Convergent evolution as an indicator for selection during acute HIV-1 infection

*Frederic Bertels[1,2], Karin J. Metzner[3,4], Roland Regoes[1]

[1]Department for Environmental Systems Sciences, ETH Zurich, Zurich, Switzerland
[2]Max Planck Institute for Evolutionary Biology, Department for Evolutionary Theory, Plön, Germany
[3]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland
[4]Insitute of Medical Virology, University of Zurich, Zurich, Switzerland

Convergent evolution describes the process of different populations acquiring similar phenotypes or genotypes. For complex organisms with large genomes it is unlikely to converge to the same genotype, although there are examples where this happens under very strong selection. In contrast, for viruses with very small genomes and high mutation rates, even without selection it is possible that different virus population acquire the same mutation. Here we measure the number of convergent mutations in a well-studied dataset of full-length HIV-1 *env* genes sampled from HIV-1 infected individuals during early infection. We compare this data to a neutral model and find an excess of convergent mutations in the *env* genes. These differences are likely to confer a selective advantage as they occur significantly more often in gp41, the part of the *env* gene involved in fusion with the target cell. Interestingly, convergent mutations are not significantly more likely to occur in genetically diverse regions of the genome and are also not significantly more likely to be synonymous compared to a neutral model, highlighting the highly non-neutral nature of many synonymous sites in the HIV-1 genome. Non-convergent mutations, on the other hand, are significantly more likely to be synonymous and the non-synonymous proportion of the mutations occur significantly more often in highly diverse regions of the *env* gene than randomly distributed mutations. Together our analysis suggests that convergent evolution can be a good indicator for positive selection in small viral genomes.

* Max Planck Institute for Evolutionary Biology, August-Thienemann-Straße 2, 24306 Plön, Germany, Tel: +49 4522 763 222, bertels@evolbio.mpg.de

Introduction

Convergent evolution is ubiquitous in nature. Popular examples for phenotypic convergence are the evolution of flight, which occurred in bats, birds and insects independently, or the evolution of the eye, which occurred many times independently across the tree of life (Serb and Eernisse 2008). We can also observe convergent evolution at the level of the genotype. This usually occurs when organisms are under extremely high selection pressure. For example, the acquisition of antibiotic resistance (Farhat et al. 2013) or the development of cancer (Miki et al. 1994). Convergent evolution in viruses is particularly common because viruses have small genomes, high mutation rates and large population sizes (Bull et al. 1997; Crandall et al. 1999; Wichman et al. 2000).

To study convergent evolution in viruses we will focus on one of the best studied human viruses: Human immunodeficiency virus type 1 (HIV-1). HIV-1 is an RNA virus with a small genome and an extremely high mutation rate. The high mutation rate allows the virus to evade the human immune response and persist for years within the host (Coffin 1995; Koenig et al. 1995; Borrow et al. 1997; Goulder et al. 1997; Wei et al. 2003; Trkola et al. 2005; Liao et al. 2013). Over this time span HIV-1 evolves and acquires a large number of mutations (Shankarappa et al. 1999; Lemey et al. 2006; Keele et al. 2008; Poon et al. 2010). The mutations that eventually fix in the viral population are not randomly distributed though. Particularly mutations that help the virus escape the immune response are similar among different viruses to the degree where a specific viral genotype is predictive of the HLA allele present in the host (Bhattacharya et al. 2007; Bartha et al. 2013; Carlson et al. 2016). Convergence is also important during the infection of a new host. Recent studies have shown that transmitted founder viruses are often more closely related to the HIV-1 population consensus sequence than to the most prevalent strain in the donor (Carlson et al. 2014; Fraser et al. 2014; Zanini et al. 2015).

This effect of convergence is noticeable as soon as viruses are detectable in an acutely HIV-1 infected individual (Wood et al. 2009). Wood et al. show that there is strong selection immediately after infection and among other evidence show that HIV-1 populations acquire identical mutations in the *env*

gene in different hosts. These mutations seem to be mainly due to the effect of APOBEC proteins, which protect against retrovirus infection by hypermutating the viral genome but, in the case of HIV-1, is inhibited in large by the vif protein, the remaining APOBEC activity has been proposed to lead to faster adaptation of the virus to the host (Sadler et al. 2010). The convergent mutations Wood et al. identified occur in CTL epitopes (probably as a response to immune selection) as well as in gp41 (Wood et al. 2009).

Wood et al. used the Keele dataset to infer nucleotide sites that are under selection by applying dN/dS. This approach makes the strong assumption that synonymous mutations do not affect the fitness of the virus in contrast to non-synonymous mutations. Using dN/dS for inferring selection strength is a very reliable method when applied to reproductively isolated species but is less reliable when applied to population sequence data (Kryazhimskiy and Plotkin 2008). Furthermore, synonymous mutations have been shown to have very significant fitness effects (Lind et al. 2010; Brandis and Hughes 2016), mostly when they occur in essential genes. Hence, synonymous mutations may have strong fitness effects in HIV-1, where every single gene is important for the survival of the virus (Zanini and Neher 2013; Zanini et al. 2017).

Here, we pursue an alternative approach to detect selection that is based on the extent of convergent evolution at different nucleotide sites. Focusing on convergent evolution allows us to infer selection pressures that are exerted uniformly on viral populations across many hosts. To detect these selection pressures we re-analyze a dataset of full length *env* genes sampled from infected individuals during early infection (Keele et al. 2008; Li et al. 2010). We find that convergent mutations are over-represented in the Keele and Li datasets compared to a neutral model. When we compare convergent mutations with private mutations, we observe a very uniform distribution across the gene for private mutations and a distribution significantly skewed towards gp41 for convergent mutations. Both results, the comparison to a neutral model and the comparison between private and convergent mutations, indicate that we observe convergent mutations in different HIV-1 populations as a result of a specific fitness advantage.

Results

We reanalyzed two datasets from previous studies on the evolution of HIV-1 during early infection of a single founder virus (Keele et al. 2008; Li et al. 2010). Amounting to a total of 95 *env* sequence alignments each from a single HIV-1 infected individual.

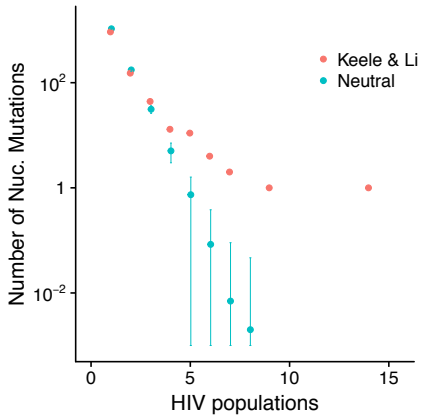

Figure 1. Convergent mutations are unusually frequent during early infection of HIV-1. Comparison between nucleotide mutations observed in one or more HIV-1 populations. Each HIV-1 population is isolated from a different infected individual. The Keele and Li data is compared to a random distribution of the same number of mutations as observed for each HIV-1 population using the same substitution rates as measured from the Keele and Li data.

Convergent mutations are unlikely under a neutral model of evolution

We observed convergent evolution for a large number of the 1530 mutations identified in the Keele and Li datasets (**Figure 1, Supplementary Data 1**). Whereas mutations occurring in one or two HIV-1 populations are underrepresented in the Keele and Li datasets, mutations occurring in three or more populations in parallel are overrepresented (compared to 1000 random distributions of the same

number and identical substitution rates as observed in the data). Mutations occurring in more than five HIV-1 populations in parallel fit particularly badly to the neutral model. The number of convergent mutations declines linear on a log scale for the neutral model. This is not the case for the HIV-1 population data. The likelihood of observing a mutation in 9 or even 14 HIV-1 populations in parallel for randomly distributed mutations is about 4×10^{-4} or 3.6×10^{-8} respectively. For lower levels of convergence the deviation between neutral model and observation is still significant. In about 70 out of 100 simulations only one mutation was observed in five populations in parallel. In contrast, there are 11 such mutations in the Keele and Li data (**Figure 1**).

In total we determined 19 mutations that are likely to have been positively selected because they occur in five or more HIV-1 populations for which we expect to see less than one mutation under a neutral model of evolution. Additionally there are about 20 mutations that occur more frequently in three or four populations than in the neutral model. For these mutations it is more difficult to distinguish between mutations that are under selection and mutations that we observe due to chance alone.

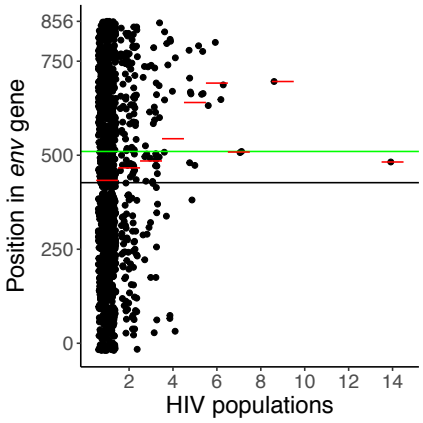

Figure 2. Convergent mutations are predominantly located in gp41. Shows the position of mutations across the *env* gene on the y-axis and the number of HIV-1 populations the mutations were observed in on the x-axis. Red bars indicate the mean of all position within one category. The black horizontal line indicates the mid point of the env gene. The green horizontal line indicates the start of gp41, a fusion protein that is part of the *env* gene.

Convergent mutations occur predominantly in gp41

Convergent mutations are also differently distributed across the *env* gene compared to non-convergent (private) mutations. Convergent mutations are more likely to be located at the end of the *env* gene (**Figure 2**). Private mutations are relatively evenly distributed across the entire *env* gene. The mean of the positions of private mutations is 433, which is very close to half the length of the *env* gene (427). However, the mean increases for mutations that occur in more than one HIV-1 population.

The end of the *env* gene encodes for gp41, a protein responsible for forming the six-helix bundle during cell fusion with the host cell (Chan et al. 1997; Skehel and Wiley 1998; Melikyan et al. 2000). The proportion of single population mutations that occur in gp41 is 0.41, which is very close to the fraction of gp41 of the whole *env* gene (0.40). Mutations that occur more than three times have a probability of 0.69 of occurring in gp41. This is a significant difference to mutations occurring in only a single individual (p-Value chi square test: 0.003).

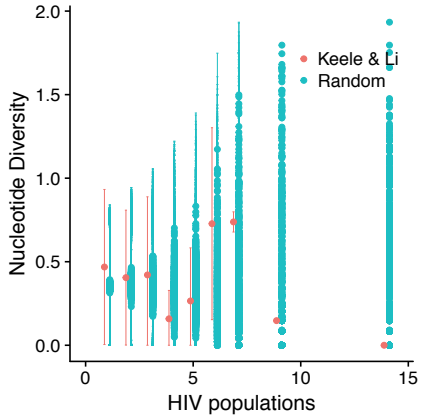

Figure 3. Convergent mutations do not occur in positions of high nucleotide diversity. Mean and standard deviation of nucleotide diversity for Keele and Li mutations (orange) as well as 100 individual random mutation distributions (turquois) that occur in a number of HIV-1 populations in parallel (x-axis). We use the Shannon entropy to measure nucleotide diversity. Hence the minimum is 0 (a single nucleotide is found in all consensus sequences at a certain reference position) and the maximum is 2 (all four nucleotides occur at equal frequencies). The same number of mutations was distributed randomly as the number of observed convergent mutations (i.e. for the mutation that occurs in 14 populations in parallel a single mutation was randomly distributed 100 times across the *env* gene).

Convergent mutations do not occur at significantly more diverse positions than randomly distributed mutations

Mutations shared across multiple hosts could arise through uniform selection on the same nucleotide across hosts, or because mutations are more likely to arise at certain positions (mutational hotspots). To test whether convergent mutations are more likely to occur at highly diverse positions in the *env* gene (such as mutational hotspots) we measured the nucleotide diversity, across all consensus sequences for all HIV populations (**Figure 3**). We assume that at mutational hotspots in the *env* gene the nucleotide diversity is higher across HIV populations. For convergent mutations there are no significant differences to randomly distributed mutations suggesting that mutational hotspots may not directly cause convergence. Interestingly, we found that in contrast private mutations occur at significantly higher rates at more genetically diverse positions in the genome. This does not necessarily mean that private mutations predominantly occur at mutation hotspots, we would also observe higher nucleotide diversity for sites that are under positive selection, or positions that evolve neutrally where drift causes diversity to increase over time.

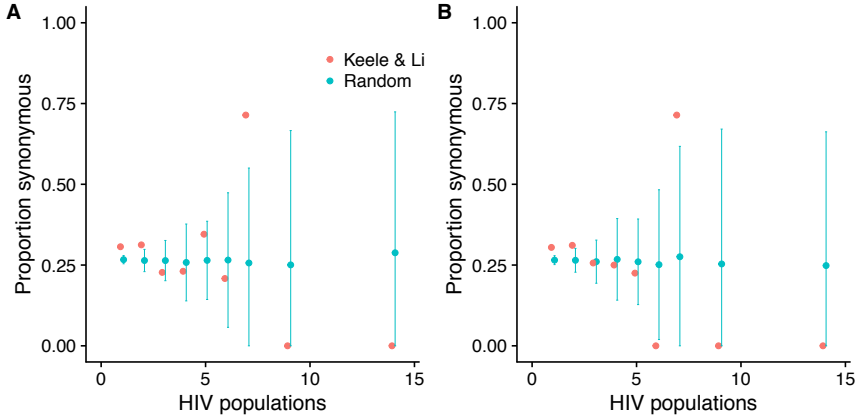

Figure 4. Convergent mutations are not significantly more likely to be synonymous. (**A**) The proportion of synonymous mutations in the Keele and Li data compared to randomly distributed mutations (100 simulations). For each randomization the number of mutations in each convergence category (x-axis) was kept constant. The error bars show standard deviations. (**B**) The same as (A) but we excluded mutations from the part of the *env* gene that overlaps with the *rev* reading frame, as this could bias our results towards synonymous mutations (synonymous mutations in *env* are likely to be non-synonymous in *rev*). We obtain a significant difference only between private mutations and randomly distributed mutations.

Convergent mutations are not more likely to cause non-synonymous changes
The proportion of synonymous mutations in the Keele and Li data do not differ significantly from the predictions of a neutral model. However, the proportion of synonymous mutations does decline for mutations that occur in three or more populations except for mutations that occur in five and seven populations in parallel (**Figure 4A**). If we exclude the portion of the *env* gene that overlaps with the *rev* reading frame the proportion of synonymous mutations declines steadily from two to five parallel mutations (**Figure 4B**). The lack of significance may be due to the decreasing number of mutations that occur in many HIV populations in parallel. On the other hand, we do find that private mutations are significantly more often synonymous mutations than randomly distributed mutations.

Non-synonymous private mutations occur in highly diverse regions of the *env* gene
The significantly higher proportion of synonymous mutations and the significantly higher nucleotide diversity for private mutations have not the same cause (i.e. the higher proportion of synonymous sites occurring in highly diverse regions of the *env* gene). We can show this by comparing the nucleotide diversity of synonymous and non-synonymous mutations and their effect sizes for randomly distributed mutations.

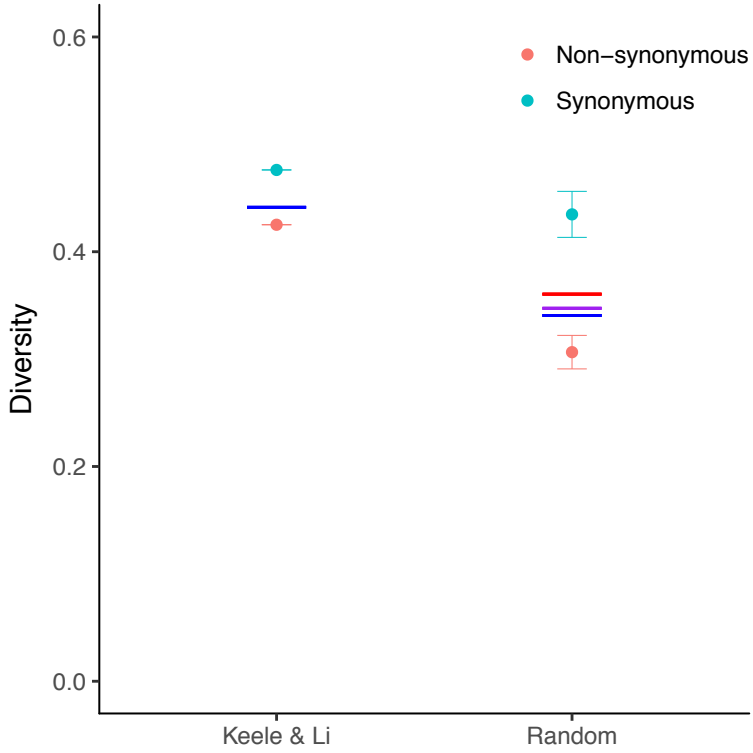

Figure 5. Nucleotide diversity of private non-synonymous and synonymous mutations in the Keele and Li data compared to randomly distributed mutations. The blue lines indicate the mean diversity of all private mutations. The purple line indicates the diversity of random mutations if we adjust the proportion of synonymous mutations to that observed in the Keele and Li data. The red line shows the mean diversity of randomly distributed mutations if the proportion of synonymous mutations is adjust as well as the diversity of synonymous mutations is set to that observed in the Keele and Li data. Error bars show standard deviations.

The nucleotide diversity of all private Keele and Li mutations is on average 0.44, the mean diversity of private mutations from a set of randomly distributed mutations is 0.34 (the maximum of all 1000 randomizations is 0.40, **Figure 5**). The nucleotide diversity at synonymous sites increases from a mean of 0.45 for randomly distributed mutations to 0.48 in the Keele and Li data. The mean diversity of non-synonymous sites increases from 0.30 in the random model to 0.43 in the Keele and Li data. This suggests that non-synonymous mutations may have a much larger effect on nucleotide diversity than synonymous mutations, which are almost identical with the diversity observed in the Keele and Li data.

We can actually calculate the effect sizes of increasing the proportion of synonymous sites in one of our random simulations. If we only increase the proportion of synonymous sites in the random distribution, then the nucleotide diversity increases from 0.340 to 0.347 (purple line in **Figure 5**). When we also increase the nucleotide diversity of synonymous mutations in the neutral model this leads to a nucleotide diversity of 0.36 (red line in **Figure 5**). If we now increase the diversity of non-synonymous mutations then the nucleotide diversity increases to the observed Keele and Li dataset. Hence increasing the proportion of synonymous mutations in our neutral model increase the mean diversity of private mutations by only 0.007 (compared to a total difference of 0.1), whereas if we plug in the high diversity of non-synonymous mutations in the Keele and Li data into our neutral model then the mean diversity increases by 0.08. Together our analyses suggest that synonymous mutations are not the main driver for the significantly higher nucleotide diversity observed for private mutations, but non-synonymous mutations occurring in more diverse regions of the *env* gene are.

Discussion

Convergent evolution, here defined as acquiring identical mutations in independent evolving populations, can be a good indicator for selection. In our example, mutations that emerge more than four times in independent HIV populations are very unlikely to occur in parallel by chance. Furthermore convergent mutations occur preferentially in gp41, which is not the case for private mutations and hence supports the hypothesis that convergent mutations are selected for.

Wood et al. identified 36 *env* positions that are potentially under strong positive selection (Wood et al. 2009) mainly using methods based on dN/dS measurements. We identified a total of 19 mutations that likely arose due to positive selection (**Supplementary Table 1**). Only six mutations are shared between our list and the list published by Wood et al. Another six of the 19 mutations are synonymous mutations that cannot be identified with a dN/dS analysis. The remaining seven mutations are novel putatively selected sites.

Our hypothesis that convergent mutations are selected for is supported by their biased distribution towards gp41. Gp41 is the C-terminal part of Env and responsible for fusion with the host cell (Chan et al. 1997; Skehel and Wiley 1998; Melikyan et al. 2000). It is conceivable that increased fusion efficiency with the host cell provides significant fitness benefits even during exponential growth in the early stages of infection by HIV-1. An example for one of these mutations is E510K. This mutation occurs at a mutation hotspot and is a reversion from a relatively rare amino acid to the database consensus (Foley et al. 2013; Davey et al. 2014). Reversions to ancestral states have been observed to continue for years after infection, consistent with our observation (Carlson et al. 2014). Another example is E648K, which increases viral replication rate for viruses with fusion inhibitor resistance (Bai et al. 2008; Davey et al. 2014). The most common mutation E482K (occurs in 14 infected individuals) is not in gp41 but has also been observed in a previous experiment as the most common mutation in a humanized mouse experiment (Ince et al. 2010). All three mutations have also been identified by Wood et al. as positively selected sites (Wood et al. 2009).

Private mutations also do not evolve entirely neutrally, as private mutations are more likely to be synonymous than expected by chance and also occur in more diverse regions of the *env* gene. These two observations may have the same cause namely that the higher nucleotide diversity observed for private mutations is caused by overrepresented synonymous mutations. We show that this is not the case because the effect size of the increased proportion of synonymous mutations is five times smaller compared to the effect size of the increased nucleotide diversity of non-synonymous mutations (0.014 compared to 0.068). Hence there is a strong shift in the distribution of non-synonymous mutations towards highly diverse regions of the *env* gene. It is possible that this shift is caused by mutational hotspots that preferentially cause non-synonymous mutations but this hypothesis is more difficult to test.

Mutational hotspots as such are also difficult to identify, because a hotspot could be a nucleotide site that changes between two, three or all four nucleotides. If it only switches between two nucleotides the maximum diversity at this position would be one. This seems to be the case for one of the convergent mutations in **Supplementary Table 1** (G7742A). The position only has a diversity of 0.69 across all studied viruses, which is higher than the median (0.2) but far away from the 95th percentile (1.36) or maximum (2). However, if we consider that this site will only likely switch between G and A the maximum diversity at this position would be 1, which makes 0.69 quite diverse.

Although an excess of convergent mutations compared to a neutral model is a good indicator of selection, the strength of selection is difficult to infer from the number of convergent mutations. A mutation that occurs in a large number of individuals may be a mutation that is beneficial in a larger number of different environments. This does not exclude the possibility that the magnitude of the fitness effect for convergent mutations is smaller than a very specialized mutation that is beneficial in only a single HIV-1 population but deleterious in the rest. Nevertheless, it is likely that the convergent mutations we identified have large fitness effects, as they need to have significantly increased in frequency since infection. .

Finally, this raises the question of why these mutations have not been present in the ancestral strain if they arose in parallel in up to 15% of all subjects. Unfortunately, we can only speculate about the answer to this question. First, the most common mutation we observed only occurred in 15% of the subjects, this leaves the possibility that the observed mutations are only beneficial in a small proportion of all hosts. Second, it is possible that these mutations are beneficial during exponential growth but do not confer benefits after the immune response takes effect and hence will be lost during later stages of infection and will therefore also not usually be present in recent infections.

Methods

Identification of mutations

The Keele and Li datasets consist of full-length *env* sequence alignments from a total of 102 and 30 infected individuals, respectively, containing on average 29 (minimum of 11 to a maximum of 63) full-length *env* sequences per infected individual amplified by single genome amplification (Keele et al. 2008; Li et al. 2010). The sequences are from viruses at different stages of early infection but each infected individual was sampled only once. In 78 of the 102 infected individuals from the Keele study and in 17 out of 30 infected individuals of the Li study the infection was likely caused by a single founder strain We only analyzed those sequences alignments (a total of 95), because it becomes almost impossible to distinguish mutations from recombination events for the other cases (Keele et al. 2008).

We defined a mutation as a change from the consensus sequence of the virus population. The consensus sequence of each HIV-1 population is likely to be identical to the most recent common ancestor or even the sequence of the founder virus (Keele et al. 2008). Between infected individuals the consensus sequences differ. Hence to be able to compare mutations between individuals we also aligned all sequences to a reference sequence.

Mutation comparison between HIV populations

To be able to compare mutations that occurred in virus populations from different infected individuals we aligned each of these 95 alignments individually to a reference sequence (HIV-1 NL4-3). Reference alignments were performed using Clustal-Omega (Sievers et al. 2011). This resulted in 95 individual reference alignments. Mutations that occur in virus populations of different infected individuals that changed from the same consensus nucleotide to the same mutated nucleotide are considered identical if they were mapped to the same reference position.

Reference sequence

The HIV-1 NL4-3 reference has been sequenced and assembled previously (Giallonardo et al. 2014). Differences to the HIV-1 NL4-3 reference strain (accession AF324493) are listed in Supplementary Table 3 of the same paper (Giallonardo et al. 2014).

Neutral mutation distribution model

To assess whether the number of observed convergent mutations is indicative of selection we compared the Keele and Li data to a neutral model. We constructed the model in the following way:

1) We determined all possible mutations for each consensus sequence individually to generate a pool of mutations from which we can later randomly draw mutations.

2) To make sure that our results are not affected by mutation biases we amplified mutations according to their respective substitution rates measured from the Keele and Li dataset. First we counted the frequency of each of the twelve possible substitutions (A to C, A to G…). We normalized the lowest value in the substitution matrix to one and the remaining values in the matrix were scaled up accordingly. Each mutation from step one was amplified by the respective value in the substitution matrix.

3) From the resulting mutations we randomly drew (without replacement) the same number of mutations that was observed for each HIV-1 population in the Keele and Li datasets.
4) We repeated step 3) 1000 times to obtain statistically robust results.

Randomly distributing mutations keeping the number of convergent mutations constant
To compare characteristics of convergent mutations to a neutral null model we randomized mutations for each convergence category (i.e. mutations occurring in different numbers of HIV-1 populations). In this randomization, we (1) maintain substitution rates; (2) maintain the origin of the mutation, i.e. mutations are randomly selected from the same HIV-1 population where they were observed in the Keele and Li dataset; and (3) only chose identical mutations if they were identical in the Keele and Li datasets.

Measuring nucleotide diversity
Here we define nucleotide diversity as Shannon entropy at a certain position i for all consensus sequences.

$$H_i = -\sum_{j \in \{A,T,C,G\}} p_j \, log_2 \, p_j$$

Where p_j is the proportion of nucleotide j in all consensus sequences at reference position i.

The diversity at a particular position in the *env* gene was calculated by calculating the Shannon entropy across all consensus sequences that do not contain a gap '-'. Hence as there are only four possible different nucleotides the minimum nucleotide diversity is 0 and the maximum is 2.

Measuring the proportion of synonymous mutations
For each mutation we can determine whether it causes a synonymous or non-synonymous change. The proportion of synonymous mutations is the number of synonymous changes divided by the number of total changes. For **Figure 3B** we excluded all positions from 8369 to 8643 in the HIV-1 NL4-3 reference sequence because this is the region that overlaps with the *rev* reading frame.

References

Bai X, Wilson KL, Seedorff JE, Ahrens D, Green J, Davison DK, Jin L, Stanfield-Oakley SA, Mosier SM, Melby TE, et al. 2008. Impact of the enfuvirtide resistance mutation N43D and the associated baseline polymorphism E137K on peptide sensitivity and six-helix bundle structure. Biochemistry 47:6662–6670.

Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, Haas DW, Martinez-Picado J, Dalmau J, López-Galíndez C, et al. 2013. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. Elife 2:e01123.

Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, Kadie C, Carlson J, Yusim K, McMahon B, Gaschen B, et al. 2007. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science 315:1583–1586.

Borrow P, Lewicki H, Wei X, Horwitz MS, Peffer N, Meyers H, Nelson JA, Gairin JE, Hahn BH, Oldstone MB, et al. 1997. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. Nat. Med. 3:205–211.

Brandis G, Hughes D. 2016. The Selective Advantage of Synonymous Codon Usage Bias in Salmonella.Ibba M, editor. PLoS Genet. 12:e1005926.

Bull JJ, Badgett MR, Wichman HA, Huelsenbeck JP, Hillis DM, Gulati A, Ho C, Molineux IJ. 1997. Exceptional convergent evolution in a virus. Genetics 147:1497–1507.

Carlson JM, Du VY, Pfeifer N, Bansal A, Tan VYF, Power K, Brumme CJ, Kreimer A, DeZiel CE, Fusi N, et al. 2016. Impact of pre-adapted HIV transmission. Nat. Med. 22:606–613.

Carlson JM, Schaefer M, Monaco DC, Batorsky R, Claiborne DT, Prince J, Deymier MJ, Ende ZS, Klatt NR, DeZiel CE, et al. 2014. HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. Science 345:1254031–1254031.

Chan DC, Fass D, Berger JM, Kim PS. 1997. Core structure of gp41 from the HIV envelope glycoprotein. Cell 89:263–273.

Coffin JM. 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science 267:483–489.

Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. 16:372–382.

Davey NE, Satagopam VP, Santiago-Mozos S, Villacorta-Martin C, Bharat TAM, Schneider R, Briggs JAG. 2014. The HIV mutation browser: a resource for human immunodeficiency virus mutagenesis and polymorphism data.Rein A, editor. PLoS Comput Biol 10:e1003951.

Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A, et al. 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nature Genetics 45:1183–1189.

Foley B, Leitner T, Apetrei C, Hahn B. 2013. HIV sequence compendium 2013. Los Alamos.

Fraser C, Lythgoe K, Leventhal GE, Shirreff G, Hollingsworth TD, Alizon S, Bonhoeffer S. 2014. Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. Science 343:1243727–1243727.

Giallonardo FD, Töpfer A, Rey M, Prabhakaran S, Duport Y, Leemann C, Schmutz S, Campbell NK, Joos B, Lecca MR, et al. 2014. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res 42:e115–e115.

Goulder PJ, Phillips RE, Colbert RA, McAdam S, Ogg G, Nowak MA, Giangrande P, Luzzi G, Morgan B, Edwards A, et al. 1997. Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. Nat. Med. 3:212–217.

Ince WL, Zhang L, Jiang Q, Arrildt K, Su L, Swanstrom R. 2010. Evolution of the HIV-1 env gene in the Rag2-/- gammaC-/- humanized mouse model. J. Virol. 84:2740–2752.

Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, et al. 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A 105:7552–7557.

Koenig S, Conley AJ, Brewah YA, Jones GM, Leath S, Boots LJ, Davey V, Pantaleo G, Demarest JF, Carter C. 1995. Transfer of HIV-1-specific cytotoxic T lymphocytes to an AIDS patient leads to selection for mutant HIV variants and subsequent disease progression. Nat. Med. 1:330–336.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. PLoS Genet. 4:e1000304.

Lemey P, Rambaut A, Pybus OG. 2006. HIV evolutionary dynamics within and among hosts. AIDS Rev 8:125–140.

Li H, Bar KJ, Wang S, Decker JM, Chen Y, Sun C, Salazar-Gonzalez JF, Salazar MG, Learn GH, Morgan CJ, et al. 2010. High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men.Douek DC, editor. PLoS Pathog. 6:e1000890.

Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, Fire AZ, Roskin KM, Schramm CA, Zhang Z, et al. 2013. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature 496:469–476.

Lind PA, Berg OG, Andersson DI. 2010. Mutational robustness of ribosomal protein genes. Science 330:825–827.

Melikyan GB, Markosyan RM, Hemmati H, Delmedico MK, Lambert DM, Cohen FS. 2000. Evidence that the transition of HIV-1 gp41 into a six-helix bundle, not the bundle configuration, induces membrane fusion. J. Cell Biol. 151:413–423.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266:66–71.

Poon AFY, Swenson LC, Dong WWY, Deng W, Kosakovsky Pond SL, Brumme ZL, Mullins JI, Richman DD, Harrigan PR, Frost SDW. 2010. Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. Mol. Biol. Evol. 27:819–832.

Sadler HA, Stenglein MD, Harris RS, Mansky LM. 2010. APOBEC3G contributes to HIV-1 variation through sublethal mutagenesis. J. Virol. 84:7396–7404.

Serb JM, Eernisse DJ. 2008. Charting Evolution's Trajectory: Using Molluscan Eye Diversity to Understand Parallel and Convergent Evolution. Evo Edu Outreach 1:439–447.

Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, et al. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. 73:10489–10502.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539–539.

Skehel JJ, Wiley DC. 1998. Coiled coils in both intracellular vesicle and viral membrane fusion. Cell 95:871–874.

Trkola A, Kuster H, Rusert P, Joos B, Fischer M, Leemann C, Manrique A, Huber M, Rehr M, Oxenius A, et al. 2005. Delay of HIV-1 rebound after cessation of antiretroviral therapy through passive transfer of human neutralizing antibodies. Nat. Med. 11:615–622.

Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Salazar-Gonzalez JF, Salazar MG, Kilby JM, Saag MS, et al. 2003. Antibody neutralization and escape by HIV-1. Nature 422:307–312.

Wichman HA, Scott LA, Yarber CD, Bull JJ. 2000. Experimental evolution recapitulates natural evolution. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 355:1677–1684.

Wood N, Bhattacharya T, Keele BF, Giorgi E, Liu M, Gaschen B, Daniels M, Ferrari G, Haynes BF, McMichael A, et al. 2009. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC.Douek DC, editor. PLoS Pathog. 5:e1000414.

Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, Neher RA. 2015. Population genomics of intrapatient HIV-1 evolution. Elife 4:13239.

Zanini F, Neher RA. 2013. Quantifying selection against synonymous mutations in HIV-1 env evolution. J. Virol. 87:11843–11850.

Zanini F, Puller V, Brodin J, Albert J, Neher RA. 2017. In vivo mutation rates and the landscape of fitness costs of HIV-1. Virus Evol 3:vex003.