The role of reassortment in the evolution of seasonal influenza

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universitat zu Köln

vorgelegt von

Mara Villa

aus Monza, Italien

Köln, 2018

Berichterstatter: Prof. Dr. Michael Lässig
Prof. Dr. Andreas Beyer

Tag der mündlichen Prüfung: 19.07.2018

Kurzzusammenfassung

Reassortment ist der Austausch von Teilen des Erbgut zwischen Viren, die gleichzeitig eine Wirtszelle befallen. Es spielt eine wichtige Rolle in der Evolution von Viren mit segmentiertem Genom, wie zum Beispiel dem Influenza Virus. Die umfangreiche Durchmischung des Genoms kann zu einer höheren Diversität in Viruspopulationen und zur Entwicklung von pandemischen Virenstämmen führen. Für das humane Influenza Virus tritt Reassortment in den meisten Fällen zwischen koexistierenden Virusvarianten des gleichen Subtyps auf. Dieser Prozess unterbricht die Genkopplung und die Fitnesskorrelation zwischen verschiedenen viralen Genomsequenzen. Der resultierende Effekt auf die virale Fitness ist jedoch unklar.

In dieser Arbeit bestimmen wir zunächst die Rate und den durchschnittlichen Selektionseffekt von Reassortmentprozessen des humanen Influenzasubtyps A/H3N2. Die Oberflächenproteine Hemmagglutinin und Neuraminidase haben Reassortmentvarianten mit einem mittleren Abstand von mindestens drei Nukleotiden im Vergleich zu ihrem Ausgangsstamm. Sie setzen sich mit einer Rate von ungefähr 10^{-2} in Einheiten der neutralen Punktmutationsrate durch. Unsere Inferenz basiert auf einer neuen Methode, die Reassortmentereignisse aus gemeinsamen Genealogien mehrerer Genomsegmente abbildet und wurde durch umfangreiche Simulationen getestet. Wir zeigen, dass Reassortment innerhalb eines Subtyps im Durchschnitt unter beträchtlicher negativer Selektion steht, deren Stärke mit dem genetischen Unterschied zum Ausgangsstamm zunimmt. Die nachteiligen Effekte von Reassortment zeigen sich auf zwei Arten: Erstens tritt Reassortment seltener auf, als man von der Nullhypothese neutralem Reassortments vorhergesagt. Zweitens haben Reassortmentstämme weniger Nachfahren als die entsprechenden Stämme ohne Reassortment. Unsere Ergebnisse deuten darauf hin, dass sich Influenza unter allgegenwärtiger Epistase weiterentwickelt, was wiederum Fitnessbegrenzungen gegen Reassortment sogar innerhalb von einzelnen Stämmen eines Subtyps mit sich bringt.

Weiterhin untersuchen wir die Dynamik von Reassortment innerhalb eines aktuellen

Influenza-Abstammungszeigs, welche wahrscheinlich die nächste Epidemie dominieren wird. Um die großen Datenmengen, die in den letzten Saisons gesammelt wurden, handhaben zu können, entwickeln wir eine heuristische Erkennungsmethode basierend auf Neuraminidase Alignment und Haemagglutinin Stammbäumen. Dieser neue Ansatz ist schneller als frühere Methoden und kann auf wesentliche größere Bäume angewendet werden. Wir beobachten vermehrt Reassortmentereignisse, was neben dem ursprünglichen Stamm zu der Koexistenz von drei weiteren Neurminidasekladen führt. Die eingebauten Neuraminidasesegmente unterscheiden sich von Varianten ohne Reassortment durch veränderte Aminosäuren an epistatischen Genpositionen. Die Ergebnisse dieser Arbeit sind ein Schritt auf dem Weg, Evolution von Influenza auf Grundlage des kompletten Genoms vorherzusagen. Die konsequente Verbesserung der Vorhersage zukünftiger Evolution ist der Schlüssel zur Entwicklung effektiver Impfstoffe.

Abstract

Reassortment, which is the exchange of genome sequence between viruses co-infecting a host cell, plays an important role in the evolution of segmented viruses, such as influenza. The large-scale genome reshuffling promotes diversity in the viral population and can lead to the emergence of pandemic strains. In the human influenza virus, reassortment happens most frequently between co-existing variants within the same lineage. This process breaks genetic linkage and fitness correlations between viral genome segments, but the resulting net effect on viral fitness has remained unclear.

In this thesis, we first determine rate and average selective effect of reassortment processes in the human influenza lineage A/H3N2. For the surface proteins hemagglutinin and neuraminidase, reassortant variants with a mean distance of at least 3 nucleotides to their parent strains get established at a rate of about 10^{-2} in units of the neutral point mutation rate. Our inference is based on a new method to map reassortment events from joint genealogies of multiple genome segments, which is tested by extensive simulations. We show that intra-lineage reassortment processes are, on average, under substantial negative selection that increases in strength with increasing sequence distance between the parent strains. The deleterious effects of reassortment manifest themselves in two ways: there are fewer reassortment events than expected from a null model of neutral reassortment, and reassortant strains have fewer descendants than their non-reassortant counterparts. Our results suggest that influenza evolves under ubiquitous epistasis across proteins, which produces fitness barriers against reassortment even between co-circulating strains within one lineage.

Second, we study the dynamics of reassortment occurring within a recent influenza clade which is likely to dominate the next epidemics. In order to handle the large amount of data collected in the last seasons, we develop a new heuristic detection method based on neuraminidase alignments and haemagglutinin phylogenies; this novel approach is faster than the joint-segment tree based algorithm and can be applied to much larger trees. We

find an increase in the frequency of reassorment events, which lead to the coexistence in the viral population of three new neuraminidase clades, in addition to the ancestral variant. These imported neuraminidase segments differ from the non-reassorted version by aminoacid changes at epistatic sites. The results reported in this work are a step forward towards predicting the evolution of influenza based on the entire genome. The consequent improvement of accuracy in anticipating future evolution is the key for designing more effective vaccines.

Acknowledgements

I would like to acknowledge Prof. Michael Lässig, for introducing me to the interesting problem of influenza evolution and for spending some of his time supervising and guiding my work over these years. Furthermore, I thank Simone Pompei, for his continuous support and valuable comments on this project, and Marta Łuksza, for sharing data and ideas, which has opened a way to fruitful collaboration. I also want to thank all the people I have met on my way to graduation, fellow PhD students, post-docs, professors and friends, who made this journey more enjoyable. Last but not least, I owe many thanks to my family, for supporting me throughout the last 30 years.

Contents

1. Introduction

1.1 Thesis outline

The study of evolutionary processes is a classic topic in biology. Huge effort has been made to investigate and describe the mechanisms responsible for the diversity of life, setting the stage for an exciting change of point of view: can the goal of evolutionary biology in turn evolve from drawing a picture of the past into predicting future evolutionary patterns? Although evolution is shaped by multiple stochastic forces, the growing amount of collected data and the continuous progresses in modeling complex dynamical systems have revealed that, for specific modes of evolution, the transition from description to prediction is no more an unattainable hope. In this context, fast evolving organisms as viruses and bacteria are of particular interest, as their evolution shows repeatable features.

Classical genetics has historically been focused on studying changes at the level of point mutations. Nevertheless a comprehensive understanding of the evolutionary dynamics cannot overlook the emergence of events occurring at larger scales. Exchange and variation of large portions of genetic material have been shown to be extremely relevant in determining the evolutionary outcomes for several biological systems. For example copy-number variation, namely duplication or deletion events affecting thousands of base pairs at once, have been recognized as having an important role on cancer predisposition, cancer gene expression and tumor genome profiling [1]. Horizontal gene transfer in bacteria and gene reassortment in viruses (see below) are further significant examples of processes involving thousands of loci altogether. Taking into account the effects of these kind of events can substantially improve our ability of forecasting evolution. The first stimulating challenge for such purposes is the mapping of the dynamics of these processes, which was found to be highly nontrivial. In this thesis we focus on evolution of the seasonal influenza virus, a common pathogen that unceasingly evolves to adapt against host immunity. In

particular we study how reassortment, namely the mixing of genes from different parent strains, affects and shapes the structure of the population of co-circulating viruses.

In this chapter we first briefly recall some fundamental concepts of population genetics, drawing the attention to the stochastic forces which drive genotype evolution. We highlight the complexity that arises when time-dependent selection is taken into account and we describe the emergence of competition between clones in the high mutational regime, typical of influenza, elucidating how evolutionary dynamics change as an effect of occasional gene reshuffling. We then introduce the biological system object of this study, seasonal influenza virus, and explain the important role of intra-subtype reassortment (especially between the two surface glycoproteins haemagglutinin (HA) and neuraminidase (NA), main targets of antibodies recognition) in shaping viral evolution. We underline that a deeper understanding of the dynamics of this process has the potential to enhance the predictive power of methods currently used to forecast influenza evolution and select candidate vaccine strains.

In chapter 2 we present a novel, robust, method to infer intra-lineage reassortment based on the genealogy of the virus. We analyze the joint evolution of haemagglutinin and neuraminidase by reconstructing the respective two-segment genealogy and spot specific mutational patterns along tree branches, which signal disjoint evolution of the two proteins. We interpret these "anomalies" on the tree as a result of gene reassortment and validate this assumption by comparison with a null case of non-reassorting segments, as well as by tests on simulated data. We show that a large fraction of the simulated reassortment events are recovered by our algorithm, and these events outweigh the rate of false positives.

In chapter 3 we apply our new detection method to a large dataset of seasonal influenza sequences and we obtain a list of reassortment events (Appendix A) that have produced new combinations of HA and NA genes, finding that the overall number of reassortments reported by our algorithm is in broad agreement with the results of previous studies [2–4]. We finally apply two independent and complementary selection inference methods to the list of detected events, identifying a consistent signal of broad negative selection on intra-lineage reassortment. We interpret this signal in terms of ubiquitous cross-protein epistasis and discuss evolutionary consequences.

In chapter 4 we propose our second method to infer reassortment, as an alternative approach to the method described in chapter 2. Instead of analyzing the genealogy built with paired HA-NA sequences in order to find inconsistencies in the pattern of mutations on the tree branches, we use phylogenetic information from one segment only, together

with the aligned sequences of the other segment. With heuristic approach, we spot on the one-segment tree the clades which are coupled with more than one variant in the other segment. We present some interesting preliminary results obtained by applying this new fast method to data collected in recent seasons, discussing how the dynamics of reassortment may have changed in the last few epidemics.

In the closing chapter we finally summarize and interpret our results. Parts of chapters 1, 2 and 3 and Appendix A have been published in reference [5].

1.2 Molecular evolution as a stochastic process

The change in the genomic content of populations is a stochastic process driven by time-dependent evolutionary forces: random mutations, selection, genetic drift and recombination act on the individuals and determine the genetic composition of a population over time. Mutations arise randomly at the level of individuals and promote variation within a population by creating differences in the genotype alleles at rates μ. Natural selection pushes the variants with high reproductive success towards achieving larger fractions in the population, while genetic drift models the stochastic effect of fluctuations in the reproduction process.

1.2.1 Diffusion evolutionary equations: a simple model

It is a helpful exercise to break down the evolutionary dynamics by considering separately the contribution that each of the above mentioned mechanisms gives on shaping the structure of a population. We model a large set of N individuals, grouped, for the sake of notational simplicity, into two sub-populations with genotypes a and b. In absence of mutations and recombination, each genotype evolves with the simple growth law [6]

$$\frac{d}{dt}N_{a\backslash b}(t) = F_{a\backslash b}(t)N_{a\backslash b}(t)\,, \tag{1.1}$$

with $F_{a\backslash b}(t)$ defined as the Malthusian fitness of the genotype a or b, respectively, and $N(t) = N_a(t) + N_b(t)$. Rewriting equation (1.1) for frequencies $x(t) = N_b(t)/N(t)$ one obtains

$$\frac{d}{dt}x(t) = \Delta F_{ab}(t)x(t)[1 - x(t)]\,, \tag{1.2}$$

with $\Delta F_{ab}(t) = F_b(t) - F_a(t)$. Assuming a constant fitness difference, the system evolves in a deterministic way towards the fixed points $x = 0$ or $x = 1$, which represent fixation of one genotype and loss of the other.

Random sampling of individuals (*genetic drift*) can be considered in the dynamics of the population by adding a Gaussian random variable $\mathcal{X}_{a\setminus b}(t)$ to equation (1.1), with $\overline{\mathcal{X}_{a\setminus b}(t)} = 0$ and $\overline{\mathcal{X}_a(t)\mathcal{X}_b(t')} = N_a(t)\delta(t-t')\delta_{a,b}$. The projection of the stochastic growth law onto the population fraction x can be written ([6, 7]) in terms of a Kimura diffusion equation for the probability distribution $P(x,t)$:

$$\frac{\partial}{\partial t}P(x,t) = \frac{1}{2N}\frac{\partial^2}{\partial x^2}x(1-x)P(x,t) - \Delta F_{ab}\frac{\partial}{\partial x}x(1-x)P(x,t)\,. \tag{1.3}$$

Although the monomorphic states $x = 0$ and $x = 1$ are still fixed points as in the purely deterministic case, fluctuations make the fixation probability of a genotype dependent not only on fitness differences, but also on the size and initial state of the population.

If transitions between genotypes are allowed through the introduction of mutations at rates $\mu_{a\to b}$ and $\mu_{b\to a}$, the correspondent Fokker-Planck equation reads [6]

$$\begin{aligned}\frac{\partial}{\partial t}P(x,t) =& \frac{1}{2N}\frac{\partial^2}{\partial x^2}x(1-x)P(x,t) - \Delta F_{ab}\frac{\partial}{\partial x}x(1-x)P(x,t) \\ &- \mu_{a\to b}\frac{\partial}{\partial x}(1-x)P(x,t) + \mu_{b\to a}\frac{\partial}{\partial x}xP(x,t)\,.\end{aligned} \tag{1.4}$$

Equation (1.4) is written under the assumption that only the systematic effects of random mutations are relevant for the evolution of the population probability, as their stochastic contribution is negligible if compared with the sampling noise $\mathcal{X}$. In the low mutational regime ($N\mu << 1$) and for constant fitness difference between genotypes the dynamics of the process described by equation (1.4) can be depicted as a series of switches (so called *substitutions*) between the unstable states $x = 0$ and $x = 1$. The substitution rates $u_{a\to b}$ and $u_{b\to a}$ that define the time scale for a new genotype to establish in the population depend on mutation rates, fitness differences in the genotypes and (effective) population size. Equivalent equations can be written for allele frequencies, without loss of information, under the hypothesis of linkage equilibrium, namely if the alleles at different loci are inherited independently at each generation.

The evolutionary equilibrium reached in the approximation of constant selection (time- and frequency-independent fitness differences) does not capture the ability of a population to adapt within a changing environment. The emergence of phenotypic adaptation is a general result of selection variation in space and time and it reflects a more realistic representation of actual populations evolving in a changing background. There are evidences from genomic data that selection itself can act as a random force, driving to macro-evolutionary changes on large time scales, comparable with mutations time scales

[8]. The development of simple models which include time-dependent selection with constant magnitude and random direction [9] has revealed that a surplus of advantageous over deleterious substitutions emerge as a consequence of long term fluctuations, in a regime where the latter are described as a quenched random process. Even though these studies are aimed to model and explain evolution at low mutation rate, the concept of parameterizing selection as a fluctuating random force at individual sites has a general validity and has been successfully applied to reproduce adaptation of organisms in the regime $\mu N >> 1$, such as influenza virus [10]. Below (see chapter 2) we include random switches in selection coefficients as a fundamental element to simulate the evolution of human influenza. More general forms of Kimura-Otha evolution equation [11] valid for finite size populations with k possible genotypes are discussed in reference [12]. It can be shown that the existence of evolutionary equilibrium under selection and mutations is conditioned to the possibility of writing the total rates of frequency change given by selection and mutations in a gradient form. While equilibrium Boltzmann-like distributions can be calculated in the case of low mutation - or high recombination (see below) rate, genomic equilibria do not exist if the evolution of the genomic sites is generically coupled by fitness interactions (epistasis) and genetic linkage.

1.2.2 Clonal interference and recombination

Most biological systems evolve approximately in the low mutational regime, however the evolution of some notable exceptions as mutator bacteria and viruses is fueled by high mutation rates. In this regime, the picture of adaptation by "periodic selection" [13], which implies beneficial mutations arising and fixing one by one, is replaced by a different dynamics: simultaneous beneficial mutations may appear in the population and survive against genetic drift. Without recombination, clones carrying alternative beneficial mutations first outcompete the wild-type separately, then engage a head to head race to take over and finally get fixed in the population. This scenario is known as *clonal interference* (CI) [14] and it has been identified as a preponderant mechanism shaping adaptation under genetic linkage in the evolution of a number of micro-organisms, including E. coli [15–18], bacteriophages [19] and the influenza virus [10]. Beside inducing competition between disjoint clones, which is associated to a slower rate of adaptation compared to periodic selection, the high mutation rate maintains fitness variation within the population and continuously produces new sequences. Beneficial mutations originating in nested clones may then get to fixation at the same time, generating "positive" interference interactions

[20–25].

The effects of clonal interference become less relevant in sexual reproduction, when the process of recombination reshuffles the genomes and promotes variability by creating novel genotypes at each generation as a mixture of parental alleles. Mathematically, recombination can be explicitly taken into account in the description of the evolution of time-dependent frequency distributions as an additive source to the deterministic changes occurring at time scales larger than a generation [12, 26]. If the rate of recombination is high and its effects turn out to be detectable on shorter time scales than selection, the evolution of different loci is no longer correlated and the genotype fitness can be described as an additive function of the single locus fitness. It has been shown [26, 27] that the competition between selection and recombination on short time scales, when the contribution of mutations to diversification is negligible, produces different phases: the quasi-linkage equilibrium (QLE) regime [28], corresponding to frequent recombination [29] and absence of coexisting clones, is disrupted by the buildup of sequence correlations (*epistasis*) with decreasing recombination rate. In this phase of "clonal condensation" [26] some genotypes grow to significant frequencies despite recombination. The transition from allele selection to genotype selection as a result of the race between these two evolutionary forces is analogous to the freezing transition in the Random Energy Model of spin glasses [30, 31].

Sexual reproduction is not the only mechanism which feeds variability via reshuffling of alleles. Horizontal gene transfer (HGT), namely the exchange of genetic material between organisms as opposed to parental inheritance (also referred to as vertical gene transfer), plays an important role in bacterial adaptation, often conferring and spreading antibiotic resistance [32–34]. Also viruses with segmented genomes, such as influenza, can be subject to the mixing of genes from different strains during replication. The dynamics and effects of the so called process of *reassortment* in the evolution of human seasonal influenza is the main subject of this thesis.

In the next sections we briefly introduce the biology of influenza virus, describing the structure of the pathogen and relating it to the mechanisms that drive its evolution, with a particular stress on the process of reassortment. Influenza is a notable example of a fast-evolving pathogen adapting in a time-dependent environment; the abundance of available sequence-data makes it a significant model for studying evolution of organisms out of equilibrium. A more comprehensive understanding of influenza evolution and the subsequent step of quantitative modeling represent a potential breakthrough in the ultimate challenge

of making predictions.

1.3 Influenza virus: a model of adaptive evolution

Influenza virus is a negative-sense single strand RNA virus. The pathogen is spread and common in every continent and still has a remarkable impact on public health, considering that seasonal influenza causes about half a million deaths per year in humans [35]. Infections are induced by three phylogenetically and antigenically distinct influenza lineages - A, B and C - that co-circulate globally. Among these lineages, influenza A shows the fastest rate of evolution [36–39], while type C viruses tend to cause less severe disease [40]. The genome of the virus is segmented into 8 RNA filaments that, by taking advantage of the replication/translation machinery of the host cell, encode 11 different proteins (Fig. 1.1). The filaments PB1, PB2 and PA encode the viral RNA polymerase (RdRP) and various polypeptides, likely responsible for inducing cell death [41] and for modulating viral pathogenicity [42]; NP encode the viral nucleoprotein, while M the matrix protein (M1) and a surface protein (M2); NS encodes both the nonstructural 1 (NS1) protein, involved in immune evasion, and the nuclear export protein (NS2), which mediates the nuclear export of viral ribonucleoprotein (complexes of RNA and RNA-binding proteins); last, HA and NA encode the dominant surface glycoproteins, haemagglutinin and neuraminidase, respectively. Antigenic epitopes, namely the primary loci of interaction with the human immune system, are mainly located in the globular head of HA (HA1 domain) [43]. Within the A lineage, the combination of these two proteins on the surface determines a further classification of the virus into subtypes; most of the 18 HA and 11 NA subtypes affect avian species, while the predominant variants circulating among humans are A/H3N2, A/H2N2, A/H1N1, A/H1N2 [44, 45]. A/H1N1 and A/H3N2, in particular, cause seasonal influenza virus epidemics; the latter subtype is the focus of the analysis presented in this thesis.

1.3.1 Evolution by antigenic changes

Within each segment, genomic evolution is a purely asexual process carried by point mutations, which are subject to genetic drift and natural selection. By the analysis of HA sequences from viral isolates, it has been shown that the high mutation rate of influenza is the fodder for the emergence of clonal interference [10]. A set of competing clones, broadly defined as a group of genetically similar (often not identical) strains, appears in the

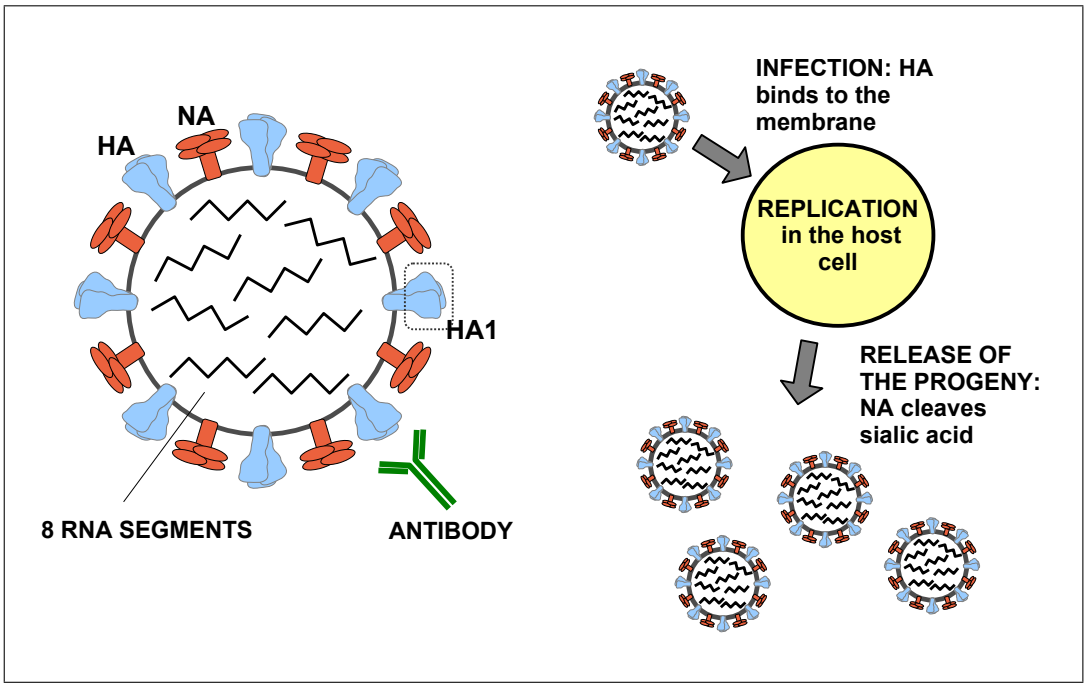


Figure 1.1: The influenza virion. The viral particles (left panel) of influenza A, B and C share similar structures; the capsid is wrapped into the viral envelope, which packages and protects the central core containing the viral RNA genome. The influenza A and B virus genome is composed by 8 gene segments, encoding for 11 proteins (see main text). Haemagglutinin (HA) protein has a trimeric structure, each monomer being composed of two domains: HA1 and HA2. HA1 contains the receptor binding site (RBS) for cell attachment and is the primary target of anti-influenza antibodies. HA is responsible for cell attachment and entry (binding of sialic acid and membrane fusion), while neuraminidase (NA) cleaves the bonds between HA and sialic acid and facilitates viral movement through the mucus [46], enabling the release of new virions from the host cell after replication (right panel).

population; some clones survive, some get eventually extinct. The rapid fixation of strongly beneficial mutations (*selective sweeps*, typical sweep time[1] being 3 - 4 years) responsible for the expansion of the successful clones has two "collateral" effects. Within sweeps, while neutral and moderately deleterious mutations often fix in the population by hitchhiking as a result of linkage, moderately beneficial changes originated in outcompeted clones are lost in the process. In this picture the viral population always remains multiclonal, as selective sweeps do reduce but not eliminate diversity. The deleterious mutation load emerging with clonal interference limits both the speed (measured by the mean fitness flux, see [12]) and the degree of adaptation [10], the latter characterizing the functionality of a gene segment. Furthermore, it has been shown that accumulation of deleterious

[1]Typical coalescent time.

mutations can explain the pattern of rising and spreading of new antigenic clusters[2] [48]. Positive selective pressure by host immunity plays an important role, in particular, in the evolution of haemagglutinin (HA), which governs viral binding and entry into host cells, and neuraminidase (NA), which drives the release and escape of new virions from the cell [49,50], thus promoting the infection of other cells. The gradual accumulation of adaptive mutations in these two proteins maintains the ability of the virus to continually evade host immunity [39,51]; this phenotypic process has been called *antigenic drift*[3] [47,52].

1.3.2 The process of reassortment

In parallel to point mutations within single proteins, the genome of the influenza virus changes by so-called reassortment processes. If the same host cell is co-infected by two or more viruses carrying distinct genomes, mixing of genomic segments within that cell may produce a hybrid genotype carrying segments from different parental strains. Fig. 1.2 shows a schematic representation of the mechanism. The evolutionary implications of

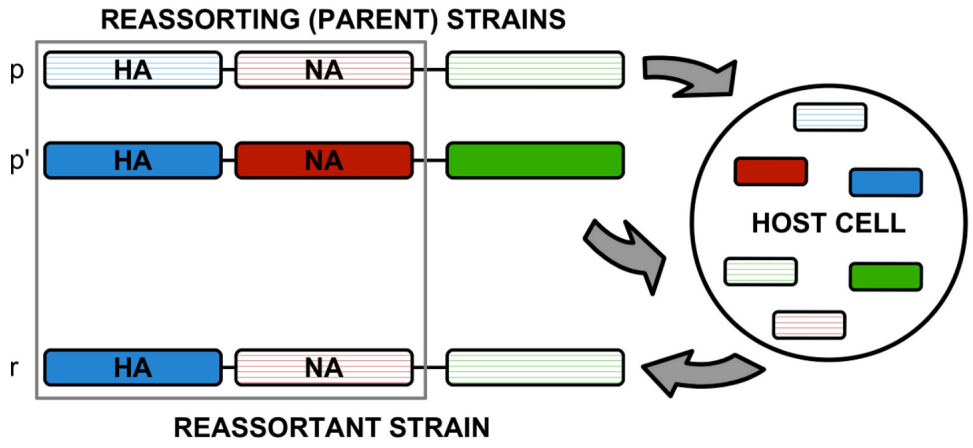


Figure 1.2: Schematic of a reassortment process. Two parent strains, p and p', co-infect a host cell and produce a reassortant strain r. Here we focus on reassortment of the two surface proteins HA (blue segments) and NA (red segments); the reassortant strain r inherits one of these segments from each parent.

these dynamics are quite complex. On the one hand, in rare cases, reassortment can lead to *antigenic shifts*, which are new combinations of haemagglutinin and neuraminidase that strongly enhance fitness [53] by escape from host immunity [54,55]. Cross-species

[2] Sets of viral strains antigenically similar one to another [47].

[3] Not to be mistaken with *genetic drift*.

infections are possible and new lineages deriving from the mixing of strains infecting different classes of living beings are often the main cause of so called *pandemics*, worldwide spread epidemics with high mortality rates among the infected population. The acquisition of new HA and NA variants by human influenza A through reassortment with avian strains, for example, has been shown to cause global pandemics in 1957 and 1968, known as the "asian" and the "Hong Kong" flu, respectively [56, 57]. Many reassortments, however, have negligible antigenic effects but may have other fitness effects. Specifically, fitness interactions between segments across lineages are observed as biases in observed pairings [58–64]. By partly randomizing such pairings, reassortment generates a fitness cost and a resulting increase of subsequent compensatory mutations [65]. In terms of statistical physics language, co-evolution results in a strongly correlated many-particle process driven by epistatic effects between proteins, while reassortment, on the other hand, acts as a randomizing factor that breaks down the correlation pattern. Broad negative selection has been postulated for reassortment between well distinct influenza B lineages [66], but the overall selective effects of intra-subtype ressortment have not been systematically analyzed so far. In this thesis we give our contribution to a better understanding of these dynamics by showing that reassortment within a given influenza lineage induces a fitness cost that increases in strength with increasing genetic distance of the parent viruses. Our finding suggests that evolution continuously produces viral proteins whose fitness depends on each other; reassortment reduces fitness by breaking up successful combinations of proteins. Thus, selection across proteins constrains viral evolution within a given lineage, and it may be an important factor in defining a viral species.

1.4 From data analysis to forecasting evolution

The continuous alteration of the antigenic phenotype of influenza virus allows reinfection of previously exposed individuals and makes frequent update of vaccines a necessary procedure. Decisions on the composition of these vaccines need to be taken nearly one year in advance and rely on early identification of novel emergent variants [67, 68]. As a first step towards an efficient surveillance and vaccine strains selection process, huge efforts are being made to collect high-quality worldwide data. As a result, large datasets of complete genome sequences of viral strains are available, making influenza one of the best documented systems of molecular evolution. The HA gene sequence, in particular, is available for several thousands strains isolated over the last 40 years [69] and has therefore been

the main target of a remarkable number of studies to investigate its primary role in the interaction with the host antibodies. Bioinformatic tools for processing raw data, such as multiple sequence alignment methods (see for example MUSCLE [70] or BLAST [71] for popular implementations of such methods), are normally used to prepare alignments[4] of RNA sequences and infer their evolutionary relationships through the reconstruction of phylogenies.

1.4.1 Inference of phylogenetic trees

It is beyond the scope of this thesis to present a detailed review on the methods currently used to build phylogenetic trees (see reference [72] for such a purpose), however we briefly mention the most common ones, since the inference of genealogies will be a relevant step in our analysis (chapter 2).

Distance based methods start from the reconstruction of the distance matrix of observed (pairwise) distances between the sequences in the alignment. Sequences close in the distance space are placed in the same "neighborhood" of the tree, as descending from a recent internal common ancestor. Algorithms based on this approach have the advantage of being extremely fast, thus become relevant in the analysis of very large datasets. Despite that, trees built with such methods may not be the "best" trees to describe the actual evolution of the data. The step of summarizing a set of sequences with a distance matrix implies loss of information and only gives an overall estimate of the relationship between trees and data.

Maximum Parsimony (MP) methods are based explicitly on the sequences, in contrast to pairwise distances, and aim at reconstructing the tree that explains the evolution of each site under the constraint of minimizing the number of evolutionary changes. Although parsimony relies on few reasonable and simple evolutionary assumptions, namely that evolutionary changes are rare, they tend to lose accuracy in the reconstruction of portions of trees in proximity of long branches [73]. Moreover, the search for the most parsimonious tree in the space of all possible trees generated by the data is generally a time-costly computational procedure.

[4] Aligning gene sequences is in general a complex problem. Lots of methods have been developed to efficiently establish homology among nucleotide positions and the choice of one approach over the other depends on the specific set of data. Producing reliable alignments of RNA sequences of influenza A single segments is however a relative simple operation, thanks both to the high quality of data and general absence of gaps eventually due to insertion and deletion events.

The latter disadvantage affects also Maximum Likelihood (ML) algorithms. Instead of preferring the tree with the fewest possible evolutionary changes, ML searches for the phylogeny that maximizes the likelihood of getting the aligned sequences under a model of sequence evolution, which evaluates the probability of particular mutations. In broad terms, low probability is assigned to trees that require more mutations at internal nodes to explain the observed phylogeny. The most parsimonious tree is also the maximum likelihood tree when the expected amount of evolutionary change is very small, which is not the case of influenza evolution. For our analysis we build the genealogies of single and paired gene sequences (chapter 2) using a ML method under a General Time Reversible (GTR) model of nucleotide substitution that incorporates rate heterogeneity among sites.[5] To obtain the larger trees analyzed in chapter 4, the ML inference is preceded by an additional step: in order to reduce the total inference time, a preliminary topology is reconstructed with a fast distance-based method (implemented in FastTree software [75]) and used to restrict the tree space for the following ML runs. Fig. 1.3 shows the ML phylogeny of A/H3N2 haemagglutinin (data from 1968 to 2015, as in chapter 2). The shape of the tree reflects the modes of evolution of the influenza virus. High supply of beneficial mutations creates competition between clones; fixation events occur on the trunk (black thick line in the figure), while the outcompeted clades - defined as sets of strains descending from recent last common ancestors - get extinct, resulting in a strongly unbalanced monophyletic tree (see Fig. 2.2 for a comparison with a ML tree built with paired HA-NA sequences and reference [10] for HA trees reconstructed with maximum parsimony principle).

1.4.2 Predictive models for influenza evolution

The genomic information coded into dense sequence alignments and respective phylogenetic trees is the base for the development of predictive models. Despite the complex and stochastic nature of the evolutionary processes at molecular scales, the specific mode of asexual evolution driven by large supply of mutations and strong selection leads to the emergence of repeatable features. Adaptation to host immunity is carried on by strongly beneficial mutations occurring in a limited amount of epitope sites; selection shrinks the space of evolutionary possibilities, thus opening the way to predictability [25]. In this picture, the massive increase of available sequence data, paired with phenotypic (antigenic)

[5] GTRCAT model implemented in RAxML software [74].

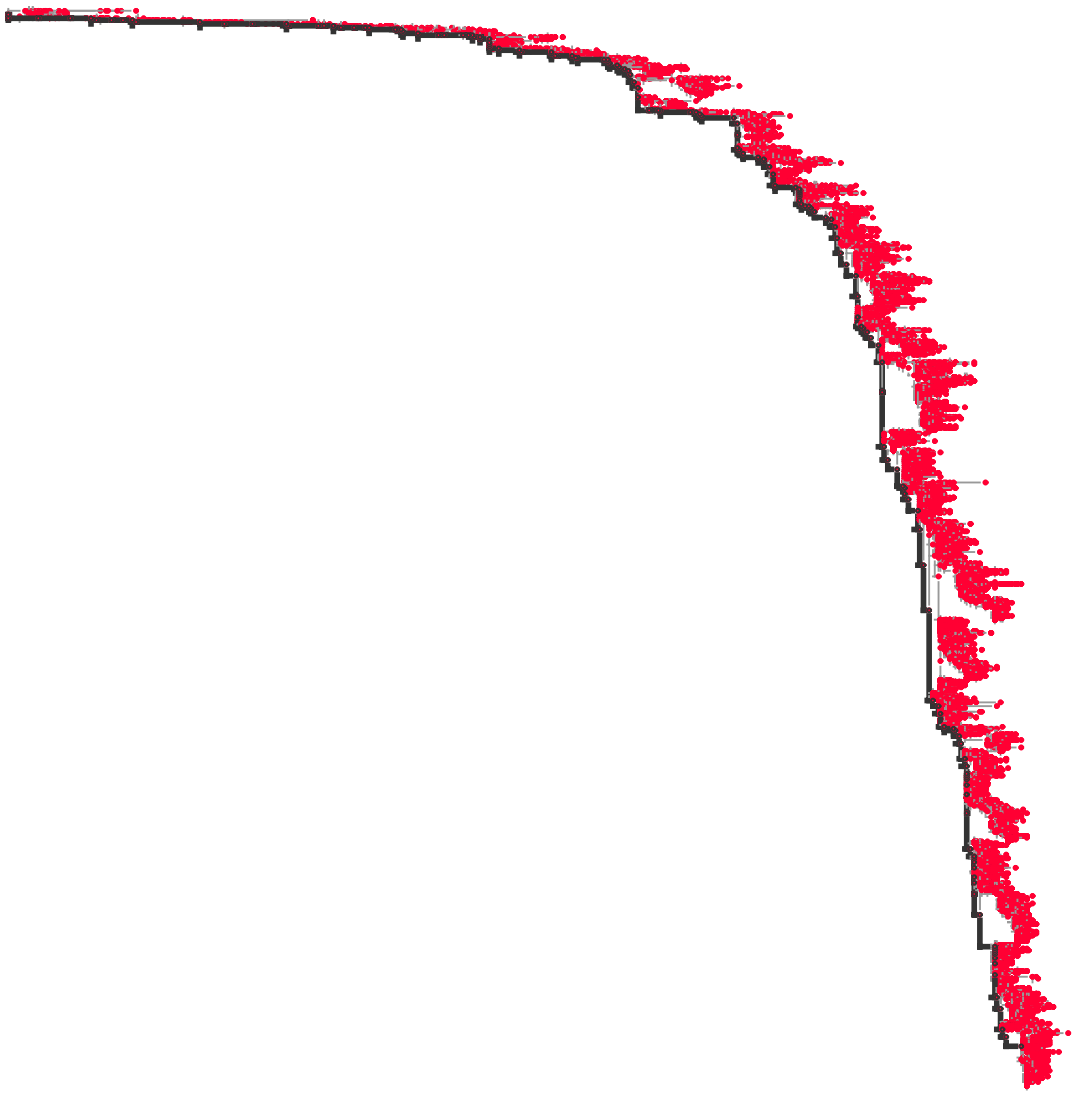


Figure 1.3: Phylogenetic tree of A/H3N2 human influenza. The evolutionary tree of the seasonal influenza virus built with haemagglutinin gene is highly asymmetric. Nodes in the tree (red dots) represent distinct HA sequences; terminal nodes, in particular, are observed strains. Distances in the horizontal axes should be read in terms of number of mutations between sequences, which represent an indirect measure of time. Selective sweeps occurring every few years drive fixation of successful clones, originating on the trunk (black thick line), and extinction of outcompeted variants. These evolutionary dynamics shape the growth of the phylogeny along one main lineage, thus determining the monophyletic aspect of the tree.

characterization of the isolates (typically obtained by haemagglutinin inhibition assays[6])

[6]Haemagglutinin inhibition assays (HI) measure the maximum dilution at which antibody-containing serum prevents a particular influenza virus from agglutinating red blood cells [76].

are the ground on which predictive models are built (see reference [68] for a recent review). The genetic history of viral clades can be tracked on the inferred trees and their frequency changes modeled by assessing fitness differences between sub-populations.

The prediction of clade success requires the establishment of a link between fitness distances and genetic data. Fitness landscapes (or time-dependent seascapes [12]) can be modeled explicitly, as in [77], however some predictive models adopt different approaches [78, 79]. The analysis of phylogenetic trees has been used, for example, as a base to develop a method for inferring directional selection on specific alleles (only nonsynonymous mutations in the epitope sites in haemagglutinin) [79, 80] and extrapolate recent growth of relevant HA clades. The mapping of HI data on the phylogeny identifies high-growth clades with relevant antigenic impact, setting a criterion for the proposal of new vaccine strains components. Although incorporating both genetic and antigenic data is potentially a successful strategy, predictions based on this method do not take into account fitness effects outside epitope mutations.

The idea that phylogenetic trees on their own contain enough information to infer fitness has brought to the development of a more general model, that does not require detailed molecular information [78]. With the assumption that the population is under persistent directional selection, fitness changes along the lineages occur gradually as a result of the continuous accumulation of small effect mutations. The fitness of most clades, then, decreases over time under the effect of weakly deleterious mutations, while few lineages adapt and remain among the fittest in the population. The fitness distribution along the tree is obtained by the observations that rapid branching is expected under fit internal nodes, and children with high fitness are likely to be recent descendants of high fitness parents. Furthermore, the ranking of nodes by fitness is only mildly dependent on the parameters required to model the posterior fitness distribution, indicating that the latter can be related to more universal quantities. In detail, high fitness internal nodes are associated to high downstream total branch length, namely, for a given number of descendants, these nodes originate star-like subtrees. At the same time, a lineage with low downstream branching is likely to have low fitness. It is therefore possible to define a heuristic measure that estimates clades growth rates relying exclusively on the local shape of the phylogeny. The so called "local branching index" $\lambda_i(\tau)$ (LBI) is defined as the total length of the tree surrounding node i, averaged with an exponential decreasing weight τ, and ranks the nodes with an accuracy comparable to ranking by calculating fitness distributions. In chapter 4 we adopt the idea of ranking nodes on the phylogeny

by heuristic algorithms and apply it for a different purpose. Instead of inferring fitness, we define a parameter to spot subtrees where reassortment is likely to be occurred.

A parallel approach to the ones described above is the explicit modeling of fitness functions, which serves to predict the evolution of genetic clades from one year to the other [77]. At a given point in time (season) t, the frequency $X_\alpha(t)$ of clade α is defined as the sum of the frequencies x_i of all the strains i part of that clade: $X_\alpha(t) = \sum_{i:\alpha,t} x_i$. The expected frequency in the next season is given by $\hat{X}_\alpha(t+1) = \sum_{i:\alpha,t} x_i \exp(f_i)$. The (Malthusian) fitness f_i of each strain i is modeled as a sum of contributions from mutations in epitope and non-epitope sites. Non-epitope mutations, often causing protein destabilization, are assigned a fitness cost $\mathcal{L}$ to model predominant negative selection.[7] Epitope mutations are predominantly under positive selection and are included with a term that describes cross-immunity C across multiple strains. The fitness of each strain depends on gene sequences $\mathbf{a}$ and takes the form

$$f_i = f_0 - \mathcal{L}(\mathbf{a}_i) - \sum_{j:t_j<t_i} x_j C(\mathbf{a}_i, \mathbf{a}_j)\,, \tag{1.5}$$

with constant f_0 ensuring normalization of strain frequencies. Here, the non-epitope load and the cross-immunity terms are approximated using Hamming distances between genetic sequences. This minimal genetic fitness model can predict the expected cross-immunity between a given strain and the circulating strains in a certain season, thus predicting the optimal vaccine strain as the one maximizing cross-immunity. Antigenic data can be included explicitly in the fitness functions and used to predict more accurately the dominating clades of the upcoming season.

Each of the predictive models described so far relies on different assumptions and explore biological details at different levels. However, they are all based on information codified in haemagglutinin protein alone. This first order approximation is justified by the primary role that HA has in the interaction with the host immune system, which made this protein the target of a remarkable number of studies. It is reasonable to assume that models built on HA alone cannot be exhaustive, since adaptive sites are actually present in other proteins [81], especially in the already cited neuraminidase. Including NA sequences into the analysis is then a crucial step in order to achieve the final goal of building a thorough fitness model for influenza that can really be predictive. Taking into

[7] Non-epitope changes that are approximately neutral, have compensatory fitness effects or contribute to the adaptive process are neglected.

account both HA and NA entails the knowledge of how the two proteins co-evolve; it is of particular interest quantifying inter-proteins linkage effects, direct result of functional coupling of antigenic changes in HA and NA [82], and quantitatively elucidating the fitness effect of reassortment (section 3.2.3). How frequent is reassortment involving these two glycoproteins among seasonal A/H3N2 influenza? Are reassortant variants fitter than non-reassortant counterparts? Can we learn from the statistics of reassortment whether the virus evolves under the constraints of fitness interactions across proteins? Giving an answer to these interesting questions means being able to draw the sketch of the implications of including reassortment dynamics in the construction of multi-protein fitness models and constitutes the main goal of this thesis.

2. Detection of reassortment: a novel genealogical method

In the previous chapter we have highlighted the importance of reassortment in shaping the evolutionary dynamics of influenza, emphasizing the existence of crucial open questions regarding selection effects. A necessary methodological basis for answering these questions is the faithful inference of intra-subtype reassortment events from sequence data. Although these dynamics have long been recognized as a potentially important mechanism for evolution [83, 84], the detection of events within the same subtype is notoriously difficult due to their weak phylogenetic signal. There is a number of current methods to infer reassortment events from a data set of viral sequences. These methods can be roughly divided into two groups: distance-based methods [85, 86] and methods based on the phylogeny [55, 83, 87–93]. As recently pointed out [94], these approaches coherently report some fraction of the reassortment events but show a substantial degree of discrepancy between their results, which can be traced to method-specific differences in sensitivity.

In this chapter, we first comment upon the methods commonly used to detect reassortment and draw the attention to the need of developing a new approach which fits the purpose of intra-subtype reassortment inference. We therefore present our novel detection method, designed and optimized for this scope, and describe the algorithm in detail. Particular care will be devoted to the validation of the approach through comparison with a non-reassortant case and tests on simulated data.

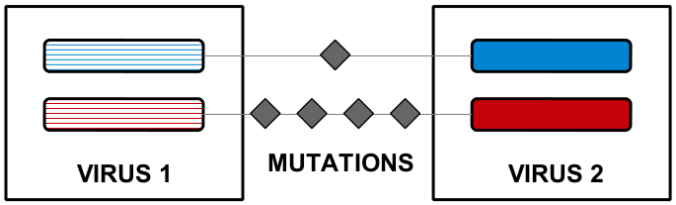
2.1 Inference of reassortment: distance-based and phylogenetic methods

Distance-based methods rely on the assumption that, for reassortant strains, high similarity between the sequences of one segment goes along with large differences in the other segment (Fig. 2.1(a)). These methods do not pass through the step of inferring the phylogeny of the virus. They are fast, can be efficiently applied to large alignments, and are insensitive to errors in tree reconstruction. Without a viral phylogeny, however, it can be hard to determine if two or more inferred events are independent. Hence, distance-based methods can generate multiple representations of the same original event between unobserved ancestral strains. Resolving false positives constitutes a major issue for this kind of algorithms.

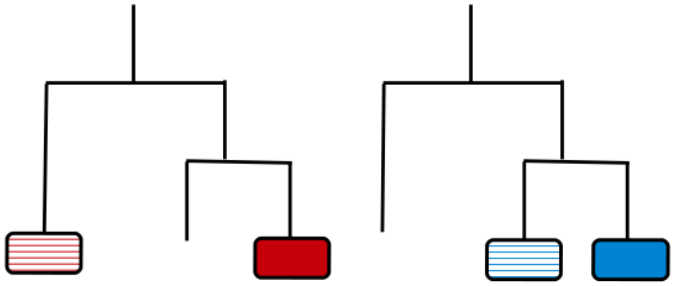
Phylogenetic methods, on the other hand, are based on the observation that reassortant strains are located in different clades of the coalescent trees built for different segments (Fig. 2.1(b)). These approaches are usually successful in detecting reassortment across different lineages of the virus, i.e. between strains with substantial genetic differences. The incompatibility between tree topologies, however, can also be a result of phylogenetic errors, so that inconsistencies in the evolutionary histories of different segments are a necessary but not sufficient condition for reassortment. Successful attempts to overcome this issue [91] have produced algorithms which are applicable only to small datasets. Since the scaling of the number of inferred events with sample sizes has not been investigated, it is not clear if the rate of reassortment is independent on the size of the trees. Hence, even if the gain in information coming from the phylogenetic trees constitutes an advantage over the distance-based algorithms, the limited resolution constrains these methods in the detection of intra-lineage events [51].

2.2 Inference of reassortment on two-segment trees

In order to fill this methodological gap, in the following sections we propose a new genealogical inference method for reassortment in fast-evolving populations with segmented genomes, such as influenza virus. We analyze the mutations arising in joint genealogies built from pairs of segments and set a simple criterion to identify reassortment events.



(a)



(b)

Figure 2.1: Current methods for inferring reassortment. (a) Distance based methods detect inconsistencies in distances between segments. Small deviations in one segment (blue) correspond to large deviations in the other one (red): the two proteins have different evolutionary history and their mixing is a product of reassortment. (b) Phylogenetic methods spot alignment mismatches between trees of different segments. The two subtrees refer to evolution in distinct segments (red, blue). Two viruses in the same clade relatively to the blue segment fall into separate subtrees in the red segment, indicating that reassortment brought a new variant which differs from the background.

2.2.1 Alignments and genealogical trees

In this study, we focus on reassortment between the haemagglutinin and neuraminidase influenza genes, because the evolution of both proteins has been linked to immune escape and functional epistasis between them affects vaccine efficacy [95–97]. Hence, we restrict the genomic analysis to the two segments carrying the HA and NA genes; each parent strain contributes exactly one of these segments to the reassortant strain (see Fig. 1.2). For our purpose, a sample of HA and NA sequences is obtained by downloading all the available A/H3N2 human strains in the EpiFlu DATABASE (http://www.gisaid.org), regardless the geographical region, which were collected between January 1968 and October 2015. Only the strains with complete HA and NA sequences are taken into consideration. From this sample, alignments of single segments are created year by year using BLAST [71].

After discarding the segments with more than three gaps,[1] a first run of RAxML [74] is performed to reconstruct the genealogy for each segment and detect clear outliers sequences (i.e. sequences that are found clearly isolated from the rest of the tree and which were likely misreported) to be excluded from the subsequent analysis. We then obtain alignments of linked HA-NA segment pairs and construct maximum-likelihood two-segment genealogies by RAxML, choosing the best-scoring ML tree out of 10 RAxML runs. Fig. 2.2 shows one HA-NA joint tree obtained by applying this routine. The Potential ambiguities in the definition of the nucleotide at a certain site in a strain are resolved by assigning the orthologous nucleotide of the closest ancestral node with an unambiguous sequence. The subsequent reassortment inference is performed on trees of subsampled data with a maximum of 600 sequences per year. This step reduces computational time and avoids over-representation of recent viruses, which are the most abundant in the database. We note that these joint genealogies differ from single-segment phylogenetic trees, because the underlying process of reassortment violates the tree topology.

Some of the isolates we use in this study were subject to passaging in cell culture before sequencing. This preliminary step amplifies the viral copy number, helping to test the features of the pathogen into a living system. It has been shown that through passaging influenza sequences can accumulate adaptive mutations [98], which constitutes a possible confounding factor in the evolutionary analysis of the virus. To exclude this eventuality, we repeat our inference of reassortment on 20 trees built from a restricted alignment of 1053 unpassaged sequences.

2.2.2 Primary inference of reassortment events

A tree representation of a joint genealogy with a reassortment process is shown in Fig. 2.3. The two parental strains p and p' appear in different sublineages, and the reassortant strain r is shown as a descendant of one of these parents (here p'). These strains define three distinct clades of descendant strains, C_p, $C_{p'}$, and C_r (grey areas in Fig. 2.3); the numbers of strains in these clades are denoted by n_p, $n_{p'}$, and n_r, respectively. We note that the "direction" of the reassortment event (here from p to p') is merely a property of the tree representation, and there is an equivalent tree with the roles of p and p' exchanged. This reassortment pattern can be readily identified in two-segment trees. Fig. 3.4 shows an

[1] This choice may appear restrictive, however it is necessary in order to guarantee the reliability of the detected events, which constitutes an essential condition to avoid uncontrollably confounding factors possibly altering the following analysis concerning selection (see chapter 3).

Figure 2.2: Typical A/H3N2 joint-HANA tree (1968-2015). Genealogical trees built pairing HA and NA segments maintain the unbalanced monophyletic structure of the single segment trees (cf. Fig. 1.3). Occasional reassortment manifests itself with long branches, mainly in peripheral regions, which isolate reassortant clades from the non reassortant background.

example of a HA-NA reassortment event in the genealogy of influenza A/H3N2, another example using tree data from simulated evolution of a population in a regime of clonal interference is shown in Fig. 2.5.

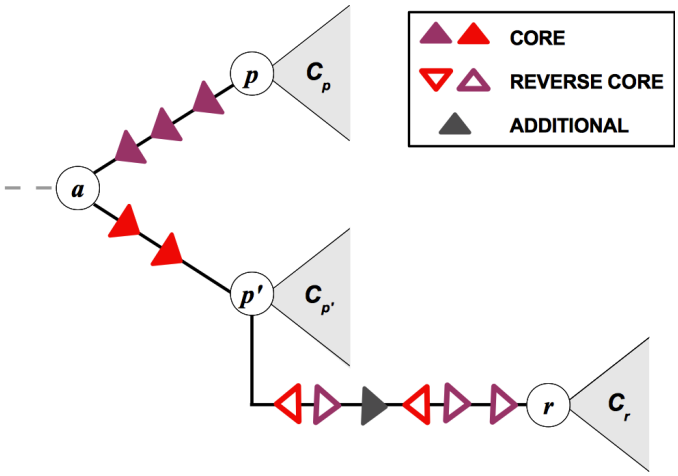


Figure 2.3: Representation of reassortment in a two-segment genealogical tree. The parent strains p and p' are in different sublineages of the tree; the reassortant strain r appears as a descendant of one of these parents (here p'; there is an equivalent tree in which r appears as a descendant of p). The strains p, p', and r are the focal nodes of the clades $C_p, C_{p'}$, and C_r, respectively (grey areas). We identify the reassortment event by its set of core mutations, $\mathcal{A}_{pp'}$, which appear on the segment that r inherits from p and generate the genetic distance between the parent strains in that segment. The core mutations appear on the branches between the nodes p and p' (filled red triangles: mutations between p and the last common ancestor a, filled purple triangles: mutations between a and p'). Their reverse mutations appear on the branch between p' and r (empty red and purple triangles), which can also contain additional mutations (grey triangles).

The representation sketched in Fig. 2.3 defines one of the two segments, referred to as the travelling segment, and a set of mutations which generate the genetic distance between the parent strains p and p' in that segment. These so-called *core mutations* appear on the branches between the nodes p and p', which pass through their last common ancestor a. We define the set $\mathcal{A}_{pp'}$ of core mutations by counting the mutations from p to a in upward direction on the tree (filled red triangles) and the mutations from a to p' in downward direction (filled purple triangles). The branch from p' to r contains a set of mutations $\mathcal{A}_{p'r}$ that includes the set of reverse core mutations, denoted by $\bar{\mathcal{A}}_{pp'}$ (open red and purple triangles), as well as additional mutations in the travelling segment (grey triangles). These mutations may reflect insufficient sampling (i.e. one of the actual parent strains is not included in the tree) and noise in the reconstruction of the phylogeny (see below), or they are point mutations unrelated to the reassortment event. Together, we obtain a criterion to detect reassortment in a two-segment tree: we parse the tree for node triplets (p, p', r)

with

$$\mathcal{A}_{p'r} \supseteq \bar{\mathcal{A}}_{pp'} \tag{2.1}$$

in a given travelling segment.

To characterize the span of a reassortment event, we use the mean genetic distance between the parent strains p and p' in both segments, $d = \frac{1}{2}(d_{\mathrm{HA}} + d_{\mathrm{NA}})$ (we evaluate these distances for nucleotides and for amino acids). The quantity d is also the mean genetic distance of the reassortant strain r from its parents. The resulting list of events must undergo further statistical analysis: false positives must be excluded and candidates representing the same reassortment event must be counted only once. In order to reduce the list of putative reassortments to a minimal set of independent events, we include internal inferred nodes as possible candidates for reassortment and cluster events with similar patterns of mutations.

2.2.3 Pruning steps: uniqueness and false positives

First, we exclude false positive events due to ambiguities in tree reconstruction by statistical comparison with a null model. Since recombination within segments does not occur in influenza, we use as null case a set of trees built from the alignments of the single segments. We decompose the sequence of the protein into two subsets of randomly chosen sites with lengths L_1 and L_2. These subsets have the appropriate ratio of lengths to mimic the segment structure in the original joint alignment, $L_1/L_2 = L_{\mathrm{HA}}/L_{\mathrm{NA}}$. In order to investigate the dependence of the number of false positive reassortments on the length of the chain, we run the detection algorithm on subsets of sites of increasing total length $L = L_1 + L_2$, maintaining a constant ratio L_1/L_2. The fidelity of our inference scheme strongly depends on the number of core mutations, $\delta = |\mathcal{A}_{pp'}|$: as detailed in chapter 3, we find an expected number $n_0(\delta)$ of false positive reassortment events that decays rapidly with increasing core distance,

$$n_0(\delta) = Ce^{-\gamma\delta}. \tag{2.2}$$

From the observation that the decay exponent γ is approximately independent of the total sequence length L (see chapter 3 for further details), we can then evaluate the expected number of false positive reassortment events in the actual data.

Second, two or more different reassortment events reported by our algorithm may represent the same biological reassortment event if they have similar core set (see Fig. 2.4). To address this source of overcounting, we compare the core sets $\mathcal{A}_{pp'}$ of putative

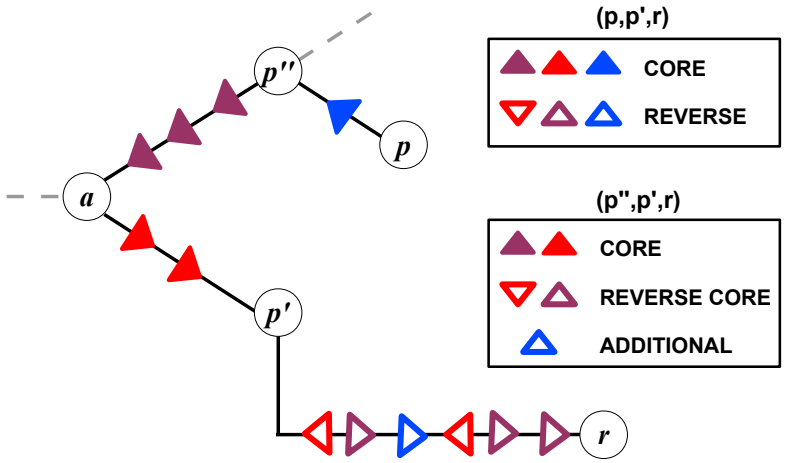
events. If these sets differ by at least 30% of their mutations, the events are considered independent; otherwise we keep only the set with the largest core distance δ. The number of pruned events turns out to be insensitive to moderate changes of the threshold number of mutations. Furthermore, we cluster the reported events with different travelling segments that have the same (or very similar) parent strains. This step is necessary to prevent that one event eventually appearing in its two equivalent representations in the joint-segment tree (p and p' have interchangeable roles and determine the artificial direction of the event) is detected twice.

2.3 Testing the inference method by simulations

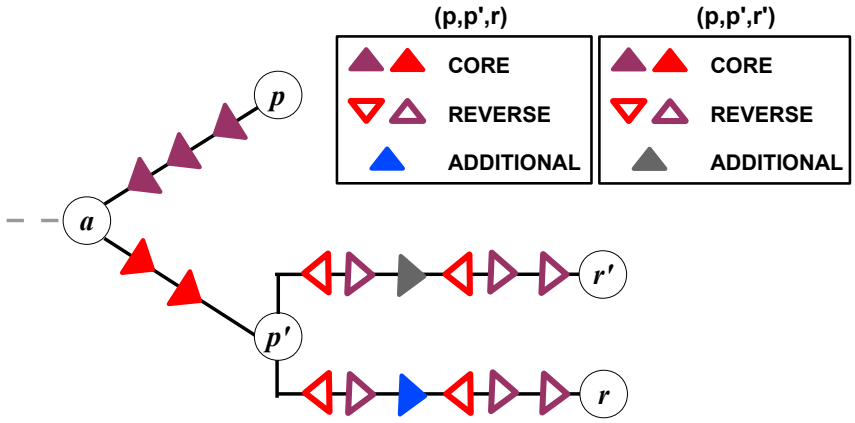
Beside verifying the faithfulness of the new method by comparison with a null case, we validate our algorithm by testing it on simulated data. We simulate the genome evolution of a population of N individuals starting at a stationary state, under the effect of mutation, genetic drift and selection, based on the model used in reference [10]. Each strain is characterized by a sequence of epitope and non-epitope sites, flanked by neutral sites. Selection on epitope sites is time-dependent and its direction fluctuates randomly at a rate γ, while non-epitope sites are modeled with time-independent direction. The time dependence of selection models the emergence of new beneficial epitopes resulting from immune escape. To these basic steps we add reassortment, which occurs at each generation with probability λ: we select randomly two individuals (the parents) and divide their genome into two parts of fixed length L_1 and L_2, then mimic the process of reassortment by creating a new individual (r) with a mixed genome. We focus on events between strains at genetic distance $d_{1,2} \geq 5$ in each segment, discarding reassortment at lower distances. The results of each simulation are a set of sampled sequences, some of them involved in a reassortment event, that we use to build up the genealogical trees, as we would do with real observed strains.

We choose the parameters of the simulations as follows:

1. The evolution of $N = 1000$ individuals is simulated for 1500 generations. Each individual has a genome of length $L = L_{ep} + L_{non-ep} + L_{neut} = 560$ ($L_{ep} = 120$, $L_{non-ep} = 160$, $L_{neut} = 300$ number of epitope, non epitope and neutral sites, respectively), selection flips the direction at rate $\gamma = 0.033$ and the mutation rate is set to $\mu = 5.8 \times 10^{-3}$ per year. With these evolution parameters, the population turns out to be in a clonal interference regime comparable to influenza [10].



(a)



(b)

Figure 2.4: Representation of events with similar core sets. (a) p differs from p'' only in one mutation (blue triangle), which appears (reversed) on the branch from p' to r. This mutation is part of the core in event (p, p', r), while it is seen as an additional mutation in event (p'', p', r). Miscounting is avoided by comparison of similar core sets; the two events are reported as a single event (p, p', r) (the triplet with the largest core is taken as the representative event). (b) The events (p, p', r) and (p, p', r') differ only in additional mutations and are counted as a single event.

2. We introduce reassortment at a rate $\lambda = 1 \times 10^{-6}$ per individual and per generation. This generates a density of reassortant variants at observable population frequencies that is comparable with the observed density in influenza A/H3N2.

With these parameters, we obtain trees that show ~ 5 coalescent events on average, corresponding to approximately 10 years of influenza evolution [99]. We apply our algorithm on each of the 100 reconstructed trees and check if the reassortment events recognizable in the sampled sequences (i.e. the ones with r and/or its offspring reaching a relevant frequency and therefore getting sampled) get detected. Out of the total 283 events generated in the simulations, 214 (76%) are correctly reported (see Fig. 2.5 for an explicit example of a detected event), with 24 false positives signaled with small cores ($\delta \leq 5$).

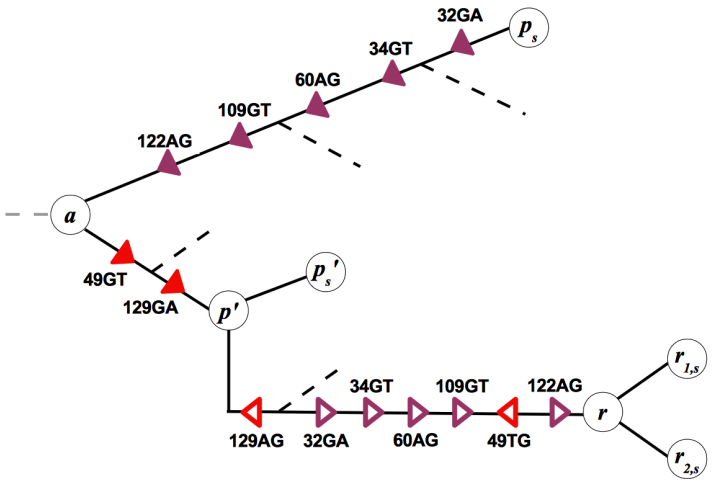


Figure 2.5: Representation of a simulated reassormtent event. The result of a simulated reassortment event on the reconstructed genealogical tree, correctly detected by the algorithm. The internal node r is inferred as the reassortant ancestor of $r_{1/2,s}$, i.e. the strains evolved from the sequence that was actually generated by reassortment between p_s and p'_s.

2.4 Remarks

In this chapter we have summarized the current state of the art regarding the approaches commonly used to infer reassortment starting from RNA viral sequences, pointing out the necessity of developing a new appropriate method to detect events occurring between strains in the same subtype. We have therefore presented a novel algorithm based on the recognition of pattern of mutations arising in joint-segment genealogies. Our method has been tuned by comparison with a null case and tested extensively on simulated data for evolving influenza-like strain populations under mutations, genetic drift, selection, and reassortment. A large fraction of the simulated reassortment events are recovered by our

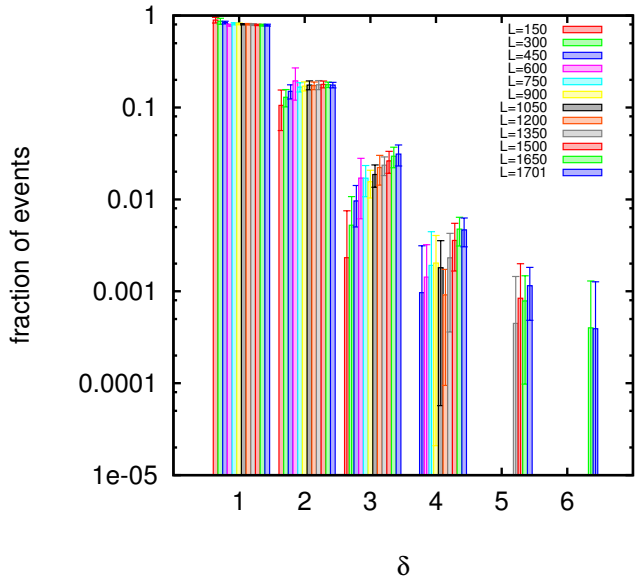
algorithm, and these events outweigh the rate of false positives, which constitutes a proof of robustness and high accuracy. In the next chapter we apply our inference routine to a large set of A/H3N2 influenza sequences; the resulting list of reassortment events are analyzed by means of two independent methods, in order to draw conclusions on selection effects.

3. Reassortment in A/H3N2 human influenza

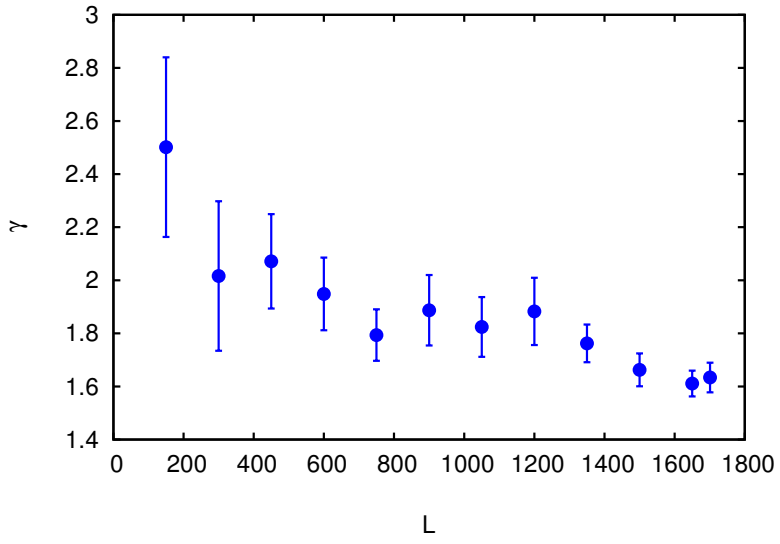
In the previous chapter we have laid the methodological foundations to reliably detect intra-subtype reassortment; here we apply our new method to a set of seasonal human influenza strains representing the evolution of the virus in a 48 years time span, from 1968 to 2015. We infer a comprehensive map of intra-lineage reassortment between the surface proteins HA and NA of influenza A/H3N2, and provide evidence that most of these events are under negative selection increasing with distance between parental strains.

3.1 Rate and genealogy of reassortment for influenza A/H3N2

In each two-segment tree built as in section 2.2.1, we map HA-NA reassortments as detailed in chapter 2: we first identify candidate events (p, p', r) by the criterion (2.1) and we notice that a large fraction of these putative events has small core distance δ. To exclude false positives we compare the distributions of these distances with the null case described in section 2.2.3. As anticipated in equation (2.2), the expected number $n_0(\delta)$ of false positive reassortment events decays exponentially as a function of the core distance (Fig. 3.1(a)), with $\gamma = 1.6 \pm 0.1$. Furthermore, the decay exponent γ does not depend critically on the length L of the segment (Fig. 3.1(b)).



(a)



(b)

Figure 3.1: Distance dependence of spurious reassortment counts in non reassorting sequences. (a) Histograms of the number of events found in a HA tree as a function of δ, for sequences of total length L. Error bars represent the standard deviation obtained from 5 different random choices of the sites for each δ. (b) The decay exponent γ is shown as a function of L (cf. equation (2.2)). The inferred values are stable for large L, allowing extrapolation to $L = L_{\mathrm{HA}} + L_{\mathrm{NA}}$.

In Fig. 3.2 we compare the number of real events as a function of δ with the null case, the latter being clearly under-represented at larger distances. Table 3.1 shows the calculated expected number of false positive reassortment events. Even if we assume that most of the counts at $\delta = 1$ are false positives (which sets the value of the constant C in equation (2.2)), the expected total number of false positive events with $\delta \geq 5$ drops below 1 (Fig. 3.2 and table 3.1). Based on these analysis, we keep only events with core distance $\delta \geq 5$.

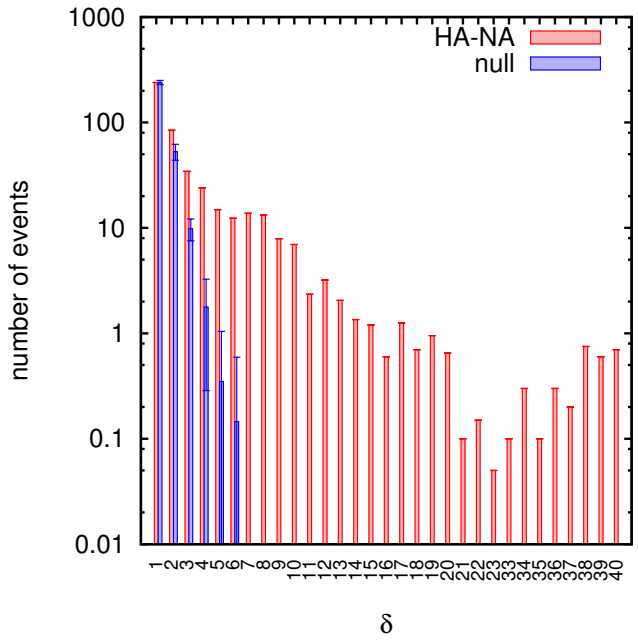


Figure 3.2: Fidelity of reassortment inference. Histograms of reported HA-NA reassortment events for different core distances δ (red bars) are compared to expected number of false positives due to ambiguities in tree reconstruction, $n_0(\delta)$, from a null model of non-reassorting sequences (blue bars; error bars reflect the statistics over different realizations of the null model). The function $n_0(\delta)$ decreases exponentially with increasing δ (cf. equation (2.2) in chapter 2); the overall amplitude is set by the conservative assumption that all counts at $\delta = 1$ are false positives. The resulting total number of false positives with $\delta \geq 5$ is below 1.

We then prune events with strongly overlapping core sets $\mathcal{A}_{pp'}$, and eliminate double-counting of events. We also note that the total number of reassortments does not depend on the number of inferred trees, indicating that our mapping exhausts the events occurring in the original dataset. Last, we verify that passaging mutations do not confound our inference (see section 2.2.1). We observe the same patterns of mutations which characterize the reassortment events reported above, as well as a very similar distance dependence of

δ	false positives
1	240
2	53
3	10
4	2
5	0.3
6	0.15

Table 3.1: Number of expected false positive reassortment counts as a function of δ (cf. Fig. 3.2)

the false positives counts (Fig. 3.3).

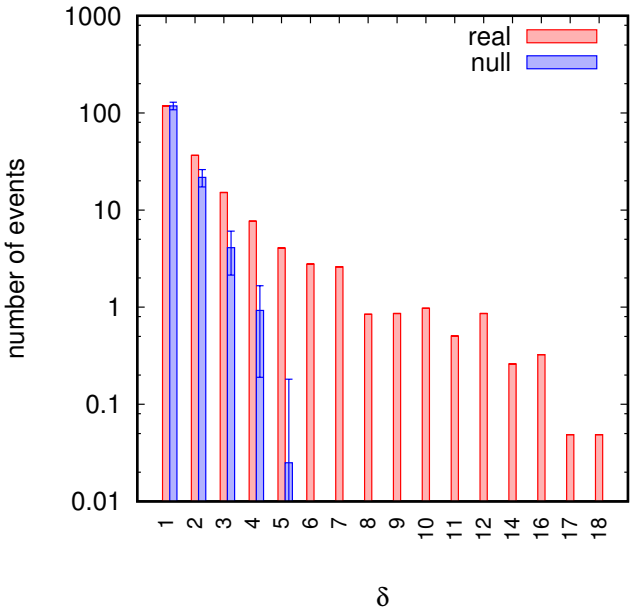


Figure 3.3: Reassortment inference between unpassaged sequences. Histograms of reported HA-NA reassortment events between unpassaged sequences for different core distances δ (red bars) are compared to expected number of false positives (blue bars), which decays exponentially with increasing δ. This result is qualitatively comparable with the distance dependence of real events and false positives that we find if we include in the analyses also strains subjected to passaging (cf. Fig. 3.2).

This procedure produces a list of 103 reliable and independent HA-NA reassortments in our data set. These events have a mean genetic distance (defined in section 2.2.2) $d \geq 3$ between reassortant and parent strains in both segments and an average $d_{\text{ave}} = 10$,

which sets them clearly apart from individual point mutations and provides the genetic basis for potentially strong and epistatic selection (see below). In table A.1, we report the genetic distance d, as well as representative strains from the reassortant clade C_r and the parent clades C_p, $C_{p'}$. From the year 2000 on, we find an average of 6 unique reassortment events per year, which is in overall agreement with other studies [2–4]. Furthermore, events detected by the majority of these studies are well represented in our list (starred events in table A.1, see Fig. 3.4 for the representation on the tree of one of these events); these include reassortments between New York strains isolated between 2000 and 2005 [55, 86, 90, 92] (see appendix A for more details). The clean statistical test used here, however, addresses the over-counting of events in a more objective way.

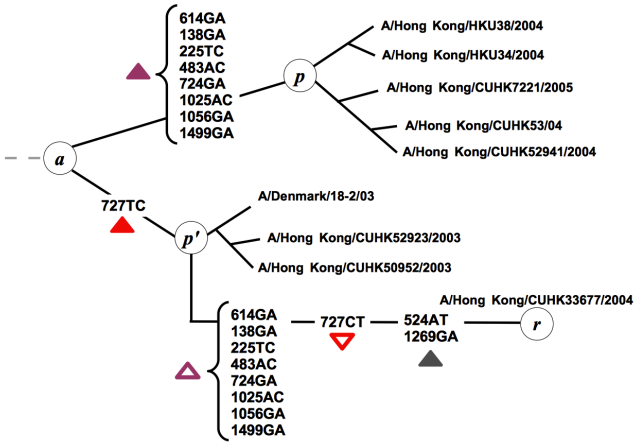


Figure 3.4: Representation of a real event in the joint genealogy. A true event (nr 1 in table A.1) detected by our algorithm on the joint HA-NA tree. Each mutation on HA segment is labeled with a number between 1 and 1701 that indicates the site. The pattern of repeated and reversed mutations (filled and empty triangles) follows the scheme in Fig. 2.3: the reassortant strain A/Hong Kong/CUHK33677/2004 is generated by an event with $\delta = 9$ between p and p' clades.

Fig. 3.5 shows the inferred reassortments since 2000 mapped on a joint HA-NA tree. These events cover the entire time interval of the tree with a slight increase in frequency in recent years, which is likely due to increased depth of the tree. In all cases, the parent strains were collected at close times, which is consistent with the fast evolutionary speed and the resulting short sojourn periods of specific genotypes in the population of circulating strains. By comparing the number of reassortment events with the number of synonymous nucleotide changes on the same tree, we estimate that reassortant variants get established at a rate of order 10^{-2} in units of the neutral point mutation rate. This establishment

rate refers to observed variants in a strain sample, which clearly depends on the sampling depth (our data set has a detection threshold frequency of order 10^{-3}). Our finding of broad negative selection on reassortment, which is reported below, suggests that the reassortment rate of individual virions is higher, but many reassortant variants are rapidly lost in the population of circulating strains.

3.2 Reassortment is under broad negative selection

As shown in Fig. 3.5, the majority of observed reassortment events are on peripheral positions of the joint HA-NA tree. This observation is broadly consistent with a neutral or, on average, deleterious process. We now turn to measuring selection on reassortment in a more quantitative way.

3.2.1 Suppression of large distance reassortment

First, we compare the distribution of distances in detected reassortment events, $P(d)$, with the corresponding background distribution sequence distances between all pairs of strains circulating in a given influenza season, $P_0(d)$ (both distributions are defined in the regime $d \geq 3$). The latter distribution represents the background pool of a priori equiprobable opportunities for reassortment. In the absence of selection, reassortment should occur with equal probability between these pairs, regardless of their genetic distance, and the distribution $P(d)$ should be similar to the background distribution $P_0(d)$. However, Fig. 3.6(a) shows significant differences between these distributions: there are far fewer actual events with larger values of d than in the background distribution. We measure the statistical significance of these differences by the Kullback-Leibler (KL) divergence $D_{KL} = \sum_{d\geq 3} P_0(d) \log(P_0(d)/P(d))$ and by the Kolmogorov-Smirnov statistics, $D_{KS} = \max |F(d) - F_0(d)|$, where $F(d)$ and $F_0(d)$ are the corresponding cumulative distributions. We find the suppression of high-d reassortment events to be significant by both tests, with $D_{\mathrm{KL}} = 0.56$ (compared to a 5% error threshold at $D_{\mathrm{KL}} = 0.1$) and $D_{\mathrm{KS}} = 0.34$ (giving a probability $p < 10^{-23}$ to find a larger distance by chance). As shown by Fig. 3.6(b), the ratio of reassortment to background counts with $d \geq d_{\min}$ decreases as a function of the lower cutoff $d_{\min}$. We attribute this effect to distance dependent deviations from neutrality: negative selection on reassortment increases in strength with distance d. The same analysis based on amino acid distances, which provide a more coarse-grained measure of genetic differences, also shows a significant suppression of large-d reassortment

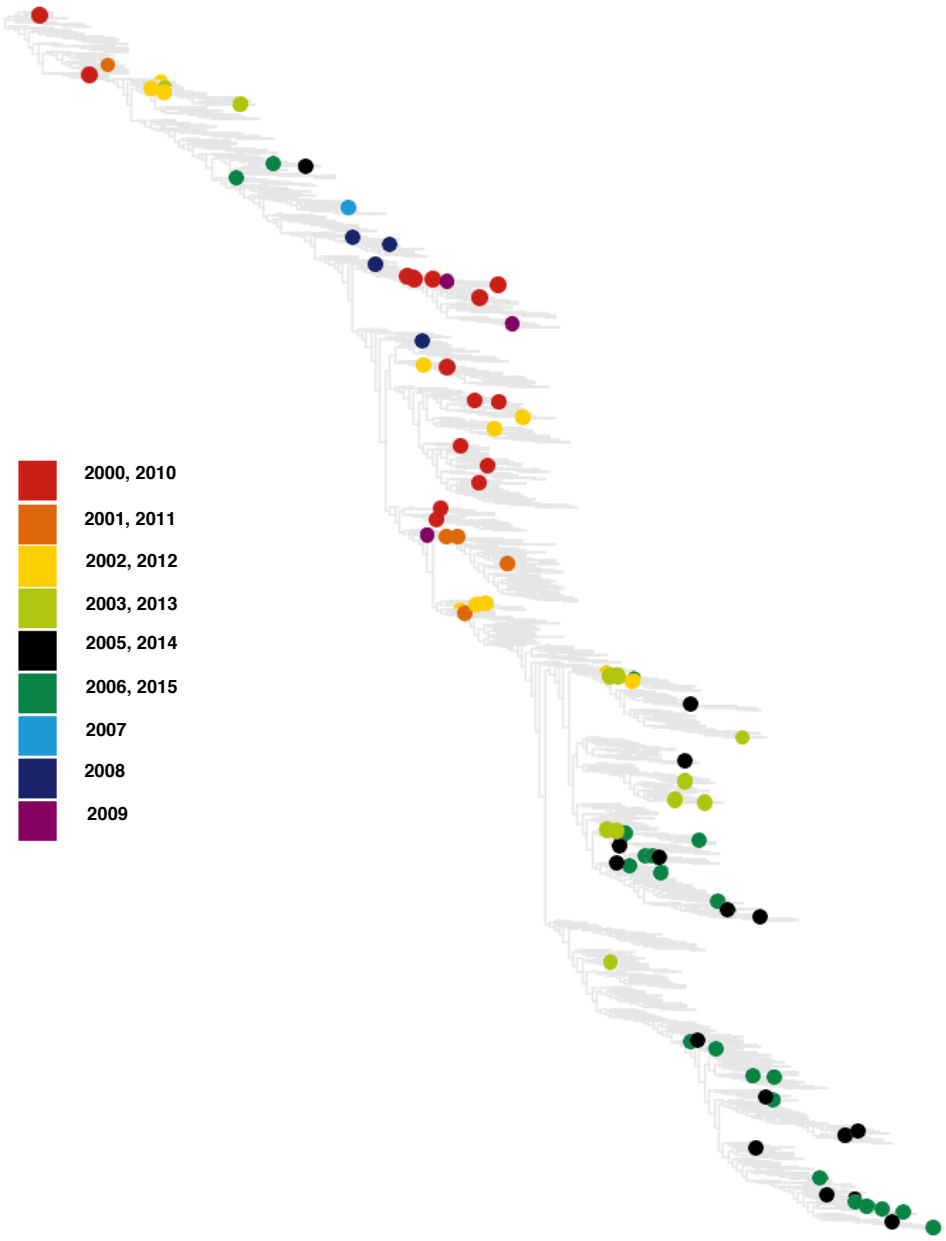


Figure 3.5: Reassortment of HA and NA in human influenza A/H3N2 from 2000 to 2015. The 95 inferred events are mapped on a joint HA-NA tree. The reassortant strain r of each event is represented by a filled circle (color-coded by year of occurrence). The events are homogeneously distributed over the tree and the reassortant clades are predominantly at peripheral positions of the tree.

events (Fig. 3.7).

A potential confounding factor for this analysis is the spatio-temporal population dynamics of the human influenza virus [54, 100]. Although influenza strains are known to

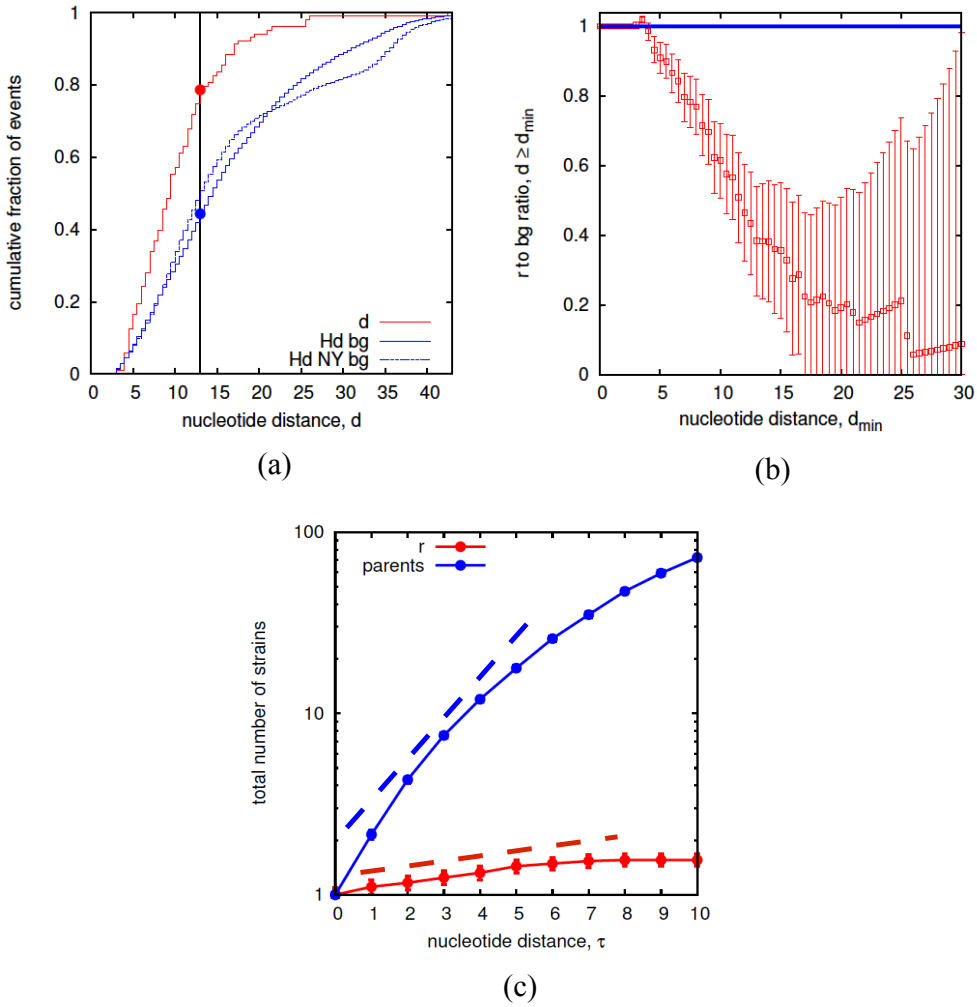


Figure 3.6: Negative selection on reassortment. (a) The cumulative distribution of mean nucleotide distances d between reassortant and parent strains for the HA-NA reassortments in influenza A/H3N2 (red line) is compared to the corresponding distribution of distances for co-circulating strains in the same influenza season (solid blue line) and from the New York area only [86] (dashed blue line). (b) The ratio of reassortment counts to background counts in the interval $d \geq d_{\min}$ (red circles) decreases with increasing lower threshold $d_{\min}$ and drops significantly below 1 (blue line). The suppression of reassortment at larger values of d signals distance-dependent negative selection. Bars show statistical errors due to the finite number of inferred reassortments. (c) The average number of strains in the reassortant clades with nucleotide distance $\leq \tau$ from the focal node, $\langle N_r \rangle(\tau)$ (red line), is compared to the corresponding average number of strains in the parent clades, $\langle N_0 \rangle(\tau)$. For $\tau \lesssim 6$, both functions increase with τ in an approximately exponential way; we estimate growth rates $f_r \approx 0.07$ and $f_0 \approx 0.5$, respectively (dashed lines; cf. equation (3.1)). The growth rate difference $\bar{s} \equiv f_0 - f_r \approx 0.4$ measures the average fitness cost of reassortment. Bars represent statistical errors due to the finite number of counts (not shown when these errors are smaller than the dot size). See Fig. 3.7 for an analogous inference based on amino acid distances.

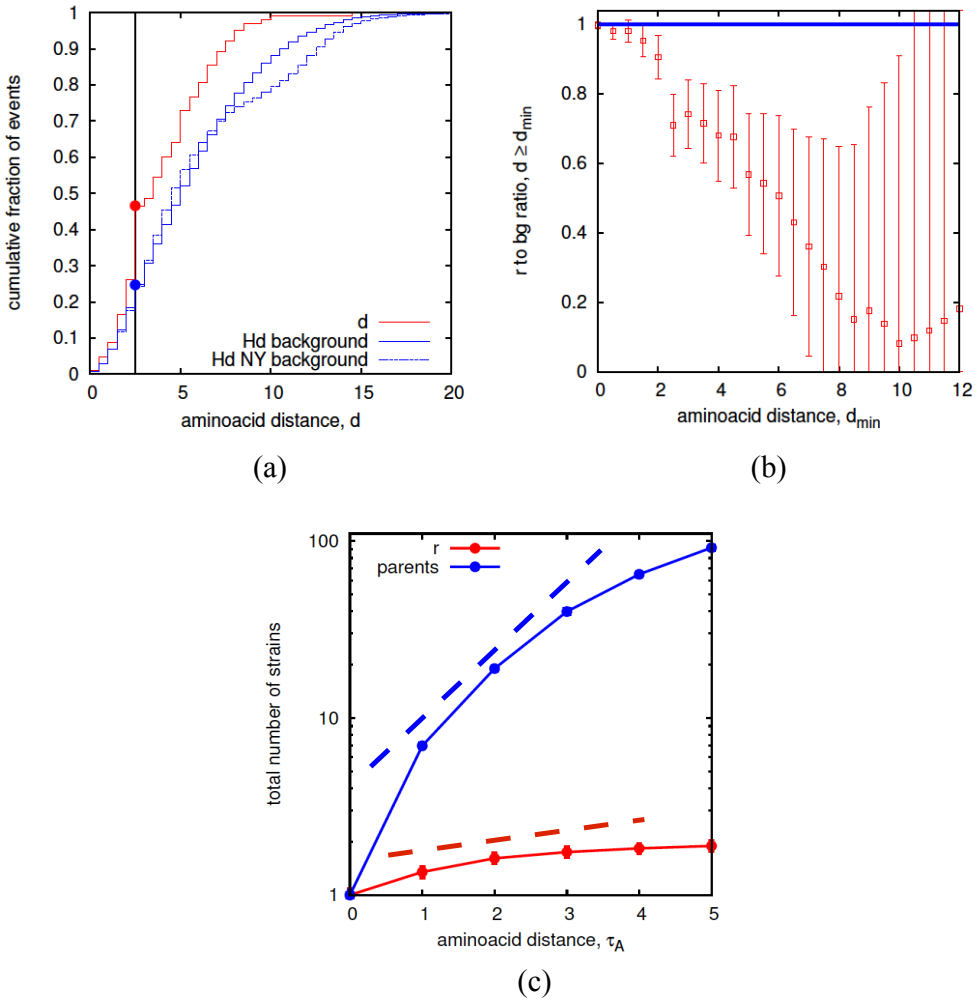


Figure 3.7: Selection inference based on aminoacid distances. (a) Comparison between cumulative distributions of mean amino acid distances d in real (red) and null (blue) case. (b) Decrease of the ratio of reassortment counts to background counts in the interval $d \geq d_{\min}$, in amino acid units. See Fig. 3.6 for the same analysis using nucleotide distances. (c) The average number of strains in the reassortant clades with aminoacid distance $\leq \tau_A$ from the focal node, $\langle N_r \rangle(\tau_A)$ (red line), is compared to the corresponding average number of strains in the parent clades, $\langle N_0 \rangle(\tau_A)$. For $\tau_A \lesssim 4$, both functions increase with τ_A in an approximately exponential way; we estimate growth rates $f_r(A) \approx 0.2$ and $f_0(A) \approx 0.7$, respectively (dashed lines; cf. equation (3.1)). The growth rate difference $\bar{s}_A \equiv f_0(A) - f_r(A) \approx 0.5$ inferred from distances in aminoacid units is similar to $\bar{s} \approx 0.4$ for nucleotide distances (Fig. 3.6).

travel rapidly, the local background distribution P_0 of genetic distances, which matters for reassortment, can in principle differ from its global counterpart used in our significance analysis. In order to estimate this effect, we restrict the background distribution to strains that circulate in the same region; specifically, we calculate an alternative distribution P_0

using isolates from New York State only, which are available from a previous study focusing on that region [86] (as before, Hamming distances are computed only between strains reported in the same influenza season). We still find that the actual reassortment events differ significantly from the local distribution P_0 (Fig. 3.6(a) and 3.7(a)), while the global and the local background distributions are statistically indistinguishable. We conclude that the suppression of large-d reassortment is not a spurious demographic effect, but is indicative of selection.

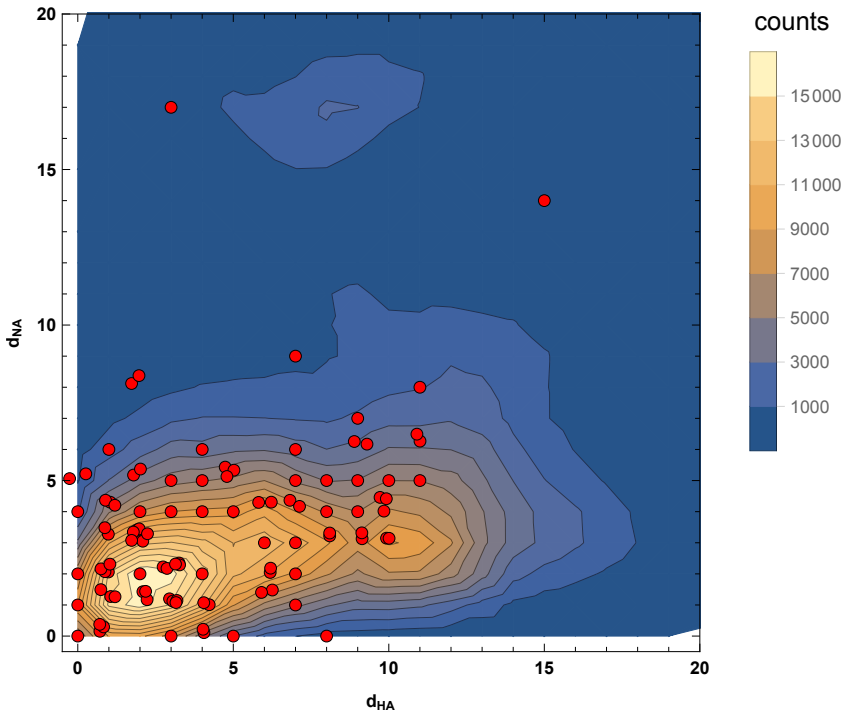
To gain some insight on how reassortment constraint is distributed on the amino acid distances in individual segments, d_{HA} and d_{NA}, we evaluate the joint background distribution $P_0(d_{\mathrm{HA}}, d_{\mathrm{NA}})$ and compare it with the amino acid distance pairs $(d_{\mathrm{HA}}, d_{\mathrm{NA}})$ of the inferred reassortment events (Fig. 3.8(a)). The joint statistics of $(d_{\mathrm{HA}}, d_{\mathrm{NA}})$ differs in the coordinates $d_{\mathrm{HA}} + d_{\mathrm{NA}}$ and $d_{\mathrm{HA}} - d_{\mathrm{NA}}$, indicating that selection is not a function of d only. In particular, the conditional distributions $P(d_{\mathrm{HA}} - d_{\mathrm{NA}}|d)$ differ between data and background (Fig. 3.8(b)), which is consistent with the expectation that reassortants similar in one protein are less selected against, even if the distance in the other protein is larger.

3.2.2 Fitness cost of reassortment and reduced tree growth

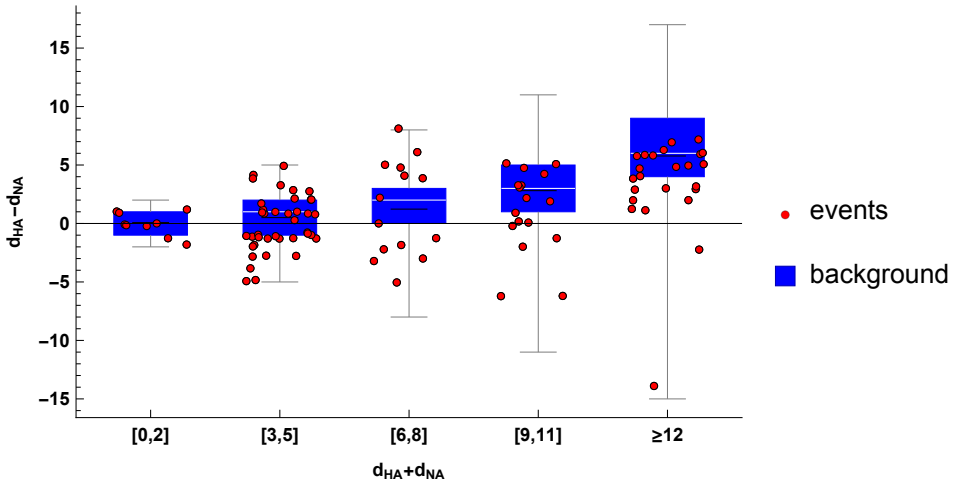
The reduced number of reassortment events with larger core distances is consistent with purifying selection that prevents some reassorted strains from reaching population frequencies detectable in our strain sample. But can we map negative selection on reassortment directly within the set of observed events? To address this question, we compare the evolution of population sizes of reassortant clades and of parent clades as a function of their age τ. We evaluate, for each reassortment event, the number $N_r(\tau)$ of strains in the reassortant clade C_r with nucleotide distance $\leq \tau$ from the focal node r, together with the mean of the corresponding numbers of strains in the parent clades, $N_0(\tau) = (N_p(\tau) + N_{p'}(\tau))/2$. We obtain these functions in the joint HA-NA tree of the full data set, counting strains with the same sequence only once. Averaging over the set of reassortment events, we can measure the expected growth of reassortant and parent clades,

$$\langle N_r \rangle(\tau) \sim \exp(f_r \tau), \qquad \langle N_0 \rangle(\tau) \sim \exp(f_0 \tau); \tag{3.1}$$

similar inference methods for clade growth are discussed in refs. [77, 78]. The functions $\langle N_r \rangle(\tau)$ and $\langle N_0 \rangle(\tau)$ for influenza A/H3N2 indeed show approximately exponential growth in the regime $\tau \lesssim 6$, which corresponds to time intervals of order one year (Fig. 3.6). The



(a)



(b)

Figure 3.8: Background distribution and reassortment events as a function of the amino acid distances d_{HA} and d_{NA} between strains. (a) The background distribution $P_0^{aa}(d_{\mathrm{HA}}, d_{\mathrm{NA}})$ (contour plot) is compared to reassortment counts (red dots). Events at the same coordinates are plotted with a small random deviation in x and y directions, in order to avoid overlapping. (b) Conditional background distributions $P_0^{aa}(d_{\mathrm{HA}} - d_{\mathrm{NA}} | d_{\mathrm{HA}} + d_{\mathrm{NA}})$ (whisker plots) are compared to reassortment counts (red dots). Whisker plots show the 0.25 quantile to the 0.75 quantile of the distribution (blue boxes); the white horizontal line represents the median, vertical bars span the dataset excluding outliers. The width of the bins is chosen to ensure a statistically relevant number of events for each bin. The reassortment data appear more spread in the coordinate $d_{\mathrm{HA}} - d_{\mathrm{NA}}$ compared to the background (red points are mainly placed outside or at the border of the blue boxes).

fitted growth rate difference estimates the average fitness cost of reassortment in our set of events,

$$\bar{s} \equiv f_0 - f_r \approx 0.4 \tag{3.2}$$

in units of the total point mutation rate in both segments, which equals approximately 5×10^{-3} per day. The same analysis performed with aminoacid distances is reported in Fig. 3.7. Statistical rank tests (Wilcoxon-Mann-Whitney) confirm that there is a significant asymmetry between the sizes of parental and reassortant clades, the latter being on average smaller ($p - value = 1.5 \times 10^{-11}$).

3.2.3 Epistasis across proteins

To further interpret the observed fitness cost of reassortment, we consider the simplest epistatic fitness model for combined (HA, NA) genotypes, $F_{\alpha\beta} \equiv f(\mathrm{HA}_\alpha, \mathrm{NA}_\beta)$, where $\alpha, \beta = +1$ denote the alleles of the parent strain p and $\alpha, \beta = -1$ the alleles of parent strain p'. The model takes the form

$$F_{\alpha\beta} = f_\alpha^{\mathrm{HA}} + f_\beta^{\mathrm{NA}} + \frac{\omega}{2}\alpha\beta, \tag{3.3}$$

where f^{HA} and f^{NA} denote single-protein fitness values and ω is the strength of cross-protein epistasis. In terms of this model, the mean fitness cost of a reassortant strain compared to its parent strains is

$$s = \frac{1}{2}(f_p + f_{p'}) - f_r = -\frac{1}{2}(f_+^{\mathrm{HA}} - f_-^{\mathrm{HA}}) - \frac{1}{2}(f_-^{\mathrm{NA}} - f_+^{\mathrm{NA}}) + \omega, \tag{3.4}$$

where we assume, without loss of generality, that the reassortant strain r inherits HA from parent p and NA from parent p'. If co-infection randomly mixes co-circulating strains, the single-protein fitness value of a reassortant strain is, on average, equal to the mean fitness of its parents. For strains observed in a sequence sample, this value can only be biased towards larger reassortant fitness; i.e., $\frac{1}{2}\langle f_+^{\mathrm{HA}} - f_-^{\mathrm{HA}}\rangle \geq 0$ and $\frac{1}{2}\langle f_-^{\mathrm{NA}} - f_+^{\mathrm{NA}}\rangle \geq 0$. Hence, the observed fitness cost (3.2) implies an average epistatic cost of reassortment, $\langle\omega\rangle > \bar{s} > 0$. Cross-protein epistasis in the observed reassortment events is of moderate strength but broadly distributed: reassortant variants are fit enough to reach population frequencies detectable in our sample, but they are, on average, less fit than their non-reassortant counterparts.

3.3 Discussion and remarks

In chapter 2 we have developed a new method to map reassortment of genomic segments in an evolving viral population. We detect reassortment events based on their trace in the genealogy of the population. On a two-segment genealogical tree, a set of *core mutations* in one of the reassorted proteins appears twice: on the branches linking the reassorting (parent) strains and, in reverse direction, on the branch to the reassortant clade (Fig. 2.3). We are interested predominantly in reassortment events above a certain minimum genetic distance d from their parent strains (here at least 3 mutations), which are clearly set apart from the dynamics of point mutations. This is also the regime in which our method allows a reliable identification of events, which is not confounded by ambiguities in tree reconstruction (Fig. 3.2).

In this chapter we have presented an application of our method to a large set of seasonal influenza sequences and we have derived a major biological result: reassortment within human influenza A/H3N2 is under broad, distance-dependent negative selection. Specifically, there are fewer large-d reassortments in our sample than expected from the distribution of co-circulating strains, and reassortant strains have fewer descendants than their parent strains (Fig. 3.6 and 3.7). These observations probe negative selection on reassortant genotypes at different scales of frequency and sojourn time in the population. The suppression of large-d reassortment signals purifying selection that prevents some reassortant variants from reaching sufficient frequencies to appear in our strain sample; the growth rate difference between reassortant and parent clades indicates moderate negative selection on the variants that do appear in the sample. Reassortment between very close sequences may well be approximately neutral, but sequence-based inference methods cannot distinguish such events reliably from point mutations. The inferred selective effects characterize the continuous evolution of a seasonal influenza lineage; they do not exclude rare large-effect reassortment events causing antigenic shifts and seeding new lineages. We stress that our results are based on statistical methods evaluating ensembles of inferred reassortments and background distributions. Any such method is subject to possible confounding factors and biases; for example, reassortment can be expected to occur preferentially in high-infection settings at the peak of seasonal epidemics. However, the consistent outcome of two distinct inference procedures gives a credible signal of selection acting on reassortment.

Reassortment can be seen as a natural “experiment” that continuously produces new

combinations of viral proteins and probes their fitness in a fast-evolving population. Hence, the statistics of reassortment is informative of key selective forces governing the evolutionary dynamics. Specifically, our result of negative selection on reassortment signals ubiquitous fitness interactions (epistasis) between viral proteins; that is, the fitness of alleles of one protein depends on the genetic background of the other proteins in the same virion. The mutation-selection dynamics in non-reassortant sublineages produces favorable combinations of protein alleles, and reassortment introduces a fitness cost by randomizing these combinations. Importantly, this cost arises between genetic variants from co-circulating strains in a given viral lineage; these variants are individually viable, differ by just a few mutations (of order 20 nucleotide changes in both reassorted segments together), and have a recent common ancestor (typically dating just one or two years back). This implies that new favorable protein combinations are continuously produced and selected for, while many random combinations incur a fitness cost. In other words, cross-protein epistasis constrains the adaptive evolutionary path of a given influenza lineage. This result could be tested, for example, by combining reassortment experiments ([58–64]) with in-vitro competitive fitness assays. Reassortant strains with substantial distances d should frequently be outcompeted by any of their parent strains.

4. Local heterogeneity in human influenza trees

4.1 Tree based measures of population dynamics and evolution

As already discussed above, the need and chance of gaining predictive power on future evolutionary changes push the development of new methods and models in evolutionary biology. The high mutation rates typical of some systems such as virus and bacteria make them optimal models for studying how evolution itself works. We can collect data at different times through the evolutionary history of these organisms and build their genealogical trees. From the trees, then, we can learn how the present populations are related and linked to their ancestors, including the extinct variants sampled in the past. The opportunity of extracting information concerning evolutionary patterns from the shape of the trees alone has become clear in the last few years [101]. Initial studies have highlighted that global statistics calculated on the phylogenies, such as the density of bifurcation events, can be informative to infer absolute exponential growth rates of populations ([102, 103]).

The extension of these principles to the study of different clades within one genealogy has lead to the development of new methods that use branching patterns to infer growth rate differences between clades [78] (see chapter 1). Neher et al. recognized that the relative fitness of sampled sequences can be linked to the shape of the tree by local statistics: in a neighborhood of a given internal node, high fitness translates into high number of offspring, and therefore corresponds to a large number of branches in the close surroundings. The so called "local branching index" (LBI), defined as the length of the tree in the surrounding neighborhood of a given node, is used to estimate growth rates. Although temporal and geographical inhomogeneous sampling may introduce biases in the

evaluation of clade growth, these kind of approaches do not require an explicit modeling of the viral fitness and can serve as general methods to predict the evolutionary trajectories of asexual populations. The concept of identifying hot spots on phylogenetic trees to estimate fitness without using any molecular information can be further generalized to measure different evolutionary parameters.

Following this intuition, we adopt the idea of ranking internal nodes of phylogenetic trees with a heuristic algorithm and apply it for a different scope: instead of relating the local properties of the genealogy to fitness, we introduce a simple parameter linked to sequence diversity. We build one-segment trees and define the neighborhood of each internal node as the set of its descendants which are homogeneous in terms of average Hamming distance between sequences in that segment (s_1, haemagglutinin in the following sections), namely strains that have similar s_1 sequences. The heterogeneity of the strains of this neighborhood calculated in a given second segment (s_2, neuraminidase) indicates that s_1 is coupled with genetically distinct s_2 versions. which correlates with reassortment. In regions of the tree where no reassortment has occurred, strains in the same neighborhood should be related to the same ancestor (or to ancestors placed back in similar times) in both segments and new mutations should accumulate at proportional rates, determining homogeneous neighborhoods in s_1 and s_2. Under the assumption that simultaneous multiple reassortments are rare events, the rank of the internal nodes automatically gives information on which are the reassortant sequences: strains which are part of a neighborhood with high heterogeneity in s_2 and, at the same time, have high Hamming distance (in the same space) to the focal node[1] are produced by reassortment. Similarly to the clade size analysis described in chapter 3, we can then evaluate if the reassortant variants grow faster or slower with respect to the non-reassortant counterparts, by counting how many leaves are flagged as reassortant or non-reassortant.

Both the methods described in this chapter and in chapter 2 rely on reconstructed trees, but with a substantial difference. We have already mentioned that joint genealogical trees should be thought as "fictitious" representations of the evolution of paired sequences. They can indeed be considered as phylogenetic trees only in the limit of no reassortment, in which both segments evolve together as a single filament. Anomalies in the topology of such trees, embodied by specific patterns of mutations and back mutations not detectable in single-segment trees, signal potential gene reassortment. This approach is valid as long as the rate at which mixing occurs is not excessively high, so that reas-

[1]The internal node associated to the neighborhood in question.

sortment can be treated as a perturbation in the otherwise monophyletic structure of the evolutionary tree. The single-segment trees inferred in this chapter, on the other hand, are "real" phylogenies describing parent-child links under asexual evolution, since there is no recombination within single influenza segments [10]. Hence this second method has the potential to be trustfully applied to systems for which no prior knowledge concerning the amount of reassortment is available. Interestingly, it can be easily extended to the study of extra-species reassortment, which occurs at large distances and for which the joint-tree method is not optimized. Last, the exponential increase of available data collected in the last few years as a result of massive surveillance initiatives necessitates the development of bioinformatical methods that can handle such a large amount of information quickly. The method that we describe in this chapter results much faster than the joint-tree method, since the computational time grows linearly with the number of nodes in the tree, instead of quadratically.

4.2 Cluster index and inference of reassortment

Our second phylogenetic method for inferring reassortment in influenza is based on the approach described above. We map reassortment between two segments s_1 and s_2 in a single-segment phylogenetic tree, using the tree itself and the alignment of sequences of the other segment. In order to map reassortment between haemagglutinin and neuraminidase, we create two separate alignments from observed RNA sequences of HA and NA (as described in section 2.2.1) and we reconstruct the Maximum Likelihood phylogenetic tree relative to HA only. Each internal node of the phylogeny is then assigned a cluster index linked to reassortment on the base of a simple assumption: in absence of genes mixing, the degree of diversity of each clade in HA should be similar to the one observed in the correspondent NA segments. Violations of this similarity, such as high diversity in NA sequences as opposed to homogeneity in HA segments, is considered as a signal of reassortment. The same procedure can be equivalently applied to an NA tree.

4.2.1 Algorithm in steps

In detail, we first define the cluster index c of internal nodes from the heterogeneity of external daughters at Hamming distance τ, then we classify both internal and external nodes into ancestral or reassortant strains.

First module: identification of clustered nodes

For each internal node i:

1. select the leaves l in the subtree of i with $\mathrm{Hd}(l,i) \leq \tau$ in HA sequence space[2]
2. fill in the neighborhood (Fig. 4.1) of $i : U_i = \{l : \mathrm{Hd}(l,i) \leq \tau \wedge f(l) = i\}$[3]
3. be $N_i = |U_i|$ the number of leaves in the neighborhood of i, calculate the cluster index based on NA distances (see below). Assign:
   - $c(i) = 0$ if $N_i = 0$
   - $c(i) = 1$ if there is one cluster C_i
   - $c(i) = 2$ if there are two (unsorted) clusters (C_i, C_i').

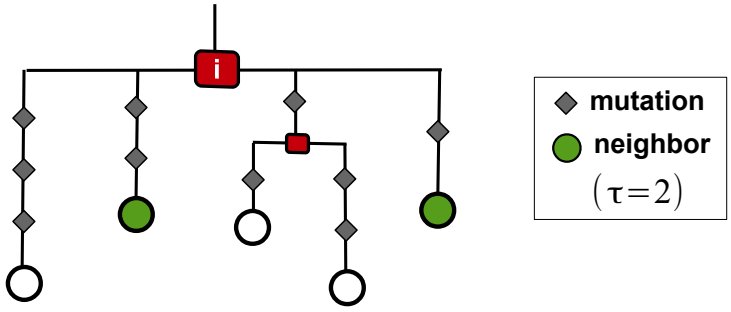


Figure 4.1: Definition of neighborhoods in HA trees. Each internal node i is associated with a neighborhood filled with the leaves (green dots) in its subtree at Hamming distance $\leq \tau$ from the focal node. The additional requirement that only direct descendants of i are included in the neighborhood prevents that the same observed strain is associated with more than one internal node (red rectangles).

If $c(i) = 2$, one group of strains in the neighborhood if i carries the NA variant inherited directly from the parent strain on the HA tree, while the second group carries a new variant, imported from a different clade. A cluster index larger than one is the primary signal for reassortment. The partition of the neighborhoods into clusters requires

[2]The abbreviation "Hd" stands for "Hamming distance", here and in the following sections.

[3]Here f is the "true" father of l, i.e. the first ancestor of l that differs from its direct father in at least one site. These constraints ensure that each leaf, if part of a neighborhood, is uniquely assigned to one internal node.

the definition of the distance between a leaf and a cluster. Given a sequence t and a cluster C, with $l_c \in C$ and $|C| = N_c$,

$$d(t, C) := \sum_{\{l_c\}} \frac{\mathrm{Hd}(t, l_c)}{N_c} \quad . \tag{4.1}$$

The neighborhoods U_i of internal nodes get clustered with an algorithm similar to k-means methods (here $k = 2$), depending on the input parameter $dmax$. At each step the sequence with the minimum distance to one of the clusters is assigned to that cluster:

1. for each j,k in U_i calculate the Hamming distance $\mathrm{Hd}(j, k)$ in the NA sequence space
2. $c(i) = 2$ if $\max(\mathrm{Hd}(j, k)) \geq dmax$, $c(i) = 1$ otherwise.
3. if $c(i) = 2$ define $(j^*, k^*) = \arg\max(\mathrm{Hd}(j, k))$ and assign $j^* \to C_i$, $k^* \to C_i'$
   - $t \in U_i \setminus (C_i \cup C_i')$
   - calculate $(t^*, C^*) = \arg\min(d(t, C_i), d(t, C_i'))$
   - assign $t^* \to C^*$
   - repeat until every $t \in U_i$ is assigned to either C_i or C_i'

If we set the parameter τ to a small value and apply the algorithm above to all the internal nodes in the HA tree, we obtain a list of "polymorphic" (also called "clustered") nodes with $c = 2$, i.e. clades homogeneous in HA space (the diameter of the neighborhood can be at most $2 \times \tau$) and heterogeneous regarding NA. The term "polymorphic" here is used to highlight the presence of distinct clusters in NA space for a specific group of similar HA segments, rather than indicating the existence of more alleles at a given locus. Nodes with $c = 1$, on the other hand, are referred to as "monomorphic". In this first module we can therefore recognize an information flux flowing from the external strains up to the internal nodes, which get classified as clustered or non-clustered accordingly. The opportunity of starting a clustering procedure in the NA space is determined by the parameter $dmax$, which sets the inferior limit for the maximum distance between two sequences in the neighborhood of a clustered node. The tuning of both τ and $dmax$ will be discussed in the sections below (4.2.2).

Second module: classification of reassortant and ancestral strains

The second part of the algorithm ranks the strains as reassortant or ancestral. Complementary to module 1, this second algorithmic unit implements a stream of information

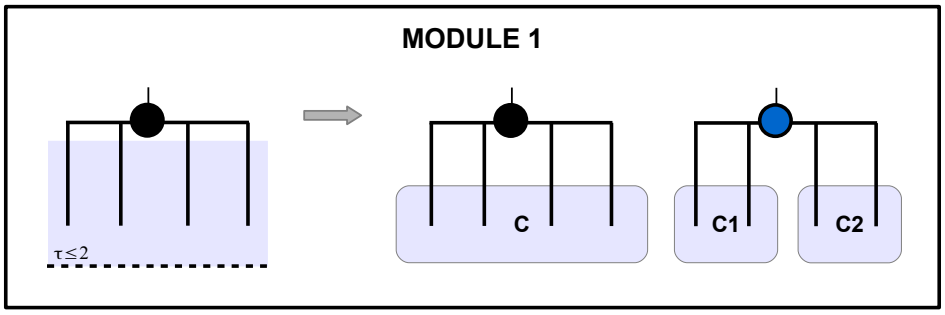


Figure 4.2: Cluster index assignment. The first module of the algorithm implements "upward" message passing, from descendant nodes to the parents: internal nodes (dots) are classified as polymorphic (clustered) or monomorphic on the base of the clustering status of their neighborhood (shaded areas). The two alternative outcomes are shown on the right-hand side of the grey arrow. Neighborhoods of polymorphic nodes ($c = 2$, represented by a blue dot) are grouped into two different clusters $C1$ and $C2$, while non-clustered nodes (black dot) are associated with homogeneous neighborhoods. Here the assignment operation (arrow) is performed at the upper level of parent nodes, which are classified and eventually colored.

from fathers to descendant nodes, which get classified (and eventually colored, see Fig. 4.3) into ancestral or reassortant. The output of the first module (black and blue nodes, right-hand side in Fig. 4.2) constitutes the input of this second module (lower level nodes, left-hand side of grey arrows in Fig. 4.3). Together, module 1 and 2 give an exhaustive map of classified strains, which is inferred on the base of information transmitted up- and downwards the phylogenetic tree, similar to the "message passing" technique described in [78] (see below). In detail, starting from the root and parsing the tree downwards, the clustered nodes get sorted, i.e. (C_i, C_i') are mapped as (A_i, R_i), ancestral and reassortant clusters, respectively. Let i be an internal node such that $c(i) = 2$:

1. define $j = f(i)$,[4] $f(i)$ first ancestor of i with $c(f) \neq 0$. [5]
2. if $c(j) = 1$ match (C_i, C_i') with A_j (or R_j) and sort (A_i, R_i)
3. if $c(j) = 2$ match (C_i, C_i') with (A_j, R_j) and sort (A_i, R_i)
4. iterate steps 1-3 to all the descendants of i with $c = 2$.

This procedure associates to each polymorphic internal node a list of external strains carrying the ancestral NA variant (populating cluster A) and of reassortant leaves (clus-

[4] We will use this notation throughout the whole section.
[5] NB ancestor internal nodes are already sorted, since we start parsing the tree from the root.

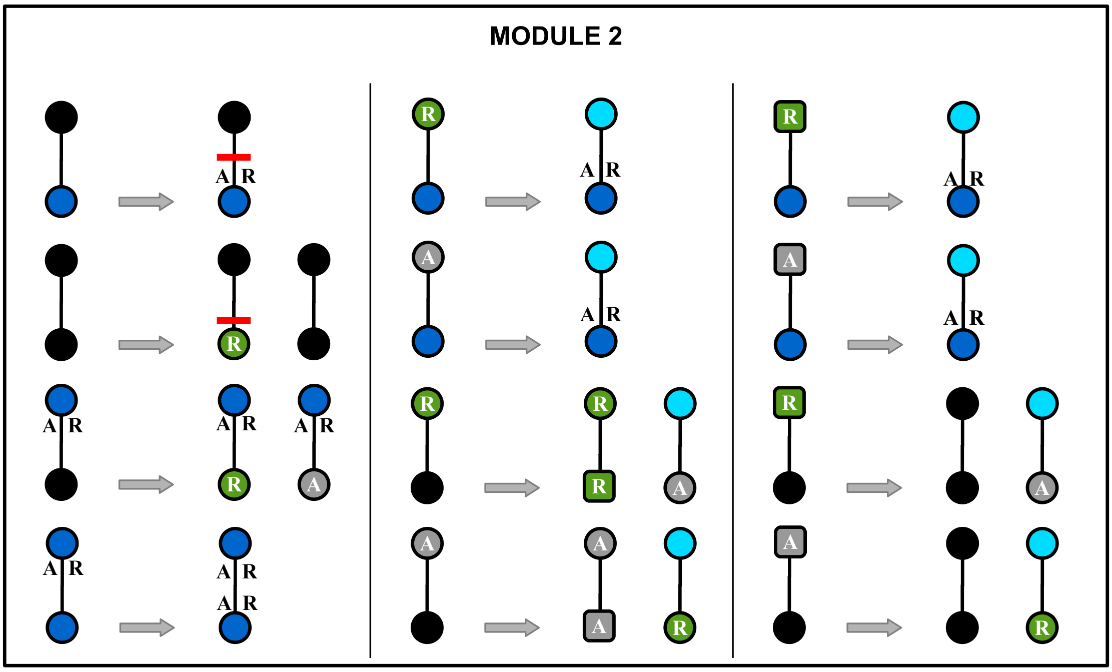


Figure 4.3: Color scheme of reassortment events on single-segment phylogenetic trees. The color assignment (represented by an arrow) of each node in the HA tree can be decomposed into single steps involving only two consecutive nodes. Each output diagram (upper and lower node connected by a vertical line) on the right-hand side of the relative arrow represents an alternative outcome of the assignment operation. Lower level input nodes (left-hand side of the arrow) are the output of module 1 (blue or black nodes, cf. Fig. 4.2). The second module represented here implements "downward" message passing, i.e. lower level nodes are classified and colored based on upper level nodes. Starting from the left panel, black parent nodes determine the beginning of a new reassortment event (red horizontal bar represents the starting point) if the child node is either polymorphic (blue node, with both A and R variants) or carries a different NA cluster (R). No assignment to the lower node is made otherwise (black-black output). Descendants of polymorphic nodes can assume three different colors: blue if polymorphic, green if monomorphic and reassortant, grey otherwise. In particular, leaves in clustered neighborhoods get green or grey color with the same criterion. In the central panel the colored nodes output from the left panel are seen as parent nodes of unassigned lower level nodes. Colored monomorphic nodes are reassigned light blue color, which shadows hidden polymorphism (see main text), if their children are either polymorphic or monomorphic carrying a different NA variant (switches). In absence of switches a one step memory is generated: squares store information on their own fathers till the following step (third panel on the right). This information gets lost if the descendants carry the same NA variant as the square, thus determining the end of the reassortment event (black-black output). On the other hand, if a switch occurs downstream a square node, the event does not stop and the square is reassigned light blue color to mimic hidden polymorphism.

tered in R). Similarly, the neighborhoods of monomorphic nodes are compared (matched) with their sorted polymorphic fathers. This basic comparison operation can be extended by performing the matching step between monomorphic i and $f(i)$ (see Fig. 4.3 for a schematic representation of the set of possible cases). If $f(i)$ directly descends from a polymorphic node, a match between its neighborhood and the one of the child strain means that one of the two variants originally present in the polymorphic ancestor has taken over in the descending clade and the ambiguity introduced by reassortment is resolved. The offspring of the lower level node, then, imports the same "label" (R or A) of its monomorphic father. If, on the other hand, the two consecutive monomorphic nodes have distant neighborhoods, the homogeneity of the subtree is only apparent, since both the reassortant and non-reassortant variants are still present in the background. More in general, any mismatch between parent-child neighborhoods of monomorphic nodes is a potential flag for "hidden polymorphism" in that region, even if the focal nodes have cluster index $c = 1$. The recognition of an actual polymorphic node in the immediate surrounding of the mismatch is expected in this case, although not required. Selection and alignment errors may influence the dynamics of "switches" in the NA variant along the tree (see the sections below for relative analysis and discussion) and a discontinuity leading from A to R cluster may represent the "starting point" of a new event (Fig. 4.3, green nodes with red bar on top[6]).

The matching steps involving polymorphic ancestors imply the comparison between clusters: the nearest cluster to the ancestral A_j (or reassortant R_j) is labeled in turn as ancestral (or reassortant). For this purpose we define the distance between two clusters $C1$ and $C2$ as follows:

$$dist(C1, C2) = \frac{\sum_{t \in C1} d(t, C2)}{N_{C1}} \tag{4.2}$$

with $d(t, C2)$ given by equation (4.1). Summarizing, the matching operation consists in checking which is the minimum distance between $dist(C1, A_j)$ and $dist(C2, A_j)$[7] and declaring the correspondent new cluster as ancestral. The second cluster will be the reassortant one. The match involving only monomorphic nodes, on the other hand, is determined by calculating the distance $dist(C_i, C_j)$ between the cluster of the parent and the child node as in equation (4.2); we declare a switch if this distance is smaller than a

[6]In Fig. 4.4 and 4.10 we use shortcuts for the beginning of an event, omitting the bar above. Green nodes remain green, while blue nodes starting an event are colored in red. In the following sections we adopt the same shortcut and refer to barred blue nodes simply as "red nodes".

[7]Or between $dist(C1, R_j)$ and $dist(C2, R_j)$ if $c(j) = 1$ and $C_j = R_j$

given threshold θ (this threshold will be discussed later in this chapter).

The final outcome of the algorithm is a list of internal and peripheral nodes, with indications about the reassortant or ancestral nature of the clade:

- internal nodes can be either polymorphic or monomorphic;
- polymorphic internal nodes are divided into primary and secondary reassortment indicators (see below and Fig. 4.3 for further details);
- monomorphic internal nodes can be reassortant, ancestral, or shading hidden polymorphisms;
- peripheral strains can be either reassortant or non-reassortant;

We finally color all the classified nodes and flag them on the HA tree, following the scheme detailed in Fig. 4.2 and 4.3.

The algorithm described above relies on the simple idea that the reassortment process leaves a trace on phylogenetic single-segment trees. Namely, the presence of clustered nodes and switches as indicators of the coexistence of viral strains with mixed HA and NA proteins constitutes evidence of reassortment. This information "travels" on the tree branches both in up and down direction: external nodes pass the message upwards to the respective focal internal nodes up to distance τ and determine if the tree is locally heterogeneous (module 1, upward message passing, Fig. 4.2); focal nodes, on the other hand, transmit the information downward to their children (module 2, downward message passing, Fig. 4.3), till one of the two variants (previously referred to as R or A) eventually takes over in the population (i.e. till at least two consecutive internal monomorphic nodes carry the same NA cluster and no further switches are observed). Fig. 4.4 shows an example of a reassortment event depicted on the actual HA tree. The first polymorphic red node is a signal of co-circulating distinct NA variants, while the persistence of the polymorphism is signaled by the nested blue node. Nodes colored in grey and green, on the other hand, represent strains (or subtrees) carrying either the ancestral or the reassortant NA variant, respectively.

4.2.2 Input parameters and preparatory steps

Definition of the neighborhoods radius: tuning τ

As introduced in the previous section, the parameter τ represents the radius of the neighborhoods in terms of Hamming distance in the HA sequence space. In order to measure

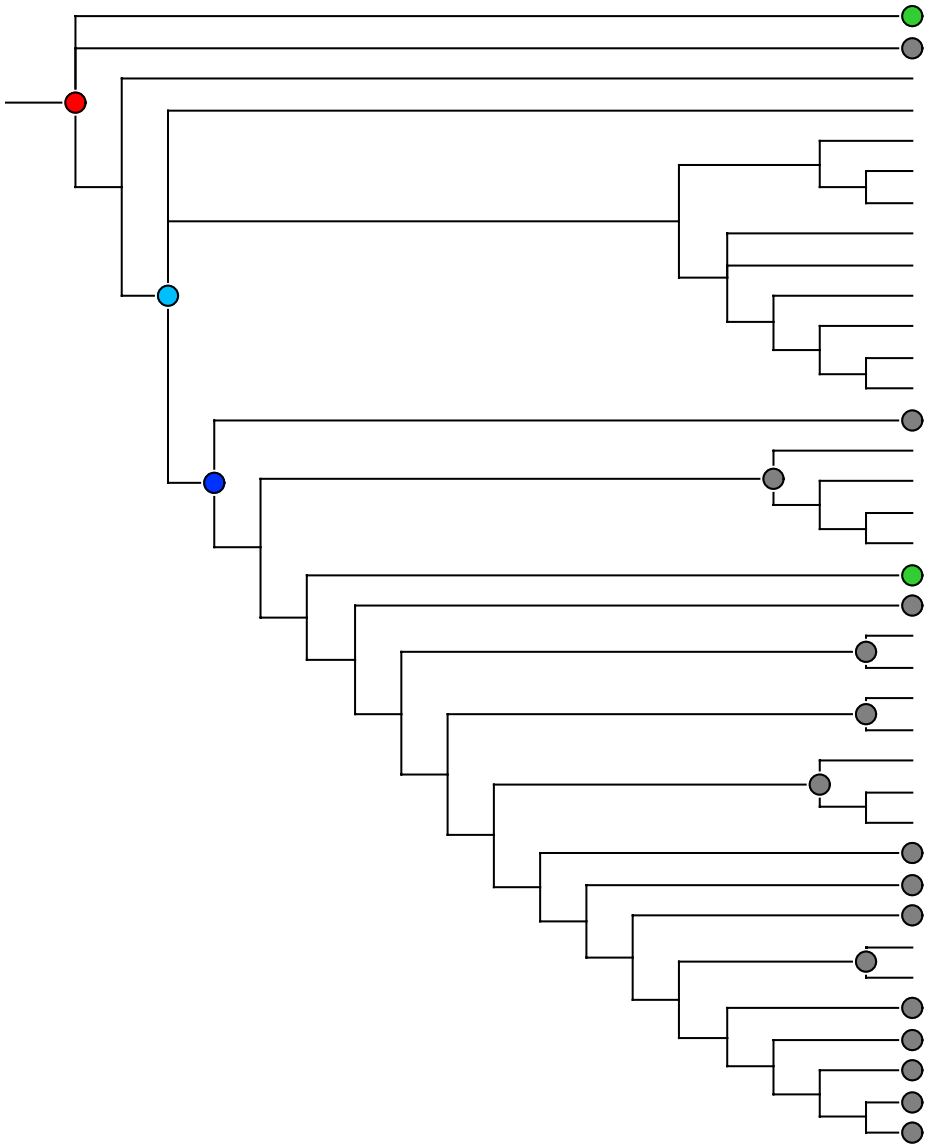


Figure 4.4: Reassortment mapped on an HA tree. Observed and reconstructed internal strains affected by reassortment are represented on the haemagglutinin tree with a colored dot. Bifurcations missing a superimposed dot may indicate internal nodes with empty neighborhood, nodes sharing the same sequence as their parent node (in both these cases they are simply skipped by the parsing algorithm) or monomorphic nodes not implying a hidden polymorphism (corresponding to black dots in Fig. 4.3). A reassortment event starts with a clustered node (red dot, referred to as "primary indicator" in the text), that represents the most ancient strain manifesting polymorphism in a given subtree (here branchlengths are not proportional to the number of mutations rising along the tree). The next polymorphic node is colored in blue ("secondary" clustered node) and is separated from the red one by a monomorphic node (light blue). External green and grey strains are direct descendants - reassortant and non-reassortant, respectively - of a clustered node. Internal nodes can also be colored in grey or green (the latter case is not represented here), indicating that all their downstream subtree shall be interpreted as ancestral or reassortant, respectively.

heterogeneity as a local feature of the tree, τ needs to be small, in the order of few point mutations. At this scope, we choose the smallest τ which meets both the following requirements:

- the fraction of internal nodes with empty neighborhood has to be small
- the majority of observed strains has to be part of a neighborhood

We set $\tau = 1, 2$ and analyze the statistics of leaves and internal nodes on the HA tree (see section 1.4.1) built with the same strains as in chapter 2, finding that only 7% ($\tau = 1$) and 5% ($\tau = 2$) of the internal nodes have no neighbors. With the same values of τ, respectively 64% and 82% of the leaves are part of a neighborhood. Although the loss of $\sim 20\%$ of represented leaves may appear significant, there are no whole areas of the tree with internal nodes included in a neighborhood exclusively for $\tau = 2$ (See Fig. 4.5). Such a homogeneous distribution of the leaves indicates that choosing $\tau = 1$ would not likely lead to the missing of entire events, however with $\tau = 2$ we ensure that the majority of external leaves get classified.

Null model: optimization of $dmax$ and θ

In order to exclude false positives in the declaration of the clusters, we perform a statistical comparison with a null model built as follows:

- We use a phylogenetic tree with the same topology of the real HA tree.
- HA sequences are divided into two parts S_1 and S_2. The sites included into S_1 or S_2 are picked randomly for each run.
- The algorithm described above (section 4.2.1) is run on S_1 and S_2, which are treated as "pseudo-" HA and NA segments.
- Diagnostic (see below) statistics are compared with the real case.

First, we analyze how the number of internal clustered nodes depends on the distance parameter $dmax$. This value shall be set so that most of the nodes in the null case result as monomorphic. Fig. 4.6 shows the number of monomorphic ($c = 1$) and polymorphic ($c = 2$) internal nodes as a function of $dmax$, for $\tau = 1, 2$, both in the null and real case: the number of polymorphic nodes in the null model is significantly smaller than in the real case, independently of $dmax$.

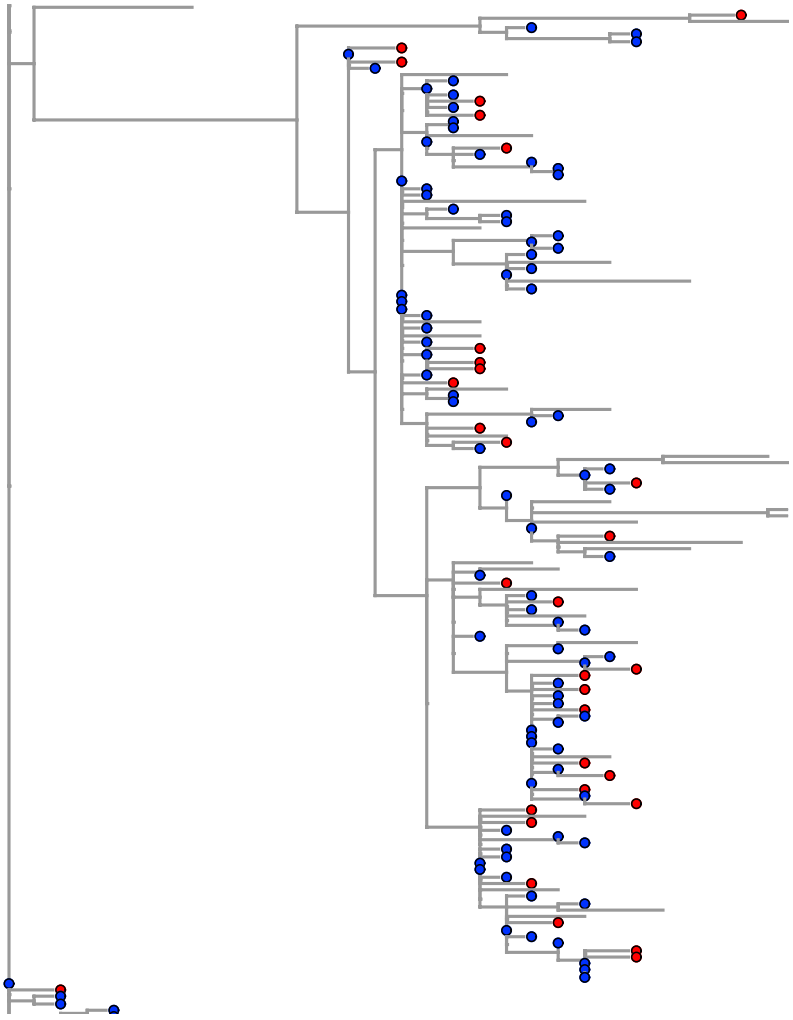
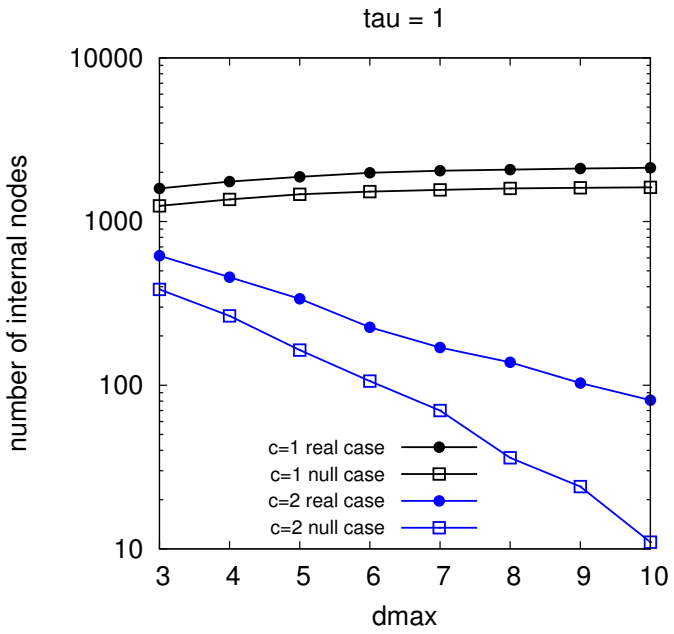
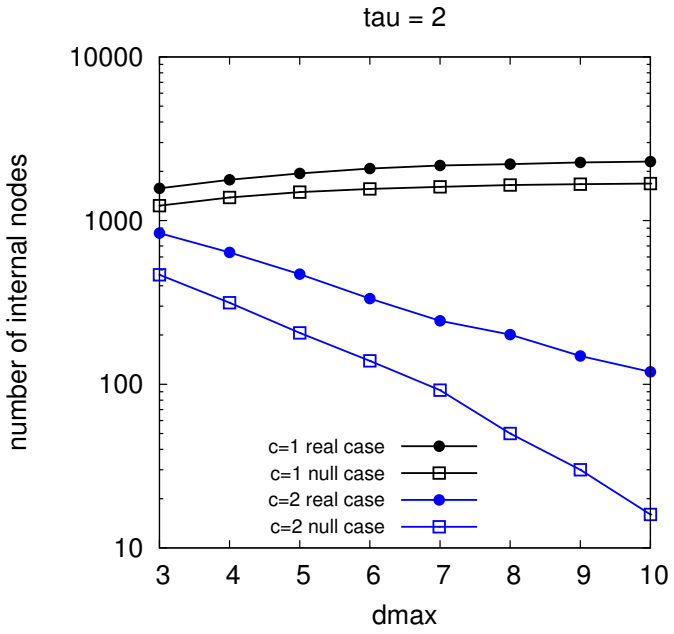


Figure 4.5: Distribution of terminal nodes included in a neighborhood for $\tau = 1, 2$. Typical subtree from A/H3N2 influenza phylogeny (haemagglutinin). Blue leaves are part of a neighborhood if $\tau = 1$, red strains are additional nodes included if $\tau = 2$, uncolored leaves have distance larger that 2 from their father node. The distribution of the colors indicates that most of the clades have representative leaves already at $\tau = 1$, meaning that with $\tau = 2$ the number of reassortment events would not change dramatically. Choosing $\tau = 2$, however, enhances the amount of classified observed strains (reassortant/non-reassortant) of about 20%.



(a)



(b)

Figure 4.6: Effect of *dmax* on the cluster index. Number of internal nodes with $c = 1, 2$ (black and blue points respectively) in the null (squares) and in the real case (dots), for $\tau = 1, 2$ ((a) and (b), respectively) as a function of the parameter *dmax*. At any distance *dmax*, the number of internal clustered nodes in the null model is small compared to the real case, suggesting that *dmax* itself is not a key parameter to discriminate real events from false positives.

In order to discriminate real events from false positives, we compare the statistics of the average distance within a cluster for nodes with $c = 2$ to the one in the null case (n_c number of couples of sequences in the cluster C),

$$d_{av}(C) = \frac{\sum_{i,j \in C} \mathrm{Hd}(i,j)}{n_c} \quad . \tag{4.3}$$

As shown in Fig. 4.7, the average diameter of a cluster in the real case is about 1.5 times the average distance in a cluster of the null model. We can therefore set a lower threshold for the average distance between clusters (equation (4.2)), based on the maximum value that this distance can assume in the null model. Fig. 4.8 shows that there are no clustered nodes in the null case with $dist(C1, C2) \geq 10$. Taking into account the scaling between real and null case found above, we can assume that there are no false positive events for $dist(C1, C2) \geq 15$. Furthermore, we can combine information on distances between and within clusters in the real and null case to retrieve events with $dist(C1, C2)$ slightly smaller than 15 and with small cluster diameter. The scatter plots in Fig. 4.9 show that, in the null case, even for $dist(C1, C2) \leq 10$ (corresponding to $dist(C1, C2) \leq 15$ in the real case) we can identify a region free from false positives (upper left corner). Based on these considerations, we finally choose to include in the list of potential reassortments also events in that range of distances, if

$$dist(C1, C2) \geq d_{av}(C) \times 0.8 + 10\,. \tag{4.4}$$

From the distribution of the distances between clusters (equation (4.2), Fig. 4.8) we can extrapolate a rough value for the third input parameter θ. We assume that the probability that two consecutive monomorphic nodes have non-compatible neighborhoods (i.e. their clusters do not match) at distances smaller than 15 (see discussion above) is small. Since this value is an approximation, following the previous debate on separation of clusters at distances ≥ 10, we run our inference algorithm for different values $10 \leq \theta \leq 15$.

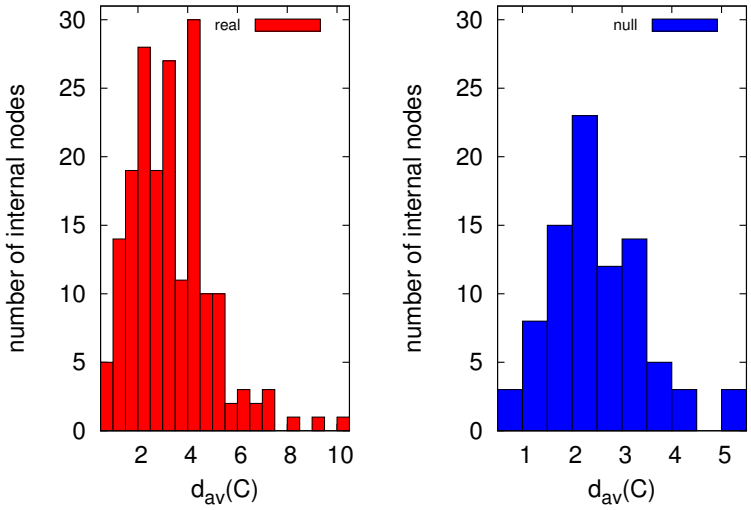


Figure 4.7: Statistics of the average distance within clusters. The histograms of average distance within the largest cluster in real (red bars) and null case (blue bars) are compared (here $\tau = 2$, $dmax = 7$). The mean of the distribution in the real case results around 1.5 times larger than the mean in the null case.

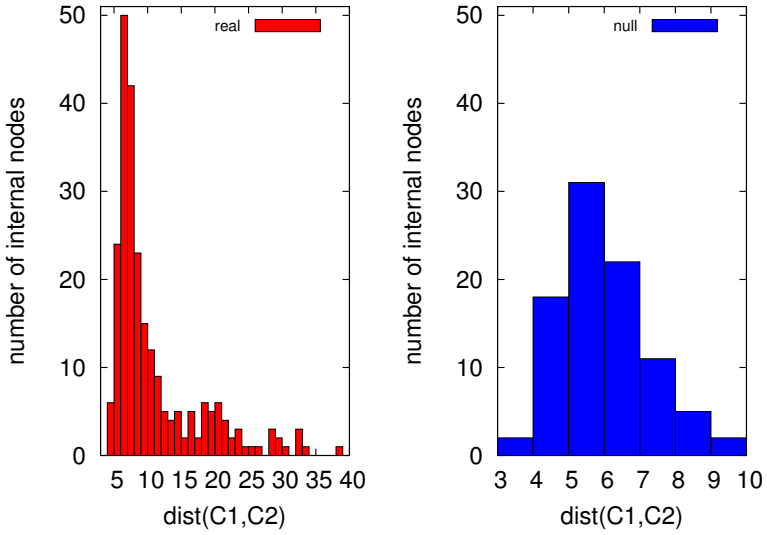
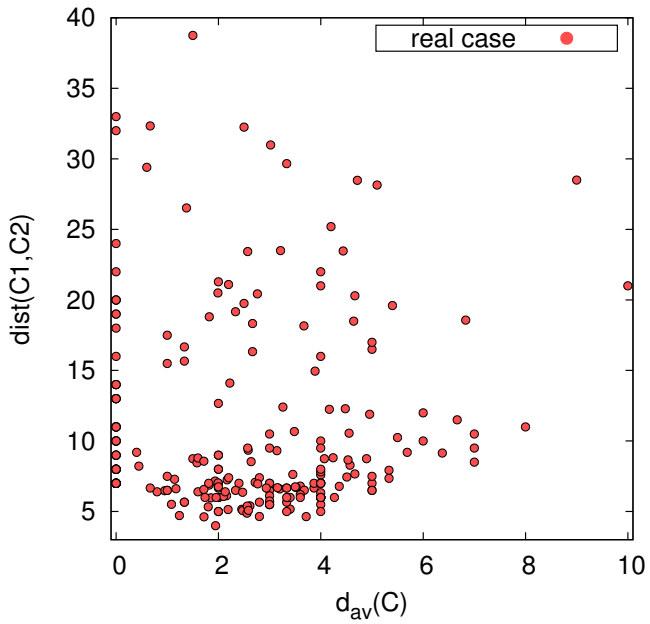
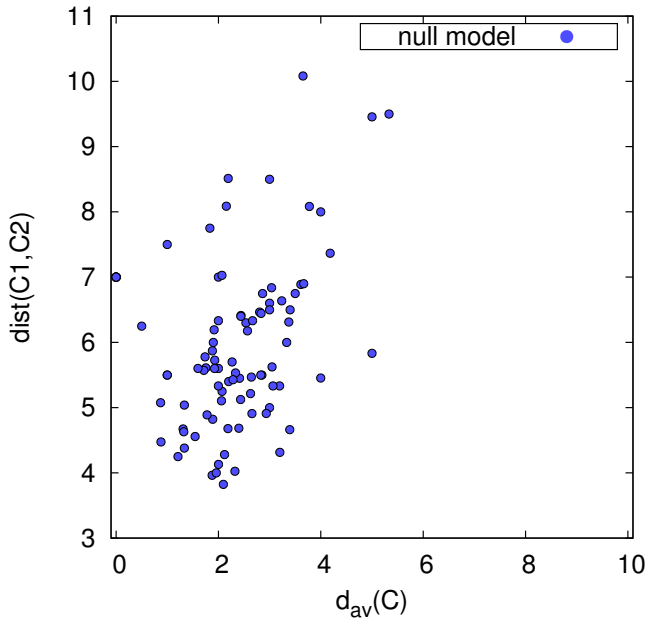


Figure 4.8: Statistics of the average distance between clusters. Histograms of average distance between clusters (equation (4.2)) in real (red bars) and null case (blue bars), with $\tau = 2$ and $dmax = 7$. At distances larger than 10 the counts for the null model drop to zero. Based on the scaling factor previously inferred (see Fig. 4.7), we conclude that there are no false positives amongst events with $dist(C1, C2) \geq 15$.



(a)



(b)

Figure 4.9: Displacement of joint ($d_{av}(C)$,$dist(C1,C2)$) in real events and in the null model. Each dot in the scatter plot represents a clustered node ($\tau = 2$, $dmax = 7$) charachterized by distances ($d_{av}(C)$,$dist(C1,C2)$), in the real (a) and null (b) case. Based on the absence of points in the upper-left corner of the null plot, we accept as real positives all the events occurring above the cut line $dist(C1,C2) \geq d_{av}(C) \times 0.8 + 10$, if $dist(C1,C2) \leq 15$ (see main text).

4.3 Preliminary results for a large set of A/H3N2 strains

Guided by the preliminary analysis, we run our algorithm on a set of 15086 A/H3N2 strains collected from 1968 to March 2018, aligned and prepared as detailed at the beginning of this chapter. We note that the improvement of computational speed compared to the joint genealogical method allows us to analyze a number of sequences about three times larger than the one analyzed before, with running times about 20 times shorter. Given that the specificity of the algorithm does not depend on $dmax$ (Fig. 4.6), any value of this parameter smaller than the cut defined by the criterion (4.4) is acceptable; the threshold on the distance between clusters is already responsible for discarding polymorphic nodes with low separation between clusters. We set $dmax = 9$. With $\tau = 2$ and $10 \leq \theta \leq 15$ we find 61 primary events (red nodes), 22 secondary clustered nodes (blue) and 9 monomorphic nodes hiding polymorphism (light blue). Further, we observe between 58 and 103 (depending on the value set for θ) switches from monomorphic nodes carrying ancestral neighborhoods to reassortant clusters and between 30 and 108 switches backwards. We find that the overall number of events is in general agreement with the number reported in chapter 3; these events are signaled either by polymorphic nodes or by switches. A switch, as already mentioned, may constitute by itself the manifestation of a reassortment event, even without further evidence of polymorphism in the surrounding area: strains with a new NA variant are grouped by the tree reconstruction algorithm in the neighborhood of a monomorphic node by chance, instead of being distributed in a neighborhood containing also the ancestral neuraminidase. Consistent with the results of chapter 2, we find, on average, that the number of reassortant external strains per event (in the neighborhood of a cluster node) is smaller than the number of non-reassortant leaves (roughly two grey ancestral leaves for each green reassortant leaf). Our results, although qualitatively in agreement with the ones reported with the joint-tree method, should not be considered as conclusive; in particular, further analysis need to be performed to test robustness under the change of the threshold θ, since switches influence the reported ratio between reassortant and non-reassortant observed strains.

4.3.1 Reassortment in recent A/H3N2 viruses

The expanding throughput of sequences sampled in the last years has revealed that the HA clade (3c.2a, according to the World Health Organization nomenclature) dominating the scene in recent seasons has diversified into several subclades in the past year [104,105].

In particular, the two subclades 3c.2a.1b and 3c.2a.2 have been predicted to be competing for taking over in the next future (see web tools [105,106], based on the models in [77,107]), with the latter likely to have exchanged the NA segment with the now extinct 3c.2a.1a clade. For notational simplicity in the following we will refer to these clades as a1b, a2 and a1a, respectively. Fig. 4.10 shows the outcome of our algorithm represented on the HA tree with the color scheme described in Fig. 4.2 and 4.3, for the specific subclade a2, with $\tau = 2$ and $\theta = 10$.

We note that in this subtree the general coloring scheme and the number of primary reassortment indicators in particular (four red nodes) are overall stable for different choices of the parameters. Our algorithm predicts multiple reassortment events, one of them involving a significantly large subtree (yellow event in Fig. 4.11). By mapping all the inferred reassortant variants on the HA and NA tree (Fig. 4.11 and 4.12), we find that there are three independent reassortment events that have introduced new NA variants within the a2 clade (red, green and yellow clusters). Consistently with the results reported in [105], a relevant fraction of HA in the a2 clade is coupled with NA imported from a1a clade (yellow cluster in Fig. 4.11 and 4.12); furthermore, we point out that the reassortant variant of this event is already present upstream the primary clustered node (lower red node in Fig. 4.10), being introduced by a switch in a monomorphic ancestor few levels higher in the tree. In this case, then, the event starts with a switch originating a whole reassortant subtree. Some of the strains in this subclade, however, are still coupled to the ancestral NA variant, which falls into the neighborhood of the polymorphic node. Interestingly, each of the reported clusters appears in conjunction with aminoacid changes at sites inferred to be epistatic by parallel analysis. Moreover, one of the mutations spreading in the NA subtree of the largest reassortment event (yellow) takes place in a site biologically relevant for enzymatic processes.[8] Last, we notice that one of the several switches introducing a reassortant variant is immediately followed by a reverse switch (grey square in Fig. 4.10) restoring the original neuraminidase version. A deeper analysis of the frequency and distribution of double switches along the whole tree occurring at different thresholds θ will be necessary to assure the robustness of the new method.

4.3.2 Discussion and remarks

In this chapter we have presented a new heuristic method to infer reassortment based on single-segment phylogenetic trees. The gain in computational speed with respect to the

[8] Namely, glycosylation process.

algorithm presented in chapter 2 makes this approach a preferable choice for the analysis of large trees, a characteristic that has been acquiring more and more relevance in the last few years, given the exponential growth of the number of collected sequences. The concept beneath the algorithm is the identification, on the phylogenetic tree of one segment, of ruptures of the flow of genetic information between fathers nodes and their children. The discontinuity generated by the introduction of a new NA variant is embodied both by large genetic distance in the second segment between the descendants of consecutive nodes (switches) and by the presence of polymorphic nodes. The output of the algorithm clearly depends on the values set for the input parameters, in particular further analysis need to be performed to check robustness to different thresholds θ. The frequency and distribution of switches along the tree strongly affect the estimate of reassortant and non-reassortant clade growth. Moreover, the recognition of certain patterns displaying switches and subsequent short time back-switches can be used as a tool to infer selection. Recurrent restoring of the ancestral variant right after a first switch may indicate the action of negative selection, but also simple alignment errors.

By applying this new method to recent A/H3N2 clades we have found an enrichment of reassortment occurring in the last seasons, which resulted in co-circulation of several distinct NA variants within the viral population. These reassortant clades contain non-synonymous mutations at epistatic sites, including adaptive and subsequent compensatory changes. As discussed in the introduction, phenotypic epistasis generates evolutionary constraints that increase predictability of future evolutionary steps. Information on reassortment can therefore be included in the design of improved models to forecast influenza evolution that take into account more than one gene: on the one hand reshuffling of genome segments introduces a fitness cost (chapter 3), on the other hand it can be associated to epistatic changes.

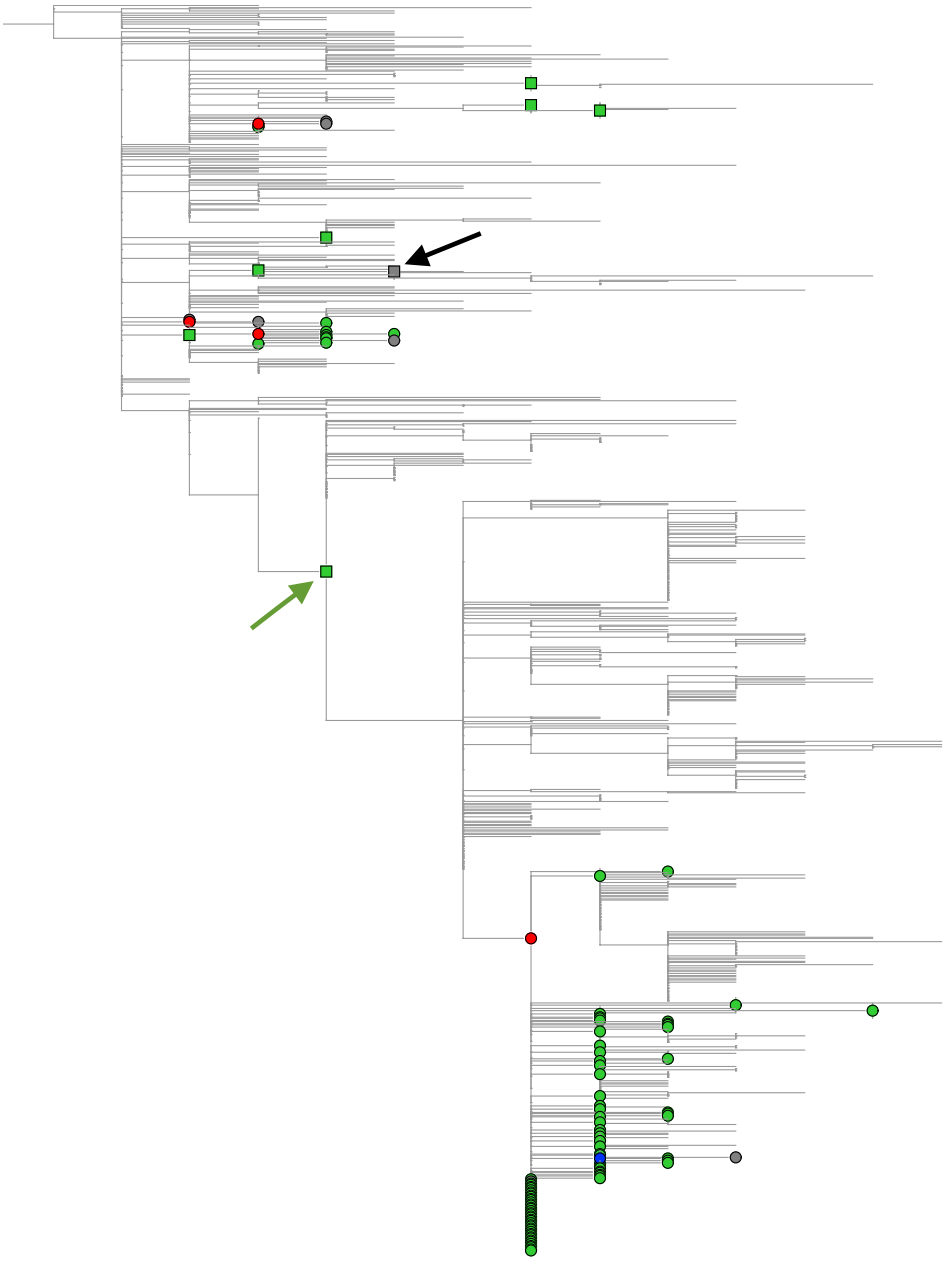


Figure 4.10: Reassortment events in recent 3c.2a.2 clade. Clade 3c.2a.2 shows a pattern of multiple reassortments ($\tau = 2$, $\theta = 10$). The presence of different NA variants is signaled by four red clustered nodes and several switches from ancestral to reassortant neighborhoods. Here we depict switches with a square, to differentiate them from the other green and grey strains. In the upper region the dominant version is the non-reassortant neuraminidase, with sporadic insertions of new NA variants and one back-switch (grey square, black arrow). A further switch (green arrow) signals a significant inversion towards a reassortant variant, which spreads in the lower part of the clade.

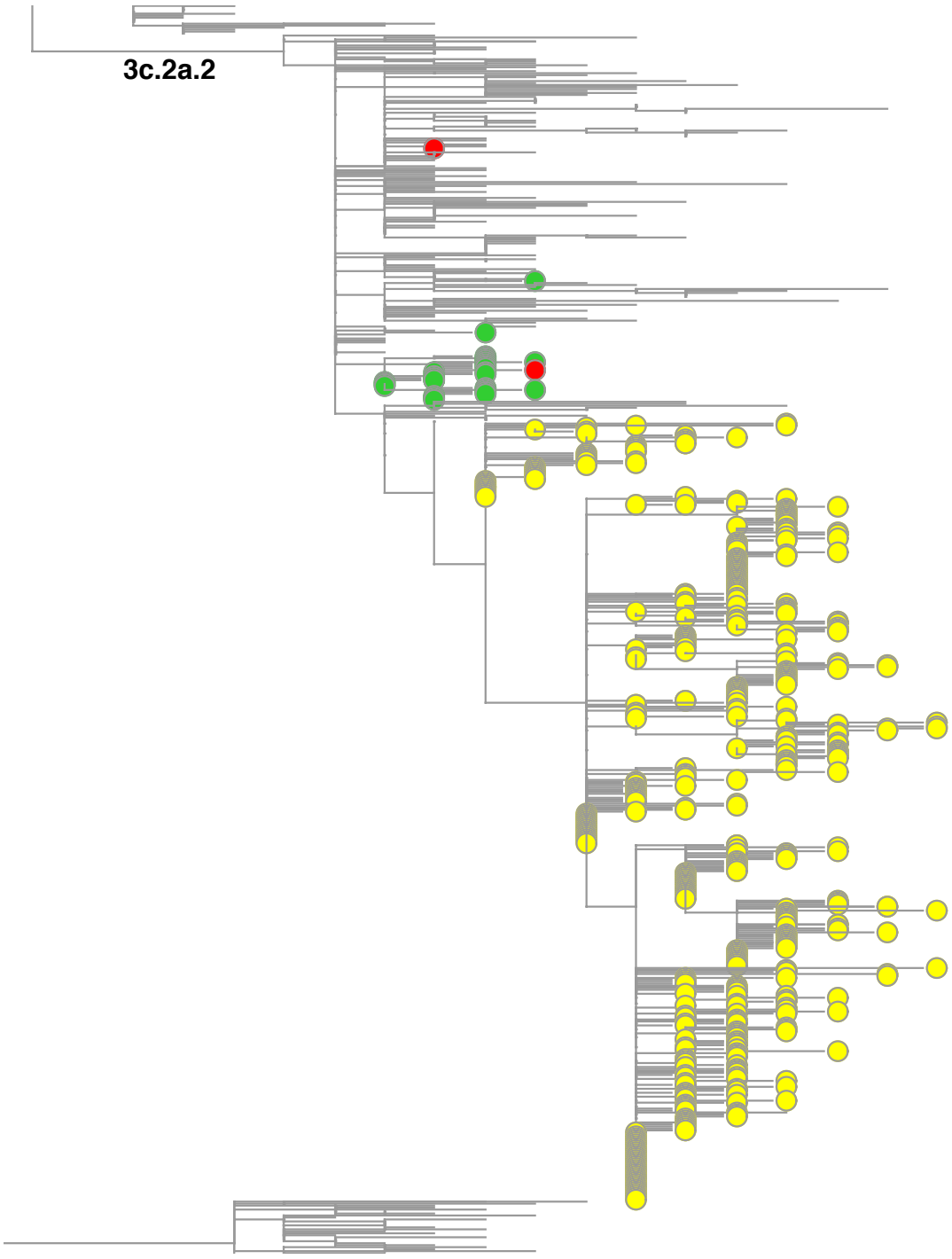


Figure 4.11: Displacement of reassortant strains on the HA tree. Each of the three reassortment events occurring in the recent a2 clade is represented with a different color. The reassortant strains produced by insertion of a new NA variant are mapped into the HA phylogenetic tree (cf. Fig. 4.12 for the correspondent mapping on the NA tree).

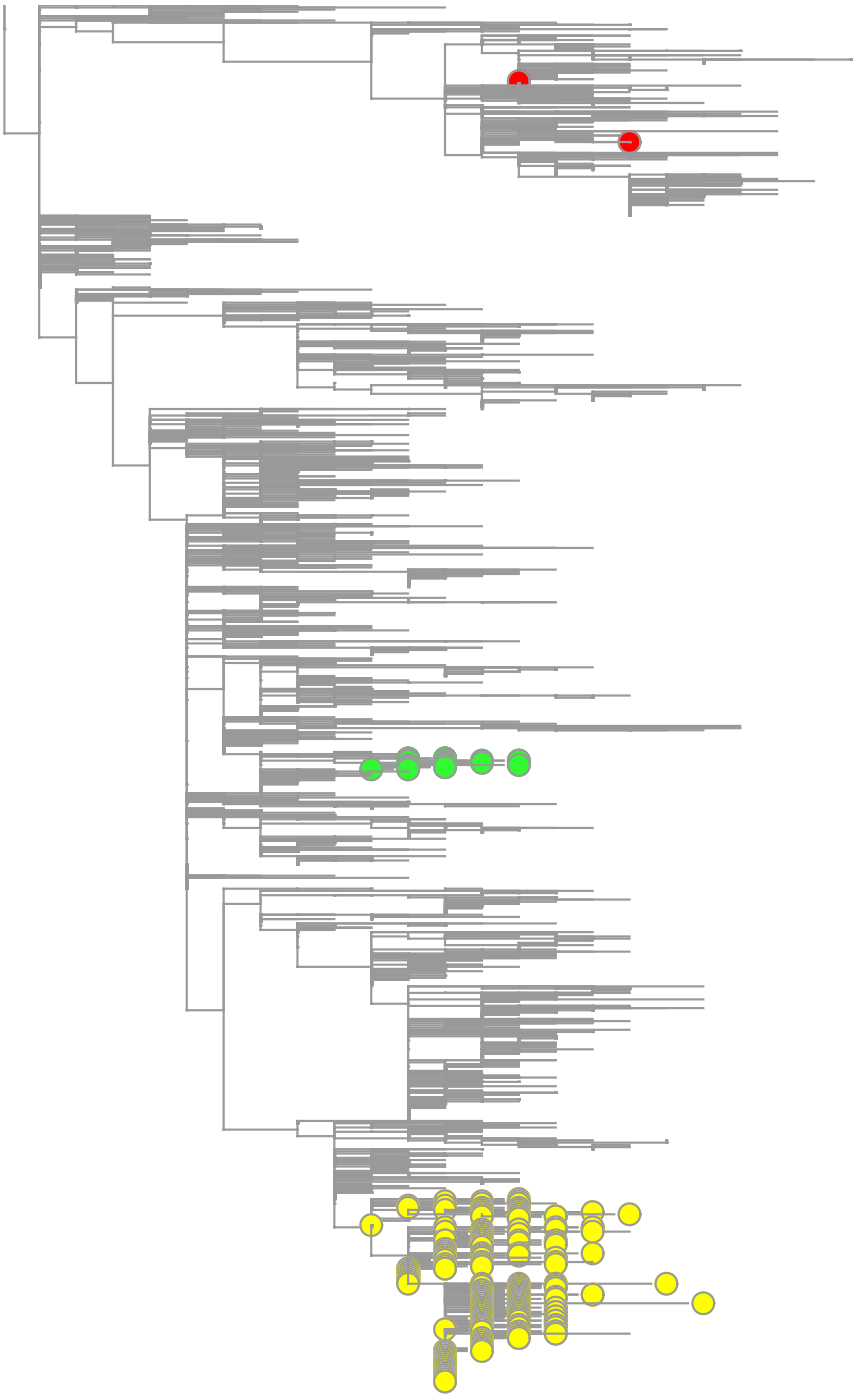


Figure 4.12: Displacement of reassortant strains on the NA tree. The reassortant variants circulating in 3c.2a.2 clade are mapped on the NA tree using the same colors and symbols as in Fig. 4.11. The distribution of the clusters indicates that there are at least three different NA reassortant variant (beside the ancestral one) paired to similar HA.

Summary and conclusions

In this thesis we have determined the net effect of reassortment on the evolution of human influenza virus, with particular focus on events building new haemagglutinin and neuraminidase combinations. The lack of a reliable method for inferring reassortment within one lineage has given us the chance to the develop a novel, robust, detection algorithm. Our inference is based on the identification of patterns of mutations occurring along the branches of joint genealogies built with paired RNA segments. With this new tool, we have drawn up a trustworthy list of reassortment events between HA and NA proteins. Starting from these events we have inferred negative selection on two different levels. First, reassortment between strains at large genetic distance is suppressed, compared to a model where gene mixing occurs randomly. The resulting negative selection on reassortment increases in strength with the genetic distance between reassortant and parent strains (Fig. 3.6). Second, gene reshuffling tends to affect peripheral strains in the genealogical tree, leading to reduced clades growth of the reassortant variants compared to their non-reassortant counterparts. Overall, mild negative selection results in a fitness cost (calculated in (3.2)) and is compatible with broadly distributed cross-protein epistasis: although reassortant variants are fit enough to reach detectable population frequency, they are less fit than the respective non-reassortant strains.

The mixing of genetic material by reassortment is somewhat similar to recombination in diploid populations or transformation in bacteria. However, recombination and transformation require physical splicing of genome segments; the rates of these enzymatic processes get strongly suppressed with increasing genetic distance of the parent sequences [108–110]. There are no corresponding physiological barriers against reassortment of viral segments. Instead, our results suggest that epistatic fitness barriers are already substantial between more distant co-circulating strains of the same lineage, which ties in with the observation of cross-lineage pairing constraints [64, 66]. Hence, on larger scales of genetic distance, such fitness barriers may be an important factor in delineating – and, thus,

defining – viral species.

The increase in number of available strain sequences in the last few seasons requires new bioinformatic tools with the ability of analyzing large amounts of data as fast as possible. In order to explore the evolutionary dynamics of recent A/H3N2 epidemics, we have designed a second efficient detection algorithm based on single-segment trees. With a heuristic approach, we identify clades on the haemagglutinin tree that carry more than one neuraminidase variant. The algorithm takes inspiration from the idea of "message passing" developed in other works [78]. First, internal nodes on the HA tree are assigned a cluster index based on heterogeneity of its descendants, calculated in the NA segment (upwards information flow). Offspring of ranked internal nodes, then, are classified as reassortant or ancestral depending on how similar their NA variants are to the ones of their parents (downwards information flow). The application of this method to influenza clades dominating the last winter infection season has revealed unusual reassortment dynamics. Three independent events have introduced in the viral population just as many new neuraminidase versions, which differ from the ancestral one by point mutations in epistatic sites. This characteristic is of particular interest in the framework of forecasting evolution. The association of reassortment with adaptive and compensatory changes opens the way to the recognition of recurrent evolutionary patterns (repeatability), a step forward to the design of predictive models based on the whole genome.

A. Comprehensive list of reassortment events in influenza A/H3N2

The table below reports the complete list of reassortment events detected with our genealogical method. Here we compare our results with reassortment published in previous works. As highlighted by Pinsent et al. [94], there are few intra-subtype reassortment events which are reported in literature with large consensus. Among these, a major event occurred between isolates co-circulating in New York in the early 2000s is well-documented. This event, which involved also strains with mixed HA and NA segments, has been related to the jump from A/Sydney/5/1997-like to A/Fujian/411/2002-like antigenic clusters (Holmes et al. [55]). Event number 3 in table A.1 coherently represents the event occurred between the main group of viruses sampled in the 1999-2000 season (A/New York/315/1999-like viruses) and the isolate A/New York/177/1999 (which appears genetically close to other two strains collected in a different geographical region: A/Memphis/59/1999 and A/Netherlands/051/2000). This event gave origin to the strains A/New York/137/1999 and A/New York/138/1999 (green colored in [55]), that diverge from the clade leading in 1999–2000 season. Event number 74 refers to the appearance of the Fujian/02-like reassortants A/New York/198/2003 and A/New York/199/2003 (referred in [55] as Clade B and colored in yellow). We report this event with a large d between the parents, which is consistent with the clear separation in NA phylogeny with respect to the main circulating clade (Clade A, light blue in [55]). The remaining events marked with a star in table A.1 refer to a consensus list of reassortments provided in [94], including two large events occurred in the Netherlands (events 98 and 99, Westgeest et al. [2]) and in Hong Kong. In addition, we have found that some of the events reported

in the referenced studies as independent (Rabadan et al. [86], [2, 94]) are grouped in our analysis as part of the same reassortment event.

We apply restrictive criteria to select the sequences that are included in our analysis, in order to both avoid over-counting and ensure that the reported events are as clean as possible. This is essential to prevent non controllable factors confounding the results on the main biological object of this study, i.e. inference of selection. As a result, some of the minor events reported in other studies include sequences which are not in our database, and therefore cannot be reported here. However, the overall agreement with the other studies, signaled by the presence of the majority of the reported events in our list, guarantees that we are able to detect at least a large fraction of the real events.

event	**distance,** d	**parent clade,** p	**parent clade,** p'	**reassortant clade,** r
1*	11.5	29489,14381,30237,14372,29491	14338,3758,14333	14358
2	8.5	154756	134337,145114,148480,144791,132668	163337
3*	15	6631,8200,166347	8978,8977,9217,5090,8981	8706,6710,6678
4	17	142809,140296,137259,160602,131807	152656,90466,98995,152503,79676	145506
5	12.5	198035	195744,195907,197988,194978,197748	198065
6	12	79650,94769,85831,86055,83861	65125,77838,93878,34977,32310	86052
7	4.5	152192,152613,152818,152498,152358	152291,152817,152255,152392,152775	152686
8	4.5	16068	97156,14327,14323,3750,14330	16000
9*	12	86052	85607,90818,94612,94642,93661	90054
10	7	194156,173986,178004,175204,179355	172767,170296,169683	170296
11	10.5	189816,170207,191802,170197,170701	145483,147970,174249,160586,132668	191800
12	5.5	137758	142809,140296,137259,160602,131807	161491
13	15.5	170266,169497	160279,140419,138011,160321,131805	175187
14	9.5	162098,161905,162096,162135,155890	145483,147970,174249,160586,132668	158821
15*	21	128700,128654	103313	192795
16	11.5	81387,81399,66705	66704,102802,102662,68571,93892	66706
17	11.5	155921	158766,159638,153710,153167,156896	163130
18	7	107846,128072,104125	99876,88040,119705	128044
19	4	194530	193338,193325,195842,195886,197932	197913
20	9	159630,162123,164260,154747	134337,145114,148480,144791,132668	164218
21	13	154746	176988,154023,174836,145111,173207	161320
22	9.5	194642	164815,165579,162145,164711,167825	191815
23	4	195846	197859,194530,197925,197917,197916	197955,197909,197942,197904,197932
24	8.5	152165,152812,152248,152363,152278	143018,83810,87896,89798,81393	152392

25*	6	6679	9235,5017,6471,9177,8630	5015,5032
26	10	15986	98707,117862,30238,14396,33845	90492
27	11	89668,91142,136596,78771,93667	143018,83810,87896,89798,81393	83166
28*	13	115460,115461	111170	115515
29	9	164985,164983,191803,168905,166842	163091,165115,152908,159630,166303	191777
30	4	197888,197905,197964	193338,191710,172747,195246,175213	197930
31	9.5	5109,6508,5696,7057,6355	14254,14247,14253,14252,14266	14249
32	7.5	119713,111411,136406,122660,106394	101916,122601,121525	152345
33	4.5	70811	62803,32221,70811,27632,23843	79673
34	9.5	20566	120392,14384,5728,90490,118003	20572
35	14	152464	152758,152192,96104,152291,90810	152439
36	9	77767,26246,23932,76127	20786,29631,23281,23270	107452
37	6	84613	29908,32258,28882,15889,19060	84614
38	6.5	160943	15483,22763,15558,19707,20684	19896
39	4	191673,194131,172496,169910,176539	193359,172580,173245,169126,170030	170667
40	12	65123	70811,62820,32252,150395,32307	79650,94769,85831,86055,83861
41	6.5	86093	86091,88045,90589	88045
42	25.5	121959,121940,122678,117434,104122	117447,99871,117446,103748,104144	119881
43	8.5	140701	88021,96020,90603,93713,90646	162151
44	5	104109	128049,94718,128048,136373,90646	104134
45	17	191710,170685,170646,191041,176677	134337,145114,157218,148480,156897	191786
46	7	175247	169932,178986,172739,173249,191695	197973
47	19.5	188727	132678,160596,153076,153306,139916	193314
48	5.5	198011,198006	178978,193175,176501,172555,170431	197908
49	17.5	191775	132678,160596,153076,153306,139916	194156

50	10.5	168140,173216,174183,172587,171390	161899,159526,173051	195875,195879
51	3	132678,160596,153076,153306,139916	161899,159526,173051	166226
52	25.5	121931,128651,128649,107858,107855	134337,145114,157218,148480,156897	159618
53	13	90835,86091,119713,109757,133061	118540,93786,85720,88963	85722
54	7	160374,160359,160405,160299	134774,148031,160490,160466,148033	148059
55	9.5	140897	143018,83810,87896,89798,81393	140901
56	11.5	131268,131358,131356,131269,133998	88021,90603,90599,94718,90646	138002
57	6.5	156905	166290,171388,158968,154764,173207	171375
58	8.5	79331	192757	172586
59	11.5	120294,99877	117447,99871,117446,103748,104144	128072
60	6	66079	62803,19055,118061,19712,23843	97483
61	9.5	143015,143014	62803,19055,118061,19712,23843	76700
62	9.5	188727	170724,188755,179340,179369,169915	192181
63	7.5	117460,117461,103744	103248,98651,117449,94724,88040	134331
64	6.5	129899,121972,107830,128027,117655	98651,121258,107852,122974,88040	104113
65	6	171931	170724,188755,179340,179369,169915	168903
66	10	123055,122961,118652,100445,100516	121258,121936,127833,122932,128053	119708
67	8.5	154022,163098,168369,165589,164250	121258,121936,127833,122932,128053	152955
68	5	168931,170309	170724,188755,179340,179369,169915	178093
69	4.5	170638,191812	193338,191710,191674,170724,169915	191813
70	8.5	13724	14410,32482	20286
71	21.5	115303,115354,114867,8696	114610	5665
72	7	173048	147970,174249,134012,192798,148094	173051
73	11	197937,197934	147970,174249,134012,192798,148094	197889
74*	42.5	5022,9235,9198,5466,6674	20810,14305,14327,14323,14321	5009,107554

75	4.5	14137	14137	14187
76	19	16064	14384,90490,154552,90544,14394	70807
77	8	93799,65120,93914,77779,93900	62803,19055,118061,19712,23843	93801
78	6	153942	147970,174249,134012,192798,148094	154770
79	10.5	83173,152458,93666,152193,152600	62803,32221,70811,19055,32251	108212
80	16	191786	168783,170701,167130	179483
81	4.5	6848,20832,25012,25013,22823	113017	8984
82	16	172559	167130,191714,169186	170711
83	5.5	191813	191674,170724,188755,179340,169915	186617
84	16	177539	168783,170701,167130	178965
85	4.5	160950	19055,19712,27632,29908,29574	60739
86	7.5	155910	147970,174249,134012,192798,148094	162113
87	17	175052,179472,174982,190933,193176	176425,173233,171740	191932
88	5	161275	147970,174249,134012,192798,148094	177541
89	14.5	191789	168897,188817,175028,167920,172766	179486
90	12	192166	147970,174249,134012,192798,148094	191968
91	12.5	86052	77951,32286,160994,87188,77762	79326
92	8	8672,8198,114249,114246	5105	114398,114396,110907
93	5	98651,99773,122942	122956,99874,99873,122999,122902	123005
94	13.5	170843	147970,174249,134012,192798,148094	191799
95	8	159133,164603,159141	147970,174249,134012,192798,148094	164216
96	12.5	161480	122956,99874,99873,122999,122902	123069
97	15.5	172583	167961,166459,189816,169089,168899,	175159
98*	14.5	7064,114014	8698,9281,8698,5106	114196
99*	26	16012,90502	9078,5007,5637,5508,9064	107634,107487,107930,107388,6659

100	7	114868	114610	111020
101	13	169946	192798,134337,131805,132579,189622	170734,170324,168100,172756,170703
102	4	85722	152656,90466,98995,152503,79676	85719
103	10.5	142676	152656,90466,98995,152503,79676	142677

Table A.1: List of inferred reassortment events from 1968 to 2015 between HA and NA segments in human influenza A/H3N2. Column 2: mean nucleotide distance d between reassortant strain and parent strains. Columns 3-5: representative observed strains in the clades of p, p' and r, respectively. Each isolate is identified by its number in the online EpiFlu DATABASE (http://www.gisaid.org) identifier (e.g. EPI_ISL_7064 is reported here as 7064). Stars indicate events which are reported in literature with large agreement.

Bibliography

[1] L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nat. Rev. Genet.*, 7:85–97, 2006. doi:10.1038/nrg1767.

[2] K. B. Westgeest and et al. Genetic evolution of the neuraminidase of influenza A (H3N2) viruses from 1968 to 2009 and its correspondence to haemagglutinin evolution. *J Gen Virol*, 93(9):1996–2007, 2012. doi:10.1099/vir.0.043059-0.

[3] K. B. Westgeest and et al. Genomewide analysis of reassortment and evolution of human influenza A(H3N2) viruses circulating between 1968 and 2011. *J Virol*, 88(5):2844–2857, 2014. doi:10.1128/JVI.02163-13.

[4] A. D. Neverov, S. Kryazhimskiy, J. B. Plotkin, and G. A. Bazykin. Coordinated evolution of influenza A surface proteins. *PLoS Gen*, 11(8):e1005404, 2015. doi:10.1371/journal.pgen.1005404.

[5] M. Villa and M. Lässig. Fitness cost of reassortment in human influenza. *PLoS Pathogens*, 13(11):e1006685, 2017. doi:10.1371/journal.ppat.1006685.

[6] M. Lässig. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics*, 8(Suppl 6):S7, 2007. doi:10.1186/1471-2105-8-S6-S7.

[7] M. Kimura. Diffusion models in population genetics. *J. Appl. Probab.*, 1(2):177–232, 1964.

[8] V. Mustonen and M. Lässig. Adaptations to fluctuating selection in Drosophila. *PNAS*, 104(7):2277–2282, 2007. doi:10.1073/pnas.0607105104.

[9] V. Mustonen and M. Lässig. Molecular evolution under fitness fluctuations. *Phys. Rev. Lett.*, 100(10):108101, 2008. doi:10.1103/PhysRevLett.100.108101.

[10] N. Strelkowa and M. Lässig. Clonal interference in the evolution of influenza. *Genetics*, 192(2):671–682, 2012. doi:10.1534/genetics.112.143396.

[11] M. Kimura and T. Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61(3):763, 1969.

[12] V. Mustonen and M. Lässig. Fitness flux and ubiquity of adaptive evolution. *PNAS*, 107(9):4248–4253, 2010. doi:10.1073/pnas.0907953107.

[13] K. C. Atwood, L. K. Schneider, and F. J. Ryan. Periodic selection in Escherichia coli. *Proc Natl Acad Sci USA*, 37:146–155, 1951.

[14] P. J. Gerrish and R. E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102/103:127–144, 1998.

[15] J. A. G. M. De Visser, C. W. Zeyl, P. J. Gerrish, J. L. Blanchard, and R. E. Lenski. Diminishing returns from mutation supply rate in asexual populations. *Science*, 283:404–406, 1999.

[16] J. A. G. M. De Visser and Lenski R. E. Long-term experimental evolution in Escherichia coli. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. *BMC Evolutionary Biology*, 2:19, 2002. doi:10.1186/1471-2148-2-19.

[17] J. A. G. M. De Visser and D. E. Rozen. Clonal interference and the periodic selection of new beneficial mutations in Escherichia coli. *Genetics*, 172(4):2093–2100, 2006. doi:10.1534/genetics.105.052373.

[18] L. Perfeito, L. Fernandes, C. Mota, and I. Gordo. Adaptive mutations in bacteria: High rate and small effects. *Science*, 317(5839):813–815, 2007. doi:10.1126/science.1142284.

[19] C. R. Miller, P. Joyce, and H. A. Wichman. Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics*, 187:185–202, 2011. doi:10.1534/genetics.110.121400.

[20] S. C. Park and J. Krug. Clonal interference in large populations. *PNAS*, 104(46):18135–18140, 2007. doi:10.1073/pnas.0705778104.

[21] M. M. Desai and D. S. Fisher. Beneficial mutationselection balance and the effect of linkage on positive selection. *Genetics*, 176:1759–1798, 2007. doi:10.1534/genetics.106.067678.

[22] S. Schiffels, G. J. Szöllősi, V. Mustonen, and M. Lässig. Emergent neutrality in adaptive asexual evolution. *Genetics*, 189(4):1361–1375, 2011. doi:10.1534/genetics.111.132027.

[23] B. H. Good, I. M. Rouzine, D. J. Balick, O. Hallatschek, and M. M. Desai. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *PNAS*, 109(13):4950–4955, 2012.

[24] M. Lässig. Chance and risk in adaptive evolution. *PNAS*, 109(13):4719–4720, 2012. doi:10.1073/pnas.1203012109.

[25] M. Lässig, V. Mustonen, and A. M. Walczak. Predicting evolution. *Nature Ecology & Evolution*, 1(0077), 2017. doi:10.1038/s41559-017-0077.

[26] R. A. Neher, M. Vucelja, M. Mezard, and B. I. Shraiman. Emergence of clones in sexual populations. *J. Stat. Mech.*, P01008, 2013.

[27] R. A. Neher and B. I. Shraiman. Competition between recombination and epistasis can cause a transition from allele to genotype selection. *PNAS*, 106(16):6866–6871, 2009. doi:10.1073/pnas.0812560106.

[28] M. Kimura. Attainment of Quasi Linkage Equilibrium when gene frequencies are changing by natural selection. *Genetics*, 52(5):875–90, 1965.

[29] R. A. Neher and B. I. Shraiman. Statistical genetics and evolution of quantitative traits. *Rev. Mod. Phys.*, 83(4):1283–1300, 2011. doi:10.1103/RevModPhys.83.1283.

[30] M. Mézard, G. Parisi, and M. A. Virasoro. Random free energie in spin glasses. *J. Physique Lett.*, 46:L217–L222, 1985.

[31] M. Mézard and A. Montanari. *Information, Physics, and Computation.* Oxford University Press, 2009.

[32] S. R. Harris and et al. Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, 327(5964):469–474, 2010. doi:10.1126/science.1182395.

[33] J. A. Lindsay. Hospital-associated MRSA and antibiotic resistance. what have we learned from genomics? *International Journal of Medical Microbiology*, 303(6):318–323, 2013. doi:10.1016/j.ijmm.2013.02.005.

[34] A. J. McCarthy and et al. Extensive horizontal gene transfer during staphylococcus aureus co-colonization in vivo. *Genome Biology and Evolution*, 6(10):2697–2708, 2014. doi:10.1093/gbe/evu214.

[35] K. Stöhr. Influenza-WHO cares. *Lancet Infect. Dis.*, 2:517, 2002. doi:10.1016/S1473-3099(02)00366-3.

[36] M. Yamashita, M. Krystal, W. M. Fitch, and P. Palese. Influenza B virus evolution: co-circulating lineages and comparison of evolutionary pattern with those of influenza A and C viruses. *Virology*, 163(1):112–122, 1988. doi:10.1016/0042-6822(88)90238-3.

[37] G. M. Air, A. J. Gibbs, W. G. Laver, and R. G. Webster. Evolutionary changes in influenza B are not primarily governed by antibody selection. *Proc Natl Acad Sci USA*, 87(10):3884–3888, 1990. doi:10.1073/pnas.87.15.6007a.

[38] E. Nobusawa and K. Sato. Comparison of the mutation rates of human influenza A and B viruses. *J Virol*, 80(7):3675–3678, 2006. doi:10.1128/JVI.80.7.3675-3678.2006.

[39] T. Bedford and et al. Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3:e01914, 2014. doi:10.7554/eLife.01914.

[40] M. C. Zambon. Epidemiology and pathogenesis of influenza. *J. Antimicrob. Chemother.*, 44(Suppl 2):3–9, 1999.

[41] W. Chen and et al. A novel influenza A virus mitochondrial protein that induces cell death. *Nat. Med.*, 7:1306–1312, 2001. doi:10.1038/nm1201-1306.

[42] B. W. Jagger and et al. An overlapping protein-coding region in influenza A virus segment 3 modulates the host response. *Science*, 337(6091):199–204, 2012. doi:10.1126/science.1222213.

[43] D. C. Wiley, I. A. Wilson, and J. J. Skehel. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, 289:373–378, 1981. doi:10.1038/289373a0.

[44] J. P. Lynch and E. E. Walsh. Influenza: evolving strategies in treatment and prevention. *Semin. Respir. Crit. Care. Med.*, 28(2):144–158, 2007. doi:10.1055/s-2007-976487.

[45] V. N. Petrova and C. A. Russell. The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, 16:47–60, 2018. doi:10.1038/nrmicro.2017.118.

[46] M. Cohen and et al. Influenza A penetrates host mucus by cleaving sialic acids with neuraminidase. *Virol. J.*, 10:321, 2013. doi:10.1186/1743-422X-10-321.

[47] D. J. Smith and et al. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, 2004.

[48] K. Koelle and D. A. Rasmussen. The effects of a deleterious mutation load on patterns of influenza A/H3N2s antigenic evolution in humans. *eLife*, 4:e07361, 2015. doi:10.7554/eLife.07361.

[49] P. Palese, K. Tobita, M. Ueda, and R. W. Compans. Characterization of temperature sensitive influenza virus mutants defective in neuraminidase. *Virology*, 61(2):397–410, 1974. doi:10.1016/0042-6822(74)90276-1.

[50] C. Liu, M. C. Eichelberger, R. W. Compans, and G. M. Air. Influenza type A virus neuraminidase does not play a role in viral entry, replication, assembly, or budding. *J Virol*, 69(2):1099 1106, 1995.

[51] M. I. Nelson and E. C. Holmes. The evolution of epidemic influenza. *Nature Reviews Genetics*, 8:196–205, 2007. doi:10.1038/nrg2053.

[52] A. W. Hampson. Influenza virus antigens and antigenic drift'. *Perspectives in Medical Virology*, 7:49–85, 2002. doi:10.1016/S0168-7069(02)07004-0.

[53] C. Li and et al. Reassortment between avian H5N1 and human H3N2 influenza viruses creates hybrid viruses with substantial virulence. *PNAS*, 107(10):4687–4692, 2010. doi:10.1073/pnas.0912807107.

[54] A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, and E. C. Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453:615–619, 2008. doi:10.1038/nature06945.

[55] E. C. Holmes, E. Ghedin, N. Miller, J. Taylor, Y. Bao, K. St. George, and et al. Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol*, 3(9):e300, 2005. doi:10.1371/journal.pbio.0030300.

[56] R. G. Webster, W. G. Laver, G. M. Air, and G. C. Schild. Molecular mechanisms of variation in influenza viruses. *Nature*, 296:115–121, 1982. doi:10.1038/296115a0.

[57] R. B. Belshe. The origins of pandemic influenza–lessons from the 1918 virus. *N Engl J Med*, 353:2209–2211, 2005. doi:10.1056/NEJMp058281.

[58] M. D. Lubeck, P. Palese, and J. L. Schulman. Nonrandom association of parental genes in influenza A virus recombinants. *Virology*, 95(1):269–274, 1979.

[59] M. Hatta and et al. Human influenza A viral genes responsible for the restriction of its replication in duck intestine. *Virology*, 295(2):250–255, 2002. doi:10.1006/viro.2002.1358.

[60] T. R. Maines and et al. Lack of transmission of H5N1 avianhuman reassortant influenza viruses in a ferret model. *PNAS*, 103(32):12121–12126, 2006. doi:10.1073/pnas.0605134103.

[61] C. Li, M. Hatta, S. Watanabe, G. Neumann, and Y. Kawaoka. Compatibility among polymerase subunit proteins is a restricting factor in reassortment between equine H7N7 and human H3N2 influenza viruses. *J Virol*, 82(23):11880–11888, 2008. doi:10.1128/JVI.01445-08.

[62] N. L. Varich, A. K. Gitelman, A. A. Shilov, Y. A. Smirnov, and N. V. Kaverin. Deviation from the random distribution pattern of influenza A virus gene segments in reassortants produced under non-selective conditions. *Arch Virol*, 153(6):1149–1154, 2008. doi:10.1007/s00705-008-0070-5.

[63] C. P. Octaviani, M. Ozawa, S. Yamada, H. Goto, and Y. Kawaoka. High level of genetic compatibility between swine-origin H1N1 and highly pathogenic avian H5N1 influenza viruses. *J Virol*, 84(20):10918–10922, 2010. doi:10.1128/JVI.01140-10.

[64] B. D. Greenbaum and et al. Viral reassortment as an information exchange between viral segments. *PNAS*, 109(9):3341–3346, 2012. doi:10.1073/pnas.1113300109.

[65] A. D. Neverov, K. V. Lezhnina, A. S. Kondrashov, and G. A. Bazykin. Intrasubtype reassortments cause adaptive amino acid replacements in H3N2 influenza genes. *PLoS Gen*, 10(1):e1004037, 2014. doi:10.1371/journal.pgen.1004037.

[66] G. Dudas, T. Bedford, S. Lycett, and A. Rambaut. Reassortment between influenza B lineages and the emergence of a coadapted PB1-PB2-HA gene complex. *Mol Biol Evol*, 32(1):162–72, 2015. doi:10.1093/molbev/msu287.

[67] A. C. McHardy and B. Adams. The role of genomics in tracking the evolution of influenza A virus. *PLoS Pathogens*, 5(10):1–6, 2009. doi:10.1371/journal.ppat.1000566.

[68] D. H. Morris, K. M. Gostic, S. Pompei, and et al. Predictive modeling of influenza shows the promise of applied evolutionary biology. *Trends in Microbiology*, 26(2):102–118, 2018. doi:10.1016/j.tim.2017.09.004.

[69] Y. Bao and et al. The influenza virus resource at the national center for biotechnology information. *J. Virol.*, 82(2):596–601, 2008. doi:10.1128/JVI.02005-07.

[70] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004. doi:10.1093/nar/gkh340.

[71] S. F. Altschul and et al. Basic Local Alignment Search Tool. *J Mol Biol*, 215(3):403–410, 1990. doi:10.1016/S0022-2836(05)80360-2.

[72] R. D. M. Page and E. Holmes. *Molecular Evolution. A phylogenetic approach*. Blackwell Science Ltd, 1998.

[73] J. Felsenstein. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410, 1978. doi:10.1093/sysbio/27.4.401.

[74] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 2014. doi:10.1093/bioinformatics/btu033.

[75] M.N. Price, P. S. Dehal, and A. P. Arkin. FastTree: Computing large minimum-evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26:1641–1650, 2009. doi:10.1093/molbev/msp077.

[76] G. K. Hirst. studies of antigenic differences among strains of influenza A by means of red cell agglutination. *J. Exp. Med.*, 78(5):407–423, 1943. doi:10.1084/jem.78.5.407.

[77] M. Łuksza and M. Lässig. A predictive fitness model for influenza. *Nature*, 507:57–61, 2014. doi:10.1038/nature13087.

[78] R. A. Neher, C. A. Russell, and B. I. Shraiman. Predicting evolution from the shape of genealogical trees. *eLife*, 3:e03568, 2014. doi:10.7554/eLife.03568.

[79] L. Steinbrück, T. R. Klingen, and A. C. McHardy. Computational prediction of vaccine strains for human influenza A (H3N2) viruses. *J. Virol.*, 88(20):12123–12132, 2014. doi:10.1128/JVI.01861-14.

[80] L. Steinbrück and A. C. McHardy. Allele dynamics plots for the study of evolutionary dynamics in viral populations. *Nucleic Acids Res.*, 39:e4, 2011. doi:10.1093/nar/gkq909.

[81] S. Bhatt, E. C. Holmes, and O. G. Pybus. The genomic rate of molecular adaptation of the human influenza A virus. *Mol. Biol. Evol.*, 28(9):2443–2451, 2011. doi:10.1093/molbev/msr044.

[82] N. J. McDonald, C. B. Smith, and N. J. Cox. Antigenic drift in the evolution of H1N1 influenza A viruses resulting from deletion of a single amino acid in the haemagglutinin gene. *J. Gen. Virol.*, 88:3209–3213, 2007. doi:10.1099/vir.0.83184-0.

[83] S. E. Lindstrom, Y. Hiromoto, R. Nerome, and et al. Phylogenetic analysis of the entire genome of influenza A (H3N2) viruses from Japan: evidence for genetic reassortment of the six internal genes. *J Virol*, 72(10):8021–8031, 1998.

[84] B. Schweiger, L. Bruns, and K. Meixenberger. Reassortment between human A(H3N2) viruses is an important evolutionary mechanism. *Vaccine*, 24(44-46):6683–6690, 2006. doi:10.1016/j.vaccine.2006.05.105.

[85] U. C. De Silva, H. Tanaka, S. Nakamura, N. Goto, and T. Yasunaga. A comprehensive analysis of reassortment in influenza A virus. *Biol. Open*, 1:385–390, 2012. doi:10.1242/bio.2012281.

[86] R. Rabadan, A. J. Levine, and M. Krasnitz. Non-random reassortment in human influenza A viruses. *Influenza Other Respir Viruses*, 2(1):9–22, 2008. doi:10.1111/j.1750-2659.2007.00030.x.

[87] I. Malijkovic Berry and et al. Frequency of influenza H3N2 intra-subtype reassortment: attributes and implications of reassortant spread. *BMC Biology*, 14:117, 2016. doi:10.1186/s12915-016-0337-3.

[88] M. I. Nelson, L. Simonsen, C. Viboud, M. A. Miller, J. Taylor, K. St. George, and et al. Stochastic processes are key determinants of short-term evolution in influenza A virus. *PLoS Pathog*, 2(12):e125, 2006. doi:10.1371/journal.ppat.0020125.

[89] M. I. Nelson, C. Viboud, L. Simonsen, and et al. Multiple reassortment events in the evolutionary history of H1N1 influenza A virus since 1918. *PLoS Pathog*, 4(2):e1000012, 2008. doi:10.1371/journal.ppat.1000012.

[90] N. Nagarajan and C. Kingsford. GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic Acids Res*, 39(6):e34, 2011. doi:10.1093/nar/gkq1232.

[91] V. Svinti, J. A. Cotton, and J. O. Mclnerney. New approaches for unravelling reassortment pathways. *BMC Evol Biol*, 13(1), 2013. doi:10.1186/1471-2148-13-1.

[92] A. Yurovsky and B. M. E. Moret. FluReF, an automated flu virus reassortment finder based on phylogenetic trees. *BMC Genomics*, 12:S3, 2011. doi:10.1186/1471-2164-12-S2-S3.

[93] Y. Suzuki. A phylogenetic approach to detecting reassortments in viruses with segmented viruses. *Gene*, 464(1-2):11–6, 2010. doi:10.1016/j.gene.2010.05.002.

[94] A. Pinsent, C. Fraser, N. M. Ferguson, and S. Riley. A systematic review of reported reassortant viral lineages of influenza A. *BMC Infectious Diseases*, 16(1):1–13, 2016. doi:10.1186/s12879-015-1298-9.

[95] A. S. Monto and et al. Antibody to influenza virus neuraminidase: an independent correlate of protection. *J Infect Dis*, 212(8):1191–1199, 2015. doi:10.1093/infdis/jiv195.

[96] R. B. Couch and et al. Antibody correlates and predictors of immunity to naturally occurring influenza in humans and the importance of antibody to the neuraminidase. *J Infect Dis*, 207(6):974–981, 2013. doi:10.1093/infdis/jis935.

[97] F. Tria, S. Pompei, and V. Loreto. Dynamically correlated mutations drive human influenza A evolution. *Scientific Reports*, 3(2705), 2013. doi:10.1038/srep02705.

[98] C. D. McWhite, A. G. Meyer, and C. O. Wilke. Sequence amplification via cell passaging creates spurious signals of positive adaptation in influenza virus H3N2 hemagglutinin. *Virus Evol*, 2(2):vew026, 2016. doi:10.1093/ve/vew026.

[99] T. Bedford, S. Cobey, and M. Pascual. Strength and tempo of selection revealed in viral gene genealogies. *BMC Evolutionary Biology*, 11:220, 2011. doi:10.1186/1471-2148-11-220.

[100] C. A. Russell and et al. The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320(5874):340–346, 2008. doi:10.1126/science.1154137.

[101] M. Lässig and M. Łuksza. Adaptive evolution: can we read the future from a tree? *eLife*, 3:e05060, 2014. doi:10.7554/eLife.05060.

[102] T. Stadler. On incomplete sampling under birthdeath models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58–66, 2009. doi:10.1016/j.jtbi.2009.07.018.

[103] T. Stadler. Sampling-through-time in birthdeath trees. *Journal of Theoretical Biology*, 267(3):396–404, 2010. doi:10.1016/j.jtbi.2010.09.010.

[104] WHO. Review of global influenza activity, October 2016-October 2017. *Weekly epidemiological record (WER)*, 92(50):761–780, December 2017.

[105] T. Bedford and R. A. Neher. Seasonal influenza circulation patterns and projections for Feb 2018 to Feb 2019. `https://nextflu.org/reports/feb-2018/`, February 2018.

[106] M. Łuksza. Plupredict. evolutionary predictions for influenza. `http://www.flupredict.uni-koeln.de/`.

[107] T. Bedford and R. A. Neher. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, 31(21):35463548, 2015.

[108] P. Zawadzki, M. S. Roberts, and F. M. Cohan. The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust. *Genetics*, 140(3):917–32, 1995.

[109] O. H. Ambur, S. A. Frye, M. Nilsen, E. Hovland, and T. Tønjum. Restriction and sequence alterations affect DNA uptake sequence-dependent transformation in neisseria meningitidis. *PLoS ONE*, 7(7):e39742, 2012. doi:10.1371/journal.pone.0039742.

[110] H. Gangel, C. Hepp, Müller S., E. R. Oldewurtel, F. E. Aas, M. Koomey, and et al. Concerted spatio-temporal dynamics of imported DNA and ComE DNA up-

take protein during gonococcal transformation. *PLoS Pathog*, 10(4):e1004043, 2014. doi:10.1371/journal.ppat.1004043.

List of Figures

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschlielich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Michael Lässig betreut worden.

Köln, May 2018

Mara Villa

Teilpublikationen:

M. Villa and M. Lässig. Fitness cost of reassortment in human influenza. *PLoS Pathogens*, 13(11):e1006685, 2017.