: the intermittent lack of selection on a trait may lead to its loss. If such a trait becomes essential from time to time, its loss implies the eventual extinction of the population. These lethal-defection-like phenomena may jeopardize a population if the timescale in which a neutral trait is lost is faster than the timescale in which such a trait becomes essential.

Mutation rate, MOI, population size and cell lifespan during persistent infections are key factors that control the aforementioned timescales. It can be expected that, in natural conditions, viruses have evolved to avoid stochastic extinction by adjusting mutation rates and other biological parameters. In doing so, each time a virus-free cell population becomes infected at a low MOI, the viral population gets rid of defective genomes.

The effect of the mutation rate and the MOI in the sensitivity of RNA viruses to lethal defection has been recently studied by Moreno et al. (2012). They treated four different viruses with small to moderate doses of the mutagen 5-fluorouracil (FU) at different MOIs. Interestingly, they found that the antiviral activity of FU was more pronounced at low MOI in the case of the negative strand RNA viruses, while there was no dependence with the MOI in positive strand picornaviruses. They interpreted such a different behavior on the basis that positive strand viruses show a lower tendency to establish interactions in *trans* (i.e. to share proteins inside the cell). Thus, an extended model of lethal defection that included coinfections, variable degrees of protein sharing, and multiple classes of defective interfering genotypes, was proposed to account for their experimental results. Such a work shows how the relatively simple idea of lethal defection can be developed into more complete models, which combined with laboratory experiments contribute to the design of novel antiviral strategies.

7.2 Optimal drug combination in antiviral therapies

An area of research where the evolutionary response of viral populations to multiple selective pressures plays a central role is the development of novel antiviral therapies. Let us see why. Single antiviral drugs impose single selective pressures that viruses face (with relative ease) by acquiring resistance mutations. In contrast, the use of additional drugs brings along a scenario with multiple selective pressures that, if properly managed, complicates viral adaptation. In such a line of thinking, the appropiateness of sequential versus combined protocols involving two antiviral drugs can be systematically explored by means of simple evolutionary models that take into account the action mechanism of each drug. Such models are especially appealing as a guide to the design of preliminary *in vitro* assays, where the absence of an immune system and structural complexities improves their predictive ability.

In experiments comparing sequential versus combination therapies, a very simple model can be derived if several conditions hold. First, replication mechanisms do not include provirus phases (as retroviruses) or latency steps (as herpesviruses), which would require a careful evaluation of time delays. Second, the viral load decreases when the therapy is applied, so that competition for resources relaxes, and resource limitation does not need to be considered in the model. Third, the multiplicity of infection (MOI) is low, which might result in diminished complementation or interference by defective genomes.

Under such conditions, intracellular viral dynamics can be modeled as a series of genome replicative cycles, with each cycle resulting in several copies made from the templates obtained in the previous one. Thus, the exact meaning of a replicative cycle depends on the replication mechanism of the virus: for single-stranded RNA (ssRNA) genomes, the replicative cycle refers to the synthesis of multiple genomic strands from each complementary strand; whereas for double-stranded DNA (dsDNA) genomes with semiconservative replication, the replicative cycle is just the semiconservative replication itself.

The hypotheses above, together with the known action of the two antiviral drugs, can be synthesized in a few dynamical equations that allow predicting the response of the viral population to different protocols and drug doses. As a first result, therapies involving two similar drugs (two inhibitors or two mutagens) are more efficient when administered in a combined way. However, if an inhibitor and a mutagen are used, the sequential protocol may be preferable depending on the drug doses and clinical criteria (maximal reduction of viral titer versus prevention of viral resistance), as schematically depicted in Fig. 7.1 (Perales et al. 2012). The root of this dose-dependent behavior lies at the double role that mutagens play. The exposure of the virus to mutagenic drugs increases the mutation rate. In lythic infections at low MOIs, no complementation takes place and defective mutants behave as lethal (no phenomenon resembling lethal defection takes place). At the same time, an increase in the mutation rate accelerates the appearance of mutants that are resistant to the inhibitor, thus leading to a nonlinear interaction between the two drugs that could yield unwanted effects (Iranzo et al. 2011).

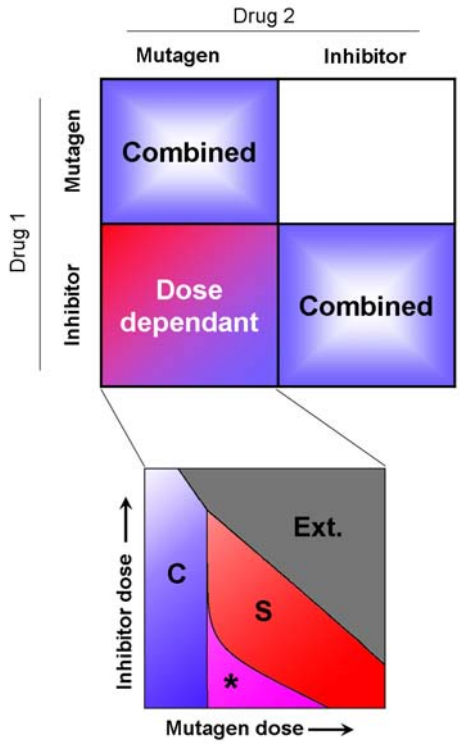


Figure 7.1: *Optimal protocol choice in multidrug therapies.* According to theoretical models, the optimal protocol for drug administration in multidrug therapies depends on the action mechanism of the drugs. If two inhibitors (or two mutagens) are to be used, their efficiency is optimized through simultaneous administration (combined protocol). However, for mixed inhibitor-mutagen therapies the optimal protocol depends on the nature of the virus and drug doses: the lower panel shows this dependence for foot-and-mouth disase virus (FMDV) with guanidine as inhibitor and ribavirin as mutagen. Blue region (C): combined protocol performs better than sequential administration; red region (S): sequential protocol—first inhibitor, second mutagen—performs the best; gray region (Ext): at high drug doses, the virus becomes easily extinct with both protocols; pink region (*): combined protocol is more effective in reducing the viral titer but it produces resistant mutants with higher probability than the sequential protocol. The sequential protocol when the mutagen is provided before the inhibitor is not considered because its performance is always worse than the others.

According to Figure 7.1, the optimal protocol for the administration of an inhibitor and a mutagen depends on drug doses. In addition, the dose combinations for which a sequential or a combined protocol is preferred vary depending on biological properties of the virus. This means that for different viruses the drug doses that make the sequential therapy more effective (red regions in Fig. 7.1) may change. In practice, a given protocol is suitable if it becomes advantageous for a wide range of drug combinations. In the particular case of a sequential inhibitor-mutagen protocol, it is expected to be more suitable when applied to viruses with a small to moderate yield and a replication mechanism that produces many copies from the same template (e.g., ssRNA viruses with replication via minus strands that each produce many plus strand RNA copies).

One of such viruses is hepatitis C virus (HCV), a leading cause of chronic hepatitis, cirrhosis, and liver cancer in the Western world (Rosen 2011). Preliminar experiments with HCV, treated with the mutagen ribavirin and a survey of available inhibitors, seem to confirm that sequential administration performs the best for a wide range of drug doses. The quantitative assessment of viral parameters and the subsequent application of a mathematical model to anti-HCV therapy is, at present, ongoing work.

Future perspectives include the translation of different viral replication mechanisms into dynamical equations similar to those in Chapter 3. It is known, moreover, that mutations that render the virus resistant to drugs may also entail a great fitness cost in the absence of such drugs (Cong et al. 2007; Sierra et al. 2007). A simple modification of the model equations shows that such a fitness cost usually makes the sequential protocol more profitable, which calls upon an evaluation of multidrug therapies in the light of resistance cost. Finally, the simulation of an *in vivo* situation certainly entails additional difficulties such as the interaction with the immune system, or environmental and individual characteristics. For the moment, the predictions of any model, once tested *in vitro*, should be taken only as a rough guide to apply one or another administration protocol and to infer minimum drug doses in *in vivo* assays.

7.3 Genome segmentation and the origin of multipartite viruses

Multipartite viruses are formed by a variable number of genomic segments packed in independent viral capsids. This fact poses stringent conditions on their transmission mode, demanding, in particular, a high MOI for successful propagation (many viral particules must enter the same cell in order to ensure that at least one representative of each segment is present). Due to their enigmatic nature and the unclear benefits of the multipartite strategy, the origin of multipartite viruses represents an evolutionary puzzle.

Experiments on viral evolution carried out by García-Arriaza et al. (2004) report an instance of spontaneous transition from a non-segmented virus to a bipartite form. Such a finding provides some clues on how multipartite viruses could have originated. A tentative mechanism can be summarized as follows. First, shortened, defective genomes spontaneously appear as a result of large deletions. Some of those defective genomes, that still conserve a subset of functional genes, constitute putative segments.

Segments are capable of replication, provided they are complemented with the genes they lack; moreover, their gene products can complement other segments. The accumulation of genomic segments gives rise to a nascent multipartite form of the virus, i.e. a set of segments that collectively code for all genes. Such a set of segments and the full virus compete for replication inside and transmission across cells. Under certain conditions, the segments reach fixation and the original virus becomes extinct: the transition from a full-genome to a multipartite virus has taken place.

The first step of the process, namely the generation of defective genomes that may constitute putative segments, is extensively documented in the literature (Bangham and Kirkwood 1990; Roux et al. 1991). On the other hand, subsequent competition between the full virus and the complementary set of segments is a key step that requires careful evaluation. The outcome of such a competition critically depends on two factors: (1) the MOI, and (2) the selective forces favouring genome segmentation. Even though the nature of the latter remains a matter of discussion, enhanced virion stability, faster genome replication, and greater mutational robustness have been proposed as reasonable advantages of a multipartite strategy. For any of them, a minimum value of the MOI is required so that segments outcompete the full-genome virus. We found in Chapter 4 that such a critical MOI dramatically increases with the number of segments, in such a way that, under realistic conditions, the aforementioned mechanism can only explain the origin of multipartite viruses with a small number –two or three– of segments.

In the light of our results, alternative hypotheses are required to explain the origin of multipartite viruses of the family *Nanoviridae*, which are composed of six or eight segments. One possible mechanism is based on the idea that capsids and genome size coevolve, the former decreasing in size as the latter becomes segmented. If the advantage of a multipartite strategy relies on a greater stability of the viral particle, it can be expected that once the capsid gets adjusted to the segment size, the chemical interaction between genome and capsid increases and so does the relative fitness advantage of further segmentation (Fig. 7.2(a)). A second mechanism would consist on the occurrence of one or two "true" segmentation events followed by the acquisition of additional segments through the capture of useful genes from other viruses (Fig. 7.2(b)). At this respect, the search for horizontal gene tranfer in the available viral genomic data raises as a complementary approach to test such a hypothesis. While the strength of our theoretical approach lies in its generality and simplicity, the combination of models, bioinformatic tools and phylogenetic analysis become appealing when it comes to unveil the evolutionary history of particular viral families.

An aspect that has not been considered yet is the reversibility of the segmentation process, i.e. the recovery of a full genome from a multipartite virus. This was indeed observed by García-Arriaza et al. (2006), in the context of the experimental transition to a bipartite virus already mentioned. In their experiment, recombination between both segments gave rise to a full genome, that could be selected for by imposing a very low MOI. It is interesting to test if the model developed in Chapter 4 is compatible with such an experimental observation, and what does it predict in natural conditions. To that end, let us denote with r the recombination rate, and assume that the evolutionary

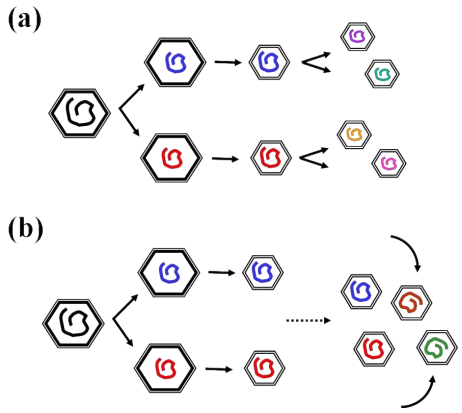


Figure 7.2: *Alternative mechanisms for the origin of highly multipartite viruses.* (a) Multipartite viruses with many segments may be the result of a coevolutionary process between capsid and genome, such that as the capsid gets adapted to the segment size, the fitness advantage of further segmentation increases. (b) Some of the segments may be not the result of a segmentation event, but useful genes captured from an external source.

conditions are largely favourable to the bipartite form. In such a case, the fraction of full genomes (resulting from recombination) in the population is $p_{wt} = \sigma r/2$ (here σ is the relative degradation of the wt). According to García-Arriaza et al. (2006), reasonable values for the parameters are $\sigma = 0.1$ and $r = 2{\cdot}10^{-4}$, which renders $p_{wt} = 10^{-5}$ (below their experimental detection limit). If the MOI is reduced to a very small value, no cell will be infected by more than one particle at a time. Therefore, all single segments will be removed, as they will not find complementing counterparts. In contrast, if at least one wt viral particle manages to infect a cell, it will grow and eventually, after several generations, it will become detectable. If the number of cells to be infected is N and we assume that exactly one virion infects each cell, the probability that at lest one wt genome is recovered becomes $1 - (1 - \sigma r/2)^N$. The number of cells used by García-Arriaza et al. (2006) in the experiments was $N = 10^6$, so that the probability of recovering the wt was higher than 99.99%. As a result, it can be concluded that reversion to the wt at low MOI is almost sure in laboratory conditions. However, the situation in nature is probably quite different, since the number of susceptible cells to be infected during natural bottlenecks is much smaller (indeed, the concept of a natural bottleneck is related to that of a small number of chances to start an infection). For instance, if N is reduced to 1000 susceptible cells, reversion probability drops below 1%. In the long term, viral capsid and genomic segments experience a coevolutionary process—the capsid becomes smaller to fit the segment size—, such that a full genome can hardly be packed into a single capsid. Due to the fact that capsid shapes are discrete (Luque et al. 2010), the coevolutionary process leading to capsid downsizing is

discontinuous. After such a process has taken place, reversion of a natural multipartite virus to a non-segmented form is no more possible.

It must be noted that once multipartite viruses have been generated, and given that the process is not reversible in natural conditions, they can persist at MOIs much smaller than those required for their fixation. This makes possible that multipartite viruses with many segments persist: even if the present framework does not explain the origin of *Nanoviridae*, once they have appeared (through whatever mechanism) their survival is guaranteed at moderate (realistic) MOIs. Still, such moderate MOIs are only attained in plants, which explains why multipartite viruses are not found infecting animals.

From a broader perspective, the framework here proposed for genome segmentation in multipartite viruses is formally equivalent to the Black Queen Hypothesis (BQH) for reductive evolution in prokaryotic genomes (Morris et al. 2012). The BQH states that some *Cyanobacteria* strains have selectively lost genes whose products are provided by other species in the microbial community. The loss of genes coding for extracellular proteins becomes favourable if there is a cost associated to their expression and somebody else already produces (and shares) the protein. Eventually, the BQH predicts a labour distribution in the microbial community that is analogous to the distribution of functions among different segments in multipartite viruses. The population of segments inside a cell is, therefore, equivalent to a community where each segment is specialized in a single function and all functions must be performed in order to complete a successful infection cycle. But the reciprocal is also true: as well as a minimum MOI is required for the multipartite virus to appear and persist, the survival of a microbial population under the BQH requires that representative samples of the population travel together when spreading to new areas. This example shows how some of the ideas developed in viral evolution may be applicable to the ecology of communities and vice-versa. It seems that even if the scales vary greatly, the relationships among the members of an ecosystem, let it be genomic, viral, microbial or eukaryotic, remain surprisingly similar.

7.4 Dynamics of transposable elements on prokaryotic genomes

Transposable elements (TEs) are widespread in genomes. Formerly considered detrimental selfish elements, nowadays they are also though to contribute to genome plasticity through promoting recombination and the interchange of genetic material (Kazazian 2004; Oliver and Greene 2009; Werren 2011; Pál and Papp 2013). At a small scale, their insertion can alter gene expression; at a greater scale, they facilitate large genomic rearrangements that change the architecture of the genome. Hence, understanding the dynamics of TEs is relevant to the study of genome evolution.

It has been postulated that TEs in a genome form a complex ecosystem, with different classes of TEs playing distinct ecological roles. For instance, transposition of defective TEs requires complementation by functional TEs, and different TEs may compete for the cellular resources –namely binding to certain DNA-interacting molecules–

needed for transposition. From this perspective, the dynamics of TEs in a genome would resemble some of the features observed in other ecosystems, such as populations of viral quasispecies. Yet, a remarkable difference arises from the fact that, since TEs live inside a genome, they may be subject to a selection pressure at the level of the host genome.

In Chapter 5 we have investigated the dynamics of 36 IS families (the simplest kind of TE in prokaryotes) by comparing their abundance distributions with the predictions of various evolutionary models. Surprisingly, we found that a very simple model with as few as two parameters was enough to reproduce the observed distributions. At odds with the *a priori* expectation, neither complex "ecological" interactions (i.e. complementation, competition, interference, etc.) nor purifying selection at the host genome level are required to explain the overall abundances of ISs. In turn, IS abundances seem to be the result of a neutral process involving duplications, deletions, and HGT.

Our results are compatible with a scenario where the IS dynamics is biased towards deletions. In such a scenario, the selfish proliferative tendency of ISs is counteracted by deletions and inactivating mutations, which take place at a greater rate than duplications. This differs from the traditional view, stating that IS explosions are prevented by purifying selection. Although further work is required to discern between both hypotheses, we point out that: (1) purifying selection alone, if weak—as our data and some authors suggest—, may be insufficient to control ISs even in the absence of HGT; and (2) the large number of nonfunctional IS elements detected in some genomes may be the hallmark of a deletion-biased dynamics.

The fit of the genomic data to the model gives us estimates for the duplication, deletion, and HGT rates. It is worth to mention that, even if their transposition mechanisms vary at the molecular level, the 36 IS families display quite similar values of the estimated dynamical rates. Such a finding suggests that some kind of stabilizing selection is behind the values we observe. Indeed, duplication, deletion, and HGT rates balance according to a critical relation that allows for long term coexistence of ISs and their hosts. This is the case for most genomes studied, although a minority of them show signs of being out of equilibrium. In the framework of a deletion-biased IS control, imbalances among the factors that govern the IS dynamics trigger transient episodes of IS expansion, which result in a punctuated dynamics for ISs.

Nothing has been said in Chapter 5 about IS diversity at the genome level. Some preliminar results on that issue are described in the next lines, although they are still a matter of further study. For the sake of simplicity, we characterize the IS diversity in a genome by the number of different IS families that the genome contains. The intragenomic abundance of each family is not considered, just their presence or absence. Therefore, the term "family abundance" within this paragraph refers to the number of genomes where a given family is present. The distribution of genomic diversity, i.e. the fraction of genomes that host a given number of different families, is plotted in figure 7.3. Randomization of the data shows that the observed distribution does not correspond to a random allocation of families (see details in Appendix C). Indeed, the observed distribution shows an increased proportion of genomes that are richer or poorer in families than expected.

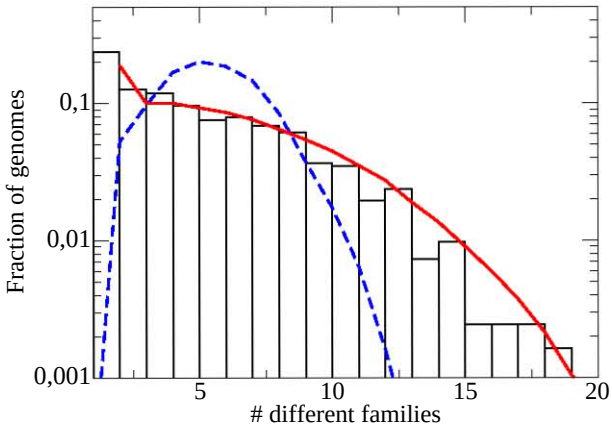


Figure 7.3: Distribution of genomic diversity: fraction of genomes that contain a given number of different IS families. The histogram corresponds to the real data. The dashed blue line is the expected distribution for a random allocation of families. The red line is the distribution generated by a preferential acquisition process with parameter $q = 0.75$.

In order to shed light into the factors that may cause the observed distribution we looked for possible correlations between genome sizes and diversity, as well as co-occurrences between pairs of IS families. If genomic diversity correlated with the genome size, larger genomes would host a greater number of families than expected by chance, thus causing a deviation in the diversity distribution. However, no significant correlation of this kind seems to be present (see Appendix C). In contrast, the study of family co-occurrences reveals a nontrivial pattern: co-occurrences among abundant families are less abundant than expected, while rare families show some significant co-occurrences (a null model is obtained through redistribution of families, while keeping constant the original family abundances and the genomic diversity distribution, as explained in Appendix C). The co-occurrence pattern is depicted in Fig. 7.4. Such a systematic pattern is somehow intriguing, since its dependence with the family abundance seems to point at a general underlying cause rather than at specific interactions between IS families.

It is remarkable that both, the diversity distribution and the family co-occurrence pattern, may be explained by a simple model of preferential acquisition of families – analogous to a preferential attachment process. In such a model, explained with detail in Appendix C, the probability that a genome receives new IS's via HGT is proportional to the number of families it hosts. Such an idea can be easily applied to the neutral model by writting $\gamma(x) = \gamma(1 + qx)/Q$, where x is the number of families in the genome and $Q = 1 + q\langle x\rangle$ is a normalization factor. Normalization assures that after averaging over all genomes the observed HGT rate remains the same as in the "basic" neutral model. Simulations of the infection process under this assumption

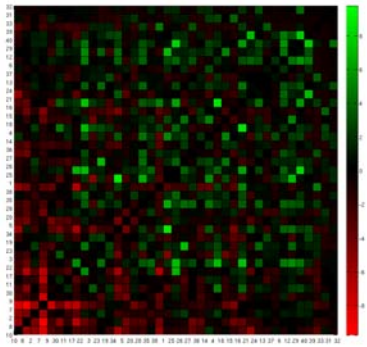


Figure 7.4: Standarized co-occurrence matrix between pairs of IS families. For each pair, a co-occurrence index is calculated based on the cosine distance (see Appendix C). By simulating a null model, the mean and standard deviation of each co-occurrence index are obtained. The figure shows the standarized values of the co-occurrence indexes, after substraction of the simulated mean and division by the standard deviation. Families are ordered according to their abundance (from more frequent to rarer), while numerical labels in the axes correspond to the alphabetical ordering of families. Red (green) indicates fewer (more) co-occurrences than expected.

produce diversity distributions that resemble the real one when q takes values around 0.75. Moreover, a co-occurrence pattern similar to the one previously described is generated. In any case, the biological interpretation of a preferential acquisition pattern, or even the possibility that it reflects some degree of selection that is not captured by the analysis of single abundance distributions, remains to be investigated.

Finally, a complete characterization of the IS dynamics would require to quantify the actual dynamical rates –duplication, deletion and HGT–, for which there are no accurate estimates yet. We circumvented this difficulty by using relative ratios instead of absolute rates. Nevertheless, a better knowledge of the absolute rates would be desirable for a series of reasons. First, it would allow a numerical estimation of the fitness cost associated to IS elements; moreover, it would inform us about the feasibility of studying the IS dynamics in the laboratory; finally, it would set a timescale for IS-related events, that could be used to detect genomes whose ISs are out of equilibrium and decouple isolated episodes of IS expansion from those due to fast genome-scale evolution. Our modelling approach to IS dynamics provides a starting point for the determination of the absolute dynamical rates, as it predicts how the IS composition of two closely related genomes will diverge in time. A future line of work will consist of comparing the divergence on IS composition with the genetic distance for a series of related prokaryotic genomes. Such an approach seems promising for connecting the

timescales of IS-related events and neutral mutations. Ultimately, that would shed light on which is the pace of transpositions in the context of the molecular clock.

7.5 Prokaryotic adaptive immunity through CRISPR-Cas system

A stochastic, agent-based mathematical model was developed in Chapter 6 to explore the coevolution of the prokaryotic adaptive immunity system, CRISPR-Cas, and lytic viruses. The analysis of the model shows tha CRISPR-Cas immunity can stabilize the virus-host coexistence, rather than lead to the extinction of the virus. In the model, CIRSPR-Cas immunity does not specifically promote viral diversity, presumably because the selection pressure on each single proto-spacer is too weak. However, the overall virus diversity in the presence of CRISPR-Cas grows due to the increase of the host and, accordingly, the virus population size. Above a threshold value of total viral diversity, which is proportional to the viral mutation rate and population size, the CRISPR-Cas system becomes ineffective and is lost by the hosts due to the associated fitness cost.

The previous modelling study carried out by Weinberger et al. (2012) has suggested that the ubiquity of CRISPR-Cas in hyperthermophiles—organisms that thrive in extremely hot environments—, which contrasts its comparative low prevalence in mesophiles—those growing best in moderate temperatures—, is due to lower rates of mutation fixation in thermal habitats. The findings exposed in Chapter 6 offer a complementary, simpler perspective on this contrast through the much larger population sizes of mesophiles compared to hyperthermophiles, because of which CRISPR-Cas can become ineffective in mesophiles.

The efficacy of CRISPR-Cas sharply increases with the number of proto-spacers per viral genome. This finding might explain the low information content of the Protospacer-Associated Motif (PAM) that is required for spacer acquisition by CRISPR-Cas because a higher specificity would restrict the number of spacers available to CRISPR-Cas and so hamper the immune response. The very existence of the PAM migth reflect the trade-off between the requirement of multiple spacers for efficient immunity and avoidance of autoimmunity. Clearly, the mechanism of self-nonself discrimination by CRISPR-Cas requires further, detailed exploration.

V

Appendices

A

Materials and Methods of Chapter 3

This Appendix contains the experimental Material and Methods relative to Chapter 3, as well as some mathematical calculations about the model exposed there. The experiments described in the first and second sections were carried out by Dr. Celia Perales at the laboratory of Prof. Esteban Domingo (Centro de Biología Molecular Severo Ochoa), with whom we collaborated in this work.

A.1 Cells and viruses

The origin of BHK-21 cells and procedures for cell growth in Dulbecco's modification of Eagle's medium (DMEM), and for plaque assays in semisolid agar have been previously described (Domingo et al. 1980; Sobrino et al. 1983). FMDV C-S8c1 is a plaque-purified derivative of serotype C isolate C1 Santa Pau-Sp70 (Sobrino et al. 1983). An infectious clone of FMDV C-S8c1, termed pMT28 was constructed by recombining into a pGEM-1 plasmid subclones that represented the C-S8c1 genome, as described (García-Arriaza et al. 2004; Toja et al. 1999). Thus, FMDV pMT28 used in the experiments is the progeny of infectious transcripts that express the standard FMDV C-S8c1. To control for the absence of contamination, mock-infected cells were cultured and their supernatants were titrated in parallel with the infected cultures; no signs of infectivity or cytopathology in the cultures or in the control plaque assays were observed in any of the experiments.

A.2 Treatment with ribavirin (R) and guanidine hydrochloride (GU)

A solution of GU in DMEM was prepared at a concentration of 50 mM, sterilized by filtration, and stored at 4°C. A solution of R in PBS was prepared at a concentration of 100 mM, sterilized by filtration, and stored at −70°C. Prior to use, the stock solutions were diluted in DMEM to reach the desired concentration. For infections of BHK-21 cells with FMDV in the presence of GU, no pretreatment of the cell monolayer with GU was performed. For infections in the presence of R, cell monolayers were pretreated during 7 h with 5 mM R prior to infection. After addition of FMDV and washing of the cell monolayers, infections were allowed to continue in the presence of a combination of [GU+R] or sequential passages, consisting of a first passage in the presence of GU and a second passage in the presence of R. For the combination treatment, the infections were carried out at an MOI of 0.4 PFU/cell. For the sequential treatment the initial infection in the presence of increasing concentrations of GU was carried out also at an MOI of 0.4 PFU/cell. The second infection in the presence of 5 mM R was carried out at an MOI of 1.1, 2.0 $\times 10^{-2}$, 2.2 $\times 10^{-4}$ and 1.2 $\times 10^{-4}$ PFU/cell, for GU 3, 6, 12 and 20 mM GU, respectively. Infections in the absence of GU, R or a combination of GU+R, and mock-infected cells were maintained in parallel; no evidence of contamination of cells with virus was observed at any time.

A.3 Complete solution of Equation 3.1

The dynamics of the model system can be written in compact form through the vector $\mathbf{n}(g)$, whose components are the number of individuals in each of the classes v and V after g replication cycles,

$$\mathbf{n}(g+1) = m\mathsf{A}\mathbf{n}(g) \tag{A.1}$$

with A being the transition matrix of the system,

$$\mathsf{A} = \begin{pmatrix} i\beta & 0 \\ i(\alpha-\beta) & \alpha \end{pmatrix} \tag{A.2}$$

and where we have defined $\alpha = 1 - w$ and $\beta = 1 - \mu - w$. This dynamical system can be exactly solved for the initial condition $\mathbf{n}(0) = \{S_0, 0\}$ to yield the population of each viral class after g cycles,

$$\begin{aligned} v(g) &= S_0(im)^g\beta^g \\ V(g) &= S_0 im^g \left(\tfrac{\alpha-\beta}{\alpha-i\beta}\right)[\alpha^g - (i\beta)^g] \end{aligned} \tag{A.3}$$

from where the exact expression for the total number of viable elements is obtained (Eq. [3.2]).

A.4 Approximate analytic expression for $C_{\mathcal{CS}}$

The expression obtained from Eq. [3.3] yields the dependence between w and i in an essential non-algebraic way, so it can be only numerically solved. However, if we assume that $\mu \ll 1$, an expansion in powers of μ yields the following approximate dependence for the points on curve $C_{\mathcal{CS}}$:

$$w_c = 1 - \frac{k\gamma}{k\gamma - i^G + m^G(1-w_0)^G\,(i^G + \gamma\mu_0/(1-w_0))}\,, \tag{A.4}$$

where $\gamma = [i + i^G(iG - i - G)]/(1-i)$. Given the amount of mutagen i, the sequential treatment causes a larger decrease in the viral titer for values of $w > w_c$, while the combined treatment is more efficient for $w < w_c$. The preferred therapy changes as well if the value of w is fixed and the amount of inhibitor increases. For values of i close to one the combined treatment is better, while the sequential treatment will be preferred above a certain amount of inhibitor. Curve [A.4] has two important limits, absence of inhibitor ($i \to 1$) and large doses of inhibitor ($i \to 0$). In the former case, both treatments become equivalent for values of $w \to 1$ (see Fig. 3.3(f)), a situation where all genomes produced under replication would be non-viable. Actually, $w = 1$ cannot be empirically tested, since the complete extinction of the population occurs at values of w below one, as will be shown. On the other hand, there is a saturation effect when the amount of inhibitor is very high, in the sense that additional decreases in w diminish the viral titer but do not change the preference for one or another therapy:

$$w_{i\to 0} \equiv \lim_{i\to 0} w_c = \frac{m^G w_0 (1-w_0)^{G-1}}{m^G w_0 (1-w_0)^{G-1} + 1}\,, \tag{A.5}$$

which is independent of i.

B

Analysis of the multipartite virus model

This appendix details the mathematical developments supporting results and statements in Chapter 4. In the first section we obtain the fixed points for the evolution equation corresponding to pure populations, to two coexisting populations, and to all three coexisting viral forms, and study the stability of these solutions. The formal expression for σ_{crit} results from a condition of stability on the point where $\Delta 1$ and $\Delta 2$ coexist and the wt is absent. Next, we develop the analytical form of σ_{crit} for the situations where the infection configuration at a given MOI follows a Poisson or a multinomial distribution. In the third section we show how the evolution equation can be generalized to the case of replication rates depending on the segment length, to the situation where segments can be lost through mutation (thus leading to an increased replication fidelity of shorter types), and to a scenario where the productivity of viral particles per cell is constant. In section B.4, we sketch how the equations can be applied to a multipartite virus by means of the three-partite case. Section B.5 discusses the mathematical form of the relationship between genome length and degradative advantage.

B.1 Solutions of the evolution equation

Let us begin by explicitely writing all terms involved in the evolution equation when differential degradation is the only selective pressure favouring single-segment mutants. We recall that

$$\mathbf{p}_{n+1} = Z^{-1}\,\mathsf{D}\sum_{a,b,c}\Pr(a,b,c\mathbf{p}_n)\,\mathsf{M}_{a,b,c}\begin{pmatrix}a\\b\\c\end{pmatrix} \tag{B.1}$$

where

$$\mathsf{D}=\begin{pmatrix}1&0&0\\0&1&0\\0&0&\sigma\end{pmatrix}\quad;\quad \mathsf{M}_{a,b,c}\begin{pmatrix}a\\b\\c\end{pmatrix}=\begin{pmatrix}\min\{a,b+c\}\\\min\{b,a+c\}\\c\end{pmatrix}\equiv\begin{pmatrix}f_{\Delta 1|a,b,c}\\f_{\Delta 2|a,b,c}\\\sigma^{-1}f_{wt|a,b,c}\end{pmatrix} \tag{B.2}$$

and Z is a normalization factor

$$Z=\left\|\mathsf{D}\sum_{a,b,c}\Pr(a,b,c|\mathbf{p}_n)\,\mathsf{M}_{a,b,c}\right\|_1 \tag{B.3}$$

The explicit form for the fitness matrices in concordance with our definitions of conditional fitness $f_{i|a,b,c}$ is

$$\mathsf{M}_{a,b,c}=\begin{cases}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} & \text{if } c>|a-b|\\ \begin{pmatrix}1&0&0\\1&0&1\\0&0&1\end{pmatrix} & \text{if } c<|a-b|\,,\ a<b\\ \begin{pmatrix}0&1&1\\0&1&0\\0&0&1\end{pmatrix} & \text{if } c<|a-b|\,,\ a>b\end{cases} \tag{B.4}$$

For the calculation of fixed points and their stability, in the remaining of this section, we will use the evolution equation in the replicator form, which reads

$$p_i^{(n+1)}=\frac{\langle f_i\rangle^{(n)}}{Z^{(n)}}\,,\quad i\in\{\Delta 1,\Delta 2,wt\} \tag{B.5}$$

and where, for the case of diferential degradation,

$$\begin{aligned}\langle f_{\Delta 1}\rangle^{(n)}&=\sum_{a,b,c}\Pr(a,b,c|\mathbf{p}_n)\min\{a,b+c\}\\ \langle f_{\Delta 2}\rangle^{(n)}&=\sum_{a,b,c}\Pr(a,b,c|\mathbf{p}_n)\min\{b,a+c\}\\ \langle f_{wt}\rangle^{(n)}&=\sum_{a,b,c}\Pr(a,b,c|\mathbf{p}_n)\sigma c\end{aligned} \tag{B.6}$$

and the $n-$th iteration of the normalization factor under the dynamics is

$$Z^{(n)} = \sum_i \langle f_i \rangle^{(n)} = \sum_{a,b,c} \Pr(a,b,c|\mathbf{p}_n)\,(\sigma c + \min\{a+b, 2b+c, 2a+c\}) \quad \text{(B.7)}$$

The sums in previous expressions are extended to all possible natural numbers a, b, c that are compatible with the multiplicity of infection m (not explicitly written, for the sake of notation simplicity). The probability of a given infection configuration (a, b, c), denoted as $\Pr(a, b, c|\mathbf{p})$, can follow a Poisson or a multinomial distribution. In principle, a Poisson distribution would account for a random infection process, which should be the typical situation. Nonetheless, some viruses are able to actively control m, what makes multinomial distributions more appropriate in those cases. For the sake of notation simplicity, we will omit the dependence on $\mathbf{p}$ when it does not lead to confusion.

The evolution equation is symmetric under the interchange of $\Delta 1$ and $\Delta 2$ (first and second components of all vectors involved). This property will be useful in the analysis of the equilibrium points of the system.

Equilibrium states are compositions such that they are fixed points of the replicator equation. Therefore, they fulfill the condition

$$p_i^* = \frac{\langle f_i(\mathbf{p}^*) \rangle}{Z}\,, \qquad \forall i \in \{\Delta 1, \Delta 2, wt\} \quad \text{(B.8)}$$

The stability of an equilibrium point can be evaluated by looking at the Jacobian matrix J of the replicator function evaluated at that point. The equilibrium point is stable if the Jacobian matrix constitutes a contractive application,

$$\mathsf{J} \equiv D\left(\frac{\langle f_i \rangle}{Z}\right)_{\mathbf{p}=\mathbf{p}^*} \quad \text{is contractive:} \quad |\lambda| < 1\,, \quad \text{(B.9)}$$

where D represents the derivation operator and λ is the largest eigenvalue in absolute value of J evaluated at the fixed point.

In the following sections we study the equilibrium points in the simplex $p_{\Delta 1} + p_{\Delta 2} + p_{wt} = 1$, which defines the possible population compositions.

B.1.1 Equilibrium points with pure populations

Pure wt population

Let us study the population defined by vector $(0, 0, 1)^T$, that corresponds to a pure wt population. If $p_{\Delta 1} = p_{\Delta 2} = 0$, only infection configurations with $a = b = 0$ will contribute to the sums in the evolution equation. Hence,

$$p_{\Delta 1} = p_{\Delta 2} = \frac{\sum_c \Pr(0,0,c) \cdot 0}{\sum_c \Pr(0,0,c)\sigma c} = 0 \quad \text{(B.10)}$$
$$p_{wt} = \frac{\sum_c \Pr(0,0,c)\sigma c}{\sum_c \Pr(0,0,c)\sigma c} = 1$$

As a fixed point, it constitutes an equilibrium state of the system. Regarding its stability, the Jacobian matrix at that point is

$$\begin{pmatrix} \frac{1}{\sigma} & 0 & 0 \\ 0 & \frac{1}{\sigma} & 0 \\ -\frac{1}{\sigma} & -\frac{1}{\sigma} & 0 \end{pmatrix} \tag{B.11}$$

Since the largest eigenvalue $1/\sigma > 1$, this is an unstable equilibrium point. This is the reason why the fragmented forms cannot become extinct in the framework of our model. The intuitive explanation is as follows. We are assuming that the population of cells is infinite and that the MOI can take values larger than one in a non-vanishing fraction of infection events (that is, co-infection occurs). On the verge of extinction, when the composition of the population is $(\epsilon, \epsilon, 1-2\epsilon)$, with $\epsilon \to 0$, the probability of co-infecting a cell with a genome in the wt class (which yields complementation) tends to one. Since the fragmented forms are more stable, then their population would increase, thus avoiding extinction. Note that this situation does not necessarily hold if the number of available cells is finite (in which case stochastic extinction of the fragmented class becomes possible) or if the MOI equals strictly one (thus preventing complementation).

Pure $\Delta 1$ or $\Delta 2$ population

A population containing only one class of segments is unable to replicate, as no complementary segment can be found. As a result, the population in the next generation is zero (total extinction) and remains no more in the simplex.

B.1.2 Equilibrium points with two coexisting populations

Coexistence of $\Delta 1$ and $\Delta 2$

Let us consider a generic composition with no presence of the wt class, $(x, 1-x, 0)^T$. In this case, only infection configurations with $c = 0$ will contribute to the equations. The evolution equation for $\Delta 1$ yields

$$p_{\Delta 1} = \frac{\sum_a \sum_{b>a} \Pr(a,b,0)\, a + \sum_b \sum_{a \geq b} \Pr(a,b,0)\, b}{\sum_a \sum_{b>a} \Pr(a,b,0)\, 2a + \sum_b \sum_{a \geq b} \Pr(a,b,0)\, 2b} = \frac{1}{2} \tag{B.12}$$

It is straightforward to see that p_{wt} remains zero in the next generation and, as a result, $p_{\Delta 2} = 1/2$. Therefore, equiabundant composition of $\Delta 1$ and $\Delta 2$ is an equilibrium point. Moreover, it is reached in just one step from every point in the border $p_{wt} = 0$.

The stability of the equilibrium point $(1/2, 1/2, 0)^T$ is determined by the Jacobian matrix

$$\begin{pmatrix} 0 & 0 & -\frac{\sigma m}{4\beta} \\ 0 & 0 & -\frac{\sigma m}{4\beta} \\ 0 & 0 & \frac{\sigma m}{2\beta} \end{pmatrix} \quad \text{where} \quad \beta = \sum_{a,b} \Pr(a,b,0|\tfrac{1}{2},\tfrac{1}{2},0)\min\{a,b\} \tag{B.13}$$

Hence, the equilibrium point $(1/2, 1/2, 0)^T$ will be stable if $\left|\frac{\sigma m}{2\beta}\right| < 1$. Since all factors are positive, the stability condition can be written in the form $\sigma < \sigma_{crit}$, where

$$\sigma_{crit} = \frac{2}{m} \sum_{a,b} \Pr(a,b,0|\tfrac{1}{2},\tfrac{1}{2},0)\min\{a,b\} \tag{B.14}$$

That condition needs to be fulfilled if a population composed only by single segments is to resist invasion by the wt virus.

Coexistence of wt and $\Delta 2$ (or $\Delta 1$)

Let us consider a population lacking the $\Delta 1$ class, with composition $(0, x, 1-x)$. The abundance of class $\Delta 2$ in the next generation can be written as

$$p_{\Delta 2} = \frac{\sum_{b,c} \Pr(0,b,c)\min\{b,c\}}{\sum_{b,c} \Pr(0,b,c)(\sigma c + \min\{b,c\})} = \frac{\beta(x)}{\sigma m p_{wt} + \beta(x)} \tag{B.15}$$

where

$$\beta(x) = \sum_{b,c} \Pr(0,b,c|0,x,1-x)\min\{b,c\} \tag{B.16}$$

For $(0, x, 1-x)^T$ to be a fixed point, and considering the symmetry of the evolution equations with respect to the change $\Delta 1 \to \Delta 2$, also $(x, 0, 1-x)^T$ should be a fixed point with identical stability properties. This leads to the following equation for the fixed point

$$\sigma m = \frac{\beta(x)}{x} \tag{B.17}$$

The Jacobian matrix for this equilibrium point is

$$\begin{pmatrix} 1/\sigma & 0 & 0 \\ \frac{\alpha_c(1-x)-x}{\sigma} - k & \frac{\alpha_b(1-x)}{\sigma} - k & \frac{\alpha_c(1-x)}{\sigma} - x - k \\ -\frac{(1+\alpha_c)(1-x)}{\sigma} + k & -\frac{\alpha_b(1-x)}{\sigma} + k & -\frac{\alpha_c(1-x)}{\sigma} + x + k \end{pmatrix} \tag{B.18}$$

where

$$k = (m-1)x(1-x) \tag{B.19}$$
$$\alpha_b = \sum_{b,c} \Pr'(0, b-1, c) \min\{b, c\}$$
$$\alpha_c = \sum_{b,c} \Pr'(0, b, c-1) \min\{b, c\}$$

and the probabilities $\Pr'$ should be calculated using a distribution with mean $m' = m$ in the Poisson case and $m' = m - 1$ in the multinomial one.

The eigenvalues of the Jacobian matrix are $1/\sigma$, $(\alpha_b - \alpha_c - \sigma)(1-x)/\sigma$, and 0. As the first one is always greater than one, this equilibrium point is unstable. The second eigenvalue is smaller than one and corresponds to the movement in the absence of the $\Delta 1$ class (the corresponding eigenvector is $(0, -1, 1)^T$). This implies that the equilibrium point $(0, x, 1-x)^T$ is indeed a saddle point and, in the absence of $\Delta 1$, the population evolves towards it. The case $x = 0$ recovers the equilibrium point for a pure wt virus.

B.1.3 Equilibrium point for all populations coexisting

The symmetry of the evolution equations suggests looking for general solutions of the form $(x/2, x/2, 1-x)$. Let us apply the conditions for a fixed point to this solution:

$$\frac{x}{2} = Z^{-1} \sum_{a,b,c} \Pr(a, b, c|\tfrac{x}{2}, \tfrac{x}{2}, 1-x) \min\{a, b+c\} \tag{B.20}$$
$$1 - x = Z^{-1} \sum_{a,b,c} \Pr(a, b, c) \sigma c = Z^{-1} \sigma m (1-x)\,,$$

where the first equation holds for $\Delta 1$ and $\Delta 2$, and where we have used the fact (in the second equation) that the mean value of c over all possible infecting configurations is mp_{wt}.

The second equation yields $Z = \sigma m$, and substituting in the first equation we obtain the condition that the fixed point fulfills:

$$\sigma m = \frac{2\,\beta(x)}{x} \tag{B.21}$$

where

$$\beta(x) = \sum_{a,b,c} \Pr(a, b, c|\tfrac{x}{2}, \tfrac{x}{2}, 1-x) \min\{a, b+c\} \tag{B.22}$$

The previous condition holds in particular for $x = 0$, corresponding to the equilibrium point $(0, 0, 1)^T$. The equilibrium point $(1/2, 1/2, 0)^T$ is also obtained if $\sigma = \sigma_{crit}$. An equilibrium point of the form $(x/2, x/2, 1-x)^T$ and $x < 1$ can only exist if

$\sigma > \sigma_{crit}$. Hence, there is no equilibrium point in the interior of the simplex while the border point $(1/2, 1/2, 0)^T$ remains stable.

The Jacobian matrix determining the stability of the interior point $(x/2, x/2, 1 - x)^T$ has an involved form that we do not reproduce here. Its eigenvalues are

$$\begin{aligned} \lambda_0 &= \frac{\alpha_a(x) - \alpha_b(x)}{\sigma} \\ \lambda_1 &= \lambda_0(1-x) + x \\ \lambda_2 &= 0 \end{aligned} \tag{B.23}$$

where

$$\begin{aligned} \alpha_a(x) &= \sum_{a,b,c} \Pr'(a-1, b, c|\tfrac{x}{2}, \tfrac{x}{2}, 1-x) \min\{a, b+c\} \\ \alpha_b(x) &= \sum_{a,b,c} \Pr'(a, b-1, c|\tfrac{x}{2}, \tfrac{x}{2}, 1-x) \min\{a, b+c\} \end{aligned} \tag{B.24}$$

For those values of x that satisfy the fixed point condition, one can check that the largest eigenvalue evaluated at those points fulfils $|\lambda_0| < 1$, and similarly for λ_1. Therefore, the equilibrium point in the interior of the simplex, when it exists, is stable.

As a summary, there is one equilibrium point containing $\Delta 1$ and $\Delta 2$, which is $y^* = (1/2, 1/2, 0)^T$, and that point is stable for $\sigma \leq \sigma_{crit}$. There is a bifurcation for $\sigma = \sigma_{crit}$, so that for $\sigma > \sigma_{crit}$ the equilibrium point y^* becomes unstable and a new stable equilibrium point z^* appears in the interior of the simplex. Moreover, as $\sigma \to \sigma_{crit}^+$ the stable equilibrium point $z^* \to y^*$.

B.2 The curve $\sigma = \sigma_{crit}$ for Poisson and multinomial MOI distributions

Here we develop the functional form of the critical curve separating coexistence of the three genomic types from fixation of the segmented forms (i.e. extinction of the wt) in the case of a multinomial or a Poisson distribution for the infection configuration at a fixed MOI.

A multinomial probability distribution reads

$$\Pr(a, b, c|\mathbf{p}) = \delta_{m-(a+b+c)} \frac{m!}{a!\, b!\, c!} p_{\Delta 1}^a\, p_{\Delta 2}^b\, p_{wt}^c \tag{B.25}$$

(here δ is Kronecker's delta function). Substituting in Eq. B.14 and using the condition $m = a + b$ (since $c = 0$ on the critical curve), σ_{crit} becomes

$$\sigma_{crit} = \frac{1}{2^m} \left[\frac{4}{m} \sum_{a < m/2} a \binom{m}{a} + \frac{m}{2} \binom{m}{m/2} \right] . \tag{B.26}$$

This equation cannot be simplified further, though for $m \gg 1$ the functional behaviour reported in Eq. 4.15 of the main text is obtained.

For a Poisson distribution, σ_{crit} reads

$$\sigma_{crit} = \frac{2e^{-m}}{m}\left[\sum_{b=1}^{\infty}\frac{1}{(b-1)!}\left(\frac{m}{2}\right)^b\left(2\sum_{a=b+1}^{\infty}\frac{1}{a!}\left(\frac{m}{2}\right)^a + \frac{1}{b!}\left(\frac{m}{2}\right)^b\right)\right] . \quad \text{(B.27)}$$

This expression can be simplified with the aid of Bessel functions. The modified Bessel function of the first kind is defined as

$$I_\alpha(x) = \sum_{n=0}^{\infty}\frac{1}{n!\Gamma(n+\alpha+1)}\left(\frac{x}{2}\right)^{2n+\alpha} , \quad \text{(B.28)}$$

such that

$$\sum_{b=0}^{\infty}\frac{1}{b!}\left(\frac{m}{2}\right)^b\frac{1}{(b+1)!}\left(\frac{m}{2}\right)^{b+1} = \sum_{b=0}^{\infty}\frac{1}{b!(b+1)!}\left(\frac{m}{2}\right)^{2b+1} = I_1(m) . \quad \text{(B.29)}$$

Using the expression for $I_0(m)$, the sums in σ_{crit} can be conveniently rewritten, such that after some algebra, one can finally obtain

$$\sigma_{crit} = 1 - e^{-m}\left[I_0(m) + I_1(m)\right] . \quad \text{(B.30)}$$

For $x \gg 1$, the Bessel functions behave as

$$e^{-x}I_\alpha(x) = \frac{1}{\sqrt{2\pi x}} , \quad \text{(B.31)}$$

so for $m \gg 1$, we obtain the same asymptotic behaviour obtained with the multinomial distribution,

$$\sigma_{crit} \sim 1 - \sqrt{\frac{2}{\pi m}} . \quad \text{(B.32)}$$

B.3 Generalizations of the evolution equation

Further generalization of the evolution equation may be introduced by means of modifications of the fitness matrices $\mathsf{M}_{a,b,c}$. First of all, if several replication cycles take place inside the cell, application of the fitness matrix will have to be iterated accordingly. Let G be the number of replication cycles, then iteration of the fitness matrix will be denoted as $\mathsf{M}_{a,b,c}(G)$. Note that, in general, $\mathsf{M}_{a,b,c}(G) \neq \mathsf{M}^G_{a,b,c}$, as the relations among a, b and c that rule the choice of the matrix may change from one to the following iteration.

B.3.1 Different replication rates

Let R be the replicative advantage of single segments with respect to the wt. The effect on the evolution equation can be reduced to a change in the fitness matrices, that now will depend on R. Let us make this dependence explicit by writing $\mathsf{M}_{a,b,c}(R,G)$. If the time scale is fixed according to the replication rate of the single segments, fitness matrices for the first replication cycle take the following form:

$$\mathsf{M}_{a,b,c}(R,1) = \begin{cases} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & R^{-1} \end{pmatrix} & \text{if } c > |a-b| \\ \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & R^{-1} \end{pmatrix} & \text{if } c < |a-b|\,,\ a < b \\ \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & R^{-1} \end{pmatrix} & \text{if } c < |a-b|\,,\ a > b \end{cases} \tag{B.33}$$

The issue is how to find $\mathsf{M}_{a,b,c}(R,G)$ for an arbitrary number of cycles G. To that end, we need to analyse how the application of matrix $\mathsf{M}_{a,b,c}(R,1)$ affects the conditions for matrix choice:

$$\begin{pmatrix} a' \\ b' \\ c' \end{pmatrix} = \mathsf{M}_{a,b,c}(R,1) \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{cases} \begin{pmatrix} a \\ b \\ cR^{-1} \end{pmatrix} & \text{if } c > |a-b| \\ \begin{pmatrix} a \\ a+c \\ cR^{-1} \end{pmatrix} & \text{if } c < |a-b|\,,\ a < b \\ \begin{pmatrix} b+c \\ c \\ cR^{-1} \end{pmatrix} & \text{if } c < |a-b|\,,\ a > b \end{cases} \tag{B.34}$$

Note that $c < |a-b|\,,\ a < b \Rightarrow c' < |a'-b'|\,,\ a' < b'$, and the same holds if $a > b$. This means that, for the last two conditions, the fitness matrix can be iterated in the simple way $\mathsf{M}_{a,b,c}(R,G) = \mathsf{M}_{a,b,c}(R)^G$. However, condition $c > |a-b|$ does not necessarily imply that $c' > |a'-b'|$. It is easy to see that when this condition holds initially the fitness matrix can be iterated for just a certain number γ of cycles, and then a change of matrix must be done. In particular, the value for γ is

$$\gamma = \text{int}\left\{ \frac{\log(c/|a-b|)}{\log R} \right\} + 1 \tag{B.35}$$

Taking into account all previous considerations a general expression for the product $\mathsf{M}_{a,b,c}(R,G)\,(a,b,c)^T$ can be obtained:

$$\mathsf{M}_{a,b,c}(R,G)\begin{pmatrix}a\\b\\c\end{pmatrix}=\begin{cases}\begin{pmatrix}a\\a+cR^{1-G}\\cR^{-G}\end{pmatrix} & \text{if } c<|a-b|\,,\ a<b\\ \begin{pmatrix}b+cR^{1-G}\\b\\cR^{-G}\end{pmatrix} & \text{if } c<|a-b|\,,\ a>b\\ \begin{pmatrix}a\\b\\cR^{-G}\end{pmatrix} & \text{if } c>|a-b|\,,\ G\leq\gamma\\ \begin{pmatrix}a\\a+cR^{1-G}\\cR^{-G}\end{pmatrix} & \text{if } c>|a-b|\,,\ G>\gamma\,,\ a<b\\ \begin{pmatrix}b+cR^{1-G}\\b\\cR^{-G}\end{pmatrix} & \text{if } c>|a-b|\,,\ G>\gamma\,,\ a>b\end{cases} \tag{B.36}$$

It is possible to write the previous result in a more compact form

$$\mathsf{M}_{a,b,c}(R,G)\begin{pmatrix}a\\b\\c\end{pmatrix}=\begin{pmatrix}\min\left\{a,b+cR^{1-G}\right\}\\ \min\left\{b,a+cR^{1-G}\right\}\\ cR^{-G}\end{pmatrix} \tag{B.37}$$

From this expression, it is straightforward to obtain the conditional fitness

$$\begin{aligned}f_{\Delta 1|a,b,c}&=\min\{a,b+cR^{1-G}\}\\ f_{\Delta 2|a,b,c}&=\min\{b,a+cR^{1-G}\}\\ f_{wt|a,b,c}&=\sigma cR^{-G}\end{aligned} \tag{B.38}$$

B.3.2 Loss of segments through mutation and replication fidelity

Let ρ be the probability that a genomic segment is lost during replication. As a wt genome contains two segments, the probability that it gives rise to other complete wt genome is $(1-\rho)^2$. With probability $\rho(1-\rho)$, replication of a wt genome will produce a mutant of class $\Delta 1$ or $\Delta 2$. On the other hand, a single-segment genome will reproduce successfully with probability $(1-\rho)$. These transition probabilities can be included in a transition matrix T

$$\mathsf{T}=\begin{pmatrix}1-\rho & 0 & \rho(1-\rho)\\ 0 & 1-\rho & \rho(1-\rho)\\ 0 & 0 & (1-\rho)^2\end{pmatrix} \tag{B.39}$$

The fitness matrices for the first replication cycle are simply the product of matrices $\mathsf{T}\cdot\mathsf{M}_{a,b,c}$

$$\mathsf{M}_{a,b,c}(\rho,1)=\begin{cases}\begin{pmatrix}1-\rho & 0 & \rho(1-\rho)\\ 0 & 1-\rho & \rho(1-\rho)\\ 0 & 0 & (1-\rho)^2\end{pmatrix} & \text{if } c>|a-b|\\ \begin{pmatrix}1-\rho & 0 & \rho(1-\rho)\\ 1-\rho & 0 & (1-\rho)(1+\rho)\\ 0 & 0 & (1-\rho)^2\end{pmatrix} & \text{if } c<|a-b|\,,\ a<b\\ \begin{pmatrix}0 & 1-\rho & (1-\rho)(1+\rho)\\ 0 & 1-\rho & \rho(1-\rho)\\ 0 & 0 & (1-\rho)^2\end{pmatrix} & \text{if } c<|a-b|\,,\ a>b\end{cases} \tag{B.40}$$

The effect of a single application of $\mathsf{M}_{a,b,c}(\rho,1)$ is the following

$$\begin{pmatrix}a'\\ b'\\ c'\end{pmatrix}=\mathsf{M}_{a,b,c}(\rho,1)\begin{pmatrix}a\\ b\\ c\end{pmatrix}=\begin{cases}\begin{pmatrix}(1-\rho)(a+\rho c)\\ (1-\rho)(b+\rho c)\\ c(1-\rho)^2\end{pmatrix} & \text{if } c>|a-b|\\ \begin{pmatrix}(1-\rho)(a+\rho c)\\ (1-\rho)[a+(1-\rho)c]\\ c(1-\rho)^2\end{pmatrix} & \text{if } c<|a-b|\,,\ a<b\\ \begin{pmatrix}(1-\rho)[b+(1-\rho)c]\\ (1-\rho)(b+\rho c)\\ c(1-\rho)^2\end{pmatrix} & \text{if } c<|a-b|\,,\ a>b\end{cases} \tag{B.41}$$

It can be checked that condition $c<|a-b|\,,\ a<b\Rightarrow c'<|a'-b'|\,,\ a'<b'$, and the same holds if $a>b$. that means that, in such cases, $\mathsf{M}_{a,b,c}(\rho,G)=\mathsf{M}_{a,b,c}(\rho,1)^G$. However, if $c>|a-b|$, the initial matrix can only be iterated for a number γ of cycles such that $d(1-\rho)^\gamma>|a-b|$. From here we can extract γ:

$$\gamma=\text{int}\left\{\frac{\log(|a-b|/c)}{\log(1-\rho)}\right\}+1 \tag{B.42}$$

By applying the previous reasoning, the general expression for the product $\mathsf{M}_{a,b,c}(\rho,G)\,(a,b,c)^T$ can be obtained

$$\mathsf{M}_{a,b,c}(\rho,G)\begin{pmatrix}a\\ b\\ c\end{pmatrix}=\begin{cases}(1-\rho)^G\begin{pmatrix}a+c\big(1-(1-\rho)^G\big)\\ a+c+c\rho(1-\rho)^{G-1}\\ c(1-\rho)^G\end{pmatrix} & \text{if } c<|a-b|\,,\\ & a<b\\ (1-\rho)^G\begin{pmatrix}a+c\big(1-(1-\rho)^G\big)\\ b+c\rho\big(1-(1-\rho)^G\big)\\ c(1-\rho)^G\end{pmatrix} & \text{if } c>|a-b|\,,\\ & G\le\gamma\\ (1-\rho)^G\begin{pmatrix}a+c\big(1-(1-\rho)^G\big)\\ a+c+c\rho(1-\rho)^{G-1}\\ c(1-\rho)^G\end{pmatrix} & \text{if } c>|a-b|\,,\\ & G>\gamma\,,\ a<b\end{cases} \tag{B.43}$$

while the two remaining cases for $a > b$ can be obtained by writing b instead of a and switching the first and second vector components.

A simpler expression can be attained by using a minimum function

$$\mathsf{M}_{a,b,c}(\rho,G)\begin{pmatrix}a\\b\\c\end{pmatrix}=(1-\rho)^G\begin{pmatrix}\min\left\{a,b+c(1-\rho)^{G-1}\right\}+c\big(1-(1-\rho)^G\big)\\ \min\left\{b,a+c(1-\rho)^{G-1}\right\}+c\big(1-(1-\rho)^G\big)\\ c(1-\rho)^G\end{pmatrix} \tag{B.44}$$

Conditional fitness can be derived directly from here and becomes

$$\begin{aligned} f_{\Delta 1|a,b,c} &= (1-\rho)^G\left[\min\{a,b+c(1-\rho)^{1-G}\}+c\big(1-(1-\rho)^G\big)\right] \\ f_{\Delta 2|a,b,c} &= (1-\rho)^G\left[\min\{b,a+c(1-\rho)^{1-G}\}+c\big(1-(1-\rho)^G\big)\right] \\ f_{wt|a,b,c} &= \sigma c(1-\rho)^{2G} \end{aligned} \tag{B.45}$$

B.3.3 Constant per-cell viral productivity

The reproductive ratio Π_i of class i is defined as the number of particles of that class produced per cell and per infecting particle of such class. The critical curve separating the regions of coexistence and extinction of the wt can be obtained from the condition that the reproductive ratio Π_{wt} of the wt and that of the segmented forms, $\Pi_{\Delta 1} = \Pi_{\Delta 2} \equiv \Pi_\Delta$ take equal values. Indeed, if for two classes i and j one has $\Pi_i > \Pi_j$, then the fraction of class i increases at the expense of class j. Let us consider a population of bipartite classes at equilibrium. If a small amount of wt virus is added, the reproductive ratios are obtained from

$$\Pi_{wt} = \frac{\langle f_{wt}\rangle}{m\epsilon} = \frac{\sigma\sum_{a,b,c} Pr\left(a,b,c|\frac{1-\epsilon}{2},\frac{1-\epsilon}{2},\epsilon\right)\frac{c}{z_{a,b,c}}}{m\epsilon} \tag{B.46}$$

$$\Pi_\Delta = \frac{\sum_{a,b,c} Pr\left(a,b,c|\frac{1-\epsilon}{2},\frac{1-\epsilon}{2},\epsilon\right)\frac{\min\{a,b+c\}}{z_{a,b,c}}}{m/2}\,, \tag{B.47}$$

where ϵ is the average fraction of wt particles infecting a cell. The critical curve is obtained in the limit $\epsilon \to 0$. Setting $\Pi_{wt} = \Pi_\Delta$ and discarding terms of order ϵ and larger, we obtain

$$\sigma_{crit} = \frac{Pr\left(a,b>0|1/2,1/2\right)}{m\sum_{a,b}\frac{Pr'(a,b|1/2,1/2)}{1+\min\{a+b,2a+1,2b+1}}}\,. \tag{B.48}$$

The probability inside the sum has to be calculated for $m' = m$ in the case of a Poisson distribution and for $m' = m-1$ in the case of a multinomial distribution. The numerical solution of Eq. (B.48) has been obtained in both cases, and represented in Figure B.1 together with the asymptotic solution calculated in Chapter 4.

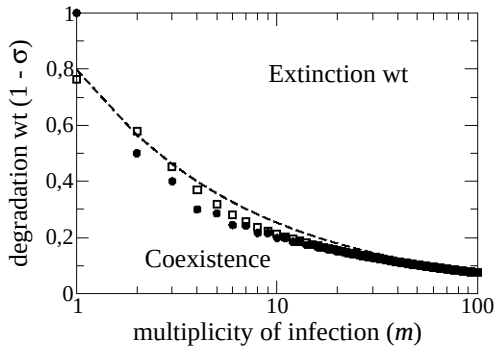


Figure B.1: *Evolutionary regions depending on selective pressures (σ and MOI) under cell-limited resources.* Solid circles: multinomial MOI distribution; open squares: Poisson MOI distribution. Dashed line: asymptotic behaviour as obtained in a situation of unlimited cellular resources. Compare these results with those represented in Figure 4.3.

B.4 Evolution with three segments

In this section we expand the model to describe genomes that can be fragmented into three segments. According to the presence or absence of each segment there are now seven possible viral classes: three single-segment classes (denoted as A_1, A_2 and A_3), three two-segment classes (denoted as B_1, B_2 and B_3, that lack first, second and third segment, respectively), and the wt virus, which contains the whole genome and thus does not require complementation. All incomplete classes can be complemented by the wt. Moreover, a two-segment class B_i can also be complemented by a different two-segment class (B_j, $j \neq i$) or by the complementary single segment A_i. Single-segment class A_i can be complemented by their complementary B_i or by a set of classes that jointly contain all segments $j \neq i$. In each cell, the offspring produced by a given class will be the minimum among the number of copies of that class that infected the cell and the total number of copies for each genomic segment that can be found inside the cell.

The infection configuration $\vec{v} = (a_1, a_2, a_3, b_1, b_2, b_3, c)$ is defined as the number of segments of each class that infect the cell (the identity of each class is the intuitive one: a_i for class A_i, b_i for class B_i and c for the wt). Degradation of each class is proportional to the number of segments they have, being zero for the single segments, $(1-\sigma)/2$ for the two-segment classes and $1-\sigma$ for the wt.

Specifically, the conditional fitness in the three segment setting takes the following form

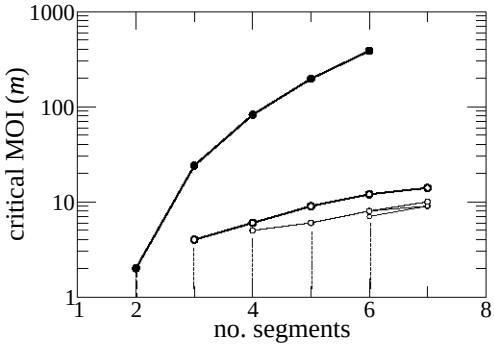


Figure B.2: *Transitions to increasingly fragmented forms for genomes with multiple segments in the case of a cubic relationship between degradative advantage and number of segments.* We represent the critical value of the MOI required for fixation of single-segment classes (thick line on the left), for $\sigma = 0.5$. Other lines indicate further transitions: from double-single coexistence to triple-double-single, and so on. Compare these results with those represented in Figure 4.4(b).

$$f_{wt|\vec{v}} = \sigma c \tag{B.49}$$

$$f_{B_i|\vec{v}} = \frac{1+\sigma}{2} \min\{b_i, a_i + \sum_{j \neq i} b_j + c\}$$

$$f_{A_i|\vec{v}} = \min\{a_i, a_j + \sum_{k \neq j} b_k + c \quad \forall j \neq i\} \tag{B.50}$$

Once conditional fitness are defined, the replicator equation that rules the evolution of the population can be expressed in the same way as in the two segment setting.

B.5 On the relationship between genome length and degradative advantage

In our analyses of genomes with multiple segments, we have assumed an inverse linear relationship between the length of a genome (the number of segments is contains) and its selective advantage. Though this might be a reasonable assumption if one believes that the speed at which the elements of a genome are degraded is constant, other scenarios are possible. In order to assess the robustness of our model under changes in this relationship, we have studied the situation where the relevant ingredient for stability is how genomes are packed inside capsids. A shorter genome occupies a smaller volume, thus reducing the interaction with the capsid. Let us further assume that there is a linear relationship between volume and stability. In this case, the selective advantage σ_n of a genome with n segments takes the form

$$\sigma_n = 1 - \frac{(n^{1/3} - 1)(1 - \sigma)}{N^{1/3} - 1}\,, \tag{B.51}$$

where we have imposed $\sigma_1 = 1$ and $\sigma_N = \sigma$, which hold by definition. We have numerically studied this scenario for genomes with different number of segments and analysed the values of MOI at which the transition to fragmented forms with an increasing number of segments occurs. The results are depicted in Figure B.2. Though some quantitative differences arise, the qualitative results are robust.

C

Bioinformatics, statistics and additional calculations for Chapter 5

This appendix contains information about the bioinformatics and statistical methods, as well as some analytical calculations and additional results concerning Chapter 5. The extraction and classification of IS data from prokaryotic genome sequences (section C.1) was carried out by Dr. Manuel J. Gómez and Dr. Francisco J. López de Saro (Centro de Astrobiología), wih whom we collaborated in this work.

C.1 IS data retrieval and classification

File collections containing orientation and coordinates of protein coding genes (*.ptt), predicted protein sequences (*.faa) and chromosomal nucleotide sequences (*.fna) of partially and completely sequenced prokaryotic genetic elements were downloaded from the bacterial section of the NCBI Genome database, on October 24, 2012, as well as a summary file containing a table that linked accession numbers, replicon type (chromosome, plasmid) and taxonomic name. The working, curated data set consisted of 2074 completely sequenced, circular, bacterial chromosomes, out of which 1811 contained at least one IS (harboured by 1685 species or strains).

The whole collection of predicted proteins from the genomic data set ($6,055,750$ sequences) was aligned with HMMER 3.0 against the Pfam 26.0 database[1] of domain profiles (Punta et al. 2012), using domain-specific score thresholds to filter the hits. The

[1] pfam.sanger.ac.uk

output of HMMER was processed with a Perl script to reconstruct protein architectures using a positional cometition strategy to assemble the predicted protein domains and allowing no overlaps. IS-related proteins were identified by comparing the new annotations against a list of 286 architectures that were considered as characteristic of proteins ecnoded by IS elements and that were composed by a restricted collection of Pfam domains . The architecture list was generated by manually extracting IS-encoded protein descriptions from the Pfam and the ISfinder databases (www-is.biotoul.fr/is.html) (Punta et al. 2012; Siguier et al. 2006). We were able to identify $82,516$ IS-associated genes. Once IS-related proteins had been identified in the set of bacterial genomes, IS elements were predicted following a strategy, articulated in four steps, that took into account that ISs can be composed of several genes and that they can appear in chromosomes as tandem insertion, difficulting the definition of their boundaries. In the first step, clusters of consecutive IS genes (separated by intergenic distance ≤ 500 bp) were identified in all genomes to calculate distance distributions for all possible pairs of IS-related gene types (as defined by the architecture of the corresponding gene products). In the second step, cluster detection was repeated, this time restricting the allowed intergenic distances to gene pair-specific distance ranges, deduced from the previous step (mean$\pm$2SD). Clusters detected in this step had ten genes at most. In the third step, the resulting collection of clusters was used to manually derive a list of 209 clusters that were accepted as representatives of the genetic organization of complete IS elements, on the basis of correspondence to described IS structures , abundance (assuming that highly abundant and distributed clusters should correspond to complete IS elements), and length (in terms of number of genes). Each of these clusters was classified as belonging to a particular IS family. Accepted clusters had three genes at most. In the fourth step, each IS gene cluster detected in the second step was decomposed into all possible collections of non-overlapping accepted sub-clusters to identify the collection that maximized the length of sub-clusters. Each sub-cluster from the optimal collection was then assigned to a particular IS family following the correspondences established in the list of accepted clusters. $57,515$ sub-clusters were detected, each of them representing a complete IS, that comprised $69,438$ (84%) of the IS-related genes.

C.2 Data pre-processing

In order to compare the genomic data with the models we assume that the dynamics of a particular IS family is similar in all genomes. Therefore, the genomic frequencies observed for a given IS family can be interpreted on the basis of the probability p_k of finding a genome with k copies in a population of independent genomes. We also assume that different IS families behave independently, so that it is possible to analyse them separately. In order to minimize the possible bias introduced by closely related strains, we restricted our analysis to a dataset composed of only one strain per species. Athough genomes from distinct species may be not completely independent, the averaging on many non-related groups compensates for that. As a confirmation, taking

one genome per genus and repeating the analysis did not change our results. The full dataset with multiple strains per species was only used to detect outliers.

Some extra pre-processing of the genomic data is required, resulting from the fact that the HGT-deletion ratio β is correlated to the fraction of genomes that contain the IS family of interest. That implies that estimation of β may be biased if the absence of an IS in some genomes is not due to the natural gain-loss dynamics but to other factors that make the IS unable to settle down in such genomes. In order to minimise that risk, we excluded from this study those genomes which do not contain any IS family at all. The remaining dataset contains 1079 bacterial chromosomes (harboured by 1014 species). As it is quite a large number, special cases of genomes that may be non-invadable by certain IS families are not expected to introduce a significant bias into the estimation of β. Alternatively, IS famiIes that are very specific to certain genomes can be detected through their poor fits.

Genomes belonging to the main phyla Proteobacteria and Firmicutes + Tenericutes were analysed separately at a first stage of the work. Since we obtained similar results in both groups, we pooled the data from all phyla in a single dataset.

C.3 Parameter estimation, goodness of fit test and model comparison

IS families that appear in fewer than 20 genomes were discarded, thus preventing unreliable estimations associated to small datasets. The following parameter estimation was done independently for each of the 36 remaining IS families. First, the frequency distribution of the family was extracted from the genomic data. Then, for each model a maximum likelihood approach was applied to determine the parameters that best fit the model to the data. As a numerical optimization algorithm, we used the simplex method implemented in *MATLAB*[2].

Some care must be taken in order to evaluate the role of selection. The key difficulty is the fact that parameter estimation in the selection model is confused by multiple local maxima in the likelihood function. Since local maxima with similar values are distributed along the whole parameter range, parameter estimation becomes strongly dependent on their initial guesses. As a result, an *a priori* estimation of some parameters is required before the selection model can be fitted to the data. Because the neutral model is a particular case of the selection model, we took α from the neutral setting and tried to refine the fit by adding selection. Alternatively, we explored the selection model by choosing a qualitatively different range of values of α, between 10^2 and 10^3 (as suggested in (Bichsel et al. 2012)); and also the case of a small (but greater than one) $\alpha = 2$.

The goodness of the fits was evaluated by means of a likelihood ratio test that compared the observed and expected frequencies for each abundance interval. This test is similar to a Chi-square test, but more suitable if any of the differences between

[2]MATLAB version 7.6.0.324 (R2008a). Natick, Massachusetts: The Mathworks Inc.

the observed and expected frequencies is greater than the expected frequency. Different abundance intervals have been defined for each IS family in such a way that at least two occurrences are expected for each interval (alternative criteria have been tried without major changes in results). The p-values associated to the test statistics have been numerically computed by simulating a sampling process on the expected distribution. Comparison between neutral and selection models was done in terms of the corrected Akaike Information Criterion (Akaike 1974), both models containing two degrees of freedom (because α is fixed in the model with selection).

The results of the fits to the neutral model are shown in Table C.1. It shows the fit parameters and the p-value associated to the goodness of fit test for two variants of the model: either with a single HGT rate β ("One HGT rate", as explained in Chapter 3) or with different HGT rates in empty and IS hosting genomes ("Two HGT rates", parameter β_0 accounts for HGT rate in empty genomes). The latter scenario would account, for instance, for IS families that provide the host genome with some kind of resistance against the entry of additional IS copies. The last column in the table contains the difference in the corrected Akaike Information Criterion for both models (with two and three degrees of freedom, respectively). Since we are dealing with nested models, the ΔAICc is equivalent to a likelihood ratio test and follows a Chi-squared distribution with one degree of freedom. Therefore, $\Delta\text{AICc} > 6.64$ implies that the "two HGT rates" model is more probable to be true at a (non-corrected by multiple comparisons) significance level of 0.01.

Table C.2 contains the results of the fit to the model with selection. The case where the value of α was taken from the neutral fit rendered selection parameters $\sigma \leq 10^{-5}$ and it is not included in the table. In the scenarios with $\alpha = 10^3$ and $\alpha = 2$, the ΔAICc values show that the fits to the model with selection are as good as—but not better than—those to the neutral model.

C.4 Detection of outlier genomes

For each IS family, outliers are genomes that contain a large copy number, so large that it cannot be explained by any of the models. Specifically, let us define P_k as the probability of having k or more copies, $P_k = \sum_{i \geq k} p_i$. The probability that a genome with k or more copies is found in a sample of G genomes is $s_k = 1 - (1 - P_k)^N$. The value of s_k is indeed the significance level, already corrected by the sample size (Šidák 1967). It can be set to the desired value in order to numerically obtain the copy threshold k_s. Thus, genomes with more than k_s copies are outliers at a corrected significance level s. Copy thresholds are different accross IS families, thus detection of outliers was carried out independently for each family. We tried $s = 0.05$ and $s = 0.01$ with similar results. As we looked for outliers in the full dataset (including more than one strain per species), we took a sample size $N = 1811$ chromosomes. That is a conservative choice, since the actual number of independent instances in the dataset may be smaller; however, similar results were obtained by setting $N = 1079$

Table C.1: **Fit of the data to the neutral models.**

		One HGT rate			Two HGT rates				
	N	α	β	p	α	β	β_0	p	ΔAICc
IS1	46	0.95	0.01	0.44	0.97	-0.18	0.02	0.48	-0.77
IS110	529	0.91	0.27	0.53	0.88	0.46	0.22	0.64	5.00
IS1182	227	0.88	0.10	0.34	0.90	0.02	0.10	0.36	-1.46
IS380	67	0.92	0.02	0.54	0.92	0.06	0.02	0.53	-2.15
IS1595	135	0.86	0.06	0.35	0.94	-0.33	0.08	0.63	**9.33**
IS1634	34	0.93	0.01	0.58	0.87	0.41	0.01	0.58	-0.48
IS200	422	0.77	0.26	0.29	0.84	0.02	0.30	0.52	5.00
IS200/IS605	105	0.78	0.05	0.58	0.72	0.35	0.05	0.58	-1.21
IS21	498	0.81	0.30	**0.02**	0.89	-0.08	0.39	0.33	**23.9**
IS256	376	0.92	0.16	0.59	0.93	0.11	0.16	0.57	-1.56
IS3	702	0.91	0.40	0.29	0.92	0.38	0.41	0.31	-1.91
IS30	243	0.89	0.10	0.46	0.91	-0.01	0.11	0.53	-0.67
IS481	63	0.87	0.03	0.51	9,86	0.09	0.02	0.51	-2.11
IS4a	57	0.91	0.02	0.68	0.91	0.05	0.02	0.67	-2.21
IS4b	40	0.90	0.02	0.50	0.82	0.42	0.01	0.52	-0.31
IS5a	264	0.92	0.10	0.53	0.94	-0.02	0.12	0.60	0.22
IS5a/b	135	0.82	0.06	0.35	0.92	-0.32	0.08	0.50	**7.46**
IS5b	285	0.93	0.11	0.61	0.90	0.33	0.09	0.48	3.83
IS5c	54	0.91	0.02	0.43	0.94	-0.18	0.02	0.57	-0.92
IS5d	53	0.95	0.02	0.60	0.96	-0.01	0.02	0.75	-2.21
IS6	103	0.87	0.04	0.54	0.86	0.12	0.04	0.45	-1.87
IS607	52	0.88	0.02	0.54	0.85	0.17	0.02	0.58	-1.86
IS630	253	0.95	0.08	0.67	0.95	0.04	0.09	0.73	-1.75
IS66a	177	0.90	0.07	0.64	0.90	0.07	0.07	0.64	-2.07
IS66b	31	0.93	0.01	0.14	0.98	-0.42	0.02	0.51	2.60
IS701	106	0.91	0.04	0.26	0.91	0.04	0.04	0.61	-2.12
IS91	86	0.82	0.04	0.43	0.92	-0.35	0.05	0.64	4.15
IS982	04	0.95	0.03	0.50	0.94	0.13	0.03	0.55	-1.56
ISAs1	91	0.92	0.03	0.48	0.88	0.26	0.03	0.54	-0.25
ISAzo13	22	0.80	0.01	0.50	0.70	0.31	0.01	0.58	-2.27
ISL3	277	0.91	0.11	0.41	0.93	-0.03	0.13	0.54	1.10
ISNCYa	166	0.86	0.08	0.23	0.64	1.00	0.05	0.54	**24.0**
ISTnp1	351	0.92	0.14	0.62	0.92	0.15	0.14	0.61	-2.03
PDDEXK	102	0.91	0.04	0.66	0.88	0.20	0.03	0.67	-0.99
Tn3	98	0.69	0.06	0.52	0.48	0.54	0.05	0.54	1.25
Tn7	64	0.42	0.05	0.55	0.31	0.22	0.05	0.56	-1.99

Parameter β_0 is the HGT rate to empty genomes. The p-values correspond to the goodness of fit tests. ΔAICc > 6.64 implies that the second model is more probable to be true at a (non-corrected) significance level of 0.01.

Table C.2: **Fit of the data to the selection models.**

	$\alpha = 10^3$			$\alpha = 2$		
	β	σ	ΔAICc	β	σ	ΔAICc
IS1	15	57	-2.6E-5	0.03	0.06	-2.9E-5
IS110	293	102	4.2E-6	0.59	0.11	3.1E-6
IS1182	112	135	1.1E-5	0.22	0.15	8.8E-6
IS1380	25	85	6.4E-6	0.05	0.09	5.3E-6
IS1595	68	164	1.9E-6	0.14	0.19	1.3E-6
IS1634	12	77	4.4E-7	0.02	0.08	-1.8E-6
IS200	333	294	8.3E-6	0.67	0.36	8.0E-6
IS200/IS605	69	283	1.8E-6	0.14	0.35	1.6E-6
IS21	367	241	5.2E-6	0.73	0.29	4.7E-6
IS256	171	87	9.8E-6	0.34	0.09	8.3E-6
IS3	436	95	1.7E-6	0.87	0.10	1.4E-6
IS30	116	122	3.2E-6	0.23	0.14	9.5E-7
IS481	29	144	1.5E-6	0.06	0.16	1.3E-6
IS4a	22	95	6.0E-6	0.04	0.10	5.3E-6
IS4b	17	114	2.7E-6	0.03	0.13	2.0E-6
IS5a	111	86	1.7E-5	0.22	0.09	1.0E-5
IS5a/b	79	225	1.7E-6	0.16	0.27	9.1E-7
IS5b	120	80	4.6E-6	0.24	0.09	1.1E-6
IS5c	22	104	1.0E-6	0.04	0.11	-1.4E-6
IS5d	17	49	-1.9E-4	0.03	0.05	-2.0E-4
IS6	49	145	3.3E-7	0.10	0.16	-3.7E-7
IS607	23	130	3.2E-6	0.05	0.14	2.8E-6
IS630	90	53	-6.8E-4	0.18	0.06	-6.8E-4
IS66a	80	118	1.9E-6	0.16	0.13	-1.6E-6
IS66b	11	76	1.7E-5	0.02	0.08	1.5E-5
IS701	44	100	2.2E-5	0.09	0.11	2.0E-5
IS91	48	217	1.6E-6	0.10	0.26	1.4E-6
IS982	31	53	-1.6E-4	0.06	0.06	-1.6E-4
ISAs1	36	93	2.5E-6	0.07	0.10	2.9E-7
ISAzo10	13	252	3.4E-6	0.03	0.30	3.3E-6
ISL3	124	100	9.3E-6	0.25	0.11	6.5E-6
ISNCYa	87	160	1.5E-6	0.17	0.18	2.6E-7
ISTnp1	155	83	1.9E-5	0.31	0.09	1.6E-5
PDDEXK	42	103	1.1E-5	0.08	0.11	8.6E-6
Tn3	83	450	5.8E-7	0.16	0.59	5.6E-7
Tn7	113	1368	2.3E-7	0.22	2.16	2.2E-7

The ΔAICc values correspond to the comparison between the neutral model and the models with selection.

(the number of different species). Notice that the correction for sample size implies that the significance threshold per genome, in all these conditions, is close to 10^{-5}.

C.5 Independent estimation of α and β

The critical relation sets an implicit constraint if a stationary abundance distribution is to be established. When it comes to study the critical relation, such a constraint may give rise to a false correlation if the fitting algorithm estimates α and β jointly. In order to avoid that, we used an alternative approach that provides an independent (although less precise) estimation of the parameters. First, the HGT-deletion ratio was estimated as $\beta = F(1)/F(0)$, where $F(1)$ and $F(0)$ are the frequencies of genomes with one and no copies, respectively. Next, we discarded genomes with no copies and estimated α only from "infected" genomes. These parameter values were used to test the critical relation. By simulating non-stationary genomes we checked that the independent estimation algorithm does not give rise to false correlations.

C.6 Derivation of the critical condition $\alpha + \gamma = 1$

Let us consider the scenario defined by the free-parameter neutral model. The copy number distribution in such a scenario is defined by the equation 5.1 in the main text. In correspondence, the average number of copies per genome for a particular IS family is equal to

$$\langle k \rangle = \frac{\beta\,\beta_0}{(1-\alpha)\left(\beta_0 + (\beta-\beta_0)(1-\alpha)^{\beta/\alpha}\right)} \tag{C.1}$$

Let us explore the possibility that the HGT rate is proportional to the abundance of the IS, expressed as its average copy number. Specifically, we write $\beta = \gamma\langle krangle$ and $\beta_0 = \gamma_0\langle krangle$, where the proportionality constants γ and γ_0 will be called the relative HGT-deletion ratios. Substitution of these assumptions in eq. (C.1) results in a fixed point equation, that after some manipulation takes the following form:

$$(\gamma_0 - \gamma)(1-\alpha)^{\gamma\langle k\rangle/\alpha} = \gamma_0\left(1 - \frac{\gamma}{1-\alpha}\right) \tag{C.2}$$

If both HGT rates are equal, then $\gamma_0 = \gamma$. In such a case, a stationary state where $\langle k \rangle$ is finite and greater than zero can only be reached if

$$0 = \gamma\left(1 - \frac{\gamma}{1-\alpha}\right)$$

what leads to the condition $\alpha + \gamma = 1$.

Therefore, in the neutral model with HGT-deletion ratio proportional to the average copy number and $\beta = \beta_0$, the relation $\alpha + \gamma = 1$ determines the critical condition for the stability of the system. If $\alpha + \gamma < 1$ the IS's will become extinct, whereas if $\alpha + \gamma > 1$ an explosive proliferation of copies will take place.

C.7 Transposition ratchet in small populations

In this section, the effect of weak selection on the IS copy number is analized. Even if selection acts against the expansion of IS's, the stochastic dynamics in finite populations of genomes may lead to the fixation of genomes with greater copy number.

Let us consider a Moran process on a population with effective size N. Moreover, we will assume that the duplication rate $r \ll N^{-1}$, and the same for the deletion rate d. Under this assumption, the copy number in the population is homogeneous and mutants with increased (or reduced) copy number either become fixed or get extinct before the next mutant appears.

Starting with an homogenous population of genomes that contain k copies, we are interested in obtaining the probabilities that the population evolves towards states with $k+1$ and $k-1$ copies. Let us denote those probabilities as ρ_k^+ and ρ_k^-, respectively. We focus on the case where the genome fitness decreases linearly with the copy number as $f_k = 1 - sk$ (here s is the fitness cost of a single copy). For the Moran process considered, the transition probabilities can be calculated analytically as the product of the mutation rate and the fixation probability:

$$\rho_k^+ = kr\frac{1 - f_k/f_{k+1}}{1 - (f_k/f_{k+1})^N} = \frac{kr}{N}\left(1 - \frac{N-1}{2}\,s\right) + \mathcal{O}(Ns)^2 \tag{C.3}$$

$$\rho_k^- = kd\frac{1 - f_k/f_{k-1}}{1 - (f_k/f_{k-1})^N} = \frac{kd}{N}\left(1 + \frac{N-1}{2}\,s\right) + \mathcal{O}(Ns)^2 \tag{C.4}$$

In conditions of weak selection, $s << N^{-1}$, the terms of higher order in Ns can be neglected. Thus, it is straightforward to obtain the ratio between both transition probabilities, that does not depend on the actual copy number.

$$\frac{\rho^+}{\rho^-} = \alpha\left(1 - (N-1)\,s\right) \tag{C.5}$$

where $\alpha = r/d$ is the duplication-deletion ratio.

An interesting consequence is the prediction of IS expansions in finite populations when selection is weak and $\alpha \gg 1$ (notice that this holds even if there is no HGT). This phenomenon may allow for discrimination between the supercritical and subcritical scenarios. In the former case (characterized by $\alpha \gg 1$) selection, even if weak, is essential for controlling the copy number. In the latter, deletion alone compensates for IS proliferation and weak selection, even if present, can be neglected without qualitative consequences.

C.8 Genomic diversity and family co-occurrences

The IS diversity of a genome is defined as the number of different IS families hosted by that genome. This definition, that does not consider the intragenomic abundances

of each IS, is motivated by the fact that the mechanism determining the presence of the IS in the genome (horizontal transfer) plays only a minor role in determining the IS abundance (mainly driven by duplication-deletion processes).

Based on the above definition of genomic diversity, we call diversity distribution to the fraction of genomes displaying a given diversity, i.e. a given number of different families. The diversity distribution of the experimental data is shown in figure 7.3 of the main text.

In order to check if the observed diversity distribution corresponds to a process of random, independent acquisition events, the following randomization procedure was applied. First, each IS family was associated a presence probability equal to the fraction of genomes where it appears. Then, for a number of empty genomes, we simulated an IS acquisition process in which every genome incorporates every IS family according to its presence probability. The diversity distribution obtained in this way is plotted as a dashed, blue line in figure 7.3 of the main text. Notice that the procedure here exposed is analogous to a random reshuffling of the IS family content among genomes, averaged over a large number of realizations.

The real diversity distribution is markedly different to the randomized one. The real distribution contains a larger fraction of genomes with diversity higher and lower than expected by chance. In trying to search for possible causes, we explored a possible correlation between diversity and genome size, as well as correlations among family occurrences themselves.

Inspired by the classic species-area relation of ecology we looked for a genomic analogue in the form of a power-law relation between the diversity and the genome size. To that end, the 1079 diversity-size points corresponding to chromosomes from different species in the dataset were logarithmically transformed. After transformation, both variables, genome size and number of families, were standardized by substraction of their means (14.93 and 1.45, respectively) and division by their standard deviations (0.54 and 0.76). Finally, a principal component analysis (PCA) was carried out on the data. Let us recall that in the context of a random model where both variables (diversity and genome size) are subject to fluctuations, a PCA rather than a correlation test is the best choice for the study of correlations. Moreover, standardization of the data is convenient as diversity and genome size are measured in different units.

A plot of the standarized data with the direction or the first principal component is represented in figure C.1. The relative weigth of the first component is $s = 0.53$, very close to the expected value $s = 0.5$ for two-dimensional datasets with no correlation. It remains the same if the data are not logaritmically transformed. Thus, we conclude that the dataset shows no linear or power-law correlations between genome size and number of different families.

The study of correlations among family occurrences was addressed by calculating a normalized similitude matrix, that contains information on the similitude of the genome location between pairs of families. Prior to the calculation, each IS family is expressed as a binary vector with 1233 elements. Each element corresponds to a genome in the dataset, takes the value one if that genome contains the family and zero otherwise. As similitude indexes for pairs of families several choices are possible,

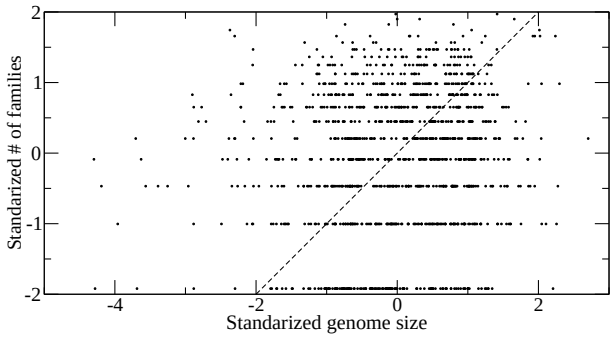


Figure C.1: Principal component analysis of the dataset containing genome sizes and number of different families per genome. The variables (genome size and number of families) were standarized by substraction of their means (14.93 and 1.45) and division by their standard deviations (0.54 and 0.76). The dashed line shows the direction of the first principal component. The relative weigth of this component is $s = 0.53$, close to the expected value $s = 0.5$ for two-dimensional datasets with no correlation.

based on the cosine associated to the vectors, the Spearman coefficient, the Hamming distance and the Jaccard index (the number of co-occurrences divided by the number of genomes where any of the families appears). The results in all cases are qualitatively the same. The similitude matrix for the real data may be biased by the non-random distribution of genomic diversity. In order to remove this bias, a null model that conserves the observed diversity distribution must be built. Such a model was obtained through a random redistribution of families within genomes, while keeping constant the original family abundances and the diversity distribution. This process, that removes any genuine co-occurrence pattern among IS families, was repeated in order to obtain 10^3 datasets of the null model. By using such datasets, the mean and the standard deviation of the similitude index for each pair of families was computed. Finally, normalized values for all similitude indexes were calculated by substracting the mean values to the observe ones and dividing by the standard deviations. The matrix built in this way (figure 7.4, main text) indicates the deviation between the observed similitude index and the expected one in units of standard deviation. Values greater (smaller) than zero correspond to pairs of families with more (fewer) co-occurrences than expected if their behaviors were independent of each other. The IS families in figure 7.4 are ordered according to their abundance, what makes evident a global pattern of negative (positive) co-occurrence among abundant (rare) families.

C.9 The preferential acquisition model

The observed diversity distribution contains a greater fraction of genomes with high and low number of different families than expected by chance. This suggests that the

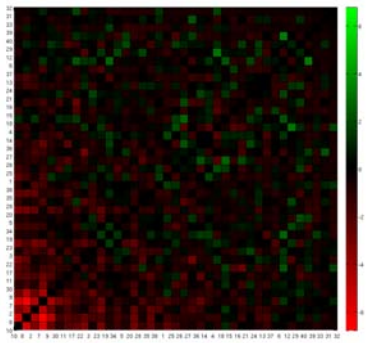


Figure C.2: Standarized co-occurrence matrix between pairs of IS families in a simulation of the preferential acquisition model ($q = 0.75$). For each pair, a co-occurrence index is calculated based on the cosine distance. By simulating a null model, the mean and standard deviation of each co-occurrence index are obtained. The figure shows the standarized values of the co-occurrence indexes, after substraction of the simulated mean and division by the standard deviation. Families are ordered according to their abundance (from more frequent to rarer), while numerical labels in the axes correspond to the alphabetical ordering of families. Red (green) indicates fewer (more) co-occurrences than expected.

acquisition of new families through HGT is easier the greater the number of families already in the genome.

In order to test that idea we have simulated the following process of IS gain, that we call the preferential acquisition model. Starting with a set of G empty genomes, we choose at every step a particular genome with probability π_x, that depends on the number of families that the genome hosts.

$$\pi_x = (1 + qx)/Q \tag{C.6}$$

Here x is the number of different IS families in the genome and Q is a normalization factor. We take a linear relation between π_x and x as the simplest possible choice. The proportionality factor q is a new parameter that tunes the effect of diversity on the acquisition probability. By setting $q = 0$ an independent acquisition process is recovered.

Once a genome has been chosen, an IS family is selected for infection. Family selection is done according to the relative family abundances in the original dataset. If the genome already contains the selected family nothing happens. Otherwise, the new family is added to the genome. The preferential acquisition process is iterated until the average number of family occurrences is the same as in the original dataset. That allows us to compare the simulated diversity distribution with the real one. The best fit

is obtained for $q = 0.75 \pm 0.1$. It corresponds to the red curve in figure 7.3 of the main text.

We have also addressed the possibility that the non-trivial co-occurrence pattern among families may be a by-product of the preferential acquisition process. To that end, we have processed the simulated datasets as described in the previous section. The result is a co-occurrence matrix where frequent families have fewer co-occurrences than expected, in a similar fashion as in the real data (compare figure C.2 with figure 7.4 in the main text).

The preferential acquisition model as described so far is not completely satisfactory as it is not stationary. That means that if the process is iterated enough times a final situation with all genomes containing all families is reached. It is easy to overcome this problem by complementing the preferential acquisition process with the subcritical neutral dynamics. Now, the subcritical duplication-deletion process removes families in the long term, thus avoiding total saturation of genomes. We have implemented this full model by writting the HGT-deletion ratio for each genome as a function of its diversity. Let us take a particular IS family with a mean abundance of $\langle k \rangle$, the HGT-deletion ratio in a genome hosting x different families becomes $\beta = \gamma \langle k \rangle (1 + qx)/Q$, where $Q = 1 + q\langle x \rangle$.

Simulations of the full model reveal that the stationary diversity distribution depends on the initial state of the population. In particular, the experimentally observed distribution is a valid stationary state of the full model with $q = 0.75$. On the other hand, if one starts with the experimental diversity distribution and sets $q = 0$, the distribution evolves until the random diversity distribution is recovered. Thus, preferential acquisition of IS families is required to maintain a diversity distribution like the experimental one.

D

Analytic calculations for the CRISPR-Cas model

This appendix details the mathematical developments supporting results and statements in Chapter 6. In the first section, we analyse the Lotka-Volterra equations that approximate our stochastic agent based predator-prey model in the limit of large populations and negligible fluctuations. Next, we present an estimation of the fraction of immune encounters, which can be expressed as a function of two key observables: the number of spacers per host genome and the total number of proto-spacers in the viral population. The last section explores the conditions required for CRISPR-Cas maintenance, by studying the stability of mean-field three species dynamics.

D.1 Lotka-Volterra dynamics in the absence of CRISPR-Cas

When the population size is large and fluctuations can be neglected, the stochastic agent based predator-prey model described in the main text can be approximated by a classical Lotka-Volterra (LV) system. Let N_b be the number of hosts and N_v the number of viruses. In the absence of CRISPR-Cas, neither adaptive immunity nor fitness cost have to be taken into account. In addition, mechanisms allowing for Cas gain/loss are not considered. The dynamical equations in the mean-field approximation read

$$\left\{ \begin{array}{rcl} \dot{N}_b & = & N_b \left(1 - b\,(1-s)\,N_v\right), \\ \dot{N}_v & = & N_v \left(b\,(M - Ms - s)\,N_b - d\right). \end{array} \right. \tag{D.1}$$

The non-trivial fixed point (N_b^*, N_v^*) of the dynamics is given by

$$(N_b^*, N_v^*) = \left(\frac{d}{b\,(M - Ms - s)}\,,\ \frac{1}{b\,(1-s)} \right). \tag{D.2}$$

The fixed point corresponds to a limit cycle in the LV dynamics. The behavior of a population close to that point can be described in terms of cycles with period $2\pi/\sqrt{d}$. It follows from Eq. (D.2) that the inverse of the encounter rate, b determines the size of the population. As a consequence, due to the dynamics of the Lotka-Volterra system, the encounter rate *per host* is independent of b.

Note that the stationary state host population size is substantially greater than unity only when $d \gg b\,(M - Ms - s)$. Also, fluctuations in discrete, finite size LV systems eventually lead to the extinction of the population. The limit cycle in the canonical LV system is marginally stable and therefore the extinction probability in such systems depends on the initial condition: the closer to the fixed point, the longer the population survives. Moreover, the survival time in finite canonical LV systems scales linearly with the system size. As a result, an inverse relation between parameter b and the survival time can be expected in the absence of CRISPR-Cas. In contrast, the mean survival time of a non-canonical LV system with a stable limit cycle is expected to grow exponentially with the system size.

Finally, if a fitness cost c is assigned to hosts the stationary population size for the virus becomes $N_v^* = \big(b\,(1-s)(1+c)\big)^{-1}$, while N_b^* remains unchanged. The period of the oscilations around the equilibrium point is $2\pi/\sqrt{d\,(1+c)}$.

D.2 Estimation of the adaptive immune probability p_c

The probability p_c of a CRISPR-mediated immune encounter can be calculated as one minus the probability that none of the spacers in the host genome coincide with the proto-spacers of the virus. Let s_i be the i-th proto-spacer in a virus genome (that contains N_s proto-spacers) and S the set of all different proto-spacers in the viral population (the cardinality if S is N_t). Host spacers will be denoted as m_i. For a genome with L spacers, the adaptive immunity can be written as

$$p_c = 1 - \prod_{i=1}^{N_s}\prod_{j=1}^{L}(1 - \text{Prob}\{s_i = m_j\}) = 1 - \prod_{i=1}^{N_s}(1 - F_{s_i})^L, \tag{D.3}$$

where F_s has been defined as the probability of finding spacer s in a random sampling of the whole set of spacers in the host population. The use of F_s in the Eq. (D.3) implies that genomic correlations between spacers are neglected. Under the assumption that $F_s \ll 1$, it makes sense to define $f_s(L) = LF_s$ as the probability that a host with a total of L spacers contains the spacer s. Note that this assumption is reasonable unless a spacer is present at a high frequency. The next step is to define the probability G_s of finding proto-spacer s in the complete set S of viral proto-spacers. With this definitiion

it can be written:

$$p_c = 1 - \prod_{i=1}^{N_s} \sum_{k \in S} \Pr\{s_i = k\}(1 - f_k(L)) = 1 - \prod_{i=1}^{N_s} \sum_{k \in S} G_k(1 - f_k(L)). \quad \text{(D.4)}$$

If genomic correlations between proto-spacers in viral genomes are neglected, the previous expression becomes:

$$p_c = 1 - \left(\sum_{k \in S} G_k(1 - f_k(L)) \right)^{N_s}.$$

If viral genomes do not contain repeated proto-spacers, the fraction of genomes that contain the proto-spacer k can be expressed as a new variable $g_k = N_s G_k$. After making the change of variables and the expression for p_c becomes

$$p_c = 1 - (1 - \chi)^{N_s} \quad \text{(D.5)}$$

The immune parameter χ can be expressed in several ways as a function of the genomic spacer distributions (f and g) or the global ones (F and G):

$$\chi = N_s^{-1} \sum_{k \in S} f_k g_k = L \sum_{k \in S} F_k G_k. \quad \text{(D.6)}$$

Further insight can be gained if the average values $\langle fg \rangle$ (and the same for F and G) are used in Eq. (D.6). The immune parameter χ then becomes

$$\chi = \frac{N_t}{N_s} \langle fg \rangle = N_t L \langle FG \rangle,$$

where N_t is the total number of distinct proto-spacers. Averages of products between f and g can be written as a function of their correlation in a simple way

$$\langle fg \rangle = C(f, g)\sigma(f)\sigma(g) + \langle f \rangle \langle g \rangle,$$

where $C(f, g)$ denotes the correlation coefficient and σ the standard deviation. Let us define the coefficients of variation (CV) for f and g as their standard deviations divided by their means. It is easy to see that $\langle f \rangle = L/N_t$ and $\langle g \rangle = N_s/N_t$. Substituting these values one gets the final expression or χ

$$\chi = \frac{L}{N_t} \big(1 + C(f, g)\, \mathrm{CV}(f)\, \mathrm{CV}(g)\big), \quad \text{(D.7)}$$

which in combination with expression (D.5) gives the estimated probability of an adaptive immune encounter for a host with L spacers. The total immunity, p, is obtained by adding the probability of an innate immune encounter

$$p = s + p_c(1 - s). \quad \text{(D.8)}$$

D.3 CRISPR-Cas conservation in a three species LV model

To derive the conditions of CRISPR-Cas maintenance in the host population, we introduce the three species model, i.e. the mean field approximation of the full stochastic agent based model which is valid when fluctuations can be ignored and the fraction of immune encounters for CAS+ hosts is taken as a constant parameter p. We allow for non-saturated horizontal transfer by adding an extra parameter K, which plays the role of a Michaelis constant for the horizontal transfer rate as a function of the host population size. The population size N_{b+} of CAS+ hosts, N_{b-} of CAS− hosts, and N_v of viruses obey the dynamic equations

$$\begin{cases} \dot{N}_{b+} &= N_{b+}\left(\frac{1-\lambda}{1+c} + \frac{\sigma N_{b-}}{K+N_{b+}+N_{b-}} - b\,(1-p)\,N_v\right), \\ \dot{N}_{b-} &= N_{b-}\left(1 - \frac{\sigma N_{b+}}{K+N_{b+}+N_{b-}} - b\,(1-s)\,N_v\right) + \frac{\lambda N_{b+}}{1+c}, \\ \dot{N}_v &= N_v\big(b\,(M-Mp-p)\,N_{b+} + b\,(M-Ms-s)\,N_{b-} - d\big). \end{cases} \tag{D.9}$$

There is a simple solution with $N^*_{b+} = 0$, corresponding to the loss of CRISPR-Cas in the population. On the other hand, the solution where CRISPR-Cas is maintained takes a complicated form, with CAS+ and CAS− hosts coexisting in the population. The condition for of CRISPR-Cas the maintenance can be obtained by looking at the stability of the $N_{b+} = 0$ solution to Eqs. (D.9). Specifically, it becomes unstable (and thus CRISPR-Cas is maintained) when

$$\frac{1-p}{1-s} < \frac{1-\lambda}{1+c} + \frac{\sigma}{1+K/N^*_{b-}}, \tag{D.10}$$

where $N^*_{b-} = d/(b(M-s-Ms))$ is the equilibrium abundance of CAS− hosts, which is identical to that in Eq. (D.2).

Therefore CRISPR-Cas is maintained when it provides an average immunity greater than,

$$p_{min} = 1 - \left(\frac{1-\lambda}{1+c} + \frac{\sigma}{1+K/N^*_{b-}}\right)(1-s). \tag{D.11}$$

A scenario with saturated horizontal transfer is recovered if $N^*_{b-} \gg K$. In such a case, the CRISPR efficacy threshold becomes $p_{min} = 1-((1-\lambda)/(1+c)+\sigma)(1-s)$. By comparing this value with the general expression in Eq. (D.11), it can be concluded that the saturated horizontal transfer scenario is the most favorable for CRISPR-Cas maintenance.

Glossary

Combination therapy: treatment that consists of the administration of two or more drugs.

Complementation: increase of viral progeny production mediated by gene products supplied by another virus (in quasispecies, supplied by closely related variants). The same concept can be applied to populations of selfish genetic elements, such as transposable elements.

Complexity of a mutant spectrum: number of mutations and genomic sequences in a viral population. It is often quantified by pairwise genetic distances, mutation frequency (calculated by dividing the number of different mutations by the total number of nucleotides sequenced), and Shannon entropy (poportion of different genomes in the population). New technologies should allow a quantitative characterization of quasispecies complexity in terms of phenotypic diversity.

Consensus sequence: in a set of aligned nucleotide or amino acid sequences, the one that results from taking the most common residue at each position.

Defective: this term has several meanings. In viral populations, it may refer to mutant genomes that can replicate either on their own or under complementation, usually in the presence of the wild type. They can interfere actively with replication of the standard virus if the latter sequester nonfunction or poorly functional *trans*-acting products expressed by the defectors. In the model of lethal defection, for instance, defectors have lost their ability to infect susceptible cells.

Error rate: term used as a synonimous of mutation rate in the context of viral replication.

Error threshold: a theoretical average error rate that sets a maximum limit for maintenance of the genetic information encoded by a replicating system. Error rates above the error threshold lead to loss of genetic information, also termed error catastrophe.

Fitness: when referred to in regard to viruses (or bacteria), fitness means the replicative capacity measured relative to some virus variant (or bacterian strain) taken as a reference. Fitness is environment-dependent.

Fixation: in population genetics, fixation is the result of any process by which an allele or genetic variant that represents a fraction of the population spreads to the whole population, while the alternative alleles disappear. The term can be generalized to refer to variants of a virus, the presence/absence of a gene, or qualitative phenotypical traits.

Horizontal gene transfer (HGT): transfer of genetic material between organisms through mechanisms other than vertical (from parent to offspring) inheritance.

Insertion sequence (IS): a class of transposable elements found in prokaryotes, characterized by their small size and the fact that they only code for proteins implicated in transposition.

Interference: this term has several meanings in biology. In this thesis, it means the capacity of viral genomes to reduce he replicative activity of higher fitness genomes through *trans*-acting interactions. It can be regarded as the converse of complementation.

Lethal mutagenesis: viral extinction achieved through an excess of mutations, often promoted by mutagenic nucleotide analogs during viral genome replication,

Master sequence: the genomic nucleotide sequence that dominates a mutant spectrum because of its superior fitness. It may or may not be identical to the consensus sequence. The most abundant genome may still be a minority relative to the ensemble of los frequency variants. Owing to the abundance of quasineutral mutations and epistatic interactions in viral genomes, there might be a large ensemble of sequences of almost identical fitness that compose a "master phenotype".

Monotherapy: treatment that consists of the administration of a single drug.

Multipartite virus: a peculiar class of virus whose genome consists of several fragments (termed segments), each of them being separately encapsidated. At least one representative of each segment must simultaneously infect a cell in order to develop a successful infection.

Multiplicity of infection (MOI): in a viral infection, average number of viral particles that enter the same cell. In experimental settings, it is computed as the ratio of infectious particles to the number of target cells.

Mutant spectrum: the ensemble of mutant genomes that compose a viral quasispecies. It is also termed mutant swarm or mutant cloud.

Mutation frequency: the proportion of mutated sites in a population of genomes. It is often calculated by dividing the number of different mutations found in a mutant spectrum by the total number of nucleotides sequenced.

Mutation rate: the frequency of occurrence of a mutation during genome replication.

Purifying selection: also termed negative selection, it is the selective removal of alleles or genetic variants that are deleterious.

Rate of evolution: the frequency of mutations that become dominant (i.e., are represented in the consensus sequence) as a function of time. In the case of viruses, it may refer to evolution within a host individual or upon epidemic expansion of the virus.

Stabilizing selection: a type of natural selection in which genetic diversity decreases as the population stabilizes on a particular trait value. It may be the result of purifying selection acting against extreme values of the trait.

Trans-acting: in molecular biology, any kind of action that comes from a different source. In this thesis, it often alludes to proteins that may act on genomes other than those coding for them. Proteins that act in *trans* may be "shared" inside the cell.

Transposable element (TE): a DNA sequence able to move along the genome that hosts it. The process by which a TE changes its genomic location is termed transposition, and it can produce a duplication of the TE. There are multiple families of TEs, differing on their size, gene content, and trasposisiton mechanism.

Viral quasispecies: a set of viral genomes that belong to a replicative unit, subject to genetic variation, competition, and selection, and which acts as a unit of selection. It has been extended to denote ensembles of similar viral genomes generated by a mutation-selection process.

Wild type (wt): term that refers to the typical form of a species (or virus) as it occurs in nature. It is commonly used in contrast to non-standard, mutant forms.

Publications

The original content of this thesis appears in the following papers:

- Chapter 2:
  - *Stochastic extinction of viral infectivity through the action of defectors*, J. Iranzo and S. C. Manrubia, *Europhys. Lett.* **85**, 18001 (2009).
- Chapter 3:
  - *Tempo and mode of inhibitor-mutagen antiviral therapies: a multidisciplinary approach*, J. Iranzo, C. Perales, E. Domingo and S. C. Manrubia, *Proc. Natl. Acad. Sci. USA* **108**, 16008-16013 (2011).
  - *The impact of quasispecies dynamics on the use of therapeutics*, C. Perales, J. Iranzo, S. C. Manrubia and E. Domingo, *Trends Microbiol.* **20**, 595-603 (2012).
- Chapter 4:
  - *Evolutionary dynamics of genome segmentation in multipartite viruses*, J. Iranzo and S. C. Manrubia, *Proc. R. Soc. Lond. B* **279**, 3812-3819 (2012).
- Chapter 5:
  - *Neutral punctuated dynamics of insertion sequences in prokaryotic genomes: insigths from a large scale genomic analysis*, J. Iranzo, M. J. Gómez, F. J. López de Saro and S. C. Manrubia, in preparation.
- Chapter 6:
  - *Evolutionary dynamics of prokaryotic adaptive immunity systems, CRISPR-Cas, in an explicit ecological context*, J. Iranzo, A. E. Lobkovsky, Y. I. Wolf and E. V. Koonin, submitted to *J. Bacteriol.* (2013).

The following papers were also written in the same period, but do not keep relation with the topics treated in this thesis:

- *The Ultimatum Game in complex networks*, R. Sinatra, J. Iranzo, J. Gómez-Gardeñes, L. M. Floría, V. Latora and Y. Moreno, *J. Stat. Mech.* P09012 (2009).

- *The spatial Ultimatum Game revisited*, J. Iranzo, J. Román and A. Sánchez, *J. Theor. Biol.* **278**, 1-10 (2011).

- *Empathy emerges spontaneously in the Ultimatum Game: small groups and networks*, J. Iranzo, L. M. Floría, Y. Moreno and A. Sánchez, *PLoS ONE* **7**, e43781 (2012).

Abstract

This thesis deals with the mathematical modeling of evolutionary processes that take place in heterogeneous populations. Its leitmotif is the response of complex ensembles of replicating entities to multiple—and often opposite—selection pressures. Even though the specific problems here addressed belong to different organizational levels—genome, population and community—all of them can be conceptualized as the evolution of a heterogeneous population—let it be a population of genomic elements, viruses or prokaryotic hosts and phages—facing a complex environment. As a result, the mathematical tools required for their study are quite similar. In contrast, the strategies that each population has discovered to perpetuate vary according to the different evolutionary challenges and environmental constraints that the population experiences.

Along this thesis, there has been a special interest on connecting theoretical models with experimental results. To that end, most of the work presented here has been motivated either by laboratory findings or by the bioinformatic analysis of sequenced genomes. We strongly believe that such a multidisciplinary approach is necessary in order to improve our knowledge on how evolution works. Moreover, experiments are a must when it comes to propose antiviral strategies based on theoretical predictions.

This thesis is structured in two main blocks. The first one focuses on studying instances of viral evolution under the action of mutagenic drugs, paying particular attention to their possible application to the development of novel antiviral therapies. Within this block, we first discuss the phenomenon of lethal defection, by which defective individuals that appear in a viral population after treatment with small doses of a mutagenic drug can lead to the stochastic extinction of the virus. We analyze the factors required for this phenomenon to occur, and find two key contitions: first, the size of the intracellular viral population must be relatively small; second, the viral infection must be persistent. If both conditions are fulfilled, lethal defection becomes possible. Next, we study the optimal way of combining mutagens and inhibitors in multidrug antiviral treatments. According to our model, that has later been experimentally tested, the optimal protocol for drug administration depends in a predictable way on the action mechanism of the drugs and the drug doses. When a mutagen and an inhibitor are selected, the best choice for most drug doses is a sequential inhibitor-mutagen therapy.

The convenience of a sequential protocol has strong implications for clinical practice, as it allows to reduce the risk of side-effects and undesired interactions among drugs.

The second block of the thesis is devoted to the study of the evolutionary forces that shape genome structure. Based on experimental observations, we propose a mechanism through which multipartite viruses could have originated. Interestingly, the pathway leading to genome segmentation shares some steps with lethal defection, but each outcome is reached at specific environmental conditions. Going deeper at the genomic scale, we dedicate a chapter to analyse the abundance distribution of transposable elements in prokaryotic genomes, with the aim of determining the key processes involved in their spreading. We explicitly explore the hypothesis that transposable elements follow a neutral dynamics, so that they entail a negligible fitness cost for their host genomes. We also propose a mechanism for explaining transient episodes of transposon proliferation (punctuations), that according to some authors would be of great relevance for understanding evolution at greater scales. In the final part of this block, a higher level of organization is studied. There, an agent based coevolutionary model based on Lotka-Volterra interactions is used to investigate the evolutionary dynamics of the prokaryotic antiviral immunity system CRISPR-Cas. We examine the environmental factors that are responsible of its maintenance of loss, concluding that there exists a critical value of viral diversity that makes CRISPR-Cas useless. According to that, CRISPR-Cas is preferentially found in prokaryotes that live in extreme environments, where phage populations are small and not very diverse.

In sum, this thesis shows how biological populations at a variety of scales share some properties that derive from their fast evolution and high adaptability, that giving rise to a series of (sometimes counterintuitive) characteristic evolutionary phenomena where opposite selection pressures come into play. A common set of mathematical and computational tools can be employed to study such populations and build models that, once experimentally validated, provide useful knowledge with applications that range from understanding genome evolution to developing novel antiviral therapies.

Resumen

Esta tesis trata de la aplicación de modelos matemáticos sencillos al estudio de los procesos evolutivos que tienen lugar en poblaciones heterogéneas. Su hilo conductor se plasma en el análisis una serie de fenómenos que surgen cuando un cojunto de entidades capaces de replicarse se enfrenta a múltiples presiones de selección. Si bien desde el punto de vista biológico los casos aquí referidos ocurren a distintos niveles de organización (genoma, población y comunidad) todos ellos responden a un mismo patrón conceptual: la dinámica evolutiva de una población heterogénea (sea ésta formada por elementos genómicos, virus o bacterias y fagos) en un contexto ambiental complejo. Dicha unidad conceptual permite utilizar un reducido número de herramientas matemáticas y computacionales para abordar una gran variedad de situaciones. En cada una de estas situaciones, por el contrario, las estrategias concretas que cada población haya desarrollado variarán en función de los distintos retos evolutivos y restricciones ambientales a los que la población se haya visto sujeta.

A lo largo de la tesis se ha prestado un interés especial por conectar modelos teóricos con observaciones experimentales. Por esta razón, la mayor parte del trabajo aquí presentado ha sido motivado bien por resultados de laboratorio, bien por el análisis bioinformático de secuencias genómicas. Creemos firmemente que una aproximación multidisciplinar es fundamental para entender mejor cómo funciona la evolución. Además, la verificación experimental es un deber cuando se pretende proponer nuevas estrategias terapéuticas a partir de predicciones teóricas.

La tesis está estructurada en dos grandes bloques. El primero se centra en el estudio de la evolución viral bajo la acción de fármacos mutagénicos, prestándose una especial atención a sus posibles aplicaciones en el desarrollo de terapias antivirales. Dentro de este bloque se discute el fenómeno denominado ”defección letal”, que consiste en la extinción estocástica de una población de virus provocada por la acción de genomas virales defectivos que se originan al exponer a la población a dosis pequeñas de mutágeno. Tras analizar las condiciones que pueden dar lugar a este fenómeno, encontramos dos requisitos esenciales: primero, que la infección viral sea persistente; segundo, que el número de genomas virales dentro de las células infectadas sea relativamente pequeño. En tales circunstancias, puede esperarse que la extinción viral por defección letal tenga lugar de manera natural. A continuación, pasamos a estudiar la

forma óptima de combinar mutágenos e inhibidores de la replicación en una terapia antiviral múltiple. Según nuestro modelo, posteriormente validado en el laboratorio, el protocolo de administración óptimo puede dictaminarse a partir del mecanismo de acción de los fármacos utilizados y de su dosis. En particular, cuando se combina un inhibidor con un mutágeno, la mejor elección para la mayoría de las dosis consiste en un tratamiento secuencial inhibidor-mutágeno. Este resultado encierra un gran valor clínico, ya que una terapia secuencial reduciría el riesgo de efectos secundarios e interacciones no deseadas entre fármacos.

El segundo bloque de la tesis está dedicado al estudio de las fuerzas evolutivas que dan forma a los genomas. Basándonos en observaciones experimentales, comenzamos este bloque poniendo a prueba un mecanismo que podría haber dado lugar a la aparición de los virus multipartitos (virus con un genoma fragmentado en varios segmentos que se empaquetan por separado). Es interesante señalar que el proceso que lleva a la segmentación del genoma viral comparte algunos pasos con la defección letal, aunque el resultado del proceso es totalmente distinto según cuáles sean las condiciones externas. Con la intención de profundizar en el estudio de la estructura y composición de los genomas, dedicamos un capítulo a analizar las distribuciones de abundancia de ciertos elementos móviles (transposones) en genomas bacterianos. Con este análisis pretendemos dilucidar cuáles son los procesos clave implicados en su proliferación, mantenimiento y desaparición. De manera explícita, examinamos la hipótesis según la cual los transposones siguen una dinámica neutral, sin que su presencia suponga un coste apreciable para el genoma que los contiene. Proponemos además un mecanismo para explicar los episodios transitorios de proliferación que se observan en algunos transposones y que, según algunos autores, podrían ser de gran relevancia para entender la evolución a escalas mayores. Para cerrar el bloque, afrontamos un sistema perteneciente a un nivel de organización superior: una comunidad de procariotas (bacterias o arqueas) y fagos en coevolución. Para este fin planteamos un modelo basado en agentes inspirado en una interacción de tipo Lotka-Volterra entra depredadores y presas, y lo aplicamos al estudio de la dinámica evolutiva del sistema CRISPR-Cas, que proporciona a los procarioatas que lo poseen inmunidad frente a los fagos. El resultado principal de este apartado es que existe un valor crítico de diversidad viral por encima del cual el sistema CRISPR-Cas se vuelve incapaz de proveer inmunidad y, puesto que su mantenimiento implica un coste, acaba por perderse. En consecuencia, y de acuerdo con las observaciones experimentales, el sistema CRISPR-Cas se encuentra presente de forma mayoritaria en procariotas que habitan ambientes extremos, en los cuales las poblaciones de fagos son pequeñas y poco diversas.

En conclusión, esta tesis muestra cómo poblaciones de elementos que se replican, pertenecientes a distintas escalas biológicas, comparten ciertas propiedades que derivan de su rápida tasa de evolución y elevada adaptabilidad, y que a su vez dan lugar a una serie de fenómenos evolutivos característicos (y a veces no intuitivos) en los que entran en juego presiones de selección opuestas. Para estudiar la dinámica evolutiva de tales poblaciones puede emplearse un conjunto de herramientas matemáticas y computacionales comunes, generando modelos que, una vez validados, pueden aplicarse

a un amplio rango de problemas en biología, desde la evolución del genoma hasta el desarrollo de nuevas terapias antivirales.

References

Aguirre J., Lázaro E., and Manrubia S. C. (2009). A trade-off between neutrality and adaptability limits the optimization of viral quasispecies. *J. Theor. Biol.* **261**, 148–155.

Akaike H. (1974). New look at statistical model identification. *IEEE T. Automa. Contr.* **19**, 716–723.

Anderson J. P., Daifuku R., and Loeb L. A. (2004). Viral error catastrophe by mutagenic nucleosides. *Annu. Rev. Microbiol.* **58**, 183–205.

Andersson A. F., and Banfield J. F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**, 1047–1050.

Ball L. A. (2007). Virus replication strategies. In Knipe D. M., and Howley P. M. (Eds.), *Fields virology*, Volume 1, pp. 119–140. Philadelphia, PA: Lippincott Williams and Wilkins.

Bandi C., Anderson T. J. C., Genchi C., and Blaxter M. L. (1998). Phylogeny of *Wolbachia* in filarial nematodes. *Proc. R. Soc. Lond. B* **265**, 2407–2413.

Bangham C. R. M., and Kirkwood T. B. L. (1990). Defective interfering particles: effects in modulating virus growth and persistence. *Virology* **179**, 821–826.

Barrangou R., Fremaux C., Deveau H., Richards M., Boyaval P., Moineau S., Romero D. A., and Horvath P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712.

Basten C. J., and Moody M. E. (1991). A branching-process model for the evolution of transposable elements incorporating selection. *J. Math. Biol.* **29**, 743–761.

Batschelet E., Domingo E., and Weissmann C. (1976). The proportion of revertant and mutant phage in a growing population, as a function of mutation and growth rate. *Gene* **1**, 27–32.

Belsham G. J. (2005). Translation and replication of FMDV RNA. *Curr. Top. Microbiol. Immunol.* **288**, 43–70.

Betancourt M., Fereres A., Fraile A., and García-Arenal F. (2008). Estimation of the effective number of founders that initiate an infection after aphid transmission of a multipartite plant virus. *J. Virol.* **82**, 12416–12421.

Bhaya D., Davison M., and Barrangou R. (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defensse and regulation. *Annu. Rev. Genet.* **45**, 273–297.

Bichsel M., Barbour A. D., and Wagner A. (2012). Estimating the fitness effect of an insertion sequence. *J. Math. Biol.* **66**, 95–114.

Biebricher C. K., and Eigen M. (2005). The error threshold. *Virus Res.* **107**, 117–127.

Bikard D., Hatoum-Aslan A., Mucida D., and Marraffini L. A. (2012). CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe* **12**, 177–186.

Blower T. R., Evans T. J., annd P. C. Fineran R. P., and Salmond G. P. (2012). Viral evasion of a bacterial suicide system by RNA-based molecular mimicry enables infectious altruism. *PLoS Genet.* **8**, e1003023.

Blower T. R., Salmond G. P., and Luisi B. F. (2011). Balancing at survival's edge: the structure and adaptive benefits of prokaryotic toxin-antitoxin partners. *Curr. Opin. Struct. Biol.* **21**, 109–118.

Bonhoeffer S., May R. M., Shaw G. M., and Nowak M. A. (1997). Virus dynamics and drug therapy. *Proc. Natl. Acad. Sci. USA* **94**, 6971–6976.

Breitbart M., and Rohwer F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**, 278–284.

Brookfield J. F. Y. (2005). The ecology of the genome - Mobile DNA elements and their hosts. *Nat. Rev. Genet.* **6**, 128–136.

Bull J. J., Meyers L. A., and Lachmann M. (2005). Quasispecies made simple. *PLoS Comput. Biol.* **1**, 450–460.

Bull J. J., Sanjuán R., and Wilke C. O. (2007). Theory of lethal mutagenesis for viruses. *J. Virol.* **81**, 2930–2939.

Carroll D. (2012). A CRISPR approach to gene targeting. *Mol. Ther.* **20**, 1658–1660.

Cases-González C., Arribas M., Domingo E., and Lázaro E. (2008). Benefitial effects of population bottlenecks in an RNA virus evolving at increased error rate. *J. Mol. Biol.* **384**, 1120–1129.

Cerveau N., Leclercq S., Leroy E., Bouchon D., and Cordaux R. (2011). Short- and long-term evolutionary dynamics of bacterial insertion sequences: insights from Wolbachia endosymbionts. *Genom. Biol. Evol.* **3**, 1175–1186.

Chakraborty S., Snijders A. P., Chakravorty R., Ahmed M., Tarek A. M., and Hossain M. A. (2010). Comparative network clustering of direct repeats (DRs) and cas genes confirms the possibility of the horizontal transfer of CRISPR locus among bacteria. *Mol. Phylogenet. Evol.* **56**, 878–887.

Chandler M., and Mahillon J. (2002). Insertion sequences revisited. In Craig N. L., Craigie R., Gellert M., and Lambowitz A. M. (Eds.), *Mobile DNA II*, pp. 305–366. Washington DC: ASM Press.

Chao L. (1991). Levels of selection, evolution of sex in RNA viruses, and the origin of life. *J. Theor. Biol.* **153**, 229–246.

Chare E. R., and Holmes E. C. (2006). A phylogenetic survey of recombination frequency in plant RNA viruses. *Arch. Virol.* **151**, 933–946.

Charlesworth B., and Charlesworth D. (1983). The population dynamics of transposable elements. *Genet. Res.* **42**, 1–27.

Childs L. M., Held N. L., Young M. J., Whitaker R. J., and Weitz J. S. (2012). Multiscale model of CRISPR-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin. *Evolution* **66**, 2015–2029.

Christie-Oleza J. A., Nogales B., Martín-Cardona C., Lanfranconi M. P., Albertí S., Lalucat J., and Bosch R. (2008). IS*Pst*9, an IS3-like insertion sequence from *Pseudomonas stutzeri* AN10 involved in catabolic gene inactivation. *Int. Microbiol.* **11**, 101–110.

Cong M.-e., Heneine W., and García-Lerma J. G. (2007). The fitness cost of mutations associated with Human Immunodeficiency Virus type 1 drug resistance is modulated by mutational interactions. *J. Virol.* **81**, 3037–3041.

Crotty S., Cameron C. E., and Andino R. (2001). RNA virus error catastrophe: direct molecular test using ribavirin. *Proc. Natl. Acad. Sci. USA* **98**, 6895–6900.

Darwin C. (1859). *On the Origins of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (1st ed.). London: John Murray.

Deveau H., Barrangou R., Garneau J. E., Labonte J., Fremaux C., Boyaval P., Romero D. A., Horvath P., and Moineau S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J./ Bacteriol.* **190**, 1390–1400.

Deveau H., Garneau J. E., and Moineau S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol.* **64**, 475–493.

Domingo E. (1989). RNA virus evolution and the control of viral disease. *Prog. Drug Res.* **33**, 93–133.

Domingo E. (2006). *Quasispecies: Concepts and Implications for Virology*, Volume 299 of *Current Topics in Microbiology and Immunology*. Springer.

Domingo E., Dávila M., and Ortín J. (1980). Nucleotide sequence heterogeneity of the RNA from a natural population of foot-and-mouth disease virus. *Gene* **11**, 333–346.

Domingo E., Escarmís C., Lázaro E., and Manrubia S. C. (2005). Quasispecies dynamics and RNA virus extinction. *Virus Res.* **107**, 129–139.

Domingo E., Grande-Pérez A., and Martín V. (2008). Future prospects for the treatment of rapidly evolving viral pathogens: insights from evolutionary biology. *Expert Opin. Biol. Ther.* **8**, 1455–1460.

Domingo E., and Holland J. J. (1992). Complications of RNA heterogeneity for the engineering of virus vaccines and antiviral agents. In Setlow J. K. (Ed.), *Genetic Engineering, Principles and Methods*, Volume 14, pp. 13–31. New York: Plenum Press.

Domingo E., Menéndez-Arias L., Quiñones-Mateu M. E., Holguín A., Gutiérrez-Rivas M., Martínez M. A., Novella I. S., and Holland J. J. (1997). Viral quasispecies and the problem of vaccine-escape and drug-resistant mutants. *Prog. Drug Res.* **48**, 99–128.

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

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

Donalson D. D., and Nisbet R. M. (1999). Population dynamics and spatial scale: effects of system size on population persistence. *Ecology* **80**, 2492–2507.

Doolittle W. F., and Sapienza C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603.

Drake J. W. (2009). Avoiding dangerous missense: thermophiles display especially low mutation rates. *PLoS Genet.* **5**, e1000520.

Drake J. W., and Holland J. J. (1999). Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci. USA* **96**, 13910–13913.

Eigen M. (1971). Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523.

Eigen M. (1993). Viral quasispecies. *Scient. Am.* **296**, 32–39.

Eigen M. (2002). Error catastrophe and antiviral strategy. *Proc. Natl. Acad. Sci. USA* **99**, 13374–13376.

Eigen M., and Schuster P. (1979). *The Hypercycle. A Principle of Natural Self-organization*. Berlin: Springer-Verlag.

Endy D., and Yin J. (2000). Toward antiviral strategies that resit viral escape. *Antimicrob. Agents Chemother.* **44**, 1097–1099.

Erdmann S., and Garrett R. A. (2012). Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol. Microbiol.* **85**, 1044–1056.

Fabre L., Zhang J., Guigon G., Le Hello S., Guibert V., Accoun-Demartin M., de Romans S., Lim C., Roux C., Passet V., Diancourt L., Guibourdenche M., Issenhuth-Jeanjean S., Achtman M., Brisse S., Sola C., and Weill F. X. (2012). CRISPR typing and subtyping for improved laboratory surveillance of Salmonella infections. *PLoS ONE* **7**, e36995.

Fernández A., and Lynch M. (2011). Non-adaptive origins of interactome complexity. *Nature* **474**, 502–505.

Ferrer-Orta C., Arias A., Escarmís C., and Verdaguer N. (2006). A comparison of viral RNA-dependent RNA polymerases. *Curr. Opin. Struct. Biol.* **16**, 27–34.

Fischer S., Maier L. K., Stoll B., Brendel J., Fischer E., Pfeiffer F., Dyall-Smith M., and Marchfelder A. (2012). An archaeal immune system can detect multiple protospacer adjacent motifs (PAMs) to target invader DNA. *J. Biol. Chem.* **287**, 33351–33363.

Fisher R. A. (1930). *The genetical theory of natural selection*. London: Clarendon Press.

Fitzgerald J. B., Schoeberl B., Nielsen U. B., and Sorger P. K. (2006). Systems biology and combination therapy in the quest for clinical efficacy. *Nat. Chem. Biol.* **2**, 458–466.

Forterre P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.* **117**, 5–16.

Forterre P., and Prangishvili D. (2009). The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Ann. NY Acad. Sci.* **1178**, 65–77.

Frank S. A. (2000). Within-host spatial dynamics of viruses and defective interfering particles. *J. Theor. Biol.* **206**, 279–290.

Friedberg E. C., Walker G. C., Siede W., Wood R. D., Schultz R. A., and Ellenberger T. (2006). *DNA Repair and Mutagenesis*. ASM Press.

Frost S. D. W., Dumaurier M. J., Wain-Hobson S., and Leigh-Brown A. J. (2001). Genetic drift and within host metapopulation dynamics of HIV-1 infection. *Proc. Natl. Acad. Sci. USA* **98**, 6975–6980.

Fuhrman J. A. (2002). Community structure and function in prokaryotic marine plankton. *Antonie Van Leeuwenhoek* **81**, 521–527.

García-Arriaza J., Manrubia S. C., Toja M., Domingo E., and Escarmís C. (2004). Evolutionary transition toward defective RNAs that are infectious by complementation. *J. Virol.* **78**, 11678–11685.

García-Arriaza J., Ojosnegros S., Dávila M., Domingo E., and Escarmís C. (2006). Dynamics of mutation and recombination in a replicating population of complementing, defective viral genomes. *J. Mol. Biol.* **360**, 558–572.

Geigenmüller-Gnirke U., Weiss B., Wright R., and Schlesinger S. (1991). Complementation between Sindbis viral RNAs produces infectious particles with a bipartite genome. *Proc. Natl. Acad. Sci. USA* **88**, 3253–3257.

Geller R., Vignuzzi M., Andino R., and Frydman J. (2007). Evolutionary constraints on chaperone-mediated folding provide an antiviral approach refractory to development of drug resistance. *Genes Dev.* **21**, 195–205.

Gibbs M. J., and Weiller G. F. (1999). Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. *Proc. Natl. Acad. Sci. USA* **96**, 8022–8027.

Gillespie D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361.

González-Jara P., Fraile A., Canto T., and García-Arenal F. (2009). The multiplicity of infection of a plant virus varies during colonization of its eukaryotic host. *J. Virol.* **83**, 7487–7494.

González-López C., Arias A., Pariente N., Gómez-Mariano G., and Domingo E. (2004). Pre-extinction viral RNA can interfere with infectivity. *J. Virol.* **78**, 3319–3324.

Goren M., Yosef I., Edgar R., and Quimron U. (2012). The bacterial CRISPR/Cas system as analog of the mammalian adaptive immune system. *RNA Biol.* **9**, 549–554.

Gould S. J., and Lewontin R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. B* **205**, 581–598.

Graci J. D., and Cameron C. E. (2004). Quasispecies, error catastrophe, and the antiviral activity of ribavirin. *Vet. Microbiol.* **98**, 103–109.

Grande-Pérez A., Lázaro E., Lowenstein P., Domingo E., and Manrubia S. C. (2005). Suppression of viral infectivity through lethal defection. *Proc. Natl. Acad. Sci. USA* **102**, 4448–4452.

Grimm V., and Railsback S. F. (2004). *Individual-based Modeling and Ecology*. Princeton University Press.

Gronenborn B. (2004). Nanoviruses: genome organization and protein function. *Vet. Microbiol.* **98**, 103–109.

Gutiérrez S., Yvon M., Thébaud G., Monsion B., Michalakis Y., and Blanc S. (2010). Dynamics of the multiplicity of cellular infection in a plant virus. *PLoS Pathog.* **6**, e1001113.

Haaber J., Samson J. E., Labrie S. J., Campanacci V., Cambillau C., Moineau S., and Hammer K. (2010). Lactococcal abortive infection protein AbiV interacts directly with the phage protein SaV and prevents translation of phage proteins. *Appl. Environ. Microbiol.* **76**, 7085–7092.

Haerter J. O., and Sneppen K. (2012). Spatial structure and Lamarckian adaptation explain extreme genetic diversity at CRISPR locus. *mBio* **3**, e00126–12.

Haerter J. O., Trusina A., and Sneppen K. (2011). Targeted bacterial immunity buffers phage diversity. *J. Virol.* **85**, 10554–10560.

Haft D. H., Selengut J., Mongodin E. F., and Nelson K. E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1**, e60.

Haldane J. B. S. (1924). A mathematical theory of natural and artificial selection. Part I. *Trans. Camb. Phil. Soc.* **23**, 19–41.

Han P., Niestemski L. R., Barrick J. E., and Deem M. W. (2013). Physical model of the immune response of bacteria against bacteriophage through the adaptive CRISPR-Cas immune system. *Phys. Biol.* **10**, 025004.

Handel A., Regoes R. R., and Antia R. (2006). The role of compensatory mutations in the emergence of drug resistance. *PLoS Comput. Biol.* **2**, e137.

Harris T. E. (1963). *The Theory of Branching Processes*. Berlin: Springer-Verlag.

He J., and Deem M. W. (2010). Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). *Phys. Rev. Lett.* **105**, 128102.

Heidelberg J. F., Nelson W. C., Schoenfeld T., and Bhaya D. (2009). Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS ONE* **4**, e4169.

Herrera M., García-Arriaza J., Pariente N., Escarmís C., and Domingo E. (2007). Molecular basis for a lack of correlation between viral fitness and cell killing capacity. *PLoS Pathog.* **3**, e53.

Herrera M., Grande-Pérez A., Perales C., and Domingo E. (2008). Persistence of foot-and-mouth disease virus in cell culture revisited: implications for contingency in evolution. *J. Gen. Virol.* **89**, 232–244.

Hinkley T., Martins J., Chappey C., Haddad M., Stawiski E., Whitcomb J. M., Petropoulos C. J., and Bonhoeffer S. (2011). A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat. Genet.* **43**, 487–489.

Ho D. D. (1995). Time to hit HIV, early and hard. *N. Engl. J. Med.* **333**, 450–451.

Hofbauer J., and Sigmund K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Holden M. G. T., Heather Z., Paillot R., Steward K. F., Webb K., Ainslie F., Jourdan T., Bason N. C., Holroyd N., Mungall K., Quail M. A., Sanders M., Simmonds M., Willey D., Brooks K., Aanensen D. M., Spratt B. G., Jolley K. A., Maiden M. C., Kehoe M., Chanter N., Bentley S. D., Robinson C., Maskell D. J., Parkhill J., and S. W. A. (2009). Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens. *PLoS Pathog.* **5**, e1000346.

Holland J. J. (1990). Defective viral genomes. In Fields B. N., and Knipe D. M. (Eds.), *Virology*, Volume 1, pp. 151–165. New York, NY: Raven.

Horvath P., Coute-Monvoisin A. C., Romero D. A., Boyaval P., Fremaux C., and Barrangou R. (2009). Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int. J. Food Microbiol.* **131**, 62–70.

Horvath P., Romero D. A., Coute-Monvoisin A. C., Richards M., Deveau H., Moineau S., Boyaval P., Fremaux C., and Barrangou R. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1401–1412.

Hu J. M., Fu H. C., Lin C. H., Su H. J., and Yeh H. H. (2007). Reassortment and concerted evolution in banana bunchy top virus genomes. *J. Virol.* **81**, 1746–1761.

Hubbell S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press.

Iranzo J., and Manrubia S. C. (2009). Stochastic extinction of viral infectivity through the action of defectors. *Europhys. Lett.* **85**, 18001.

Iranzo J., Perales C., Domingo E., and Manrubia S. C. (2011). Tempo and mode of inhibitor-mutagen antiviral therapies: a multidisciplinary approach. *Proc. Natl. Acad. Sci. USA* **108**, 16008–16013.

Jansen R., Embden J. D., Gaastra W., and Schouls L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575.

Jardim A. C., Bittar C., Matos R. P., Yamasaki L. H., Silva R. A., Pinho J. R., Fachini R. M., Carareto C. M., de Carvalho-Mello I. M., and Rahal P. (2013). Analysis of HCV quasispecies dynamic under selective pressure of combined therapy. *BMC Infect. Dis.* **13**, 61.

Jorth P., and Whiteley M. (2012). An evolutionary link between natural transformation and CRISPR adaptive immunity. *mBio* **3**, e00309–12.

Kaplan N., Darden T., and Langley C. H. (1985). Evolution and extinctino of transposable elements in Mendelian populations. *Genetics* **109**, 4459–480.

Karev G. P., Wolf Y. I., Rzhetsky A. Y., Berezovskaya F. S., and Koonin E. V. (2002). Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evol. Biol.* **2**, 18.

Kazazian H. H. (2004). Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632.

Kim K. H., Narayanan K., and Makino S. (1997). Assembled coronavirus from complementation of two defective interfering RNAs. *J. Virol.* **71**, 3922–3931.

Kimura M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Kirkwood T. B. L., and Bangham C. R. M. (1994). Cycles, chaos and evolution in virus cultures: a model of defective interfering particles. *Proc. Natl. Acad. Sci. USA* **91**, 8685–8689.

Kleckner N. (1989). Transposon Tn10. In Berg D. E., and Howe M. M. (Eds.), *Mobile DNA*, pp. 227–268. Washington D.C.: American Society for Microbiology.

Kleckner N. (1990). Regulating Tn10 and IS10 transposition. *Genetics* **124**, 449–454.

Koelle K., Cobey S., Grenfell B., and Pascual M. (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science* **314**, 1898–1903.

Komarova N. L., Barnes E., Klenerman P., and Wodarz D. (2003). Boosting immunity by antiviral drug therapy: a simple relationship among timing, efficacy, and success. *Proc. Natl. Acad. Sci. USA* **100**, 1855–1860.

Koonin E. V. (2011). Are there laws of genome evolution? *PLoS Comp. Biol.* **7**, e1002173.

Koonin E. V., and Makarova K. S. (2009). CRISPR-Cas: an adaptive immunity system in prokaryotes. *F1000 Biol. Rep.* **1**, 95.

Koonin E. V., and Makarova K. S. (2013). CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.* **10**.

Koonin E. V., Senkevich T. G., and Dolja V. V. (2006). The ancient Virus World and evolution of cells. *Biol. Direct* **1**, 29.

Koonin E. V., and Wolf Y. I. (2009). Is evolution Darwinian or/and Lamarckian? *Biol. Direct* **4**, 42.

Koonin E. V., and Wolf Y. I. (2012). Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Front. Cell. Infect. Microbiol.* **2**.

Kun A., Santos M., and Szathmáry E. (2005). Real ribozymes suggest a relaxed error threshold. *Nat. Gen.* **37**, 1008–1011.

Kupczok A., and Bollback J. P. (2013). Probabilistic models for CRISPR spacer content evolution. *BMC Evol. Biol.* **13**, 54.

Kuss S. K., Etheredge C. A., and Pfeiffer J. K. (2008). Multiple host barriers restrict poliovirus trafficking in mice. *PLoS Pathog.* **4**, e1000082.

Lan R., and Reeves P. R. (2002). *Escherichia coli* in disguise: molecular origins of *Shigella*. *Microbes Infect.* **4**, 1125–1232.

Langley C. H., Brookfield J. F. Y., and Kaplan N. (1983). Transposable elements in Mendelian populations: I. Theory. *Genetics* **104**, 457–471.

Lauring A. S., and Andino R. (2010). Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog.* **6**, e1001005.

Lauring A. S., and Andino R. (2011). Exploring the fitness landscape of an RNA virus by using a universal barcode microarray. *J. Virol.* **85**, 3780–3791.

Lazarowitz S. D. (2007). Plant viruses. In Knipe D. M., and Howley P. M. (Eds.), *Fields virology*, Volume 1, pp. 641–706. Philadelphia, PA: Lippincott Williams and Wilkins.

Le Rouzic A., Boutin T. S., and Capy P. (2007). Long-term evolution of transposable elements. *Proc. Natl. Acad. Sci. USA* **104**, 19375–19380.

Leclercq S., and Cordaux R. (1997). Selection-driven extinction dynamics for group II introns in *Enterobacteriales*. *J. Virol.* **71**, 3636–3640.

Lee C. H., Gilbertson D. L., Novella I. S., Huerta R., Domingo E., and Holland J. J. (1997). Negative effects of chemical mutagenesis in the adaptive behavior of vesicular stomatitis virus. *J. Virol.* **71**, 3636–3640.

Lefeuvre P., Lett J. M., Varsani A., and Martin D. P. (2009). Widely conserved recombination patterns among single-stranded DNA viruses. *J. Virol.* **83**, 2697–2707.

Lenski R. E., Barrick J. E., and Ofria C. (2006). Balancing robustness and evolvability. *PLoS Biol.* **4**, e428.

Leplae R., Geeraerts D., Hallez R., Guglielmini J., Dreze P., and Melderen L. V. (2011). Diversity of bacterial type II toxin-antitoxin systems: a comprehensive search and functional analysis of novel families. *Nucleic Acids Res.* **39**, 5513–5525.

Levin B. R. (2010). Nasty viruses, costly plasmids, population dynamics, and the conditions for establishing and maintaining CRISPR-mediated adaptive immunity in bacteria. *PLoS Genet.* **6**, e1001171.

Levin B. R., Moineau S., Bushman M., and Barrangou R. (2013). The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. *PLoS Genet.* **9**, e1003312.

Levy M. S., Balbinder E., and Nagel R. (1993). Effect of mutations in SOS genes on UV-induced precise excision of Tn10 in *Escherichia coli*. *Mutat. Res.* **293**, 241–247.

Lobkobsky A. E., Wolf Y. I., and Koonin E. V. (2013). Gene frequency distributions reject a neutral model of genome evolution. *Genom. Biol. Evol.* **5**, 233–242.

Loeb L. A., Essigmann J. M., Kazazi F., Zhang J., Rose K. D., and Mullins J. I. (1999). Lethal mutagenesis of HIV with mutagenic nucleoside analogs. *Proc. Natl. Acad. Sci. USA* **96**, 1492–1497.

López-Sánchez M. J., Sauvage E., Da Cunha V., Clermont D., Ratsima Hariniaina E., González-Zorn B., Poyart C., Rosinski-Chupin I., and Glaser P. (2012). The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Mol. Microbiol.* **85**, 1057–1071.

Loverdo C., Park M., Schreiber S. J., and Lloyd-Smith J. O. (2012). Influence of viral replication mechanisms on within-host evolutionary dynamics. *Evolution* **66**, 3462–3471.

Luque A., Zandi R., and Reguera D. (2010). Optimal architectures of elongated viruses. *Proc. Natl. Acad. Sci. USA* **107**, 5323–5328.

Lynch M., and Conery J. S. (2003). The origins of genome complexity. *Science* **302**, 1401–1404.

Mahillon J., and Chandler M. (1998). Insertion sequences. *Microbiol. Mol. Biol. R.* **62**, 725–774.

Makarova K. S., Grishin N. V., Shabalina S. A., Wolf Y. I., and Koonin E. V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct* **1**, 7.

Makarova K. S., Haft D. H., Barrangou R., Brouns S. J., Charpentier E., Horvath P., Moineau S., Mojica F. J., Wolf Y. I., Yakunin A. F., van der Oost J., and Koonin E. V. (2011). Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9**, 467–477.

Makarova K. S., Wolf Y. I., and Koonin E. V. (2013). Comparative genomics of defensive systems in archaea and bacteria. *Nucleic Acids Res.* in press.

Makarova K. S., Wolf Y. I., Snir S., and Koonin E. V. (2011). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* **193**, 6039–6056.

Manrubia S. C. (2012). Modelling viral evolution and adaptation: challenges and rewards. *Curr. Opin. Virol.* **2**, 531–537.

Manrubia S. C., Domingo E., and Lázaro E. (2010). Pathways to extinction: beyond the error threshold. *Philos. Trans. R. Soc. Lond. B* **365**, 1943–1952.

Manrubia S. C., and Lázaro E. (2006). Viral evolution. *Phys. Life Rev.* **3**, 65–92.

Manrubia S. C., Lázaro E., Pérez-Mercader J., Escarmís C., and Domingo E. (2003). Fitness distribution in exponentially growing asexual populations. *Phys. Rev. Lett.* **90**, 188102.

Marraffini L. A., and Sontheimer E. J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190.

May R. M. (2001). *Stability and Complexity in Model Ecosystems*. Princeton University Press.

McGraw J. E., and Brookfield J. F. Y. (2006). The interaction between mobile DNAs and their hosts in a fluctuating environment. *J. Theor. Biol.* **243**, 13–23.

Mira A., Ochman H., and Moran N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–596.

Mira A., Pushker R., and Rodríguez-Varela F. (2006). The Neolithic revolution of bacterial genomes. *Trends Genet.* **14**, 200–206.

Mojica F. J., Díez-Villaseñor C., García-Martínez J., and Almendros C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733–740.

Mojica F. J., Díez-Villaseñor C., García-Martínez J., and Soria E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174–182.

Montgomery E. A., and Langley C. H. (1983). Transposable elements in Mendelian populations: II. Distribution of *copia*-like elements in natural populations. *Genetics* **104**, 473–483.

Moody M. M. (1988). A branching process model for the evolution of transposable elements. *J. Math. Biol.* **26**, 347–357.

Moran N. A., and Plague G. R. (2004). Genomic changes following host restriction in bacteria. *Curr. Opin. Genet. Dev.* **14**, 627–633.

Moreno H., Tejero H., de la Torre J. C., Domingo E., and Martín V. (2012). Mutagenesis-mediated virus extinction: virus-dependent effect of viral load on sensitivity to lethal defection. *PLoS ONE* **7**, e32550.

Morris J. J., Lenski R. E., and Zinser E. R. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio* **3**, e00036–12.

Müller V., and Bonhoeffer S. (2008). Intra-host dynamics and evolution of HIV infection. In Domingo E., Parrish C. R., and Holland J. J. (Eds.), *Origin and Evolution of Viruses*, pp. 279–302. London: Academic Press.

Mullins J. I., Heath L., Hughes J. P., Kicha J., Styrchak S., Wong K. G., Rao U., Hansen A., Harris K. S., Laurent J. P., Li D., Simpson J. H., Essigmann J. M., Loeb L. A., and Parkins J. (2011). Mutation of HIV-1 genomes in a clinical population treated with the mutagenic nucleoside KP1461. *PLoS ONE* **6**, e15135.

Nee S. (1987). The evolution of multicompartmental genomes in viruses. *J. Mol. Evol.* **25**, 277–281.

Nijhuis M., van Maarsevee N. M., and Boucher C. A. (2009). Antiviral resistance and impact on viral replication capacity: evolution of viruses under antiviral pressure occurs in three phases. In Kräusslich H.-G., and Bartenschlager R. (Eds.), *Handbook of Experimental Pharmacology*, Volume 189, pp. 299–320. Berlin: Springer-Verlag.

Novak M., Pfeiffer T., Lenski R. E., Sauer U., and Bonhoeffer S. (2006). Experimental evidence for an evolutionary trade-off between growth rate and yield in E.c̃oli. *Am. Nat.* **168**, 242–251.

Novella I. S., Reissig D. D., and Wilke C. O. (2004). Density-dependent selection in vesicular stomatitis virus. *J. Virol.* **78**, 5799–5804.

Nuzhdin S. V. (2000). Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica* **107**, 129–137.

Ojosnegros S., García-Arriaza J., Escarmís C., Manrubia S. C., Perales C., Arias A., García Mateu M., and Domingo E. (2011). Viral genome segmentation can result from a trade-off between genetic content and particle stability. *PLoS Genet.* **7**, e1001344.

Oliver K. R., and Greene W. K. (2009). Transposable elements: powerful facilitators of evolution. *BioEssays* **31**, 703–714.

O'Neill F. J., Maryon E. B., and Carroll D. (1982). Isolation and characterization of defective simian virus 40 genomes which complement for infectivity. *J. Virol.* **43**, 18–25.

Orgel L. E., and Crick F. H. C. (1980). Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607.

Paez-Espino D., Morovic W., Sun C. L., Thomas B. C., Ueda K., Stahl B., Barrangou R., and Banfield J. F. (2013). Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat. Commun.* **4**, 1430.

Pál C., and Papp B. (2013). From passengers to drivers: impact of bacterial transposable elements on evolvability. *Mobile Genetic Elements* **3**, 1–4.

Palmer K. L., and Gilmore M. S. (2010). Multidrug-resistant enterococci lack CRISPR-cas. *mBio* **1**, e00227–10.

Palukaitis P., and García-Arenal F. (2003). Cucumoviruses. *Adv. Virus Res.* **62**, 241–323.

Parera M., Fernández G., Clotet B., and Martínez M. A. (2007). HIV1 protease catalytic efficiency effects caused by random single amino acid substitutions. *Mol. Biol. Evol.* **24**, 382–387.

Pariente N., Airaksinen A., and Domingo E. (2003). Mutagenesis versus inhibition in the efficiency of extinction of foot-and-mouth disease virus. *J. Virol.* **77**, 7131–7138.

Pariente N., Sierra S., and Airaksinen A. (2005). Action of mutagenic agents and antiviral inhibitors on foot-and-mouth disease virus. *Virus Res.* **107**, 183–193.

Perales C., Agudo R., and Domingo E. (2009). Counteracting quasispecies adaptability: extinction of a ribavirin-resistant virus mutant by an alternative mutagenic treatment. *PLoS ONE* **4**, e5554.

Perales C., Agudo R., Tejero H., Manrubia S. C., and Domingo E. (2009). Pontential benefit of sequential inhibitor-mutagen treatments of RNA virus infections. *PLoS Pathog.* **5**, e100658.

Perales C., Iranzo J., Manrubia S. C., and Domingo E. (2012). The impact of quasispecies dynamics on the use of therapeutics. *Trends Microbiol.* **20**, 595–603.

Perales C., Mateo R., Mateu M. G., and Domingo E. (2007). Insights into RNA virus mutant spectrum and lethal mutagenesis events: replicative interference and complementation by multiple point mutants. *J. Mol. Biol.* **369**, 985–1000.

Plague G. R. (2010). Intergenic transposable elements are not randomly distributed in bacteria. *Genom. Biol. Evol.* **2**, 584–590.

Pleckaityte M., Zilnyte M., and Zvirbliene A. (2012). Insights into the CRISPR/Cas system of *Gardnerella vaginalis*. *BMC Microbiol.* **12**, 301.

Punta M., Coggill P. C., Eberhardt R. Y., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G., Clements J., Heger A., Holm L., Sonnhammer E. L. L., Eddy S. R., Bateman A., and Finn R. D. (2012). The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301.

Rankin D. J., Bichsel M., and Wagner A. (2010). Mobile DNA can drive lineage extinction in prokaryotic populations. *J. Evol. Biol.* **23**, 2422–2431.

Ribeiro R. M., and Bonhoeffer S. (2000). Production of resistant HIV mutants during anti-retroviral therapy. *Proc. Natl. Acad. Sci. USA* **97**, 7681–7686.

Richman D. D. (Ed.) (1996). *Antiviral Drug Resistance*. New Tork: Wiley.

Rohwer F., and Truber R. V. (2009). Viruses manipulate the marine environment. *Nature* **459**, 207–212.

Roosinck M. J. (1997). Mechanisms of plant virus evolution. *Annu. Rev. Phytopathol.* **35**, 191–209.

Rosen H. R. (2011). Chronic hepatitis C infection. *N. Engl. J. Med.* **364**, 2429–2438.

Roux L., Simon A. E., and Holland J. J. (1991). Effects os defective interfering viruses on virus replication and pathogenesis *in vitro* and *in vivo*. *Adv. Virus Res.* **40**, 181–211.

Ruiz-Jarabo C. M., Ly C., Domingo E., and de la Torre J. C. (2003). Lethal mutagenesis of the prototypic arenavirus lymphocytic choriomeningitis virus (LCMV). *Virology* **308**, 37–47.

Saakian D. B., and Hu C.-K. (2006). Exact solution of the Eigen model with general fitness functions and degradation rates. *Proc. Natl. Acad. Sci. USA* **103**, 4935–4939.

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* **365**, 1975–1982.

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

Sardanyés J., Solé R. V., and Elena S. F. (2009). Replication mode and landscape topology differentially affect RNA virus mutational load and robustness. *J. Virol.* **83**, 12579–12589.

Schneider D., and Lenski R. E. (2004). Dynamics of insertion sequence elements during experimental evolution of bacteria. *Res. Microbiol.* **155**, 319–327.

Schrenl M. O., Kelley D. S., Delaney J. R., and Baross J. A. (2003). Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. *Appl. Env. Microbiol.* **69**, 3580–3592.

Schuster P., and Stadler P. F. (1994). Landscapes: complex optimization problems and biopolymer structures. *Comp. & Chem* **18**, 295–324.

Seed K. D., Lazinski D. W., Calderwood S. B., and Camilli A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489–491.

Semenova E., Jore M. M., Datsenko K. A., Semenova A., Westra E. R., Wanner B., van der Oost J., Bronus S. J., and Severinov K. (20011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci. USA* **108**, 10098–10103.

Šidák Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Statist. Assoc.* **62**, 626–633.

Sierra M., Airaksinen A., González-López C., Agudo R., Arias A., and Domingo E. (2007). Foot-and-Mouth Disease Virus mutant with decreased sensitivity to ribavirin: implications for error catastrophe. *J. Virol.* **81**, 2012–2024.

Siguier P., Pedrochon J., Lestrade L., Mahillon J., and Chandler M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36.

Slotkin R. K., and Martienssen R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285.

Smouse P. E. (1976). The implications of density-dependent population growth for frequency and density-dependent selection. *Am. Nat.* **110**, 849–860.

Sobrino F., Dávila M., Ortín J., and Domingo E. (1983). Multiple genetic variants arise in the course of replication of foot-and-mouth disease virus in cell culture. *Virology* **128**, 310–318.

Stein P. R., and Waterman R. S. (1978). On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Math.* **26**, 261–272.

Steinhauer D. A., Domingo E., and Holland J. J. (1992). Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene* **122**, 281–288.

Stern A., Keren O., Wurtzel G., Amitai G., and Sorek R. (2010). Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet.* **26**, 335–340.

Stern A., and Sorek R. (2011). The phage-host arms race: shaping the evolution of microbes. *Bioessays* **33**, 43–51.

Stich M., Briones C., and Manrubia S. C. (2007). Collective properties of evolving molecular quasispecies. *BMC Evol. Biol.* **7**, 110.

Suttle C. A. (2007). Marine viruses - major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812.

Swetina J., and Schuster P. (1982). Self-replication with errors: a model for polynucleotide replication. *Biophys. Chem.* **16**, 329–345.

Szathmáry E. (1992). Natural selection and dynamical coexistence of defective and complementing virus segments. *J. Theor. Biol.* **157**, 383–406.

Szathmáry E. (1993). Co-operation and defection : playing the field in virus dynamics. *J. Theor. Biol.* **165**, 341–356.

Takeuchi N., and Hogeweg P. (2007). Error-threshold exists in fitness landscapes with lethal mutants. *BMC Evol. Biol.* **7**, 15.

Takeuchi N., Poorthuis P. H., and Hogeweg P. (2005). Phenotypic error threshold; additivity and epistasis in RNA evolution. *BMC Evol. Biol.* **5**, 9.

Thompson D. W. (1917). *On Growth and Form* (1st ed.). Cambridge University Press.

Toja M., Escarmís C., and Domingo E. (1999). Genomic nucleotide sequence of a foot-and-mouth disease virus clone and its persistent derivatives. Implications for the evolution of viral quasispecies during a persistent infection. *Virus Res.* **64**, 161–171.

Torella J. P., Chait R., and Kishony R. (2010). Optimal drug synergy in antimicrobial treatments. *PLoS Comput. Biol.* **6**, e1000796.

Touchon M., and Rocha E. P. C. (2007). Causes of insertion sequence abundance in prokaryotic genomes. *Mol. Biol. Evol.* **24**, 969–981.

Treangen T. J., Abraham A.-L., Touchon M., and Rocha E. P. C. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.* **33**, 539–571.

Turner P. E., and Chao L. (1998). Sex and the evolution of intrahost competition in RNA virus $\phi 6$. *Genetics* **150**, 523–532.

van der Oost J., Jore M. M., westra E. R., Lundgren M., and Brouns S. J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem./ Sci.* **34**, 401–407.

Vasu K., and Nagaraja V. (2013). Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.* **77**, 53–72.

Venner S., Feschotte C., and Biémont C. (2009). Dynamics of transposable elements: towards a community ecology of the genome. *Trends Genet.* **25**, 317–323.

Vignuzzi M., Wendt E., and Andino R. (2008). Engineering attenuated virus vaccines by controlling replication fidelity. *Nat. Med.* **14**, 154–161.

Volvok I., Banavar J. R., Hubbell S. P., and Maritan A. (2003). Neutral theory and relative species abundance in ecology. *Nature* **424**, 1035–1037.

von Kleist M., Menz S., Stocker H., Arasteh K., Schütte C., and Huisinga W. (2011). HIV quasispecies dynamics during pro-active treatment switching: impact on multi-drug resistance and resistance archiving in later reservoirs. *PLoS ONE* **6**, e18204.

Wagner A. (2006). Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Mol. Biol. Evol.* **23**, 723–733.

Wagner A. (2008). Neutralism and selectionism: a network-based reconciliation. *Nat. Rev. Genet.* **9**, 965–974.

Wagner A. (2009). Transposable elements as genomic diseases. *Mol. BioSyst.* **5**, 32–35.

Wagner A. (2011). *The Origins of Evolutionary Innovations*. Oxford University Press.

Wagner E. K., and Hewlett M. J. (2004). *Basic Virology* (2nd ed.). Blackwell Publishing.

Weinberger A. D., and Gilmore M. S. (2012). CRISPR-Cas: to take up DNA or not-that is the question. *Cell Host Microbe* **12**, 125–126.

Weinberger A. D., Sun C. L., Pluciński M. M., Denef V. J., Thomas B. C., Horvath P., Barrangou R., Gilmore M. S., Getz W. M., and Banfield J. F. (2012). Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput. Biol.* **8**, e1002475.

Weinberger A. D., Wolf Y. I., Lobkovsky A. E., Gilmore M. S., and Koonin E. V. (2012). Viral diversity threshold for adaptive immunity in prokaryotes. *mBio* **3**, e00456–12.

Werren J. H. (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proc. Natl. Acad. Sci. USA* **108**, 10863–10870.

Werren J. H., and Bartos J. D. (2001). Recombination in *Wolbachia*. *Curr. Biol.* **11**, 431–435.

Wessner D. R. (2010). The origins of viruses. *Nature Education* **3**, 37.

Westra E. R., Swarts D. C., Staals R. H., Jore M. M., Brouns S. J., and van der Oost J. (2012). The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* **46**, 311–339.

Whitaker R. J., and Banfield J. F. (2006). Population genomics in natural microbial communities. *Trends Ecol. Evol.* **21**, 508–516.

Wiedenheft B., Sternberg S. H., and Doudna J. A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331–338.

Wilke C. O. (2005). Quasispecies theory in the context of population genetics. *BMC Evol. Biol.* **5**, 44.

Wilke C. O., and Novella I. S. (2003). Phenotypic mixing and hiding may contribute to memory in viral quasispecies. *BMC Microbiol.* **3**, 11.

Wilmes P., Simmons S. L., Denef V. J., and Banfield J. F. (2009). The dynamic genetic repertoire of microbial communities. *FEMS Microbiol. Rev.* **33**, 109–132.

Woo H.-J., and Reifman J. (2012). A quantitative quasispecies theory-based model of virus escape mutation under immune selection. *Proc. Natl. Acad. Sci. USA* **109**, 12980–12985.

Yeh P., Tschumi A. I., and Kishony R. (2006). Functional classification of drugs by properties of their pairwise interactions. *Nat. Genet.* **38**, 489–494.

Zeh D. W., Zeh J. A., and Ishida Y. (2009). Transposable elements and an epigenetic basis for punctuated equilibria. *BioEssays* **31**, 715–726.

Zeldovich K. B., Berezovsky I. N., and Shakhnovich E. I. (2007). Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol.* **3**, e5.