Population genomics of intrapatient HIV evolution

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der EBERHARD KARLS UNIVERSITÄT TÜBINGEN
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Fabio Zanini
aus Trento, Italien

Tübingen
2015

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	24.9.2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Dr. Richard Neher
2. Berichterstatter:	Prof. Dr. Daniel Huson

Abstract

The Human Immunodeficiency Virus 1 (HIV-1) is a rapidly evolving human retrovirus. HIV-1 nucleic acid sequences have been sampled from many patients, with mostly one sequence per patient, to characterize HIV-1 genetics and epidemiology. Ultimately, however, HIV-1 replicates and evolves during single infections that last for several years. In my doctorate I performed whole-genome longitudinal deep sequencing on several HIV-1 patients and developed experimental, theoretical, and computational methods to (i) characterize HIV-1 evolution within single infections, (ii) organise and share the collected genomic data with the research community, and (iii) simulate evolution of rapidly adapting organisms like HIV-1 *in silico*. First, I quantified a number of central properties of intrapatient HIV-1 evolution such as genetic diversity, evolutionary rate, linkage disequilibrium, mutation rate, strength and prevalence of positive and purifying selection, and influence of RNA secondary structures. Second, exploiting modern web technologies, I realized a web application that gives other researchers the chance to perform specific analyses on the same data set. Third, I coded a computer package, FFPopSim, to simulate the evolution of populations under selection; via a novel algorithm and a cross-language design, it has proven an ideal tool to bridge theoretical predictions and experimental results.

Zusammensfassung

Das Humane Immundefizienz-Virus 1 (HIV-1) ist ein schnell evolvierendes menschliches Retrovirus. HIV-1 Nukleinsäuresequenzen wurden von vielen Patienten – normalerweise eine Sequenz pro Patienten – analysiert, um die Genetik und Epidemiologie von HIV-1 zu charakterisieren. Jedoch reproduziert sich und evolviert HIV-1 innerhalb jeder einzelnen, jahrelang andauernden Infektion. Während meiner Promotion führte ich genomweites, longitudinales deep sequencing von mehreren durch HIV-1 infizierten Patienten durch. Ich entwickelte experimentelle, theoretische und bioinformatische Methoden, um (i) die Evolution von HIV-1 während der einzelnen Infektionen zu charakterisieren, (ii) die gesammelten genetischen Daten zu organisieren und mit der Forschungsgemeinschaft zu teilen und (iii) die Evolution von schnell adaptierenden Organismen wie HIV-1 *in silico* zu simulieren. Ich quantifizierte zentrale Eigenschaften der HIV-1 Evolution innerhalb der Patienten, unter Anderem genetische Diversität, Evolutionsrate, linkage disequilibrium, Mutationsrate, Stärke und Häufigkeit positiver und reinigender Selektion und der Einfluss sekundärer RNS-Strukturen. Außerdem habe ich anhand moderner Webtechnologien eine Webanwendung entwickelt, die es anderen Forschern ermöglicht, spezifische Analysen auf demselben Datensatz durchzuführen. Darüber hinaus programmierte ich ein Computerpaket, FF-PopSim, um die Evolution von Populationen unter Selektion zu simulieren; dank eines neuen Algorithmus und eines mehrsprachigen Designs hat es sich als ideales Mittel erwiesen, um theoretische Vorhersagen und experimentelle Ergebnisse zu vereinen.

Abstract

Il Virus dell'Immunodeficienza Acquisita 1 (HIV-1) è un retrovirus umano che evolve rapidamente. Sequence di acidi nucleici di HIV-1 sono state raccolte da molti pazienti, solitamente una sequenza per paziente, per caratterizzare la genetica e l'epidemiologia dell'HIV-1. Ciononostante, l'HIV-1 si replica ed evolve durante singole infezioni che durano molti anni. Nel mio dottorato, ho effettuato deep sequencing longitudinale e sul genoma completo su alcuni pazienti di HIV-1 ed ho sviluppato metodi sperimentali, teorici, e computazionali per (i) caratterizzare l'evoluzione dell'HIV-1 durante singole infezioni, (ii) organizzare e condividere i dati genomici raccolti con la comunità scientifica e (iii) simulare l'evoluzione di organismi che si adattano rapidamente, come l'HIV-1, *in silico*. Innanzitutto, ho quantificato molte proprietà centrali dell'evoluzione intrapaziente dell'HIV-1, come la diversità genetica, il rate di evoluzione, il linkage disequilibrium, il rate di mutazione, l'ampiezza e frequenza della selezione positiva e purificante e l'influenza delle strutture secondarie dell'RNA. Inoltre, utilizzando moderne tecnologie web, ho realizzato una applicazione web che fornisce ad altri ricercatori la possibilità di condurre specifiche analisi sullo stesso set di dati. Infine, ho programmato un pacchetto al computer, FFPopSim, per simulare l'evoluzione di popolazioni sotto l'effetto della selezione; grazie ad un nuovo algoritmo e un design multilingue, si è rivelato come uno strumento ideale per connettere le predizioni teoriche con i risultati sperimentali.

Acknowledgements

Many people have supported me during my doctorate.

Professionally, my advisor Richard Neher was always very encouraging, spectacularly clever, and understanding. Jan Albert, Lina Thebo, Johanna Brodin and other colleagues in Stockholm were both excellent collaborators and caring hosts during my two short stays in Sweden. I thank Jan for his remarkable appreciation of both clinical and biological aspects of HIV research, and his committment to our common project. I hope I learned some of the qualities that make both him and Richard outstanding scientists. I would also like to stress my appreciation to Lina Thebo (KI) and Christa Lanz (MPI) for their boundless passion, patience, and teaching skills, without which my samples would have never hit the short read archive. My co-supervisor Daniel Huson, the members of my Thesis Advisory Committee, my lab members, my colleagues and mentors at the Max-Planck Institute and University of Tuebingen were very helpful and dedicated to advancing research first and foremost. Julia Kamenz, Ole Herud, Stephanie Heinrich, Iuliia Boichenko, Andrey Fadeev, Diep Tran, Prateek Mahalwar, and Giovanna Capovilla deserve special thanks for teaching me lab techiques I could not believe I would ever learn. Julia Kamenz, Vadim Puller, and Emmanuel Bénard commented on preliminary versions of the thesis; thank you for that. Thanks also to the many, semi-anonymous members of the open source software community who have coded the tools that made my work possible at all.

Personally, I would like to mention my family members, friends, and loved ones for their unexhausted support and affection. Thank you.

Ringraziamenti

Molte persone mi hanno sostenuto durante il mio dottorato.

Professionalmente, il mio relatore Richard Neher è stato sempre molto incoraggiante, spettacolarmente intelligente, e comprensivo. Jan Albert, Lina Thebo, Johanna Brodin e gli altri colleghi a Stoccolma sono stati eccellenti collaboratori e perfetti ospiti durante le mie due brevi visite in Svezia. Vorrei ringraziare Jan per la sua impressionante padronanza degli aspetti sia clinici sia di base della ricerca sull'HIV, e per la sua dedizione al nostro progetto in comune. Spero di aver fatto tesoro di almeno alcune delle qualità che rendono sia lui sia Richard degli ottimi scienziati. Vorrei anche dedicare un grazie a Lina Thebo (KI) e Christa Lanz (MPI) per la loro passione illimitata, pazienza, e abilità come insegnanti, senza le quali i miei campioni non avrebbero mai raggiunto l'archivio delle sequenze. Il mio corelatore Daniel Huson, i membri del mio TAC, i membri del mio gruppo di ricerca, i miei colleghi e mentori all'MPI e all'Università di Tuebingen sono stati sempre di grande aiuto e dedicati al progresso della ricerca scientifica prima di tutto. Julia Kamenz, Ole Herud, Stephanie Heinrich, Iuliia Boichenko, Andrey Fadeev, Diep Tran, Prateek Mahalwar e Giovanna Capovilla meritano un ringraziamento speciale per avermi insegnato delle techiche di laboratorio che non avrei mai sospettato potessi imparare. Julia Kamenz, Vadim Puller, and Emmanuel Bénard hanno commentato delle versioni preliminari della tesi; grazie. Grazie anche ai molti membri semi-anonimi della comunità di programmi open source che hanno programmato gli strumenti che hanno reso il mio lavoro possibile.

Personalmente, vorrei menzionare la mia famiglia, i miei amici e cari per il loro sostegno inesauribile ed il loro affetto. Grazie.

Contents

List of Symbols

HIV	human immunodeficiency virus
SIV	simian immunodeficiency virus
AIDS	acquired immunodeficiency syndrome
MPI	Max Planck Insitute (for Developmental Biology)
LANL	Los Alamos National Laboratory
WHO	world health organization
RNA	ribonucleic acid
DNA	desoxyribonucleic acid
cDNA	complementary DNA
PCR	polymerase chain reaction
RT	reverse transcription/transcriptase
RT-PCR	reverse transcription followed by PCR
kb	kilobase
gag	Gag gene (HIV-1)
pol	Pol gene (HIV-1)
env	Envelope gene (HIV-1)
rev	Rev gene (HIV-1)
RRE	*rev* response element
LTR	long terminal repeat
CTL	cytotoxic T lymphocyte
MHC	major histocompatibility complex
LD	linkage disequilibrium
ν	single nucleotide variant frequency
μ	mutation rate per site
ρ	recombination rate per site
s	fitness effect (benefit or cost)

S	Shannon entropy (DNA site by site)
HTML	hypertext markup language
URL	uniform resource locator
AJAX	asynchronous JavaScript and XML
CSS	cascading style sheets
API	application programming interface
REST	representational state transfer
DOM	document object model

Chapter 1

Introduction

This chapter contains the motivation for the study, a brief introduction to HIV and to the basic population genetical concepts underlying the thesis' results. It also presents a summary of previous studies on intrapatient HIV evolution.

1.1 Motivation

A Human Immunodeficiency Virus (HIV) infection is an interesting research topic in evolution for two reasons. First, HIV is remarkably good at thriving via mutation, recombination and selection in the face of a sophisticated enemy, the human adaptive immune system. The characterization of intrapatient HIV evolution described in this thesis represents a step forward towards a better understanding of the coevolutionary dynamics between pathogen and immune system.

Second, HIV can be taken as a model organism to study the evolution of rapidly adapting populations. Short generation time, high mutation rate, short genome, complex natural (immune system) and artificial selection (drugs): all these features make HIV an ideal organism to test theoretical models of population genetics against. In particular, the large sequencing data set collected for this thesis is a unique resource for the evolutionary theory community by virtue of its longitudinal nature and sequencing depth.

1.2 HIV

The Human Immunodeficiency Virus (HIV) is a lentivirus related to the family of Simian Immunodeficiency Viruses (SIVs) that infect several species of monkeys. Fig. 1.1, which shows a phylogenetic tree of HIV and SIVs,

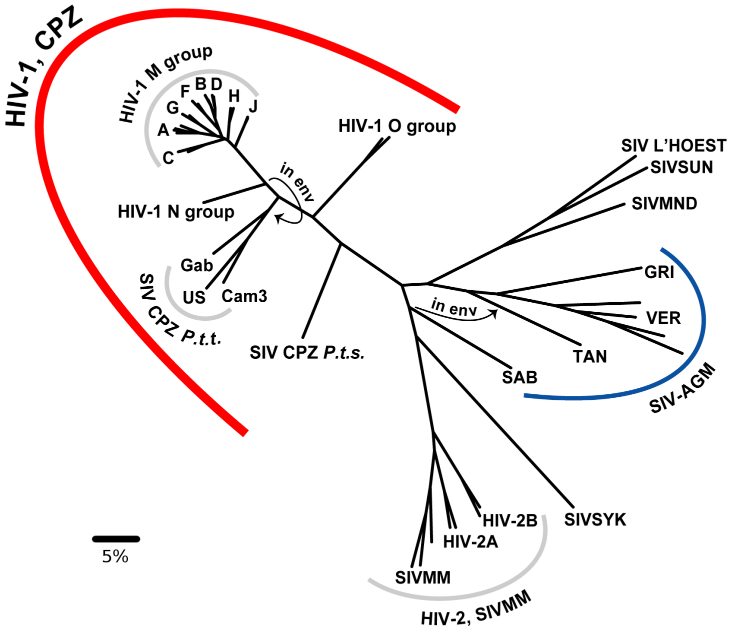


Fig. 1.1: Phylogenetic tree of HIV and SIV [1]. HIV-1 group M, in the top-left corner, is the focus of this thesis.

gives an impression of the range of genetic diversity spanned by this viral family [1]. Within the HIV-1 group M – the clade I focused on during my doctorate – any two genomes differ on average by around 10%.

HIV is the causative agent of the acquired immune deficiency syndrome (AIDS). It has been intensely studied since its discovery in the early 1980s [2, 3], when it spread across the globe into a pandemic that still kills around 2 million people every year according to the WHO [4]. Most infections are caused by HIV-1 group M viruses, with subtype B being prevalent in western countries [5]. An excellent account of the history of HIV has been recently published by Pepin [6].

In a susceptible host, HIV-1 establishes a systemic infection that cannot be cleared by the immune system. In untreated patients – the focus group of this thesis –, HIV-1 infection causes extensive immune dysregulation that after 5-10 years results in deep immunosuppression (AIDS) and eventually death. The reasons for failed clearance by the immune system are complex, but one key aspect of the problem is the ability of HIV-1 to rapidly diversify within the host into a genetically diverse population that includes *escape mutants*. These mutants are characterized by weaker affinities towards immune surveillance agents – antibodies or cytotoxic T-cells (CTLs) – and allow HIV-

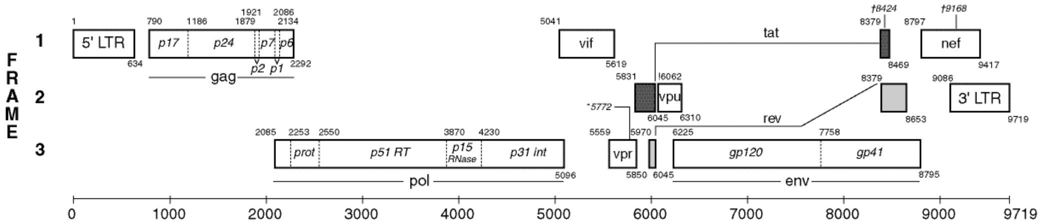


Fig. 1.2: Genome map of the HIV-1 reference strain HXB2. From [8].

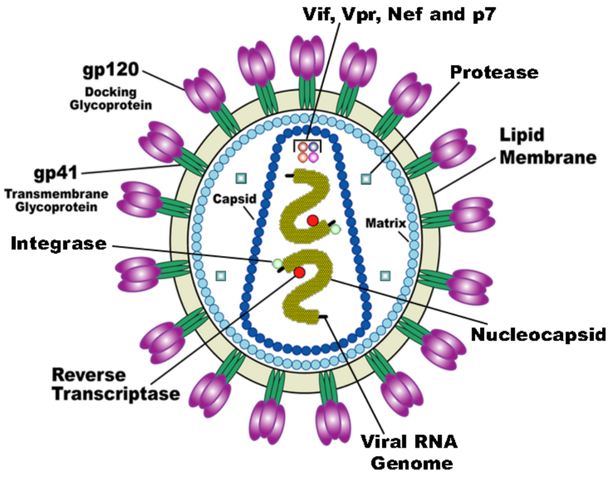


Fig. 1.3: Schematic illustration of an HIV virion showing the location of the genome and viral proteins. (Source: public domain.)

1 to persist despite the hostile host environment. **The process of genetic change that characterizes an HIV-1 infection is called intrapatient evolution and is the main topic of this thesis.**

The HIV-1 genome, represented in Fig. 1.2, is a single, positive-stranded RNA chain. It is 10kb long and codes for nine genes. The polyproteins *gag*, *pol* and *env* perform the basic functions – structural, enzymatic, and membrane respectively (see Fig. 1.3 for an illustration). The other six genes have accessory functions that suppress the host immune system and optimize gene expression [7].

A central feature of the HIV-1 genome is its functional density. Virtually the whole genome is occupied by exons, long terminal repeats (LTRs), or essential RNA structures, with frequent overlaps between them. As a consequence, the same genomic stretch may well serve different functions at the RNA, DNA, and protein level. How HIV-1 is able to accumulate mutations to escape immune recognition despite such a compact genome is one central question of this thesis. The immune surveillance processes driving intrapatient HIV-1 evolution, i.e. T-cell epitope and antibody binding, only affect viral proteins and have no direct interaction with viral RNA/DNA.

1.3 Population genetics of rapid adaptation

Population genetics studies the changes of allele frequency in a reproducing population under the effect of mutation, recombination, and selection. (Although more or different evolutionary forces may be relevant in general, the three listed above are the central ones as far as HIV-1 evolution is concerned.)

From a population genetics perspective, HIV-1 is an example of a rapidly adapting population. Adaptation in this context means: during an infection, the viral population survives by deploying escape mutants with better phenotype against the immune system (lower binding of antibodies or CTLs).

Population genetics of rapidly adapting populations has been attracting much interest especially since high-throughput sequencing allows experimental validation of theoretical models, filling the gap between mathematical theory and biological realizations of evolution. The same basic mathematical framework is applied to a variety of organisms, including but not restricted to pathogens (from HIV-1 to influenza, Hepatitis C, pathogenic bacteria) and model organisms for genetics (*S. cerevisiae*, *E. coli*) [9,10]. Applications within the HIV-1 field include better understanding of virus-host dynamics and rational design of vaccines [11,12].

1.3.1 Types of mutations and their selective effects

The key feature of rapidly adapting populations is the presence of genetic loci under *positive selection*. This means that a mutation at that locus increases fitness of the individual – such a mutation is termed *beneficial*. In HIV-1, escape mutations are generally considered to be beneficial.

Once it has appeared, the dynamics of a beneficial mutation has two phases. Let s be the selection coefficient (benefit) of the mutation. Early, the mutation is present in one or very few viruses and can be lost by stochastic effects, i.e. variation in offspring number due to noisy environmental factors. Later, if the mutant allele reaches around $1/s$ individuals (*establishment time* t_0), stochastic effects become negligibly small and the mutation frequency ν in the population starts to grow logistically:

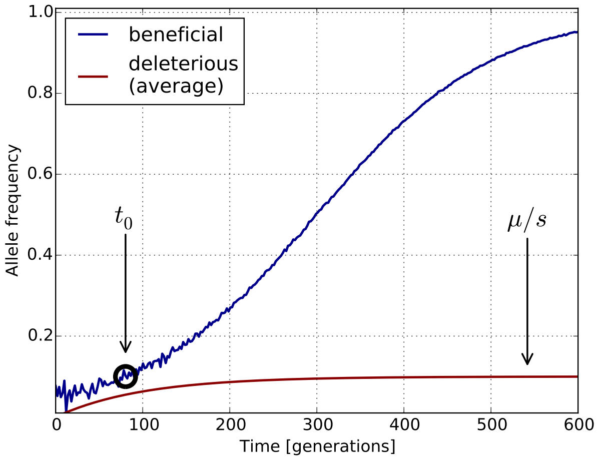


Fig. 1.4: Trajectories of mutant alleles with different fitness effects. Blue line: allele frequency trajectory for a typical beneficial mutation with selection coefficient of 1%. Red line: average trajectory for deleterious mutations with a cost of 1% and a mutation rate of 0.001.

$$\dot{\nu} = s\,\nu\,[1-\nu] \quad \Rightarrow \quad \nu(t) = \frac{1}{1 + e^{-(t-t_0)}\,\left[\nu_0^{-1} - 1\right]}, \tag{1.1}$$

where $\nu_0 := 1/sN$ is the *establishment frequency* in a population with N individuals [13]. From this moment, the rise of the mutant allele cannot be stopped and the mutation will reach fixation. See Fig. 1.4 for an illustration.

Although models with only beneficial mutations are interesting mathematically, most mutations in an actual HIV-1 population are not beneficial, because antibodies and CTLs target only few of the 10k possible genomic sites. Most mutations are unrelated to immune escape and come with a *fitness cost* instead, because they impair, to different degrees, correct biological functions such as enzymatic activity, folding, or biomolecular interactions. Such a *deleterious* mutation with a fitness cost s is mostly found at low frequencies. Because noise in offspring number is large at low frequencies, it is not possible to predict the dynamics of a single allele; the average frequency ν over many deleterious alleles, however, follows a simple mutation/selection balance:

$$\dot{\nu} = \mu - s\,\nu \quad \Rightarrow \quad \nu(t) = \frac{\mu}{s}\left[1 - e^{-st}\right], \tag{1.2}$$

where μ is the mutation rate, i.e. the rate of production of the mutant allele from the wildtype population, and time is calculated from a monomorphic population [13] – in HIV-1, start of the infection [14]. See Fig. 1.4 for an illustration.

Some mutations have little effect on fitness in either way and are termed *neutral*. They play an important role in phylogenetics because neutrality is a common assumption of coalescent models underlying tree reconstruction algorithms. Mathematically, neutral mutations follow a similar dynamics as slightly deleterious ones insofar as stochastic noise is the driving force of allele frequency changes. No deterministic approximation can be made for neutral mutations, but the long-term fate of the mutant allele is fixation or loss according to the following remark:

R 1. The **fixation probability** of a neutral allele at frequency ν is ν.

The reason is simple: in the distant future, only one of the currently extant lineages will survive. The probability that our neutral allele is present in that lineage is ν, and the presence of the mutation does not affect the choice of which lineage is actually surviving.

Beneficial mutations become fixed more frequently than neutral ones, deleterious mutations more rarely [13].

1.3.2 Recombination

HIV-1 is a facultatively recombining organism [15]. Recombination plays an important role for rapidly adapting populations, because it allows alleles under positive selection to increase in frequency more freely than in asexual populations. If the immune system deploys two antibodies at once, for instance, HIV-1 needs a double escape mutant to survive. Without having to wait for both mutations to happen on the same lineage, recombination opens up a new way to get a double mutant, thereby increasing the speed of adaptation [16].

A phenomenon that occurs in the presence of positive selection and low recombination is *hitchhiking*. This term indicates the rapid raise in frequency of neutral or slightly deleterious alleles in physical proximity of a beneficial mutation at the time of a selective sweep. Hitchhiking happens whenever these quasi-neutral alleles are already present in a lineage that acquires the beneficial mutation; because recombination is low, the frequencies of the beneficial allele and of the hitchhikers remain correlated over long times.

Pervasive hitchhiking, also called *genetic draft* [17, 18], produces stochastic noise on quasi-neutral allele frequencies which is different from the usual *genetic drift*, i.e. short-tailed offspring number variation. Nonetheless, both the average dynamics of deleterious alleles (eq. 1.2) and the fixation probability of neutral alleles (R1) are unaffected.

1.4 State of the field

1.4.1 Studies on intrapatient HIV-1 evolution

In order to study HIV-1 evolution during an infection, one needs two main ingredients. First, it is most convenient to work on longitudinal data, i.e. sequence information from samples taken at different times. (It is possible to use static snapshots as well, with more modest results.) Second, a reliable sequencing technology is required. Between the discovery of HIV-1 in 1983 and the first high-throughput sequencing machine in 2005, research in this field was limited to few sampling times per patient and few or even one HIV-1 sequence per time point. In a remarkable article from 1999, Shankarappa *et al.* followed 9-11 patients over several years of infection and clarified several basic aspects of intrapatient HIV-1 evolution [19]. They found that genetic divergence from the founder strain increases steadily whereas genetic diversity saturates after a few years. In the C2-V5 region of *env* they focused on, the average divergence rate was measured as 1% per year. The authors also used time-colored phylogenetic trees of intrapatient sequences to visualize evolutionary changes, and searched for correlations between genetic observables and clinical ones, such as CD4+ cell counts. The Shankarappa study has been a key data source for intrapatient HIV-1 evolution ever since, its main limitations being few sequences per sample and that only a small fraction of the HIV-1 genome was sequenced. Similar studies have been published along the same lines using low-throughput sequencing technology and suffered from similar limitations [20].

After the invention of high-throughput sequencing [21], several other studies using the new technology were published [22–24]. Despite great sequencing depth, however, they were limited in other respects: either short followup time, or very few patients. Researchers often focused on specific aspects of HIV-1 infection, such as tropism switch [23]; although interesting by themselves, those data can hardly be analyzed *a posteriori* as an unbiased sample of HIV-1 infections. The most complete study was performed by Henn *et al.* in 2012, with several time points and genomewide coverage; only one patient was followed though, and sampling was mostly restricted to the first year of infection [24].

1.4.2 Data sources

The Los Alamos National Laboratories HIV Database (LANL-HIV) is the *de facto* standard resource in the field [1]. Other public databases such as the Stanford HIV Database [25] and private collections from pharmaceutical companies [26] are mainly targeted at drug resistance testing; they only contain small parts of the HIV-1 genome (the enzymes) and have almost no longitudinal information. LANL-HIV aims mainly at cross-sectional data, i.e. at collecting sequences from many patients across the world. Although it does give access to the few published longitudinal data sets (e.g. the Shankarappa *et al.* data [19]), it provides little infrastructure for browsing the data along the time axis: one has to download the sequences first and then code some specific analysis and visualization software.

1.4.3 Software for rapidly adaptating populations

Computer simulation of population genetical evolution is a useful tool to link theory on abstract population with experiments on biological organisms. Especially under the assumption of rapid adaptation, many mathematical models are hard to solve analytically, so *in silico* exploration of the relevant parameter space is widely used either as a preliminary exploration tool or for validation. If experimental sequence data on rapidly adapting populations are available, which is rare, computer simulations can be used to distinguish generic evolutionary properties that can be captured by models from secundary biological specificities of the organism at hand.

Several simulation packages have been developed, with different goals in mind. Because positive selection strongly affects the shape of phylogenetic trees, coalescent approaches under neutral assumption are not appropriate (e.g. MS [27]).

Forward simulations are better suited for simulating HIV-1 evolution, at the price of increased computational cost. Valuable extant software that performs this task includes simuPop [28] and Nemo [29]. Both packages focus on diploid organisms and feature-completeness rather than speed. This is an issue for HIV-1 evolution because of two costly requirements, viz. large population sizes ($N \gtrsim 10^5$) and intermediate recombination that cannot be approximated by either asexual reproduction nor free recombination.

1.5 Motivation

My doctoral work has been directed at collecting data, developing tools, and analyzing sequences to improve our understanding of intrapatient HIV-1 evolution beyond the state of research outlined above.

First, I developed a new simulation software, FFPopSim, that can address the recombination problem with a much faster algorithm based on Fast Fourier transforms [30]. FFPopSim's approach is also insensitive to population size.

Second, I collected a new longitudinal sequence data set from 11 untreated HIV-1 infected patients that is by far the most complete data source on intrapatient evolution [31]. I analyzed these data and previously published ones to characterize various aspects of intrapatient HIV-1 evolution, including an in-depth analysis on fitness costs of synonymous mutations [32].

Third, I created a web application that exposes the new data set to the public using many different visualizations and data compilations, to allow efficient online browsing of the data features without need to download them [31].

Fourth, I developed support software tools to complement the above work and contributed to standard, open-source packages for bioinformatical data analysis, including matplotlib [33], pandas [34], and Biopython [35].

Chapter 2

Materials and methods

In my doctorate I have combined several approaches:

1. molecular biology experiments to collect sequencing data
2. computer analysis of HIV-1 sequences
3. computer simulation of rapidly adapting populations
4. web programming for HIV-1 longitudinal deep sequencing data

These methods are outlined in the following sections. For details about a specific published work, the reader is referred to my publications [30–32].

2.1 Molecular biology experiments

The experimental part of my doctorate was aimed at collecting longitudinal, deep sequencing data on intrapatient HIV-1 evolution. The protocol consists of several steps, the most relevant of which are illustrated in Fig. 2.1:

1. choice of appropriate patients
2. choice of primers for cDNA synthesis and amplification
3. collection of frozen plasma samples from blood banks
4. extraction of total RNA from patient plasma
5. one-step cDNA synthesis and PCR amplification of HIV-1 RNA
6. preparation of sequencing libraries
7. sequencing

Each of the steps will be briefly explained in the following sections.

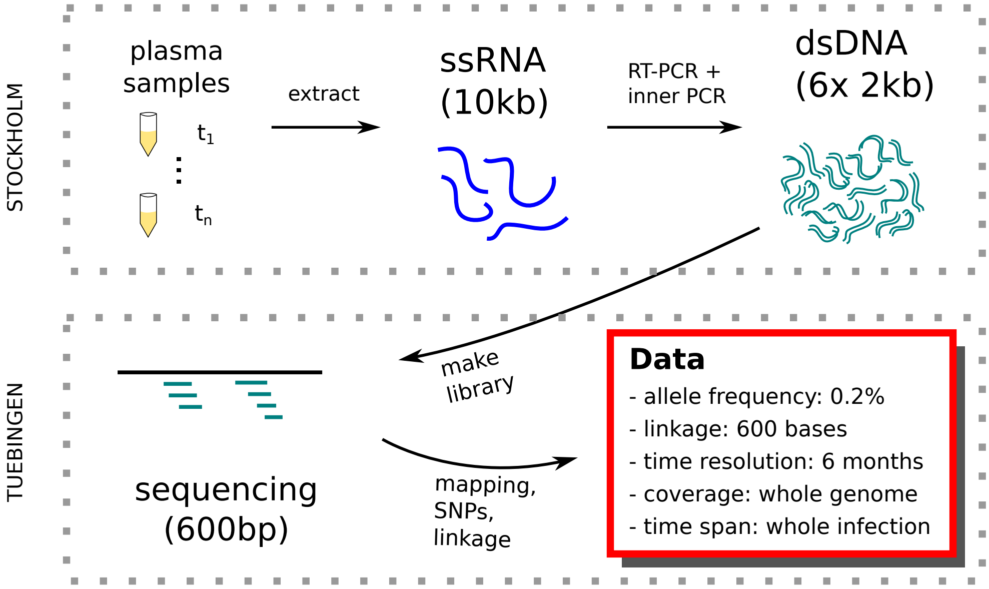


Fig. 2.1: Main steps of the sample preparation protocol for longitudinal HIV-1 high-throughput sequencing to study intrapatient evolution.

Choice of patients: Nowadays (2015), most HIV-1 positive individuals are subject to antiviral treatment early on, as it has been suggested that this preserves their immune system to some extent [36]. Patients under successful treatment, however, have extremely low viral titers in the blood, making treated patients a poor choice to study evolution [37]. However, historical samples are available; we identified patients with

- a long followup period without treatment;
- a relatively well defined time of infection;
- no special circumstances such as superinfection;
- successful therapy at the end of the study.

The list of patients is shown in the Resuts chapter, Table 3.1.

Primer design: As amplification of blood-borne HIV-1 RNA is required for sequencing, suitable primers had to be designed. Because the HIV-1 sample sequence is not known ahead of time, however, partially degenerate primers targeting conserved genomic regions at regular intervals must be carefully selected. This task is complicated by the sheer genetic diversity of HIV-1; good candidates are chosen among gene overlaps across the genome (e.g. *gag/pol*) and conserved RNA structures (e.g. *RRE*) [38]. The initial choice of primers was done by J. Brodin and J. Albert, collaborators at the Karolinska Institute. Upon poor performance in the RT-PCR, I redesigned a few of the primers. The complete list of primers is available upon request.

Collection of frozen plasma samples from blood banks: This part of the protocol was performed by J. Albert at the Karolinska Institute and involves careful choice of time points to sequence and empirical evaluation of the sample tubes to ensure *bona fide* good preservation of the genetic material. The time between two consecutive samples was typically 6-12 months, and patient follow-up was 5-8 years since infection. Ethical approval for the usage of patient samples was obtained and is available upon request.

Extraction of total RNA from patient plasma: For each sample, 400 μl of plasma (if available) was divided into two 200 μl aliquots. Total RNA was extracted using RNeasy Lipid Tissue Mini Kit (Qiagen Cat. No. 74804). Each aliquot was eluted twice with 50 μl RNase free water to maximize HIV RNA recovery. The four eluates were pooled giving a total volume of 200 μl of RNA per sample.

cDNA synthesis and PCR: The RNA was divided into twelve 14 μl aliquots for duplicate one-step RT-PCR with the outer primers for fragments 1 to 6 and Superscript III One-Step RT-PCR with Platinum Taq High Fidelity Enzyme Mix (Invitrogen, Carlsbad, California, US). The one-step RT-PCR was started with cDNA synthesis at 50°C for 30 min and denaturation step at 94°C for 2 min followed by 30 PCR cycles of denaturation at 94°C for 15 sec, annealing at 50°C for 30 sec and extension at 68°C for 90 sec and a final extension step at 68°C for 5 min. The amplification protocol is a compromise between high fidelity and high processivity. On the one hand, early misincorporations during RT-PCR represent an important source of errors; on the other, HIV-1 RNA contains RNA secondary structures that are hard to copy for slow, very high-fidelity polymerases [39]. Two reactions are run in parallel to reduce bias and the products mixed for downstream analysis. Aliquots from all PCR reactions are run on a gel for a qualitative check on PCR efficiency (presence of the expected band) and specificity (absence of additional bands).

Library preparation: A protocol based on the Tn5 retrotransposase – using Illumina's Nextera XT kit [40] – is chosen because it accepts tiny amounts of input DNA (100 pg to 1 ng). After attachment of the sequencing primers, however, the DNA is size selected for inserts between 400 and 700 bp via SageScience's BluePippin instrument. This allows for larger inserts than standard Nextera XT, improving the homogeneity and quality of the library and providing more linkage information.

Sequencing: Sequencing is performed on Illumina's MiSeq instrument, which yields relatively long reads (250 to 300 bp). Using the "paired end" option, each insert is sequenced from both ends, so that reads come in overlapping pairs of length 400 to 600 bp. PacBio's RS II instrument was tested but required more input DNA and had a much lower throughput.

2.2 Computer analysis of HIV-1 sequences

During my doctorate, I have analyzed both population (shallow) and deep sequencing data, with different pipelines. I have also extensively analyzed cross-sectional data sets. Most of the analyses are coded in Python 2.7. Speed-critical parts are coded in C/C++. I made extensive use, both at the Python and C level, of standard packages including the following:

- numpy/scipy [41]
- matplotlib [33]
- biopython [35]
- pandas [34]
- pysam/samtools [42]

The number of lines of code for the results of this thesis is about 10^5. The code is available on request but not simply public, because it contains personal data related to the patients.

2.2.1 Population sequencing

I analyzed previously published population sequencing data for one of my publications [32]. This kind of data is relatively straightforward to analyze, because one obtains only a few sequences per time point. I typically made use of the following kinds of data structures:

1. consensus sequences
2. allele frequency trajectories
3. multiple sequence alignments
4. correlation matrix trajectories

Consensus sequences: For every sample, a consensus sequence was calculated, to be used as a reference to count mutations from. The consensus of the earliest sample in each patient plays a special role, because it is the closest observed sequence to the HIV-1 founder strain in that patient.

Allele frequency trajectories: For every nucleotide plus gaps, for every time point, and for every genomic site of interest, the frequency of that nucleotide is stored. The data shape is a three dimensional array. Linkage information between sites is largely lost, but the data can be efficiently sliced by position, providing speed and statistical power to the analyses (as far as positions can be considered independent of each other).

Multiple sequence alignments: This data structure is used mainly to represent the evolution of entire genomic stretches of length between 10 and 10k bases (whole-genome). It keeps the whole linkage information, but it is prone to errors: every sequencing error generates in principle a new sequence (haplotype).

Correlation matrix trajectories: Similar to the allele frequency trajectory cube, but for pair of alleles. This structure keeps some linkage information but is not very sensitive to errors.

2.2.2 Deep sequencing

Data structures

The basic data structure for my high-throughput sequencing data is the *read pair*. The sequencing library protocol includes a fragmentation step (the Tn5 transposase activity) that is unspecific in terms of genomic position. After sequencing, I obtained (in total) around 10^8 read pairs that start at random positions in the HIV-1 genome. Between the two reads within pairs there is variable coverage, between 200 bp overlap to 200 bp gap, corresponding to an insert size distribution from 350 bp to 700 bp. The amount of overlap depends on the random cutting of the transposase as well as on the quality scores of the read ends. Because I call minor alleles as rare as 0.2%, phred quality scores above 30 are accepted and the read ends often need trimming, reducing the overlap. The broad overlap distribution also forbids a simple merge of the reads in the pair.

Because of the sheer amount of reads, the analysis must be performed efficiently. Most software tools for next-generation sequencing analysis are designed under the basic assumption that the genome is much longer than the average coverage, and that mutations are rare. The *samtools* pipeline [42], for instance, includes commands to query reads that are partially overlapping with a genomic region. A key command in long-genome SNP analyses (e.g. human), it is virtually useless for HIV-1, because the genome is only 10 kb long and at every single position many mutations are observed. By the same token the VCF format is not well suited for deep sequencing analyses such as the subject of this thesis.

For the reasons above, I performed the data analysis using the following basic data structures (i) lists of read pairs; (ii) the matrix-based formats also used for the population sequencing analyses. The latter were precomputed from the read pairs and saved to disk.

Data preparation

A characteristic of high-throughput data is that a small fraction of outlier sequences, either contaminants or otherwise error-rich, is always present in the raw data. Such problems were especially hard to solve in my doctorate work because HIV-1 has a high genetic diversity by itself, and this diversity is strongly dependent on the genomic region of interest.

I have therefore developed a data cleaning pipeline that ensures high-fidelity of the read pairs. The basic steps of the pipeline are the following:

1. preliminary rough mapping to a reference HIV-1 sequence, HXB2;

2. quality trimming and assignment of each read pair to its PCR of origin;

3. consensus building within each PCR amplicon;

4. remapping of all assigned reads against their consensi;

5. semi-automatic filtering by distance from the consensus;

6. remapping against the patient's initial consensus (founder strain);

7. thorough cross-contamination filtering.

This pipeline identified many sources of foreign reads that would have invalidated our scientific conclusions, for instance:

- cross-contamination during cDNA synthesis, PCR, or library prep;
- illegitimate recombination between the overlapping PCR amplicons;
- artifacts of the sequencing platform.

The mapping was performed using the published software Stampy [43], which is a probabilistic mapper that works well with highly variable organisms such as HIV-1. Although not the fastest mapper available, it does not require strict thresholds in terms of maximal number of mismatches. Other mappers such as BWA [44] were tested and yielded similar results, but it was hard to decide on the mismatch thresholds as the level of genetic diversity was unknown; it was actually one of the research questions addressed by my study.

The consensus building algorithm was not trivial because our data combined several peculiarities:

- large genetic variation, including short indels;
- no full coverage by a single read pair;
- extreme coverage fluctuations

In particular, coverage could typically oscillate, along the genome, from 0 to 10^5 and back several times within the same sample, based on cDNA accidents, PCR efficiency, retrotransposase preference, and sequencing quality. The final algorithm makes use of the rough mapping information to extract small subsample of reads (around 30) fully covering a genomic sliding window of around 150 bp in size; these reads are trimmed and multiple sequence aligned; a consensus within each window is obtained; a new window is set 50 bp downstream; subsequent overlapping windows are pair aligned and a final consensus is produced.

The code for the data cleaning pipeline amounts to 50k lines of mainly Python 2 and is available upon request.

2.3 Web programming

I created a web application providing access to results, informative plots, and data of my deep sequencing study on intrapatient HIV-1 evolution. The data

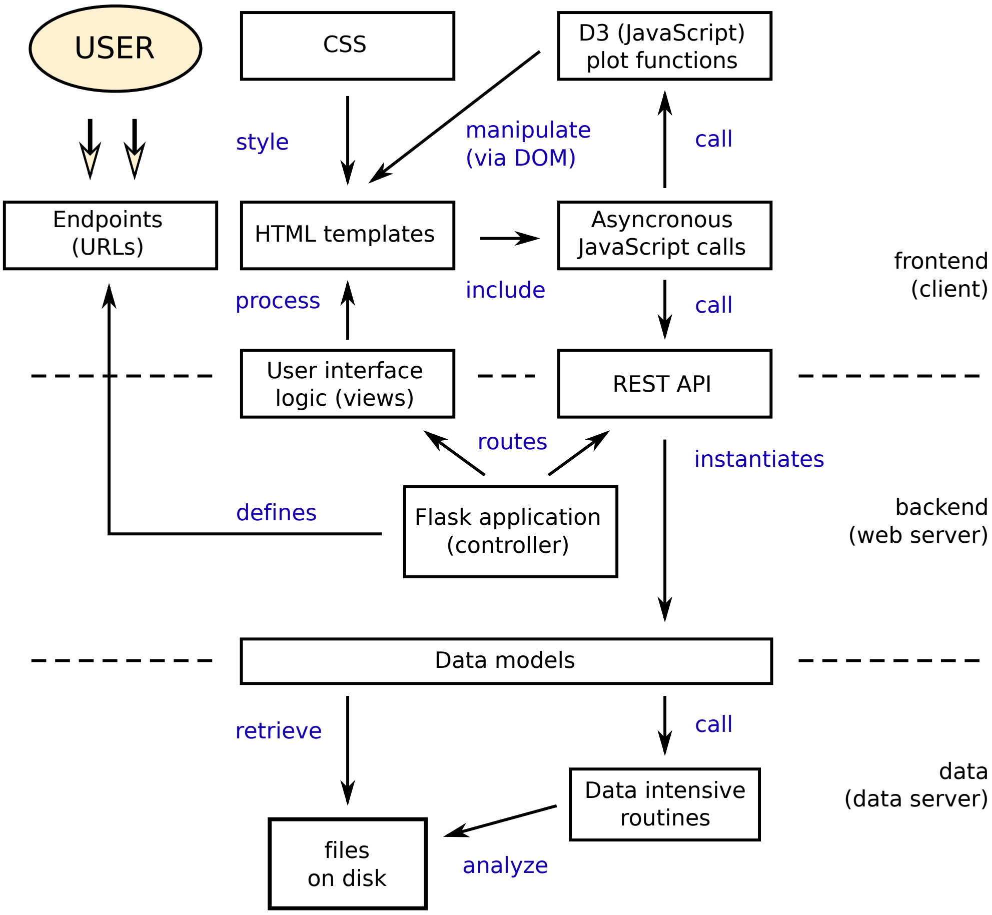


Fig. 2.2: Schematic block view of the web application. The black blocks describe the various components, the blue verbs the actions between them. The arrows indicate the code flow as the user requests a resource.

is stratified in several ways, notably by patient, by sample, by PCR amplicon. For this reason, I needed to create dynamic web pages that generated the content based on what part of the data set is being looked at, and an interactive frontend that enables intuitive browsing across the strata.

I designed the web application using a layered approach centered around the model-view-controller pattern. The layer stack is shown in a schematic view in Fig. 2.2. The application backend framework is Flask [45], a standard lightweight web framework for Python developers. In order to insulate different parts of the application for faster development, I defined a number of blueprints, i.e. plugins for the application that live in small containers (in terms of both files and namespaces). This choice made the collaboration with a master student in the lab, Bianca Regenbogen, much easier to set up.

One of the main strengths of the application are the interactive plots of various observables, such as viral load, allele frequency trajectories, and phylogenetic trees. They are created via a convenient JavaScript plotting library, D3 [46]. Although quite low-level, D3 is very flexible and provides support data structures and functions for both tabular and tree-like plot data. Technically, the charting routines are implemented as closure objects. Attributes of such an object include the chart size, color scheme, and plot representation (e.g. radial or horizontal for trees). The use of closure objects makes it easier to recycle charts for different web pages as compared to simple charting functions [47].

Another central node in the web application is the REST API, which implements a standard interface between the client and the actual data. The API is used internally within the HTML templates via AJAX calls, but is also publicly documented and open for the user to retrieve data directly, in JSON format.

As far as data manipulation and serving itself is concerned, the web application was designed for few users at a time, so there was no need for parallel-efficient solutions, such as a fully asyncronous framework (e.g. nodejs [48] or go [49]) or a queueing system for the data-intensive routines (e.g. RabbitMQ [50]). Such backend solutions could be swapped in the future, if necessary, without much affecting the frontend architecture. For the time being, the web server is being used as a data server.

2.4 Simulation of recombining, rapidly adapting populations

There are two basic ways of simulating evolution forward in time. Either all extant individuals are tracked one by one (individual-based simulations) or the whole distribution of possible genotypes is tracked together with the number of individuals for each genotype (distribution-based simulations). In FFPopSim, the software I developed for simulating HIV-1 populations, I made use of both approaches [30]. The program is designed in an object-oriented fashion and contains two main classes to represent an evolving population, one for individual-based (haploid highd) and one for distribution-based simulations (haploid lowd).

2.4.1 Individual-based simulations

The individual-based class represent a very fast simulation software for HIV-1 like populations when many genetic loci have to be simulated. The runtime complexity scales roughly like $\mathcal{O}(N\ L)$ per generation, where N is the population size and L the genome length. Thanks to its efficient dual-language design (C++/Python), FFPopSim can simulate realistic HIV-1 populations with $L = 10^4$ and $N = 10^6$ while keeping simple at the user end.

In order to optimize the usage of memory resources, two strategies are applied:

1. individuals are grouped in monomorphic clones, and a new clone is only made when a new mutant or recombinant lineage is started;

2. because the survival chances of a new clone are small (genetic drift), the pointer for a clone that goes extinct is recycled for a new mutant/recombinant.

2.4.2 Distribution-based simulations

Distribution-based simulations are better suited for large populations in which only a few genetic loci need to be followed. Given a biallelic genome, the population is represented by a distribution of abundances on the binary hypercube. This is illustrated, together with the operations of mutation and recombination, in Fig. 2.3.

To simulate evolution, occupation of each of the 2^L possible genotypes is recorded in time. The main challenge in simulating recombining, rapidly adapting populations is the runtime efficiency of the routine computing the

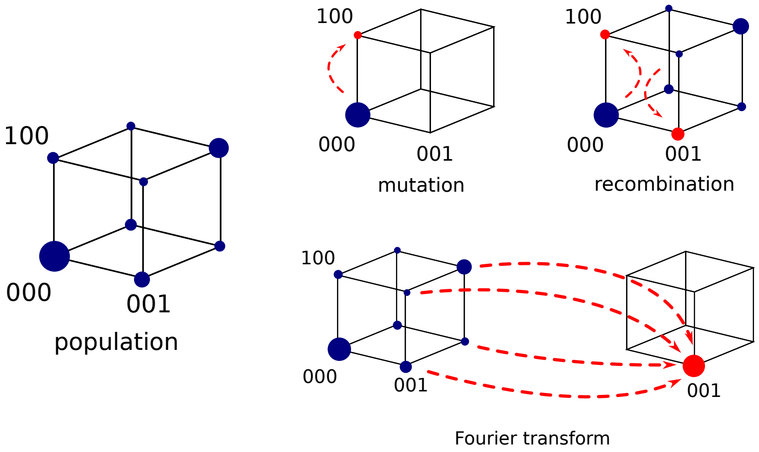


Fig. 2.3: Left panel: schematic illustration of a population on a 3D hypercube (cube). Each vertex is a biallelic genotype, and the balls indicate different number of individuals with different genotypes. Top panel: examples of mutation and recombination on the cube. Bottom panel: part of the Fourier transformation, condensing the right side of the cube into a single coefficient.

new recombinants at every generation. The runtime complexity of the naive algorithm for a biallelic genome of length L is $\mathcal{O}(8^L)$:

- 2^L possible fathers
- 2^L possible mothers
- given a pair of parents, there are 2^L possible inheritance patterns (each locus may come from either parent)

As explained in the paper, the distribution $R(g)$ of recombinant gametes would naively be computed as follows [30]:

$$R(g) = \sum_{\xi} \sum_{g'} C(\xi) P(g^m) P(g^p), \tag{2.1}$$

where g is the recombinant genotype (binary vector), ξ specifies the particular way the parental genomes are combined: $\xi_i = 0$ (resp. 1) if locus i is derived from the mother (resp. father). The genotype g' is summed over; it represents the part of the maternal (g^m) and paternal (g^p) genotypes that is not passed on to the offspring.

FFPopSim adopts an approach based on Fast Fourier transforms that reduces this complexity to $\mathcal{O}(3^L)$. An illustration of the Fourier transformation on the binary hypercube is shown in Fig. 2.3. The basic insight comes from the observation that, at every recombination event, any parent allele that is not picked by the offspring need not be calculated. For instance, given an inheritance pattern in which only the first locus is inherited from the mother (the other $L-1$ come from the father), we need not iterate over the distribution of all 2^L possible mothers, but only over the *marginal* at the first locus, i.e. over 2 mother-marginals, saving $2^L - 2$ computations – for $L = 15$ loci, a save of 32766 out of 32768 computations.

Summary of the Fast Fourier algorithm

The key change is to swap the order of the sums over inheritance patterns with the sums over parent genotypes. We can decompose each parent into successful loci that made it into the offspring and wasted loci, as follows: $g^p = \xi \wedge g + \overline{\xi} \wedge g'$ and $g^m = \overline{\xi} \wedge g + \xi \wedge g'$, where $\wedge$ and a bar over a variable indicate respectively the elementwise AND and NOT operators (i.e., $\overline{\xi_i} := 1-\xi_i$). The function C assigns a probability to each inheritance pattern. Depending on whether the entire population undergoes sexual reproduction or only a fraction r of it, the entire population or a fraction r is replaced with $R(g)$. The central ingredient for the efficient computation of $R(g)$ is the Fourier decomposition of genotypes:

$$P(g) = f^{(0)} + \sum_i t_i f_i^{(1)} + \sum_{i<j} t_i t_j f_{ij}^{(2)} + \cdots \tag{2.2}$$

where $t_i = 1$ if g is mutated at site i or $t_i = -1$ if g has the wildtype allele. There are $\binom{L}{k}$ coefficients $f^{(k)}_{i_1 \ldots i_k}$ for every subset of k loci out of L loci, so in total 2^L coefficients [51] . A coefficient $f^{(k)}_{i_1 \ldots i_k}$ is uniquely specified by

$$f^{(k)}_{i_1 \ldots i_k} = 2^{-L} \sum_g t_{i_1} \ldots t_{i_k} P(g). \tag{2.3}$$

Note that the Fourier transform of any function in this space can be achieved in $L\ 2^L$ computations via the Fast Fourier algorithm and is not entering the final runtime complexity of the recombination algorithm.

The generic Fourier coefficient of $R(g)$ is given by

$$r^{(k)}_{i_1 \ldots i_k} = 2^{-L} \sum_g t_{i_1} \ldots t_{i_k} \left(\sum_\xi \sum_{g'} C(\xi) P(g^m) P(g^p) \right) \tag{2.4}$$

Just as g^p and g^m can be expressed as a combination of g and g', we can invert the relation and express the generic t_i as a function of g^p and g^m, as follows: $t_i = \xi_i t_i^m + \overline{\xi_i}\, t_i^p$. Using this new basis and exchanging the order of summations, we obtain

$$r^{(k)}_{i_1 \ldots i_k} = 2^{-L} \sum_\xi C(\xi) \sum_{g^m, g^p} (\xi_{i_1} t^m_{i_1} + \overline{\xi_{i_1}}\, t^p_{i_1}) \ldots (\xi_{i_k} t^m_{i_k} + \overline{\xi_{i_k}}\, t^p_{i_k}) P(g^m) P(g^p). \tag{2.5}$$

Notice that $C(\xi)$ can be pulled out of the two inner sums, because the odds of inheriting a certain locus by the mother/father is independent of what their genetic makeup looks like. Next we expand the product and introduce new labels for compactness,

$$\begin{aligned} r^{(k)}_{i_1 \ldots i_k} = 2^{-L} \sum_\xi C(\xi) \sum_{g^m, g^p} P(g^m) P(g^p) \\ \sum_{l=0}^{k} \sum_{\{j_i\},\{h_i\}} \xi_{j_1} \ldots \xi_{j_l} \overline{\xi_{h_1}} \ldots \overline{\xi_{h_{k-l}}}\, t^m_{j_1} \ldots t^m_{j_l} t^p_{h_1} \ldots t^p_{h_{k-l}}, \end{aligned} \tag{2.6}$$

where l is the number of loci inherited from the mother among the k in $(i_1, \ldots, i_k)$. l runs from 0 (everything happens to be contributed by the father) to k (everything from the mother). $\{j_i\}$ and $\{h_i\}$ are all (unordered) partitions of i into sets of size l and $k-l$, respectively. Now we can group all ξ_i in the inner sum with $C(\xi)$, all t_i^m with $P(g^m)$, and all t_i^p with $P(g^p)$. The three sums (over ξ, g^m, and g^p) are now completely decoupled. Moreover, the two sums over the parental genotypes happen to be the Fourier decomposition of $P(g)$. Hence, we have

$$r^{(k)}_{i_1\dots i_k} = \sum_{l=0}^{k} \sum_{\{j_i\},\{h_i\}} C^{(k)}_{j_1\dots j_l, h_1\dots h_{k-l}} p^{(k)}_{j_1\dots j_l} p^{(k-l)}_{h_1\dots h_{k-l}}. \tag{2.7}$$

The quantity

$$C^{(k)}_{j_1\dots j_l, h_1\dots h_{k-l}} = \sum_{\xi} C(\xi) \xi_{j_1} \dots \xi_{j_l} \overline{\xi_{h_1}} \dots \overline{\xi_{h_{k-l}}} \tag{2.8}$$

can be calculated efficiently, for each pair of partitions $(\{j_i\}, \{h_i\})$, by realizing that (a) for $k = L$, there is exactly one term in the sum on the right that is non-zero and (b) all lower-order terms can be calculated by successive marginalizations over unobserved loci. For instance, let us assume that $k = L-1$ and that the only missing locus is the m-th one. We can compute

$$C^{(L-1)}_{j_1\dots j_l, h_1\dots h_{L-1-l}} = C^{(L)}_{j_1\dots j_l\, m, h_1\dots h_{L-1-l}} + C^{(L)}_{j_1\dots j_l, h_1\dots h_{L-1-l}\, m}. \tag{2.9}$$

There are $\binom{L}{k}$ ways of choosing k loci out of L, which can be inherited in 2^k different ways (the partitions in j and h in Eq. (2.8)) such that the total number of coefficients is 3^L. Note that these coefficients are only calculated when the recombination rates change. Furthermore, this can be done for completely arbitrary recombination patterns, not necessarily only those with independent recombination events at different loci [30].

Chapter 3

Results and discussion

The results of my work as a PhD candidate have been or are being published on peer-reviewed scientific journals, in the following articles, of which I am first and main author (sorted by publication date):

- FFPopSim: An efficient forward simulation package for the evolution of large populations, *Bioinformatics* (2012) [30];

- Quantifying Selection against Synonymous Mutations in HIV-1 env Evolution, *Journal of Virology* (2013) [32];

- Longitudinal whole-genome deep sequencing of HIV-1 reveals mutational and selective processes during infection, *submitted* (2015) [31].

The last project includes the web application for presenting the deep sequencing data. The results of my doctorate are presented in the following sections, divided by topic.

3.1 Intrapatient HIV-1 evolution

3.1.1 Deep sequencing data set

During my doctorate, I collected whole-genome deep sequencing data from longitudinal serum samples from 11 untreated patients. See the Methods chapter for details on the data collection and sequencing protocol, reads mapping, and filtering. The only similar previously published data, by Henn *et al.*, is limited to one patient and focuses on acute infection [24].

These data allowed me to reach the conclusions about the evolution and biology of HIV-1 outlined below. Beyond these, however, the sequencing data stands as a result by itself. It appears likely that other researchers will

Patient	Gender	Transmission route	Subtype	Age* [years]	Fiebig stage*	BED* [ODn]
p1	F	HET	01_AE	37	IV	0.41
p2	M	MSM	B	32	V	0.17
p3	M	MSM	B	52	VI	0.89
p4	M	MSM	B	29	V	0.17
p5	M	MSM	B	38	III-IV	n.a.
p6	M	HET	C	31	IV	0.29
p7	M	MSM	B	25	V-VI	0.95
p8	M	MSM	B	35	V	0.15
p9	M	MSM	B	32	VI	0.27
p10	M	MSM	B	34	II	0.10
p11	M	MSM	B	53	VI	1.22

Patient	No. of samples	First sample [days]	Last sample [years]	HLA type		
				A	B	C
p1	12	49	8.0	02/02	08/15	03/06
p2	6	74	5.5	01/24	08/39	07/12
p3	10	104	8.3	02/11	15/44	03/16
p4	8	78	8.3	02/24	27/40	01/03
p5	7	132	5.9	03/33	14/58	03/08
p6	7	46	7.0	02/02	44/51	05/16
p7	10	2248	15.9	02/02	15/27	01/03
p8	7	64	6.0	03/32	07/40	02/07
p9	8	106	8.1	25/32	07/44	04/07
p10	9	18	6.1	32/32	44/50	06/16
p11	7	167	5.5	02/32	39/44	05/12

Table 3.1: Summary of patient characteristics. Sample times from estimated date of infection. *, at time of first sample; MSM, men who have sex with men; HET, heterosexual.

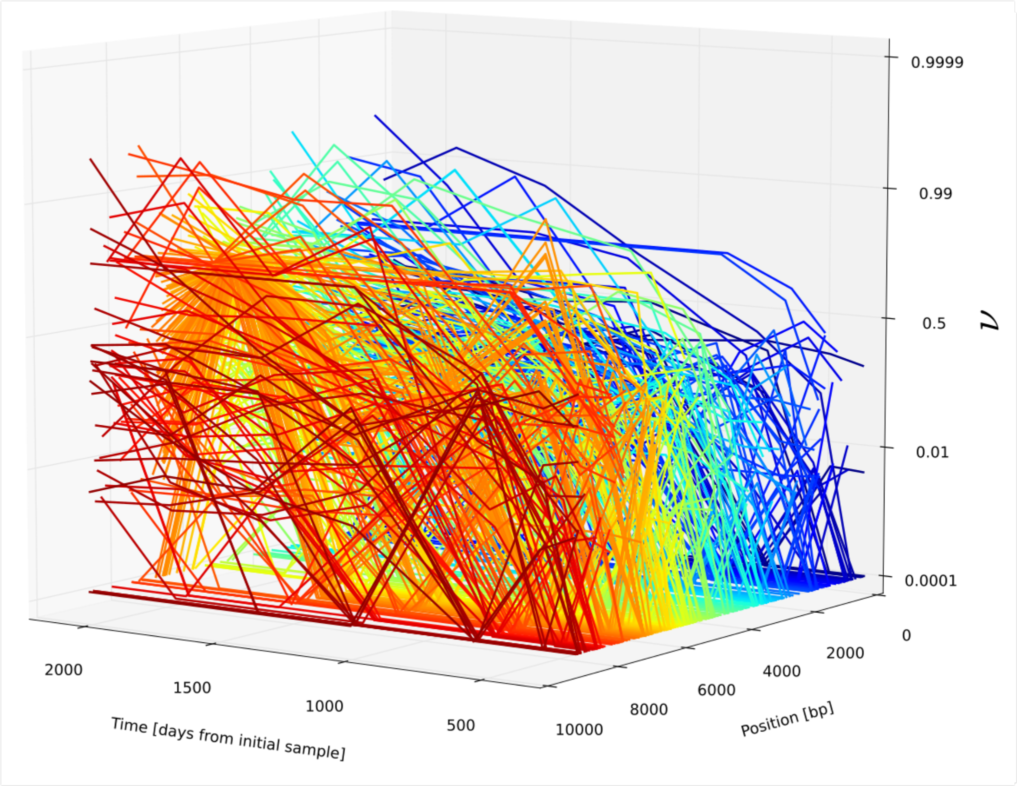


Fig. 3.1: Allele frequency trajectories for patient p11. Each line indicates a mutant nucleotide at a single site in the HIV-1 genome, with colors going through a rainbow spectrum from blue to red according to the position of the site in the HIV-1 genome. Note the logit vertical scale that expands the dynamic range in the neighborhood of 0 and 1. Polymorphisms are widespread all across the HIV-1 genome during the whole infection.

use the data to study specific aspects of HIV-1 evolution that are beyond the scope of my thesis, for instance evolution of RNA structure and immune epitopes. To foster research in this direction, I created the web application presented below.

3.1.2 Allele frequency trajectories and phylogenetic trees

A basic intuition on the evolutionary processes at work during an HIV-1 infection can be gained by considering a simple representation of the sequencing data: **allele frequency trajectories**. An illustration for patient p11 is shown in Fig. 3.1. For each position in the HIV-1 genome, any of the

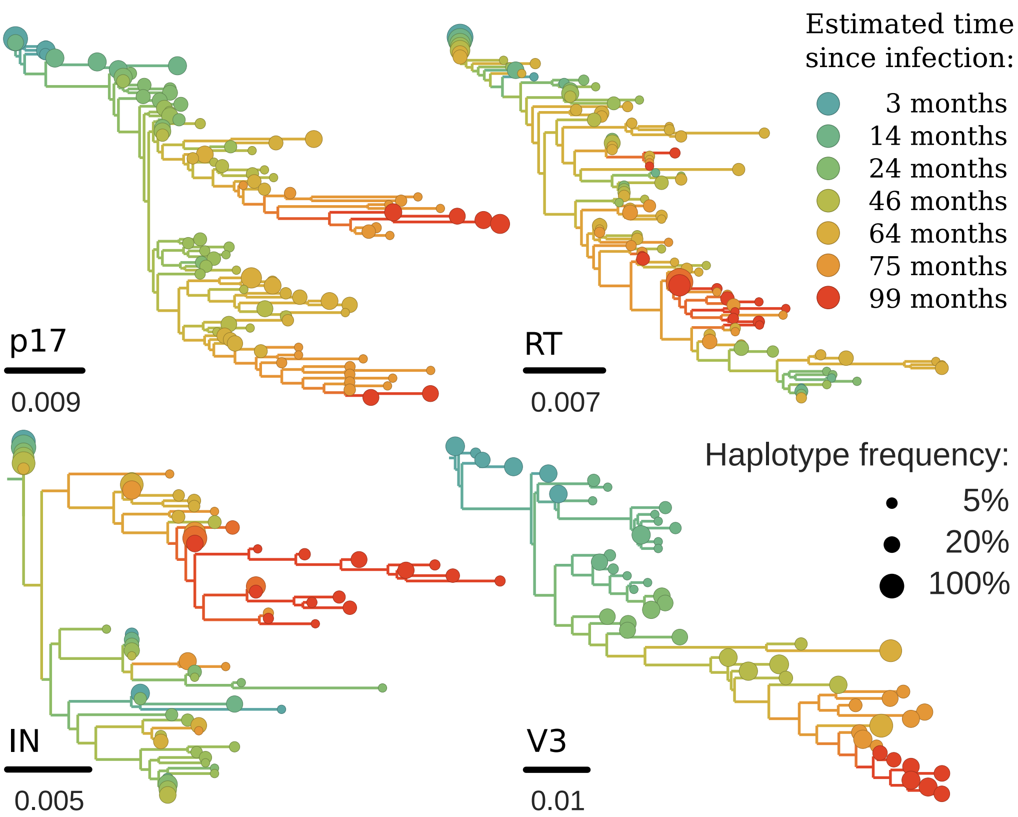


Fig. 3.2: Phylogenetic trees from patient p3 and different genomic regions: p17 (top left), reverse transcriptase (first 350 bp, top right), integrase (position: 351 to 700 bp, bottom left), V3 loop in env (bottom right). Clearly, different regions evolve according to different evolutionary processes. The trees are computed with FastTree [52].

four nucleotides (**alleles**) can in principle be found as time passes. The majority nucleotide at the first time point is called **ancestral allele**, any other one **derived allele**. At any time, an allele can be found in a certain fraction of the viral population, i.e. of the sequencing reads, between 0 and 100%: this fraction is called **allele frequency** and often indicated with the greek letter ν. The longitudinal collection of frequencies for a single allele is called **allele frequency trajectory**. The figure shows one line (allele frequency trajectory) for each derived allele that reaches a frequency of 10% at least once during the infection – alleles that are rarer than that are not shown for clarity.

A complementary representation of the genetic diversity of HIV-1 is given by phylogenetic trees. Whereas allele frequency trajectories focus on single nucleotides, a tree aims at establishing a global mathematical description of the evolution extended stretches of the HIV-1 genome. The key advantage is that linkage information is preserved. The insert sizes of our sequences are only 500-600 bp, so I could not reconstruct a whole-genome phylogenetic tree. This should not be considered a drawback, because HIV-1 recombines with a rate around 10^{-5} per base per day [53, 54], so that linkage is preserved only over $\sim$ 100 bp over a time scale of years (see also LD plot below). Having genome-wide data is nonetheless key because it enables comparisons between genomic regions. In Fig. 3.2, I show several trees reconstructed from different genomic regions in patient p3, built with FastTree [52].

Each unique sequence in such regions, of length around 350 to 400 bp, is termed **haplotype**, and its frequency in the viral population, i.e. in the reads, at any time point is called **haplotype frequency**. The timeline of haplotype frequencies is called, in parallele to single nucleotide alleles, **haplotype frequency trajectory**, and an illustration for a few genomic regions in patient p3 is shown in Fig. 3.3.

From the allele frequency trajectories plot, it appears obvious that the whole HIV-1 genome is undergoing many changes during the infection. There is great variability between these trajectories, which is suggestive of the different biological and evolutionary processes underpinning them. As shown by the different shapes of phylogenetic trees and the haplotype frequency trajectories, the type and density of genetic changes differ at different genomic regions. As shown in Fig. 3.3, for instance, the RRE is a genomic region that hosts few changes across the infection. The most parsimonious explanation of such high longitudinal conservation is to assume that most mutations within the RRE have deleterious fitness effects; this explanation fits well with the high degree of cross-sectional conservation and with the known fact that the RRE has not one, but two biological functions, i.e. (i) *rev* binding partner as RNA and (ii) part of the *gp41* protein as amino acids. Other genomic regions such as the variable loop V3 or *p17* in *gag* are changing much more rapidly (see Fig. 3.3), an indication that a higher number of mutations are tolerated.

The different shape of the trees from different genomic regions also indicates that HIV-1 evolution is strongly influenced by recombination, such that *HIV-1 adaptation can proceed at many genomic loci in parallel* – I will come back to this point in the section on positive selection and immune escape below.

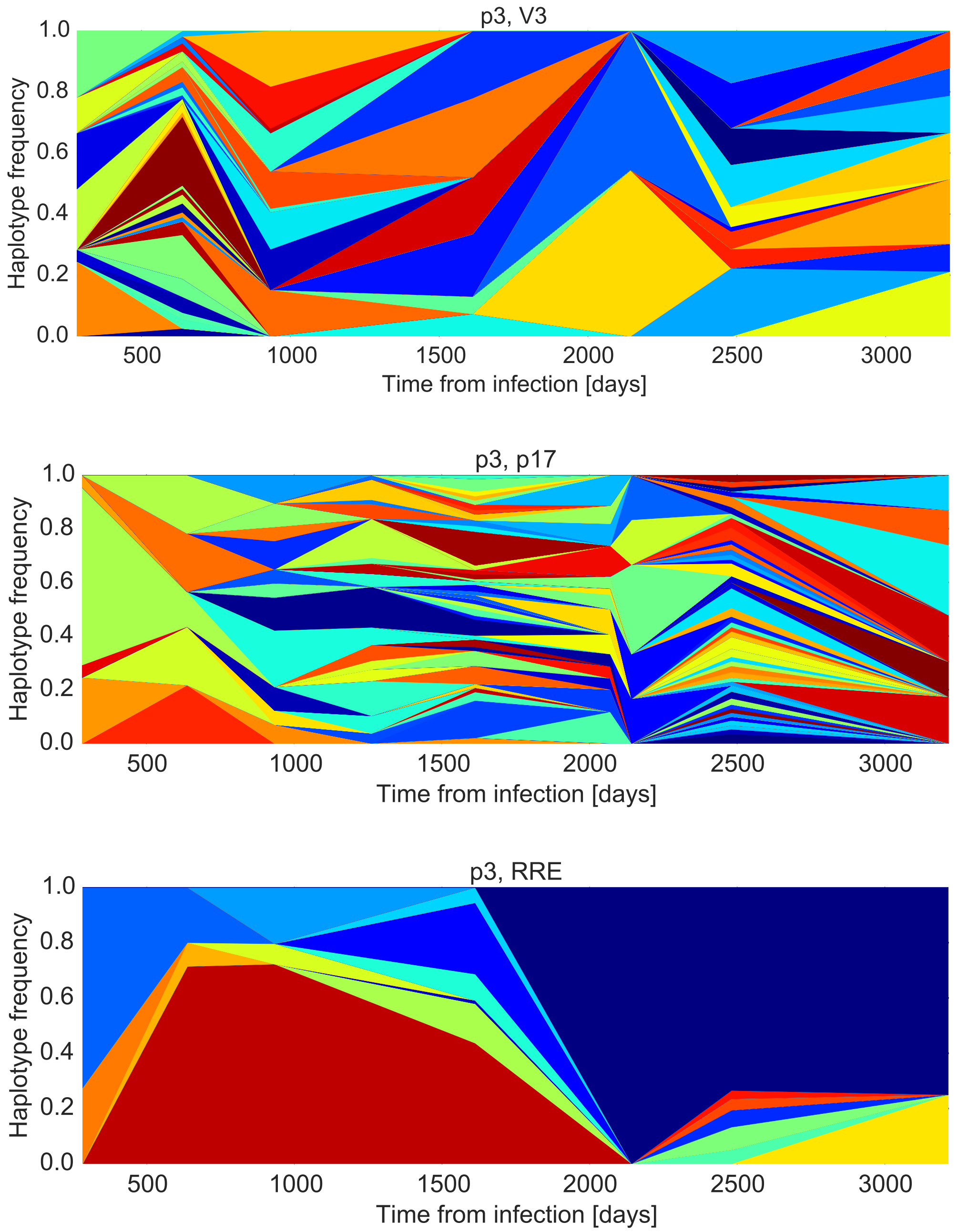


Fig. 3.3: Haplotype frequency trajectories for the V3 loop (top), the capsid protein p17 (middle), and the conserved RNA structure RRE (bottom) from patient p3. Each color (chosen randomly) indicates the frequency of a certain unique sequence over time.

3.1.3 Substitution/divergence rates and diversity

Considering the heterogeneity shown by different genomic regions in terms of phylogenies and allele/haplotype frequency trajectories, a basic question about HIV-1 evolution is: how fast is HIV-1 evolving at different genomic loci?

The answer to this question is shown in Fig. 3.4 (top panel). First, **genetic divergence** measures, for a single time point, how often one finds any derived allele, averaged over a certain genomic region:

$$\text{divergence}(t) := \left\langle \sum_{\alpha \in \{A,C,G,T\}}^{\alpha \text{ not ancestral}} \nu_\alpha(t) \right\rangle_{\text{genomic region}} , \tag{3.1}$$

where angular brackets indicate averaging. Divergence increases in time as the viral population starts to mutate away from the founder virus, and the rate of increase of divergence is the operational definition of **evolutionary rate** in use throughout this thesis. (Note that more coarse-grained definitions are sometimes used in the literature, especially from a macroevolutionary perspective in which minor alleles in a population are not observed but a great number of related species is sampled.)

Evolutionary rates change by more than one order of magnitude across the genome (computed as a sliding window of 300 bp). For comparison, the average fold change between patients is only by 0.6 ± 0.2 in logs of 2 (excluding patient p9, in which evolution seems to be slower for unclear reasons). This suggests that, in general, *a certain genomic locus of HIV-1 experiences similar evolutionary processes during any two completely independent infections.*

In the middle panel of Fig. 3.4, I plot all sites at which a derived allele has fixed during infection, for the same patients. Many of these sites mark **selective sweeps**, possibly related to immune escape. Whereas some regions such as V3 harbour substitutions in virtually all patients, larger variation is found in more conserved regions such as integrase (IN). I will discuss substitutions and selective sweeps more in depth below, in the section on positive selection and immune escape.

I also quantified divergence and **diversity**, i.e. the average distance between any two sequences, for synonymous and nonsynonymous mutations and different classes of genomic regions (see Fig. 3.5). For better data coverage, diversity is actually defined in this context assuming independent sites, i.e. by:

$$\text{diversity}(t) := \left\langle \sum_{\alpha \in \{\text{A,C,G,T}\}} \nu_\alpha(t) \left(1 - \nu_\alpha(t)\right) \right\rangle_{\text{genomic region}} , \tag{3.2}$$

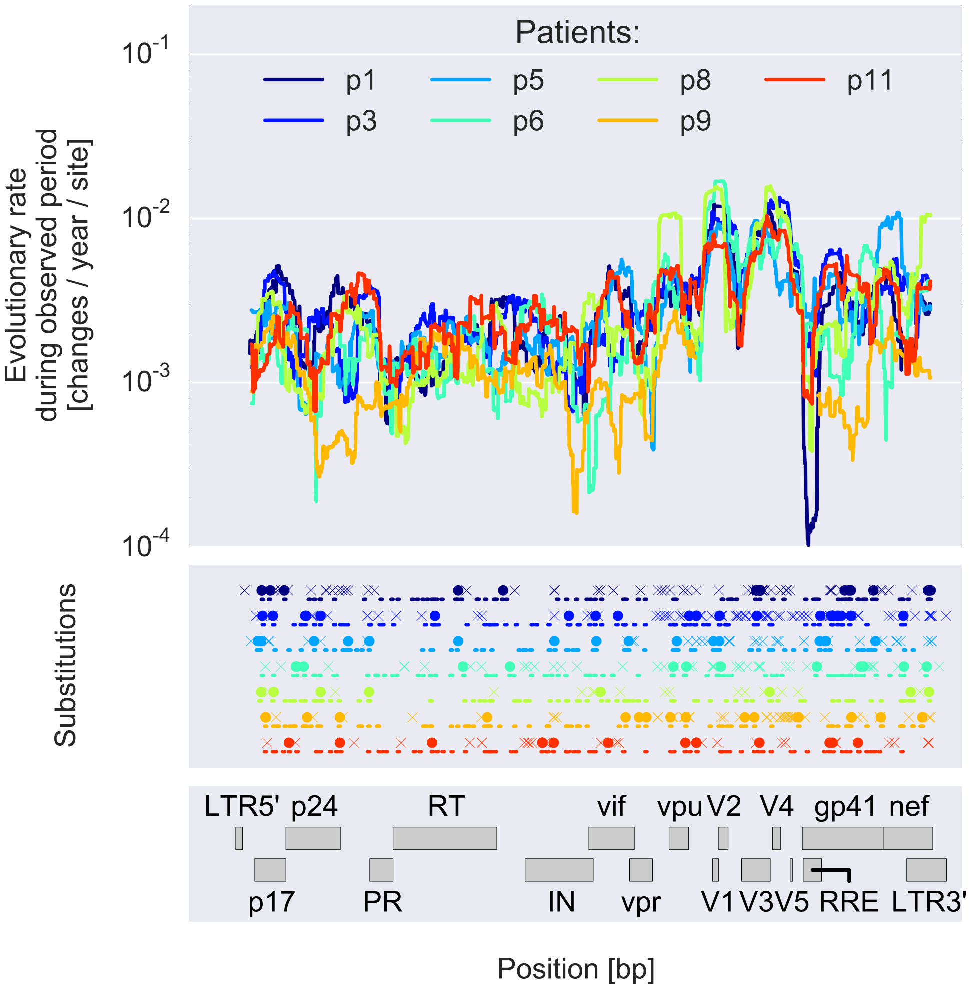


Fig. 3.4: Top: evolutionary rate of HIV-1 during infection in several patients, in a sliding window of 300 bp along the genome. Variation of the evolutionary rate across the genome is large, often larger than between patients at the same genomic location. Middle: Map of substitutions and predicted CTL epitopes. Substitutions within epitopes are indicated by circles, outside any epitope by crosses. Epitopes are predicted using the MHCi web service (see main text). Bottom: genomic features of HIV-1, for reference.

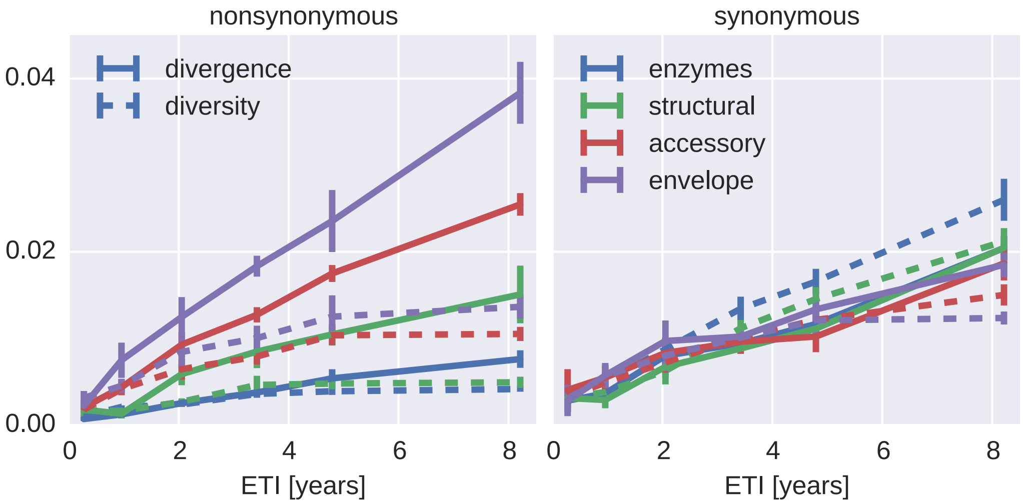


Fig. 3.5: Divergence and diversity for nonsynonymous (left panel) and synonymous (right panel) mutations in different classes of genomic regions (enzymes, structural proteins, accessory genes, and envelope).

where the angular brackets indicate averaging.

Nonsynonymous diversity increases more slowly than divergence and saturates earlier. This result is consistent with previous reports [19]. There are also differences along the genome: nonsynonymous divergence and diversity are especially low in regions of higher cross-sectional conservation such as the enzymes. The large number of nonsynonymous changes in envelope is connected to the stronger selective pressure caused by antibody targeting, and to fewer evolutionary constraints.

Synonymous divergence is similar for all regions, and synonymous diversity increases more steadily than nonsynonymous diversity. In particular, synonymous diversity is lowest in envelope and the accessory genes. This might be the effect of a higher density of selective sweeps in those genomic regions, tilting the selection/recombination balance driving HIV-1 evolution: while positive selection reduces diversity around every sweep (and there are many, as shown in Fig. 3.4), recombination maintains diversity by spreading the beneficial allele onto many genetic backgrounds. Overall, *HIV-1 keeps a high level of diversity throughout the infection, despite the many substitutions all across the genome.*

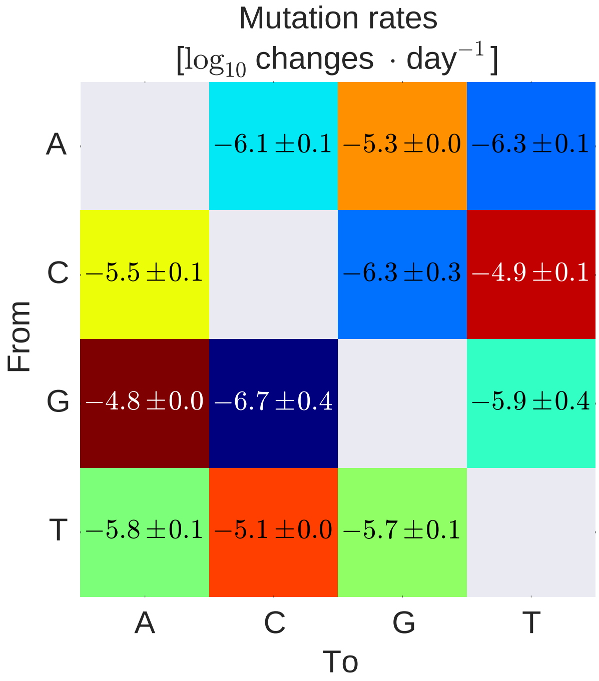


Fig. 3.6: Mutation rate matrix estimated from the *in vivo* longitudinal genetic data set. The Pearson correlation coefficient with the most recently published *in vitro* study [55] is 89% on a log scale (94% on a linear scale). Error bars are standard deviations on 100 patient bootstraps, ±0.0 means the error is less than 0.05.

3.1.4 Mutation

The high genetic diversity that characterizes HIV-1 and allows it to escape immune surveillance is generated by mutation during several steps of the viral life cycle, in particular reverse transcription by the HIV-1 RT and forward transcription by the host cell RNA polymerase. In my doctorate, I chacterized this process by analyzing the longitudinal genetic data I collected.

There are two approaches to measure the mutation rate of HIV-1. The first is to perform single-replication *in vitro* experiments and directly count the number of genetic changes of each kind (e.g. $A \to G$). This strategy has been applied and provides useful estimates, the average rate being around $1.5 \cdot 10^{-5}$ changes per generation per base [55,56]. The main drawback of this method is that the experiments were performed in an artificial cell culture environment and on non-native nucleic acid substrates (LacZ operons).

As an orthogonal approach, I took HIV-1 sequences from the longitudinal data set and counted the number of times each mutation from the viral founder is observed, for each kind. Because of recurrent mutation, the frequency of mutated alleles increases steadily in time according to the mutation rate. By simply fitting a linear increase in average frequency across the infection, I could estimate the mutation rates themselves. This basic idea works well, and the data are collected *in vivo*, which is an advantage over cell-culture systems.

In addition to recurrent mutation, however, another process influences the frequency of derived alleles when observed across long periods of time: selection. In order to unmask the underlying mutational process from the influence of selection, I developed two separate models. First, I restricted the observed alleles to synonymous mutations at nonconserved genetic sites, outside of known regions under purifying selection such as RNA secondary structures (RRE, psi element). The resulting matrix is shown in Fig. 3.6.

Second, I made a composite model that accounts for purifying selection instead of trying to avoid it. This model will be explained in the next section.

In both cases, the resulting matrix of mutation rates agrees quite well with previous experimental studies, with Pearson correlation coefficients of 80 to 90%. In other words, *the mutation rate matrix* in vivo *is close to published* in vitro *estimates.*

3.1.5 Fitness landscape

Once new alleles are generated by mutation, the driving force for their frequency in a population is selection. Alleles under positive selection increase the fitness of the virus and tend to increase rapidly in the population; alleles

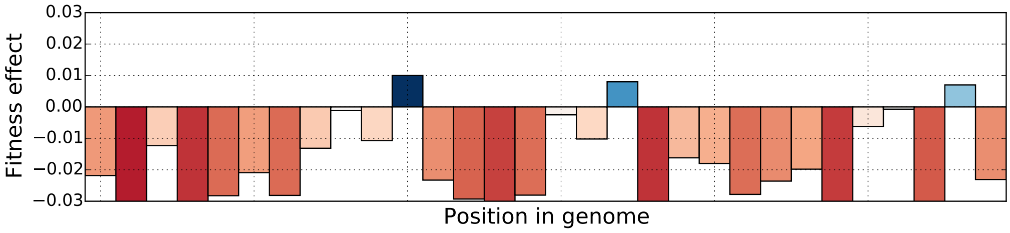


Fig. 3.7: Illustration of an additive fitness landscape on a biallelic genetic model. Each bar indicates the cost (red) or benefit (blue) of a mutation at that site. Beneficial/deleterious ratio and absolute amplitudes are approximately realistic for HIV-1, measured in 1/days or 1/generations (generation time for HIV-1 is around 2 days [57]).

under negative or purifying selection cause a fitness cost and are suppressed.

In principle, the fitness effect of a certain mutation depends not only on the location and kind of mutation itself (say $A \to C$ at site 90 in the protease), but also on the genetic background on which it happens. In practice, however, it is very hard to obtain much information about this kind of complex fitness dependence, or epistasis, from experimental data. For the sake of predictability, I adopted a simpler model that basically ignores genetic interactions and assigns to each mutation a certain fitness effect: the **additive fitness landscape** (see Fig. 3.7 for an illustration).

Given this key simplification, I estimated the effects of both purifying and positive selection acting on the HIV-1 genome by analysing the longitudinal data set.

Purifying selection

Most mutations in the HIV-1 genome are under purifying selection, that is they have a **fitness cost**. This is not surprising: introducing random mutations in a very compact genome will most likely result in a slowly or non-replicating virus.

The difficult question is *how* costly different mutations are. In order to address this question, I considered, like for the mutation rate estimate, the increase in frequency of mutated alleles during the infection. Whereas that analysis was restricted to *bona fide* neutral mutations, however, in this case I divided the genetic sites along the HIV-1 in seven categories, based on cross-sectional conservation in subtype B, and analyzed the allele frequency dynamics of each category to assign a fitness cost to it.

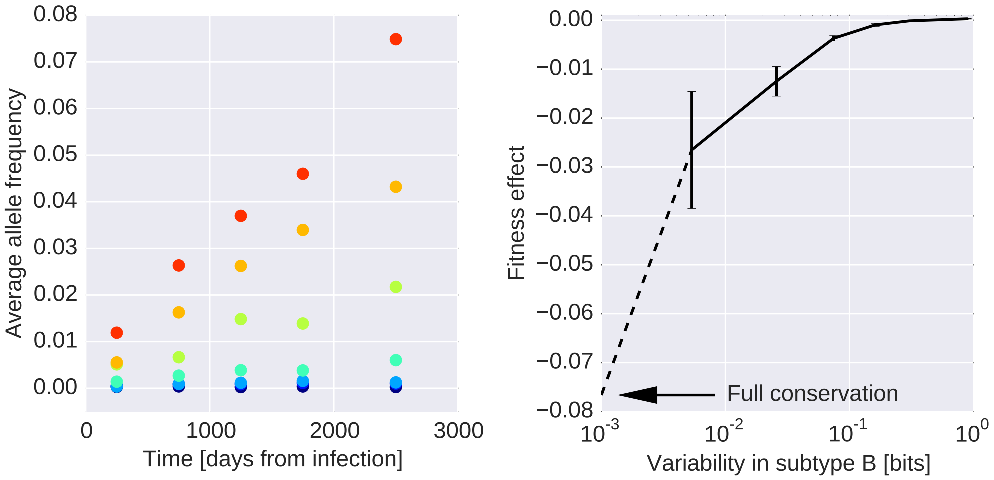


Fig. 3.8: Estimate of fitness costs for the seven-categories coarse-grained model of HIV-1. The left panel shows the data points for the substitution $G \to A$, with increasing subtype variability color coded in a rainbow from blue to red; the right panel indicates the fit result for the cost itself, as a function of site variability. Error bars are standard deviations over 100 patient bootstraps.

Ideally, one would like to be more specific than this and assign a fitness cost to every single position in the genome, as shown in Fig. 3.7 and as exemplified, using cross-sectional data, by Ferguson *et al.* [11]. This is a daunting task as it requires, for each genomic site, a large sample of allele frequency trajectories to obtain enough statistics to make a reliable estimate, and 11 patients are not enough for this level of detail. The coarse-grained model I developed, grouping all mutations at sites with similar cross-sectional conservation, greatly improves the statistics of the fit, providing a basic understanding of purifying selection in HIV-1 evolution.

The categories are established by calculating the Shannon entropy of each base of the HIV-1 genome in a subtype B alignment, and dividing them in seven equally populated quantiles. For each category, purifying selection suppresses mutated alleles to a different degree. The frequency trajectory of the average over all alleles follows the mutation/selection balance equation (1.2):

$$\nu(t) = \frac{\mu}{s}\left[1 - e^{-st}\right], \tag{3.3}$$

where μ is the mutation rate, s the fitness cost, and t the time since the beginning of the infection, in viral generations. Because mutation rates are very different (e.g. transitions much higher than transversions), each kind of mutation within a conservation class follows a curve with different μ but the same s. Once the μ rates are known, this equation can be fitted to the average of the allele frequency trajectories directly – with a single fit parameter, s.

In my case, because the data allowed an internal estimate of μ, I applied this basic idea in two variations. First, I took the mutation rates from the most recent publication by Abram *et al.* [55] and estimated fitness costs directly. Second, I made a model that jointly estimates both the mutation rate matrix and the fitness costs. Both versions yielded similar results: the result of the latter method is shown in Fig. 3.8.

The basic finding of this model is a quantitative assessment of fitness cost from strong ones, i.e. 0.01 or larger, to the very slight ones, of 10^{-4} or less. Whereas the former mutations impair viral replication enough to render the site almost fully conserved throughout the subtype, the latter costs are so small that their effect is basically invisible during the 10 years of the typical infection. Nonetheless, very small fitness costs remain relevant for long-term evolution, as they affect phylogenetic reconstructions on cross-sectional data that span decades of epidemics. One example of such tiny but relevant fitness costs is described in the next section [32].

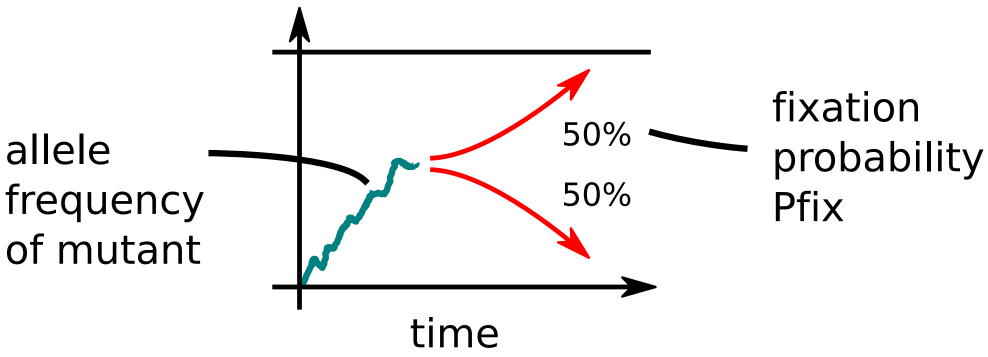


Fig. 3.9: Schematic illustration of fixation probability. A neutral allele that is found in half the current sequences will fix in half the cases in the future (if the surviving lineage carried the allele).

Purifying selection on synonymous mutations and RNA structure

Because the HIV genome is very compact, most sites are expected to be subject to some kind of natural selection, either to warrant biological function or to escape immune pressure. Of all mutations, synonymous ones are *a priori* candidates for neutral evolution, because they (i) are independent of the HIV-immune system interaction (as peptide CTL epitopes and antibody binding of surface proteins), and (ii) do not affect protein function – be it enzymatic or structural.

In my PhD, I considered whether selection is acting at synonymous sites and, if so, how strong it is and what molecular mechanisms might be underpinning it [32]. This question is not only relevant in terms of HIV biology *per se*, but also for modeling evolution. For instance, any phylogenetic analysis is based either on neutral models of evolution – this is the typical case – or on models with known selective pressures.

In order to assess selection on synonymous mutations, I analyzed published longitudinal sequence data [19, 20] on HIV-1 *env*. Because linkage disequilibrium over up to 100 bp is expected based on previous estimates of the *in vivo* recombination rate of HIV-1, I developed an analysis that does not assume independent genetic sites. The codebase used for the analysis is available online at:

`http://git.tuebingen.mpg.de/synmut.git`.

The basic observable is the **fixation probability**, illustrated schematically in Fig. 3.9. For each synonymous mutant allele (as opposed to the ancestral

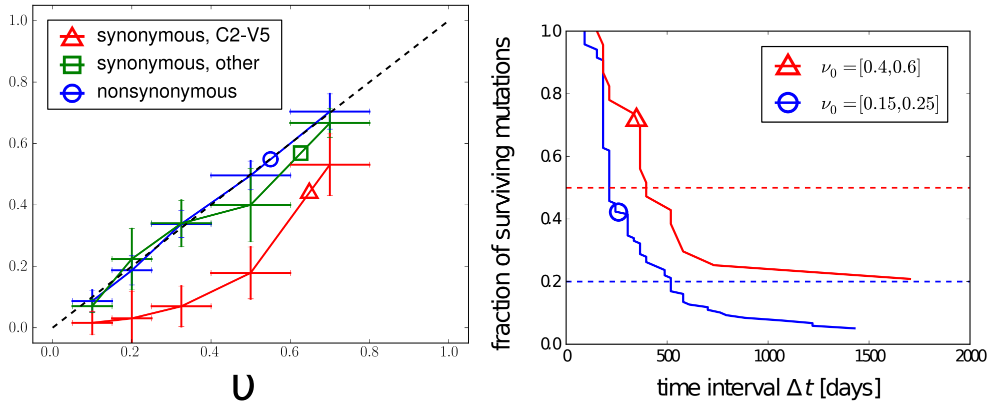


Fig. 3.10: Fixation probability (left) and survival time (right) in HIV-1 env indicate weak purifying selection on synonymous mutations in C2-V5. Left panel: the dashed line indicates the diagonal, i.e. the expectation from neutral models. Right panel: cumulative distribution of survival times for synonymous polymorphic alleles in C2-V5 indicate an average survival of around 500 days.

allele of putative founder viral sequence, i.e. the majority consensus allele at the first time point), I record its frequency at an early time and then ask, at a later time, whether it is fixed, extinct, or polymorphic. Few alleles keep polymorphic for very long times, so they can be ignored for all practical purposes. Of the two remaining categories, the fraction of fixed alleles is the fixation probability P_{fix}.

In a neutral model of evolution, the null model of synonymous mutations, P_{fix} of an allele found at frequency ν is equal to ν itself. This prediction is robust against details of the model, e.g. independent or linked sites, recombination, magnitude of genetic drift and hitchhiking. Synonymous alleles from the C2-V5 region of HIV-1, however, show a systematic trend $P_{fix}(\nu) < \nu$, as shown in the left panel of Fig. 3.10. The other lines indicate synonymous mutations in the rest of HIV-1 env (other) and nonsynonymous mutations. Both of them are compatible with a neutral model, but that is a negative result, as there is a number of factors that attract P_{fix} towards the diagonal line even in presence of selection.

The depression in P_{fix} indicates a fitness cost of those mutations. In order to quantify this cost, I analyzed the **survival time** of the mutant synonymous alleles that, after becoming polymorphic, eventually disappear again. The result, shown in the right panel of Fig. 3.10, indicates that those alleles keep polymorphic for approximately 500-1000 days before disappearing. Based on this information, I estimate an average fitness cost of 0.1-0.2% per

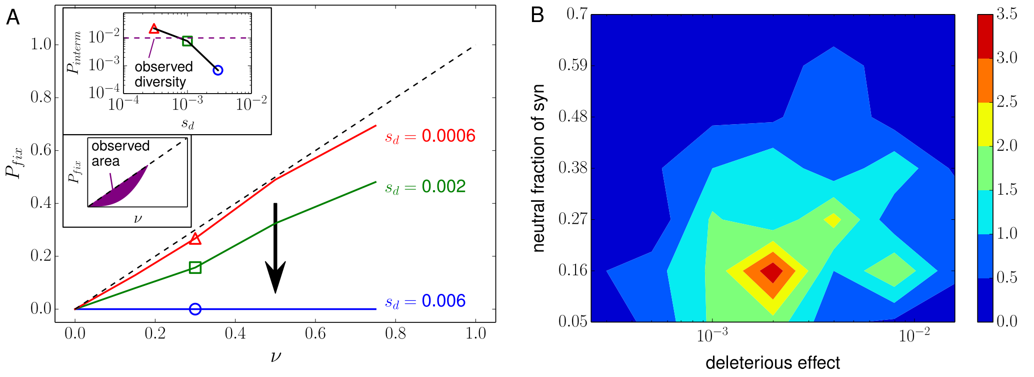


Fig. 3.11: Simulations can reproduce the allele dynamics observed in the HIV-1 samples for some parameter combinations. Left panel: increasing the fitness cost of synonymous mutations reduces both P_{fix} and genetic diversity. Both observables become compatible with the HIV-1 C2-V5 values for costs of the order of 0.2%. Right panel: number of simulations that are compatible with the data in the parameter space, as filled contours. The fraction of synonymous changes that are still neutral is quite small, around 10-20%.

day. In order to test this estimate, I used my own computer package, FFPopSim [30], to simulate the evolution of populations with similar properties to HIV-1, in terms of mutation and recombination rates, population size, fitness landscape of nonsynonymous mutations. I explored a two-dimensional parameter space:

- the fraction of synonymous changes that are neutral VS deleterious;
- the average effect of deleterious mutations.

I found that only a small region of this space is compatible with the evolution of HIV-1 in C2-V5, in terms of fixation probability and genetic diversity, as shown in Fig. 3.11. This not only confirmed a fitness cost around 0.2%, but also indicated that, of all synonymous mutations in C2-V5, the deleterious ones must be the vast majority, around 80-90%.

Given that so many synonymous mutations break the null expectation of neutrality, I proceeded to ask what the biological reason for this observation might be. I considered two general mechanisms, codon bias and RNA secondary structures. The former idea elaborates as follows: since HIV-1 is using the tRNA pool of the host cell to replicate, switching to less common codons might have a negative impact on replication capacity. Despite prolonged effort, I could not find any hint of a significant effect of codon bias

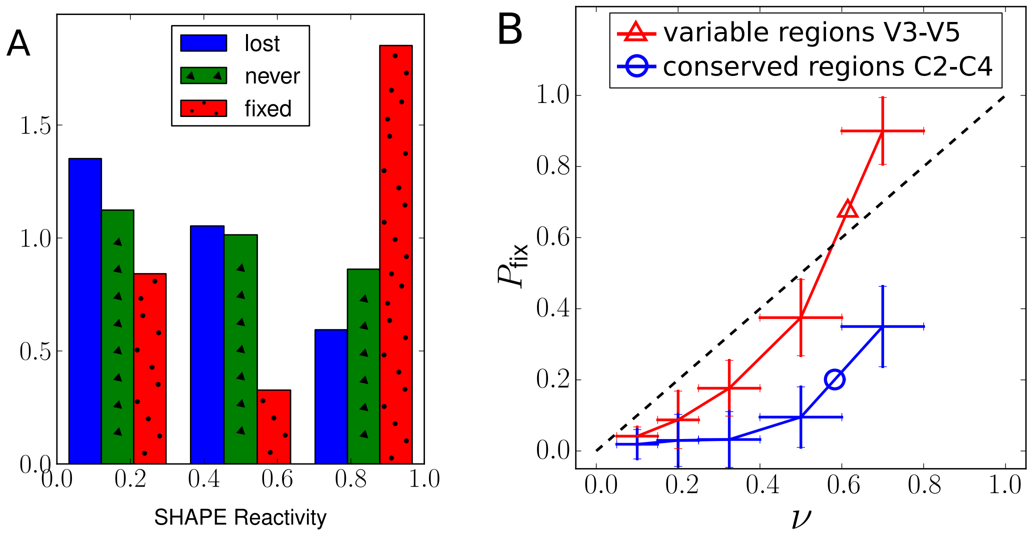


Fig. 3.12: RNA structure affects fitness in C2-V5. Left panel: fraction of fixed, lost, and polymorphic synonymous mutant alleles for three SHAPE categories. Right panel: the conserved regions C2-C4 have a larger depression in P_{fix} than the variable loops V3-V5. They also have widespread pairing propensity, suggesting again a role for RNA structure in the fitness landscape of HIV-1.

on the evolution of C2-V5 within the time range of an infection. This is expected *a priori* because HIV-1 has a very skewed codon bias to start with, if compared to human cells [58], and this skewedness does not seem to be decreasing during the last few decades [59].

The second idea, of RNA structures, is based on the realization that HIV-1, as a compact, single stranded genome, is using RNA secondary structures for various biological functions [60]. A well-known structure, located in *env* downstream of V5, is the Rev Response Element (RRE), a stable 350 bp long hairpin that is targeted by Rev during nuclear export of specific RNA transcripts [61]. In general, it is hard to infer RNA structures from the sequence reliably, as predictions tend to contain high false positive rates and are often missing pseudoknots. To overcome this vagueness, I exploited published data, collected with the SHAPE assay to chemically probe the propensity of each nucleotide in the HIV-1 genome to form a pair [60]. This kind of information is much less noisy than computer predictions, if less complete – we do not know what pairs form, only what single sites are occupied. No RNA structure was suggested by the authors of [60] in the C2-V5 region, but from their results it was not excluded either. In order to probe the data for a statistical signal, I divided the SHAPE reactivities in three categories and queried the fraction of synonymous mutations that are fixed, lost, or long-term polymorphic for each of them, as shown in Fig. 3.12. I found that highly reactive sites, i.e. low-pairing-propensity sites, fix significantly more often than low-reactive ones (KS P-value 0.002). Moreover, the P_{fix} of the conserved regions C2-C4, which contain many low-reactive sites, is much lower than the variable loops V3-V5. Both results indicate that RNA structures in C2-V5, although not yet identified experimentally, are at least one source of the fitness costs observed in the longitudinal sequence data.

My results are consistent with another publication that, using totally different methods, suggests a specific RNA architecture for the C2-V5 region of the HIV-1 genome, which the authors termed *insulating stems* [62]. The concept is the following: (i) HIV-1 needs to provide variable loops to escape a diverse range of human antibodies; (ii) in order to hedge the high risk of RNA misfolding connected to large structural variation in the loops, semi-rigid RNA hairpins are encoded in the C2-4 regions between the loops. Albeit only a suggestion, such a modular genetic organisation would be plausible in the highly compact yet evolutionarily flexible genome of HIV-1.

Positive selection

Positive selection is an obvious property of intrapatient HIV-1 evolution. Even with population sequencing, i.e. only a few sequences per time point, it appears clear that some mutant alleles increase in frequency and fix rapidly. In fact, because this increase is often much faster than genetic drift would ever cause, it must be due to positive selection, a process termed **selective sweep**.

There are two subtle points as far as selective sweeps during HIV-1 evolution are concerned. First, because the recombination rate of HIV-1 is limited (see below), mutant alleles neighbouring a sweep can spread rapidly as well, if they lie on the right genetic background – this process is called **hitchhiking**. It is not possible to distinguish a selected allele from a hitchhiker with certainty using purely sequence information, although statistical conclusions can be made, as I explain below.

Second, assuming an allele is actually under positive selection, there are two sources of positive selection on HIV-1: immune escape and general biological function.

Immune escape is a common process during HIV-1 evolution. The adaptive immune system of the host realizes the presence of viral proteins via either (i) antibody recognition of surface patches or (ii) exhibition of MHC class I bound viral epitopes by infected cells, which epitopes are subsequently recognized by specific CD8$^+$ Cytotoxic T-Cells (CTLs). Of the many viral genetic variants present at any time, the ones that carry mutations that inhibit recognition by either mechanism are obviously advantageous and get positively selected.

Not all selective sweeps, however, are caused by immune escape. Because each infection starts from one or very few virions [14], and that virus has been adapting to the immune system of the previous host, it carries a number of alleles that make it suboptimal in terms of replication capacity [14]. Whenever any of these mutations reverts to the optimal state, a selective sweep is observed. This argument hides a negative feedback: (i) because immune escape exists, it must have happened in the donor host; (ii) the transmitted virus is likely to sweep-revert (some of) those changes in the recipient host, because they bring no benefit anymore (different immune system) but most likely impair replication in some way; (iii) hence not all selective sweeps are immune escapes.

The locations of substitutions in some of our whole-genome sequenced patients is shown above in the middle panel of Fig. 3.4. The exact location of the substitution is not shared across patients.

It is not easy to quantify accurately the fitness advantage of putatively selected alleles, because our data is sampled only every 6-15 months and the allele is often observed directly jumping from very low $< 10\%$ to very high $> 90\%$ frequency from one sample to the next. A rough estimate by logistic regression, according to equation (1.1), yields fitness benefits of the order of 1% per day, in agreement with previous estimates that used independent data [53,54]. This indicates that sweeping times are of the order of 100 days.

The number of substitutions ranges between 24 and 98 in these subjects, with a median of 57, with about 10 to-be substitutions rapidly increasing in frequency at the same time. Although this is a small fraction of all polymorphisms, around 1%, it is still a large number considered that a single virion has to collect all of them onto the same genetic background within the typical sweeping time, i.e. 100-400 days. At a generation time of 2 days [57], this means finding a new, correct substitution about every 10 generations.

As of what fraction of the substitutions are actual selective sweeps, it is difficult to test without experiments. It appears that a sizeable fraction of the substitutions is synonymous ($\sim 30\%$), but those cases also show a slightly slower increase in frequency compared to nonsynonymous alleles.

By the same token, if one knew all CTL and antibody epitopes across the whole HIV-1 infection, it would be possible, at least in principle, to distinguish immune escape from biological optimization for each single mutation. For my PhD study, this information was not nearly available, mainly because of the cost and effort required at present to characterize immune repertoires [63]. However, I did obtain HLA typing of the subjects of the main whole-genome longitudinal sequencing project. I used the patient HLA information and the founder viral sequence to collect predicted CTL epitopes, using two separate approaches. First, I queried the LANL immunology database for a list of experimentally verified epitopes (using the "B list") [1]. This list is expected to contain few false positives (e.g. peptides that are listed but not actually exposed on MHC class I complexes), but many false negatives (peptides that are exposed and binding but not listed because nobody ever validated them experimentally). As a complementary appoach, I used an online prediction tool, MHCi [64, 65], to collect an extensive list of computationally predicted epitopes, ranked by their degree of (predicted) binding affinity. The full list contains thousands of peptides and fully covers the HIV-1 genome, but it is obvious that only few of them, most likely the high-affinity ones, represent actual epitopes. I hence took the top k peptides for various values of k and tested statistically whether substitutions in HIV-1 are enriched within predicted epitopes. This test was highly significant for nonsynonymous mutations, for a threshold $k \in \sim (50, 100)$, and mostly so for $k = 80$ (ratio 1.9, $P < 10^{-5}$). The same test for synonymous mutations yielded no enrichment. I expect this computationally predicted list to contain both false negatives and false positives, however I deem it a good starting point for more sophisticated analysis of immune targeting that are outside the scope of my thesis. The list of epitopes is available together with the rest of the data set in the publication [31].

The situation for antibody targets is more difficult, as the tertiary structure of both the antibody and the viral envelope, in addition to post-translational modifications thereof (e.g. glycosylation, structural rearrangements during cell entry) are known to be crucial for binding and neutralization [66]. Existing longitudinal studies on HIV-1-antibody coevolution are few, typically limited to a single patient, and very labour intensive [67,68]. Development of higher-throughput pipelines for investigating this aspect of intrapatient HIV-1 evolution is an open research question that I am planning on answering in the near future.

3.1.6 Recombination

The different looks of the phylogenetic trees reconstructed from different genomic regions of HIV-1 (see Fig. 3.2) are a clear indication that recombination plays an important role in intrapatient HIV-1 evolution. (If evolution proceeded asexually, all trees would look the same, given enough phylogenetic signal and not too high a level of recurrent mutation.) Previous estimates of the recombination rate lie around 10^{-5} per day per base [53,54]. Considering

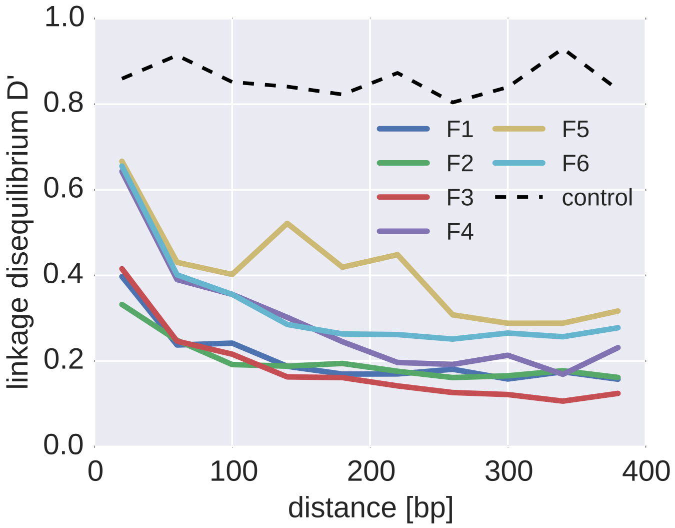


Fig. 3.13: LD in patient HIV-1 samples from each of the six amplicons and a control sample with two HIV-1 laboratory strains mixed at a 50:50 ratio.

the average time it takes HIV-1 to accumulate genetic diversity, which is of the order of 1000 days (see Fig. 3.5), one expects linkage disequilibrium (LD) to extend over:

$$d \approx 10^3 \text{days}/10^{-5} \text{per day per base} = 100\text{bp}. \qquad (3.4)$$

I have measured LD in my whole-genome, longitudinal sequencing data set, see Fig. 3.13. The black line, which refers to a control sample with a 50:50 mix of two distinct HIV-1 strains, underlines that *in vitro* recombination was not a problem for this experimental protocol. The patient data shows a decay of LD over approximately 100 bp, just like expected from the simple calculation outlined above. The exception to this behaviour is the amplicon F5, which includes the most rapid substitutions but also the structurally variable loops. Preliminary analyses indicate that this longer LD is not a simple artifact, but further investigations are necessary to explore the details and consequences of this observation.

The main consequence of the relatively short LD, as mentioned above, is that HIV-1 is able to collect adaptive mutations onto the same genetic background much faster than if it were replicating asexually. Clonal interference, a widespread phenomenon in asexual organisms such as bacteria or influenza virus (within a segment) [69], is likely to play only a minor role in HIV-1 evolution.

3.2 Web application

The web application developed to share the data and results of my whole-genome longitudinal HIV-1 sequencing project proved to be a great resource. I used it for two separate purposes: (i) for exploring the data set from several different perspectives, whenever more than one plot at a time was required to develop an intuition on the research question at hand; (ii) for discussing the results of the study with other scientists without any need for coding. The latter point was especially remarkable during meetings, seminar talks, conferences, visits to collaborators, and whenever time was limited and brainstorming most fruitful.

The web app has four key pages: home, patient, genomic region, and data download. It is available at the address:

http://hiv.tuebingen.mpg.de.

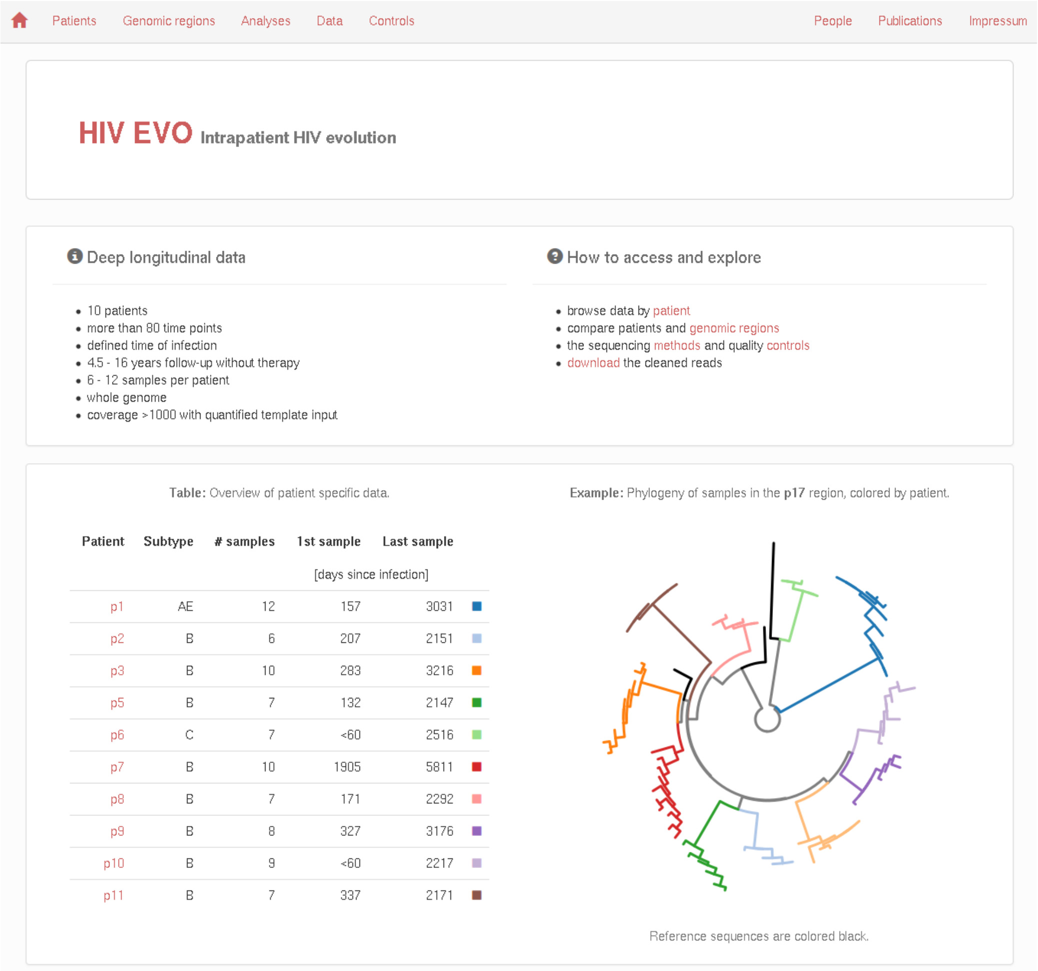
I will present them shortly in the following sections.

3.2.1 Home

The home page (see Fig. 3.14) contains the basic information needed to understand the data set: number and types of samples, a brief description of the patients, links to the more sophisticated parts of the website. It also features an interactive phylogentic tree of all samples together, to highlight the breath of the data set and the benefit of responsive web technology.

3.2.2 Patient

The patient page (see Fig. 3.15) includes a number of interactive plots that allow the user to browse the data set without any coding. Examples thereof are coverage, single nucleotide variant frequencies, diversity and divergence along the genome and as a function of time, phylogenetic trees, and the viral load and CD4$^+$ counts. It was challenging to combine these different, interactive plots on one page, because JavaScript and HTML provide poor encapsulation facilities and pollution of the global namespace is the norm more than the exception, but a combination of descriptive CSS and cautious DOM manipulation solved most issues rather satisfactorily.



HIV EVO Intrapatient HIV evolution

Deep longitudinal data

- 10 patients
- more than 80 time points
- defined time of infection
- 4.5 - 16 years follow-up without therapy
- 6 - 12 samples per patient
- whole genome
- coverage >1000 with quantified template input

How to access and explore

- browse data by patient
- compare patients and genomic regions
- the sequencing methods and quality controls
- download the cleaned reads

Table: Overview of patient specific data.

Patient	Subtype	# samples	1st sample	Last sample	
			[days since infection]		
p1	AE	12	157	3031	🟦
p2	B	6	207	2151	🟦
p3	B	10	283	3216	🟧
p5	B	7	132	2147	🟩
p6	C	7	<60	2516	🟩
p7	B	10	1905	5811	🟥
p8	B	7	171	2292	🟥
p9	B	8	327	3176	🟪
p10	B	9	<60	2217	🟪
p11	B	7	337	2171	🟫

Example: Phylogeny of samples in the **p17** region, colored by patient.

Reference sequences are colored black.

Fig. 3.14: Home page of the HIV web application. Notice the interactive phylogenetic tree on the right.

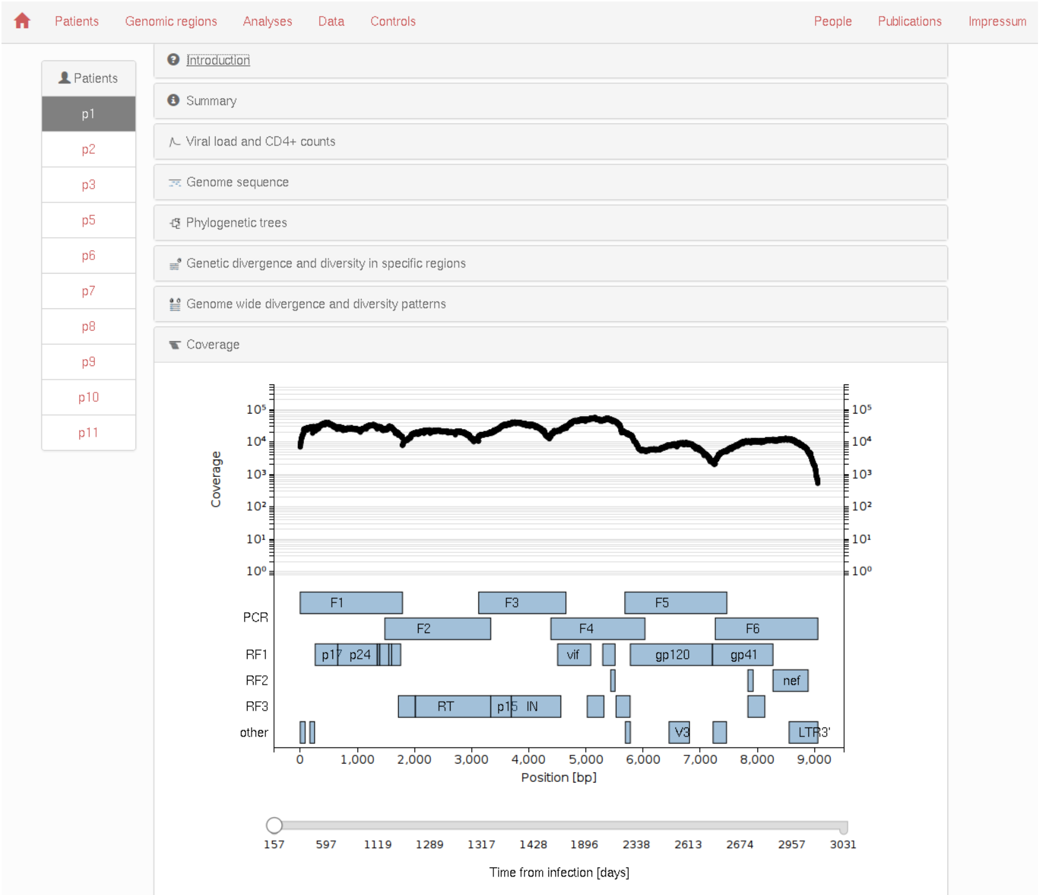
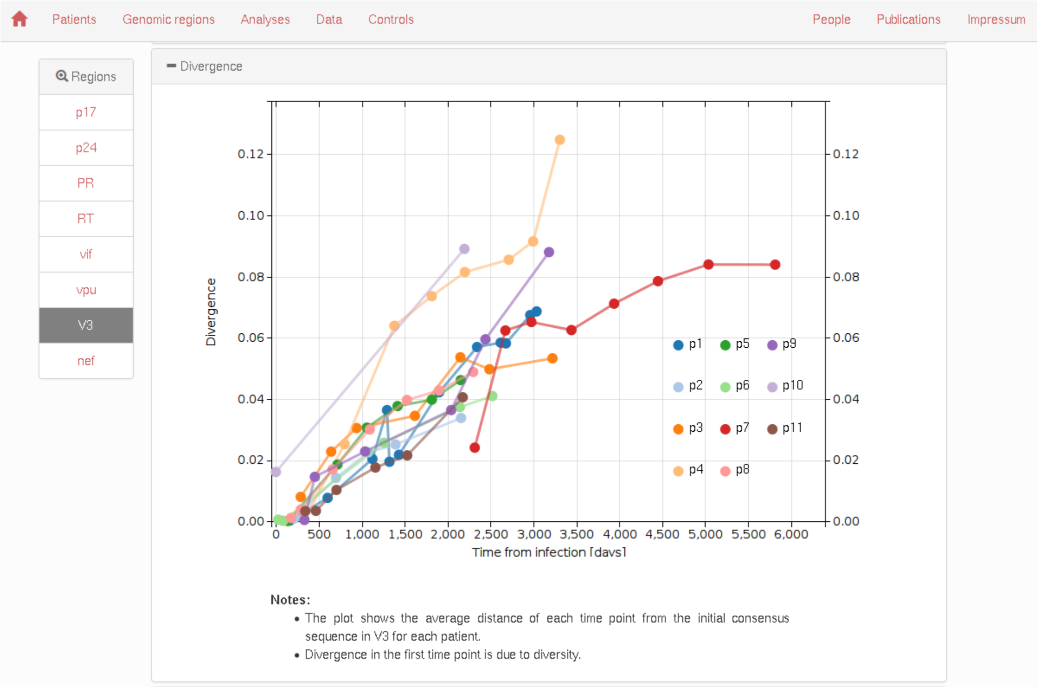


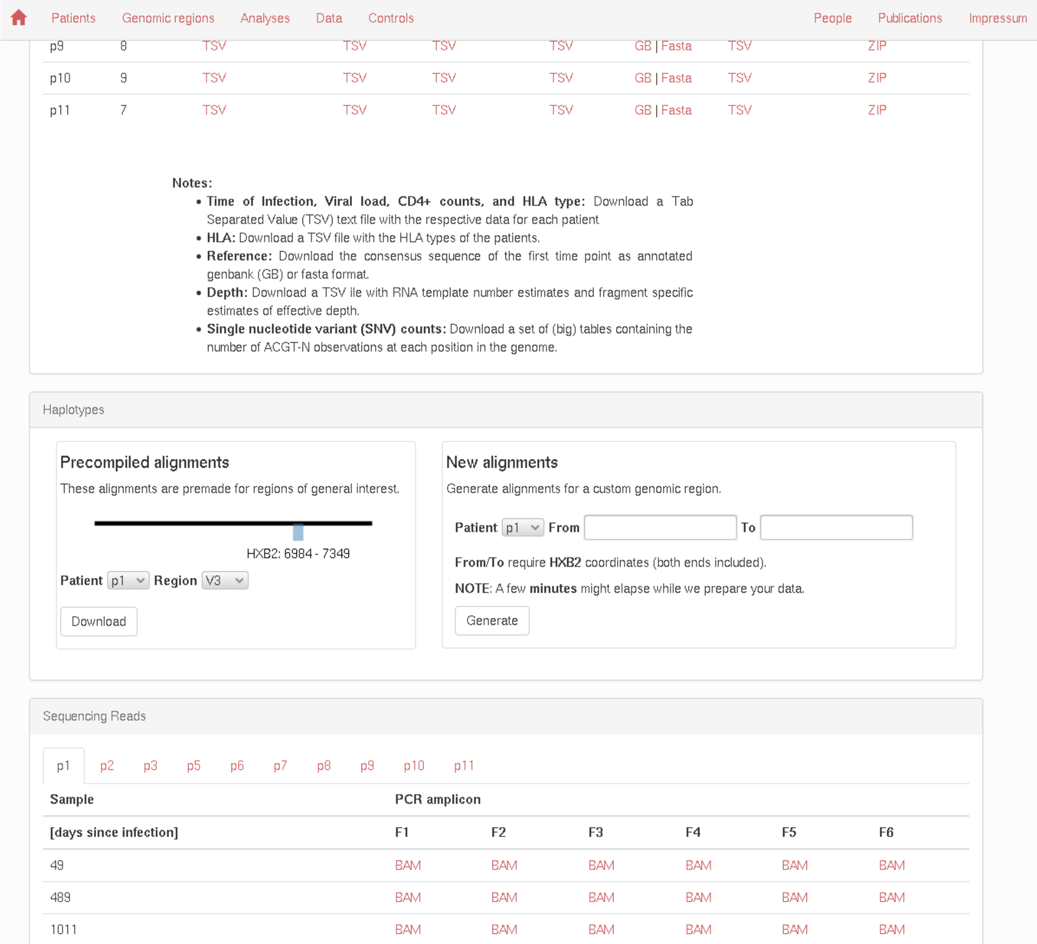
Fig. 3.15: Patient page of the HIV web application. An example of the interactive plots, the coverage plot, is shown here. Clicking on genomic regions of HIV-1 zooms into them for deeper analysis.



Notes:

- The plot shows the average distance of each time point from the initial consensus sequence in V3 for each patient.
- Divergence in the first time point is due to diversity.

Fig. 3.16: Genomic region page of the HIV web application. It is useful to compare patients to each other.



p9	8	TSV	TSV	TSV	TSV	GB | Fasta	TSV	ZIP
p10	9	TSV	TSV	TSV	TSV	GB | Fasta	TSV	ZIP
p11	7	TSV	TSV	TSV	TSV	GB | Fasta	TSV	ZIP

Notes:

- **Time of Infection, Viral load, CD4+ counts, and HLA type:** Download a Tab Separated Value (TSV) text file with the respective data for each patient
- **HLA:** Download a TSV file with the HLA types of the patients.
- **Reference:** Download the consensus sequence of the first time point as annotated genbank (GB) or fasta format.
- **Depth:** Download a TSV ile with RNA template number estimates and fragment specific estimates of effective depth.
- **Single nucleotide variant (SNV) counts:** Download a set of (big) tables containing the number of ACGT-N observations at each position in the genome.

Haplotypes

Precompiled alignments

These alignments are premade for regions of general interest.

HXB2: 6984 - 7349

Patient p1 **Region** V3

Download

New alignments

Generate alignments for a custom genomic region.

Patient p1 **From** **To**

From/To require **HXB2** coordinates (both ends included).

NOTE: A few **minutes** might elapse while we prepare your data.

Generate

Sequencing Reads

p1 p2 p3 p5 p6 p7 p8 p9 p10 p11

Sample	**PCR amplicon**					
[days since infection]	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**
49	BAM	BAM	BAM	BAM	BAM	BAM
489	BAM	BAM	BAM	BAM	BAM	BAM
1011	BAM	BAM	BAM	BAM	BAM	BAM

Fig. 3.17: Data download page of the HIV web application. Notice the haplotype generation factory in the center right of the screenshot.

3.2.3 Genomic region

The genomic region page (see Fig. 3.16) features plots on a specific genomic region, e.g. the V3 loop, merging data coming from all patients at once. It is most useful to assess how much interpatient variation there is in a certain observable, such as genetic divergence, and to spot outliers.

3.2.4 Data

The data download page (see Fig. 3.17) provides access to all information (genetic and not) obtained in the study. Coherent, computer-friendly file formats ensure an expedited downstream analysis, and a REST API infras-

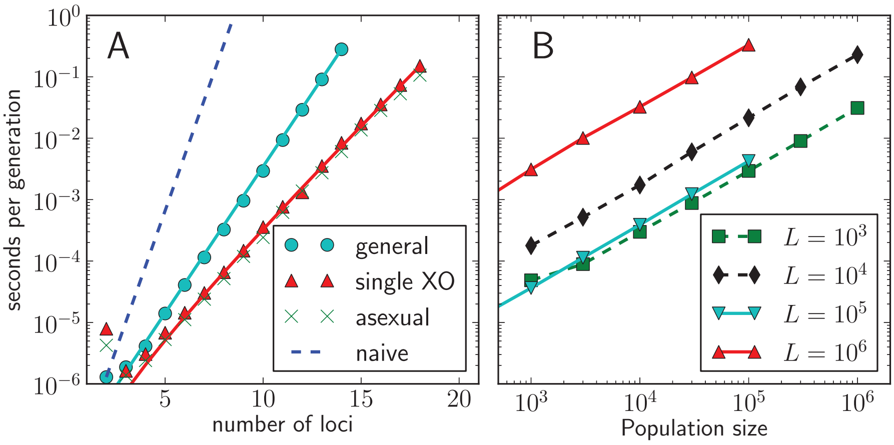


Fig. 3.18: Runtime complexity of FFPopSim scales like expected. Panel A shows the time required to simulate a single generation as a function of the number of loci, using the class haploid lowd. Panel B shows the run times of the individual-based simulations as a function of the population size for different genome sizes L using haploid highd. Solid lines correspond to crossover and mutations rates $\rho = \mu = 10^{-8}$ typical of the human genome, dashed lines to outcrossing with rate $r = 0.01$, and $\mu = 10^{-5}$, $\rho = 10^{-3}$ typical for viral evolution. Runtimes were determined on a 2.93 GHz Intel CPU.

tructure increases automation. The clean mapped reads are available for the most technically advanced users, simpler allele frequency matrices are a more compact data format whenever linkage information is not required. Even when linkage is required, the web application gives access to alignments of reads over certain genomic regions as an easier format than mapped reads. The alignments, which include minor genetic variants, are either precomputed or, if the user requires specific genomic windows, can be generated on the fly by the web server. The latter operation is by far the most computationally intensive of the whole web application and could be outsourced to an external queueing system if user load became too high.

3.3 Simulation of populations

FFPopSim is available online at the address:

https://github.com/neherlab/ffpopsim.

The main innovation of the FFPopSim package was a faster recombination algorithm that scales like $\mathcal{O}(3^L)$ instead of the naïve $\mathcal{O}(8^L)$, where L is the genome length. That this is indeed the case is shown in the left panel of Fig. 3.18. The population size has a negligible effect for the few-loci part of the library, whereas the many-loci class, haploid highd, scales linearly with population size as expected.

Since its publication, FFPopSim was used in a number of occasions both for published articles (e.g. [32, 70, 71]) and as an exploratory tool. I maintained the package and updated it for modern environments (e.g. SWIG 3.0, Python 3.4). A number of functions and refinements have been added since, making it an even better tool for modelling evolution.

Chapter 4

Concluding remarks

Intrapatient HIV-1 evolution, the subject of this thesis, is an interesting topic in terms of both biology and evolution. Because of this interdisciplinary nature of the topic, experimental, computational, and mathematical efforts were necessary for the study. My doctorate work resulted in two main products.

First, I generated a set of tools to study intrapatient HIV-1 that are useful for the research community beyond my own. These include:

- a computational simulation package, FFPopSim [30]
- protocols for HIV-1 sample preparation from plasma to the sequencing reads [31]
- a complex data analysis pipeline to filter and organize the reads in a context of great biological sequence variation [31]
- a whole-genome, longitudinal, deep sequencing data set that can be analyzed by other researchers [31]
- a web application that exposes the data set in useful ways and sets a standard for similar efforts in the field [31].

Second, I obtained a number of novel results on the biology and evolution of HIV-1 itself:

- a quantification of the effects and origins of purifying selection on synonymous mutations [32]
- a whole-genome analysis of patterns of genetic divergence, diversity, and evolutionary rates [31]
- an *in vivo* estimate of the mutation rate matrix [31]

- an estimate of fitness costs on sites at different levels of conservation [31]
- an estimate of the abundance and strength of positive selection and immune escape [31].

The most interesting extension of this work, which I will pursue in the near future, is the simultaneous characterization of the host adaptive immune system. The additional knowledge on epitopes and immune response will be key to unveil the co-evolutionary dynamics that shape the viral infection.

Bibliography

[1] Brian Foley, Thomas Leitner, Cristian Apetrei, Beatrice Hahn, Ilene Mizrachi, James Mullins, Andrew Rambaut, Steven Wolinsky, and Bette Korber. HIV sequence compendium 2013. *Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LAUR*, pages 13–26007, 2013.

[2] R. C. Gallo, P. S. Sarin, E. P. Gelmann, M. Robert-Guroff, E. Richardson, V. S. Kalyanaraman, D. Mann, G. D. Sidhu, R. E. Stahl, S. Zolla-Pazner, J. Leibowitch, and M. Popovic. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):865–867, May 1983.

[3] F. Barre-Sinoussi, J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–871, May 1983.

[4] WHO. Global HIV/AIDS Overview, 2015.

[5] Stéphane Le Vu, Yann Le Strat, Francis Barin, Josiane Pillonel, Françoise Cazein, Vanina Bousquet, Sylvie Brunet, Damien Thierry, Caroline Semaille, Laurence Meyer, and Jean-Claude Desenclos. Population-based HIV-1 incidence in France, 2003–08: a modelling analysis. *The Lancet Infectious Diseases*, 10(10):682–687, October 2010.

[6] Jacques Pepin. *The Origins of AIDS.* Cambridge University Press, September 2011.

[7] Michael Emerman and Michael H. Malim. HIV-1 Regulatory/Accessory Genes: Keys to Unraveling Viral and Host Cell Biology. *Science*, 280(5371):1880–1884, June 1998.

[8] LANL. HIV Databases, 2015.

[9] Jeffrey E. Barrick and Richard E. Lenski. Genome dynamics during experimental evolution. *Nature Reviews Genetics*, 14(12):827–839, December 2013.

[10] David Gresham, Michael M. Desai, Cheryl M. Tucker, Harry T. Jenq, Dave A. Pai, Alexandra Ward, Christopher G. DeSevo, David Botstein, and Maitreya J. Dunham. The Repertoire and Dynamics of Evolutionary Adaptations to Controlled Nutrient-Limited Environments in Yeast. *PLoS Genet*, 4(12):e1000303, December 2008.

[11] Andrew L. Ferguson, Jaclyn K. Mann, Saleha Omarjee, Thumbi Ndung'u, Bruce D. Walker, and Arup K. Chakraborty. Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. *Immunity*, 38(3):606–617, March 2013.

[12] Morgane Rolland, David C Nickle, and James I Mullins. HIV-1 Group M Conserved Elements Vaccine. *PLoS Pathog*, 3(11):e157, November 2007.

[13] Warren John Ewens. *Mathematical population genetics*. Springer-Verlag, 1979.

[14] George M. Shaw and Eric Hunter. HIV Transmission. *Cold Spring Harbor Perspectives in Medicine*, 2(11):a006965, November 2012.

[15] Jianbo Chen, Olga Nikolaitchik, Jatinder Singh, Andrew Wright, Craig E. Bencsics, John M. Coffin, Na Ni, Stephen Lockett, Vinay K. Pathak, and Wei-Shau Hu. High efficiency of HIV-1 genomic RNA packaging and heterozygote formation revealed by single virion analysis. *Proceedings of the National Academy of Sciences*, 106(32):13535–13540, August 2009.

[16] Daniel B. Weissman and Nicholas H. Barton. Limits to the Rate of Adaptive Substitution in Sexual Populations. *PLoS Genet*, 8(6):e1002740, June 2012.

[17] Richard A Neher and Boris Shraiman. Genetic Draft and Quasi-Neutrality in Large Facultatively Sexual Populations. *Genetics*, 188(4):975–996, 2011.

[18] J H Gillespie. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics*, 155(2):909–19, July 2000.

[19] R. Shankarappa, J.B. Margolick, S.J. Gange, A.G. Rodrigo, David Upchurch, Homayoon Farzadegan, Phalguni Gupta, C.R. Rinaldo, G.H. Learn, X. He, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology*, 73(12):10489, 1999.

[20] E.M. Bunnik, Linaida Pisas, A.C. Van Nuenen, and Hanneke Schuitemaker. Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. *Journal of virology*, 82(16):7932, 2008.

[21] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B. Dewell, Lei Du, Joseph M. Fierro, Xavier V. Gomes, Brian C. Godwin, Wen He, Scott Helgesen, Chun He Ho, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. I. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum Kim, James R. Knight, Janna R. Lanza, John H. Leamon, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L. Lohman, Hong Lu, Vinod B. Makhijani, Keith E. McDade, Michael P. McKenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan Fredrik Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, Pengguang Yu, Richard F. Begley, and Jonathan M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, July 2005.

[22] Athe M N Tsibris, Bette Korber, Ramy Arnaout, Carsten Russ, Chien-Chi Lo, Thomas Leitner, Brian Gaschen, James Theiler, Roger Paredes, Zhaohui Su, Michael D Hughes, Roy M Gulick, Wayne Greaves, Eoin Coakley, Charles Flexner, Chad Nusbaum, and Daniel R Kuritzkes. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PloS one*, 4(5):e5683, January 2009.

[23] Evelien M Bunnik, Luke C Swenson, Diana Edo-Matas, Wei Huang, Winnie Dong, Arne Frantzell, Christos J Petropoulos, Eoin Coakley, Hanneke Schuitemaker, P Richard Harrigan, and Angélique B van 't Wout. Detection of Inferred CCR5- and CXCR4-Using HIV-1 Variants and Evolutionary Intermediates Using Ultra-Deep Pyrosequencing. *PLoS pathogens*, 7(6):e1002106, June 2011.

[24] Matthew R. Henn, Christian L. Boutwell, Patrick Charlebois, Niall J. Lennon, Karen A. Power, Alexander R. Macalalad, Aaron M. Berlin, Christine M. Malboeuf, Elizabeth M. Ryan, Sante Gnerre, Michael C. Zody, Rachel L. Erlich, Lisa M. Green, Andrew Berical, Yaoyu Wang, Monica Casali, Hendrik Streeck, Allyson K. Bloom, Tim Dudek, Damien Tully, Ruchi Newman, Karen L. Axten, Adrianne D. Gladden, Laura Battis, Michael Kemper, Qiandong Zeng, Terrance P. Shea, Sharvari Gujja, Carmen Zedlack, Olivier Gasser, Christian Brander, Christoph Hess, Huldrych F. Günthard, Zabrina L. Brumme, Chanson J. Brumme, Suzane Bazner, Jenna Rychert, Jake P. Tinsley, Ken H. Mayer, Eric Rosenberg, Florencia Pereyra, Joshua Z. Levin, Sarah K. Young, Heiko Jessen, Marcus Altfeld, Bruce W. Birren, Bruce D. Walker, and Todd M. Allen. Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PLoS Pathog*, 8(3):e1002529, March 2012.

[25] Carla Kuiken, Bette Korber, and Robert W. Shafer. HIV Sequence Databases. *AIDS reviews*, 5(1):52–61, 2003.

[26] Trevor Hinkley, João Martins, Colombe Chappey, Mojgan Haddad, Eric Stawiski, Jeannette M Whitcomb, Christos J Petropoulos, and Sebastian Bonhoeffer. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature genetics*, 43(5):487–490, March 2011.

[27] Richard R. Hudson. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, February 2002.

[28] Bo Peng and Marek Kimmel. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687, September 2005.

[29] Frédéric Guillaume and Jacques Rougemont. Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, 22(20):2556–2557, October 2006.

[30] Fabio Zanini and Richard A. Neher. FFPopSim: An efficient forward simulation package for the evolution of large populations. *Bioinformatics*, October 2012.

[31] Fabio Zanini, Lina Thebo, Johanna Brodin, Christa Lanz, Goran Bratt, Jan Albert, and Richard Neher. Whole-genome longitudinal deep sequencing of HIV-1 from acute throughout chronic infection. 2015.

[32] Fabio Zanini and Richard A. Neher. Quantifying Selection against Synonymous Mutations in HIV-1 env Evolution. *Journal of Virology*, 87(21):11843–11850, November 2013.

[33] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[34] Wes McKinney. pandas: a Foundational Python Library for Data Analysis and Statistics. 2011.

[35] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and. *Bioinformatics*, 25(11):1422–1423, June 2009.

[36] Short-Course Antiretroviral Therapy in Primary HIV Infection. *New England Journal of Medicine*, 368(3):207–217, January 2013.

[37] Alan S. Perelson, Paulina Essunger, Yunzhen Cao, Mika Vesanen, Arlene Hurley, Kalle Saksela, Martin Markowitz, and David D. Ho. Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature*, 387(6629):188–191, May 1997.

[38] Astrid Gall, Bridget Ferns, Clare Morris, Simon Watson, Matthew Cotten, Mark Robinson, Neil Berry, Deenan Pillay, and Paul Kellam. Universal Amplification, Next-Generation Sequencing, and Assembly of HIV-1 Genomes. *Journal of Clinical Microbiology*, 50(12):3838–3844, December 2012.

[39] Francesca Di Giallonardo, Osvaldo Zagordi, Yannick Duport, Christine Leemann, Beda Joos, Marzanna Künzli-Gontarczyk, Rémy Bruggmann, Niko Beerenwinkel, Huldrych F. Günthard, and Karin J. Metzner. Next-Generation Sequencing of HIV-1 RNA Genomes: Determination of Error Rates and Minimizing Artificial Recombination. *PLoS ONE*, 8(9):e74249, September 2013.

[40] Andrew Adey, Hilary G. Morrison, Asan, Xu Xun, Jacob O. Kitzman, Emily H. Turner, Bethany Stackhouse, Alexandra P. MacKenzie, Nicholas C. Caruccio, Xiuqing Zhang, and Jay Shendure. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, 11(12):R119, December 2010.

[41] S. van der Walt, S.C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, 13(2):22–30, March 2011.

[42] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[43] Gerton Lunter and Martin Goodson. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–939, June 2011.

[44] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[45] Armin Ronacher. Flask, 2015.

[46] Mike Bostock. D3.js, 2015.

[47] Mike Bostock. Towards Reusable Charts, February 2012.

[48] Joyent, Inc. Node.js, 2015.

[49] The Go Authors. The Go Programming Language, 2012.

[50] Pivotal Software. RabbitMQ - Messaging that just works, 2015.

[51] E. D. Weinberger. Fourier and Taylor series on fitness landscapes. *Biological Cybernetics*, 65(5):321–330, September 1991.

[52] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490, March 2010.

[53] R.A. Neher and Thomas Leitner. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol*, 6(1):e1000660, January 2010.

[54] Rebecca Batorsky, Mary F Kearney, Sarah E Palmer, Frank Maldarelli, Igor M Rouzine, and John M Coffin. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14):5661–6, April 2011.

[55] Michael E Abram, Andrea L Ferris, Wei Shao, W Gregory Alvord, and Stephen H Hughes. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *Journal of virology*, 84(19):9864–78, October 2010.

[56] L M Mansky and H M Temin. Lower In Vivo Mutation Rate of Human Immunodeficiency Virus Type 1 than That Predicted from the Fidelity of Purified Reverse Transcriptase. *Journal of virology*, 69(8):5087–5094, 1995.

[57] a S Perelson, a U Neumann, M Markowitz, J M Leonard, and D D Ho. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science (New York, N.Y.)*, 271(5255):1582–6, March 1996.

[58] Gareth M Jenkins and Edward C Holmes. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Research*, 92(1):1–7, March 2003.

[59] Aridaman Pandit and Somdatta Sinha. Differential Trends in the Codon Usage Patterns in HIV-1 Genes. *PLoS ONE*, 6(12):e28889, December 2011.

[60] Joseph M. Watts, Kristen K. Dang, Robert J. Gorelick, Christopher W. Leonard, Julian W. Bess Jr, Ronald Swanstrom, Christina L. Burch, and Kevin M. Weeks. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 460(7256):711–716, August 2009.

[61] Jason Fernandes, Bhargavi Jayaraman, and Alan Frankel. The HIV-1 rev response element: An RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNA Biology*, 9(1):4–9, January 2012.

[62] Rafael Sanjuán and Antonio V. Bordería. Interplay between RNA Structure and Protein Evolution in HIV-1. *Molecular Biology and Evolution*, 28(4):1333–1338, April 2011.

[63] George Georgiou, Gregory C. Ippolito, John Beausang, Christian E. Busse, Hedda Wardemann, and Stephen R. Quake. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology*, 32(2):158–168, February 2014.

[64] Claus Lundegaard, Kasper Lamberth, Mikkel Harndahl, Søren Buus, Ole Lund, and Morten Nielsen. NetMHC-3.0: accurate web accessible

predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Research*, 36(suppl 2):W509–W512, July 2008.

[65] Claus Lundegaard, Morten Nielsen, and Ole Lund. The validity of predicted T-cell epitopes. *Trends in Biotechnology*, 24(12):537–538, December 2006.

[66] Iliyana Mikell, D. Noah Sather, Spyros A. Kalams, Marcus Altfeld, Galit Alter, and Leonidas Stamatatos. Characteristics of the Earliest Cross-Neutralizing Antibody Response to HIV-1. *PLoS Pathog*, 7(1):e1001251, January 2011.

[67] Hua-Xin Liao, Rebecca Lynch, Tongqing Zhou, Feng Gao, S. Munir Alam, Scott D. Boyd, Andrew Z. Fire, Krishna M. Roskin, Chaim A. Schramm, Zhenhai Zhang, Jiang Zhu, Lawrence Shapiro, NISC Comparative Sequencing Program, James C. Mullikin, S. Gnanakaran, Peter Hraber, Kevin Wiehe, Garnett Kelsoe, Guang Yang, Shi-Mao Xia, David C. Montefiori, Robert Parks, Krissey E. Lloyd, Richard M. Scearce, Kelly A. Soderberg, Myron Cohen, Gift Kamanga, Mark K. Louder, Lillian M. Tran, Yue Chen, Fangping Cai, Sheri Chen, Stephanie Moquin, Xiulian Du, M. Gordon Joyce, Sanjay Srivatsan, Baoshan Zhang, Anqi Zheng, George M. Shaw, Beatrice H. Hahn, Thomas B. Kepler, Bette T. M. Korber, Peter D. Kwong, John R. Mascola, and Barton F. Haynes. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature*, 496(7446):469–476, April 2013.

[68] Xueling Wu, Zhenhai Zhang, Chaim A. Schramm, M. Gordon Joyce, Young Do Kwon, Tongqing Zhou, Zizhang Sheng, Baoshan Zhang, Sijy O'Dell, Krisha McKee, Ivelin S. Georgiev, Gwo-Yu Chuang, Nancy S. Longo, Rebecca M. Lynch, Kevin O. Saunders, Cinque Soto, Sanjay Srivatsan, Yongping Yang, Robert T. Bailer, Mark K. Louder, James C. Mullikin, Mark Connors, Peter D. Kwong, John R. Mascola, Lawrence Shapiro, Betty Benjamin, Robert Blakesley, Gerry Bouffard, Shelise Brooks, Holly Coleman, Mila Dekhtyar, Michael Gregory, Xiaobin Guan, Jyoti Gupta, Joel Han, April Hargrove, Shi-ling Ho, Richelle Legaspi, Quino Maduro, Cathy Masiello, Baishali Maskeri, Jenny McDowell, Casandra Montemayor, Morgan Park, Nancy Riebow, Karen Schandler, Brian Schmidt, Christina Sison, Mal Stantripop, James Thomas, Pam Thomas, Meg Vemulapalli, and Alice Young. Maturation and Diversity of the VRC01-Antibody Lineage over 15 Years of Chronic HIV-1 Infection. *Cell*, 0(0), 2015.

[69] Natalja Strelkowa and Michael Lässig. Clonal Interference in the Evolution of Influenza. *Genetics*, July 2012.

[70] Stefan Nowak, Johannes Neidhart, Ivan G. Szendro, and Joachim Krug. Multidimensional Epistasis and the Transitory Advantage of Sex. *PLoS Comput Biol*, 10(9):e1003836, September 2014.

[71] Timothy G. Vaughan and Alexei J. Drummond. A Stochastic Simulator of Birth–Death Master Equations with Application to Phylodynamics. *Molecular Biology and Evolution*, 30(6):1480–1493, June 2013.