# 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 Microbial Population Biology, 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. Complex organisms with large genomes only rarely and only under very strong selection converge to the same genotype. In contrast, independent virus populations with very small genomes often acquire identical mutations. Here we test the hypothesis of whether convergence in early HIV-1 infection is common enough to serve as an indicator for selection. To this end, 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. Convergent mutations are not evenly distributed across the *env* gene, but more likely to occur in gp41, which suggests that convergent mutations provide a selective advantage and hence are positively selected for. In contrast, mutations that are only found in an HIV-1 population of a single individual are significantly affected by purifying selection. Our analysis suggests that comparisons between convergent and private mutations with neutral models allow us to identify positive and negative selection in small viral genomes. Our results also show that selection significantly shapes HIV-1 populations even before the onset of the adaptive immune system.**

* 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. Phenotypic convergence is the repeated and independent evolution of a particular phenotype. Textbook examples are the evolution of flight, which occurred in bats, birds and insects independently, or the evolution of the eye, which arose many times independently across the tree of life (Serb and Eernisse 2008). We can also observe convergent evolution at the level of the genotype, which usually occurs when organisms are under extremely high selection pressure. Well known examples are, the evolution of antibiotic resistance through the acquisition of mutations in the same gene or even at the same position in that gene (Farhat et al. 2013; Lindsey et al. 2013) or the evolution of cancer, which usually involves a driver mutation in a typical cancer or cancer suppressor gene (Miki et al. 1994). Convergent evolution in viruses is particularly common due to the small and functionally constrained virus genomes (Bull et al. 1997; Crandall et al. 1999; Wichman et al. 2000; Xue et al. 2017).

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).

Some of the most frequently observed mutations are reversions to the population consensus (Leslie et al. 2004). These reversions and other adaptive mutations are usually the result of strong positive selection when the virus population escapes from the immune system. Hence, during early infection of HIV-1 (where immune escape does not seem to play a major role for HIV-1 evolution), HIV-1 evolution has been modeled as a neutral process (Lee et al. 2009; Giorgi et al. 2013). In line with these findings, diversification rates of the virus population have been shown to largely adhere to a molecular clock (Herbeck et al. 2011; Park et al. 2016). However, the adherence to a molecular clock does not mean that there is no selection during early HIV-1 infection. It is more likely that the fast accumulation of neutral mutations obscures selective footprints. Selection is more visible once viral phenotypes are taken into account. For example diversification rates correlate with set point viral load a disease parameter that affects disease severity (Gounder et al. 2015). Similarly, high evolutionary rates have

been shown to correlate with fast disease progression and strong selection has been shown to correlate with slower disease progression (Boutwell et al. 2010; Garcia-Knight et al. 2016).

Selection can be measured in various different ways: (1) Probably the most common way of determining selected nucleotide sites is by comparing the evolutionary rates of non-synonymous nucleotide sites to those of synonymous sites (dN/dS) on a phylogenetic tree (Kosakovsky Pond et al. 2008; Wood et al. 2009; Boutwell et al. 2010; Yoshida et al. 2011). (2) If population sequence data that spans multiple time points is available then it is possible to identify positively selected sites by assessing the change in mutant frequency over time (Henn et al. 2012). (3) More recently it has been demonstrated that it is possible to determine nucleotide sites under selection by analyzing the distribution of those sites across a gene (Zhang and Townsend 2009; Zhao et al. 2017).

Here we will focus on yet another way of determining nucleotide sites affected by selection: measuring the frequency of convergent mutations across different HIV-1 populations from different hosts (i.e. we define an HIV-1 population as all viruses from the same infected individual). Convergent mutations are mutations that occur in independent HIV-1 populations in parallel. More specifically convergence requires that two populations that share the same nucleotide at a specific position in the genome acquire the same mutation. However, due to the small HIV-1 genome it is possible that most convergent mutations are the result of chance and not selection. Appropriate null models can be exceedingly helpful to distinguish adaptive from neutral mutations (Stayton 2015). Once we have made the distinction between selected and accidental convergence we reserve the term convergent mutation for mutations that occur in parallel in different populations more often than one would expect it under a neutral model.

By comparing mutations from HIV-1 populations to a random null model, we identify convergent mutations that are positively selected for. To this end we reanalyze 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. The convergent mutations we observe are significantly skewed towards gp41. This biased distribution of mutations indicates that convergent mutations are positively selected for. In contrast, the biased distribution of private mutations (mutations that are found in a single HIV-1 population only) towards high diversity sites in the *env* gene suggests that private mutations are strongly affected by purifying selection. Hence our results show that HIV-1 is under a range of selection pressures immediately after infection of a new host even before the adaptive immune system decimates the viral population.

**Results**

We reanalyzed two datasets from previous studies on the evolution of HIV-1 during early infection (samples were estimated to have been taken <50 days after 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.

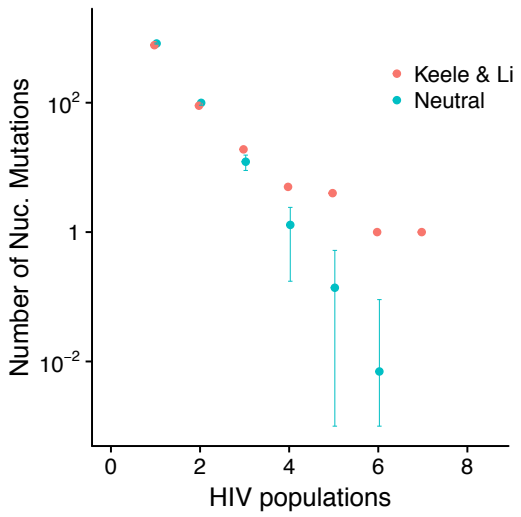


**Figure 1. Convergent mutations are unusually frequent during early infection of HIV-1.** Comparison between nucleotide mutations observed in one or more HIV-1 populations. Each HIV-1 population is isolated from a different infected individual. The Keele and Li data is compared to a neutral model of 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. Data generated under the neutral model show the mean and standard deviations of 1000 simulations.

**Convergent mutations are unlikely under a neutral model of evolution**

We observed convergent evolution for a large number of the 1065 mutations identified in the Keele and Li datasets (**Figure 1, Supplementary Data 1**). Mutations occurring in three or more populations in parallel are overrepresented in the dataset (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. We do not see a linear decline for the Keele and Li data. Instead the decline levels off for highly convergent mutations (**Figure 1**). For example, there is one mutation that occurs in seven individuals in parallel, the likelihood to observe this in a neutral model is about 0.001. For lower levels of convergence the deviation between neutral model and observation is still significant. For example, in our simulations we observe on average one mutation that occurs in four HIV populations in parallel. In contrast, there are five such mutations in the Keele and Li data (**Figure 1**).

In total we determined 11 mutations that are likely to have been positively selected because they occur in four or more HIV-1 populations for which we expect to see less than one mutation under a neutral model of evolution (**Table 1**). Mutations that occur in three populations in parallel are also overrepresented. We expect about 12 such mutations under our neutral model but observe 19 in the Keele and Li data. Hence, we probably see seven out of 19 mutations as the result of selection. Identifying these seven mutations however is difficult.

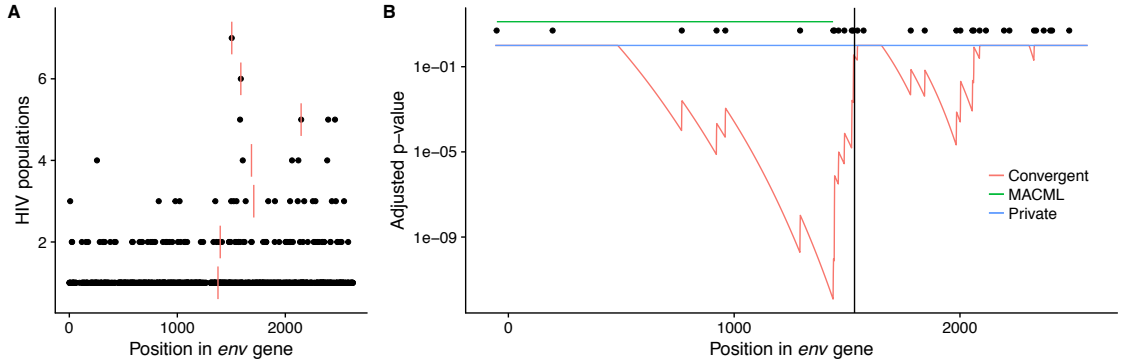



**Figure 2. Convergent mutations are predominantly located in gp41. (A)** Shows the positions of mutations across the *env* gene on the x-axis and the number of HIV-1 populations the mutations were observed in on the y-axis. Red bars indicate the mean of all positions of the mutations that occur in the same number of HIV-1 populations. (B) Shows whether there is a significant difference (adjusted p-value on y-axis) in the distribution of mutations before a certain position in the *env* gene compared to after this position. There is no significant signal for the distribution of private mutations (blue line). The red line shows the adjusted p-values for the distribution of mutations that occur in at least three HIV populations in parallel. The lowest p-value is found at nucleotide position 1438 in the *env* gene (position 7663 in HXB2, 479 in the *env* protein). A low mutation density region was identified by MACML (Zhang and Townsend 2009) and is indicated in green. The end of this region is also the minimum of the adjusted mutation distribution p-value from above. The black vertical line indicates the start of the Gp41 protein, a fusion protein that is part of the *env* gene. Black dots above the blue line indicate the positions of mutations that occur in three or more HIV-1 populations.

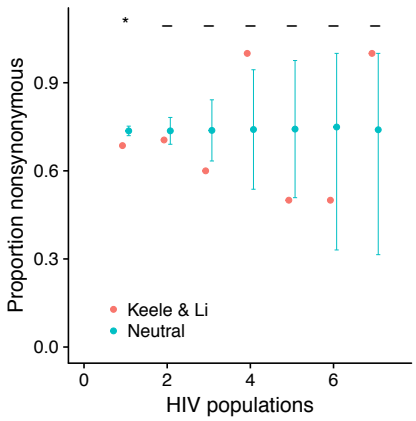
**Convergent mutations occur predominantly in gp41**

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 439, which is very close to half the length of the *env* gene (428). However, the mean increases for mutations that occur in more than one HIV-1 population.

155 **Table 1. Convergent mutations occurring in four or more HIV-1 populations in parallel.**

| Nuc. Mut. | AA Mut. | Freq. | Subtype freq.* of mutation | *env* part | Observed in other studies | Identified by Wood et al. |
|---|---|---|---|---|---|---|
| **C6420T** | H66Y | 4 | 0 | gp120 | --- | Y |
| **A7770G** | M516V | 4 | 0 | gp41 | --- | N |
| **A8216G** | S668G | 4 | 1 | gp41 | Confers neutralization sensitivity (O'Rourke et al. 2012) | N |
| **G8552A** | L776L | 4 | 155 | gp41 | --- | N |
| **G8285A** | M687I | 4 | 20 | gp41 | --- | N |
| **G8624A** | L800L | 5 | 113 | gp41 | --- | N |
| **A7745G** | Q507Q | 5 | 49 | gp41 | --- | N |
| **G8561A** | T779T | 5 | 25 | gp41 | --- | N |
| **G8311A** | R696K | 5 | 6 | gp41 | --- | Y |
| **G7742A** | E510K | 6 | 162 | gp120 | Reversion from E to the database consensus K. | Y |
| **G7668A** | E482K | 7 | 1 | gp120 | Most common mutation observed in an *in vivo* experiment with humanized mice (Ince et al. 2010). | Y |

156 *Frequency of mutant in subtype consensus out of 169 sequences
157
158 When dividing the *env* gene at every possible position and comparing the number of mutations
159 observed before and after this position then we find, again, that there is not a single position that shows
160 a significant difference between the distribution of mutations for the 5' and 3' end of the gene after
161 adjusting for multiple testing (**Figure 2B**). However, mutations that occur in three or more HIV
162 populations show very significant differences in their distribution. We found the largest difference at
163 position 479 (adj. p-value 1e-12), separating the 5' part of the *env* gene with few mutations from the 3'
164 part with many mutations. The 3' part of the *env* gene encodes for gp41, a protein responsible for
165 forming the six-helix bundle during cell fusion with the host cell (Chan et al. 1997; Skehel and Wiley
166 1998; Melikyan et al. 2000).
167




168
169 **Figure 3. Private mutations show lower proportions of non-synonymous mutations than expected**
170 **by chance.** The proportion of synonymous mutations in the Keele and Li data compared to a neutral
171 model (randomly distributed mutations,1000 simulations). For each randomization the number of
172 mutations in each convergence category (x-axis) was kept constant. A "—" means there is no
173 significant difference between the neutral model and the Keele and Li data, whereas a "*" means that

the proportion of non-synonymous mutations in the Keele and Li data was lower in at least 950 out of 1000 simulations. Mutations in overlapping reading frames are excluded. The error bars show standard deviations.

**Private mutations are more likely to cause synonymous changes**

The proportion of synonymous mutations in the Keele and Li data differ significantly from the predictions of a neutral model only for private mutations and mutations that occur in five populations in parallel (**Figure 3**). In both cases we observe more synonymous mutations in the Keele and Li data than expected from the neutral model. The observation of convergent synonymous mutations is probably an indication for the fact that synonymous mutations are not always selectively neutral in viruses (Crandall et al. 1999; Cuevas et al. 2012; Zanini and Neher 2013). We also observe higher proportions of private synonymous mutations than expected by chance. In contrast to convergent mutations, purifying selection has probably caused the disappearance of non-synonymous mutations due to lethal or highly deleterious effects. The disappearance of lethal and highly deleterious mutations then leads to an overrepresentation of synonymous mutations. Generally synonymous mutations are more likely to have less deleterious effects because they do not change the amino acid sequence of proteins.

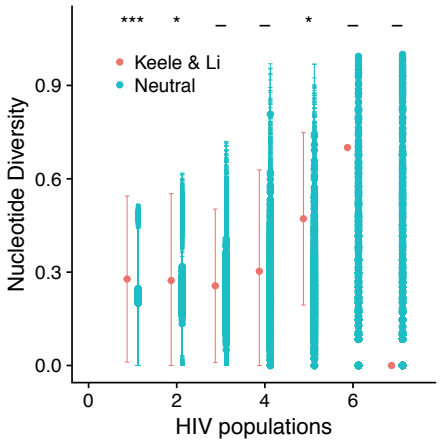



**Figure 4. Private mutations occur in positions of high nucleotide diversity.** The figure shows the nucleotide diversity (mean and standard deviation) at nucleotide positions, for which Keele and Li mutations were identified (orange) as well as the positions for 1000 individual random mutation distributions (turquois). The positions of the same number of mutations we observe in *N* HIV-1 populations was randomized in the neutral model. The diversity was determined at each position in the *env* consensus sequence alignment of all 95 HIV-1 populations (see Methods). A diversity of 1 means that all nucleotides observed at a certain position occur at equal quantities. "***" indicates that there was not a single simulation that showed a mean diversity higher or equal to that observed in the Keele and Li data among 1000 simulations. "*" indicates there were fewer than 50 simulations that showed a diversity equal or greater than the Keele and Li data. "—" indicates that there was no significant difference between the diversity observed in the Keele and Li data and 1000 random distributions of mutations.

**Private mutations are found significantly more frequently at highly diverse nucleotide positions**

Different regions of the *env* gene are under different selective pressures. Conserved regions of the *env* gene are under purifying selection whereas highly diverse regions are either under positive selection or under relaxed selection. We can identify conserved regions by measuring the nucleotide diversity in a consensus sequence alignment of all 95 HIV-1 populations. A single virus recently founded each of the 95 HIV-1 populations. Hence the consensus sequences are independent sequence samples from the HIV-1 population as a whole. Measuring the nucleotide diversity at each position in such an alignment allows us to determine regions of the *env* gene that are likely affected by purifying selection.

We observe that, in line with our analysis of non-synonymous and synonymous sites, private mutations and mutations occurring in two and five populations in parallel are found in more diverse regions of the *env* gene than compared to neutral simulations (**Figure 4**). Because of our initial convergence analysis (**Figure 1**) we postulate that the cause for the high diversity at sites that we observe in five populations in parallel is a result of positive selection. In contrast, the reason we find private mutations in regions

of high diversity is probably not because private mutations are all positively selected for. Instead we propose that the pattern we see for private mutations is caused by the absence of mutations in highly conserved regions of the *env* gene. We assume that mutations in conserved regions of the *env* gene are likely to cause more severe deleterious effects than mutations in highly diverse regions of the *env* gene. Deleterious mutations will be much less likely to be present in our dataset than neutral or beneficial mutations, because the viruses carrying these mutations will be purged before being sampled. Absence of mutations in regions of low diversity will increase the average diversity of private mutations. Therefore we propose that the higher diversity we observe for private mutations is probably a result of purifying selection.

**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 understand the causal relationship by comparing the nucleotide diversity of synonymous and non-synonymous mutations and their effect sizes for randomly distributed mutations.

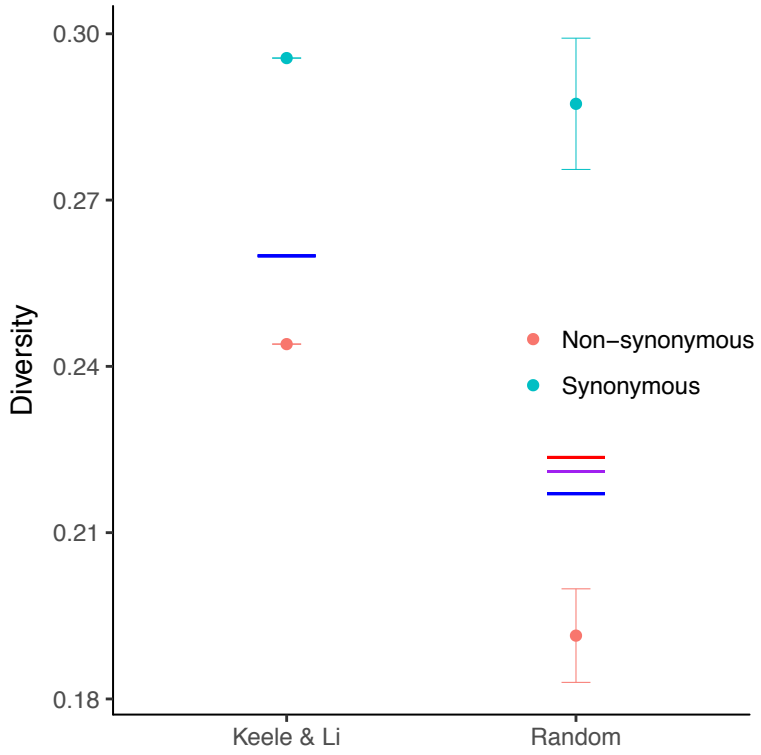



**Figure 5. Nucleotide diversity of private non-synonymous and synonymous mutations in the Keele and Li data compared to randomly distributed mutations.** The blue lines indicate the mean diversity of all private mutations. The purple line indicates the diversity of random mutations if we adjust the 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 and 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 0.26, the mean diversity of private mutations from a set of randomly distributed mutations is 0.22 (the maximum of all 1000 randomizations is 0.24, **Figure 5**). The nucleotide diversity at synonymous sites increases from a mean of 0.29 for randomly distributed mutations to 0.3 in the Keele and Li data. The mean diversity of non-synonymous sites increases from 0.19 in the randomization to 0.24 in the Keele and Li data. Hence, non-synonymous mutations deviate much more from the neutral expectation than synonymous mutations, which are almost identical to the neutral expectation.

If we set the proportion of synonymous sites in one of our random simulations to the value observed in the Keele and Li data then the nucleotide diversity increases only slightly from 0.217 to 0.221 (purple

line in **Figure 5**). If we instead increase the diversity of non-synonymous mutations to that observed in the Keele and Li data, then the mean diversity increases by 0.04, a ten-fold difference in effect size compared to increasing the proportion of synonymous mutations. This demonstrates that the increased proportion of synonymous mutations is not the main driver for the significantly higher nucleotide diversity of private mutations compared to our neutral simulation. Instead, the differences in diversity can be explained by non-synonymous mutations that are more likely to occur at highly diverse sites of the *env* gene (see also **Supplementary Figure 1**).

If purifying selection is the cause for the observed distribution of non-synonymous mutations then non-synonymous mutations do not preferentially occur in high diversity regions of the gene (although mutational hot spots could also cause part of the signal). Instead non-synonymous mutations in conserved regions are absent from the data because these mutations are likely to cause highly deleterious changes, which immediately disappear from the population. We do not see this bias in synonymous mutations because synonymous mutations are more likely to cause neutral or only slightly deleterious changes. Hence, it is more likely for us to observe the complete spectrum of synonymous mutations.

Therefore we can infer that non-synonymous mutations in low diversity regions have very strong negative effects on viral fitness. As one would expect these effects are much smaller for synonymous mutations where the difference between neutral model and Keele and Li data is much smaller (**Figure 5**). Nevertheless, there is a difference (although small) between the neutral model and Keele and Li data for synonymous mutations and we also identify synonymous convergent mutations (**Supplementary Table 1**) indicating that synonymous mutations do have non-neutral effects on HIV-1 evolution.

**Discussion**

Convergent evolution, here defined as acquiring identical mutations in independent evolving populations, can be a good indicator for selection when a large number of independent populations can be sampled. In our example, mutations that emerge more than three 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.

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 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). The most common mutation E482K (occurs in seven 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). Both mutations have also been identified by Wood et al. as positively selected sites (Wood et al. 2009).

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 11 mutations that likely arose due to positive selection (**Supplementary Table 1**). Only four mutations are shared between our list and the list published by Wood et al. We identify seven novel putatively selected sites four of which are synonymous sites and hence could not have been identified with dN/dS. There are several reasons for the small overlap between the mutations identified by us and by Wood et al. First, because we wanted to exclude mutations caused by APOBEC as well as sequences that are unlikely to replicate due to a high mutational load, we excluded all sequences that contained three or more mutations compared to the consensus sequence. Second, Wood et al. have not included the Li dataset and hence only used 81 infected individuals instead of 95 in their analysis. Third, it is possible that some non-convergent mutations are also positively selected. These mutations may be beneficial only in a particular host and hence it would be impossible for our analysis to identify such mutations. Fourth,

316 we have not included mutations that occur in three HIV-1 populations in parallel in our list, although
317 about seven of those mutations are potentially the result of positive selection. Wood et al also identified
318 one of the 19 mutations that occur in three populations in parallel (E322K) as under positive selection.
319
320 Although an excess of convergent mutations compared to a neutral model is a good indicator of
321 selection, the strength of selection is difficult to infer from the number of convergent mutations. A
322 mutation that occurs in a large number of individuals may be a mutation that is beneficial in a large
323 number of different environments. Such a mutation may still have a smaller fitness effect than a
324 mutation that increases the viral fitness only in a single environment but deleterious in the rest.
325 Nevertheless, it is likely that the convergent mutations we identified have large fitness effects, as they
326 need to have significantly increased in frequency since infection to be present in the Keele and Li
327 dataset. To infer fitness effect distributions it is possible to explicitly simulate the evolution of HIV-1
328 and infer model parameter by for example fitting the distribution of convergent mutations (Bons et al.
329 2017).
330
331 Finally, our work also raises the question of why these mutations have not been present in the ancestral
332 strain if they arose in parallel in up to 7% of all subjects. Unfortunately, we can only speculate about
333 the answer to this question. First, the most common mutation we observed only occurred in 7% of the
334 subjects. It is possible that the observed mutations are only beneficial in a small proportion of all hosts.
335 Second, it is possible that the convergent mutations we observed are beneficial during exponential
336 growth but do not confer benefits after the immune response takes effect and hence will be lost during
337 later stages of infection. Third, mutations that are beneficial in most environments (human hosts) have
338 probably already swept through the entire HIV-1 population and are now part of the consensus
339 sequence. Early on during the HIV-1 epidemic while the virus was still adapting to the human host one
340 would expect higher levels of convergence to have occurred than the maximum of 7% we observe here.
341
342 **Methods**
343
344 *Identification of mutations*
345 The Keele and Li datasets consist of full-length *env* sequence alignments from a total of 102 and 30
346 infected individuals, respectively, containing on average 29 (minimum of 11 to a maximum of 63) full-
347 length *env* sequences per infected individual amplified by single genome amplification (Keele et al.
348 2008; Li et al. 2010). The sequences are from viruses at different stages of early infection but each
349 infected individual was sampled only once. In 78 of the 102 infected individuals from the Keele study
350 and in 17 out of 30 infected individuals of the Li study the infection was likely caused by a single
351 founder strain. We only analyzed those sequence alignments (a total of 95), because it becomes almost
352 impossible to distinguish mutations from recombination events for the other cases (Keele et al. 2008).
353
354 We defined a mutation as a change from the consensus sequence of the virus population. The
355 consensus sequence of each HIV-1 population is likely to be identical to the most recent common
356 ancestor or even the sequence of the founder virus (Keele et al. 2008). Between infected individuals the
357 consensus sequences differ. Hence to be able to compare mutations between individuals we also
358 aligned all sequences to each other.
359
360 *Mutation comparison between HIV populations*
361 To be able to compare mutations that occurred in virus populations from different infected individuals
362 we aligned all of 95 alignments to each other as well as a reference sequence (HXB2). Alignments
363 were performed using Clustal-Omega with standard settings (Sievers et al. 2011). Mutations that occur
364 in virus populations of different infected individuals that changed the same consensus nucleotide to the
365 same mutated nucleotide are considered identical.
366
367 *Neutral mutation distribution model*
368 To assess whether the number of observed convergent mutations is indicative of selection we compared
369 the Keele and Li data to a neutral model. We constructed the model in the following way:
370 1) We determined all possible mutations for each consensus sequence individually to generate a pool of
371 mutations from which we can later randomly draw mutations.
372 2) To make sure that our results are not affected by mutation biases we amplified mutations according
373 to their respective substitution rates measured from the Keele and Li dataset. First we counted the
374 frequency of each of the twelve possible substitutions (A to C, A to G…). We normalized the lowest
375 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.

$$H_i = -\sum_{j \in \{A,T,C,G\}} p_j \, log_{\#bases\ at\ pos\ i}\, p_j$$

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 determined by calculating the Shannon entropy across all positions of a consensus sequence alignment. To normalize the diversity value to a range from 0 to 1, we determined the number of different bases present in all consensus sequences at each position and used this number as the base for the logarithm. Hence a diversity of 1 means that all observed nucleotides at a certain position occur in equal quantities. If only a single nucleotide occurred at a given position then the nucleotide diversity was set to 0.

*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 3** we excluded all positions from 8379 to 8653 in the HXB2 reference sequence because this region overlaps with the *rev* reading frame.

*Measuring differences in the distribution of mutations across the* env *gene*
We performed a chi-square test that tested whether the number of mutations before a given positions is significantly different from the number of mutations after this position for the corresponding sequence lengths. This means we performed a total of 2623 chi-square tests. We then adjusted the p-values from these test for multiple tests with the Bonferroni Correction. All tests were used as implemented in R (Team).

Figures were created using cowplot and ggplot2 (Wickham 2016; Wilke 2016).

**References**

Bons E, Bertels F, Regoes RR. 2017. Estimating the Mutational Fitness Effects Distribution during early HIV infection. bioRxiv:185678.

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.

Boutwell CL, Rolland MM, Herbeck JT, Mullins JI, Allen TM. 2010. Viral evolution and escape during acute HIV-1 infection. J. Infect. Dis. 202 Suppl 2:S309–S314.

Bull JJ, Badgett MR, Wichman HA, Huelsenbeck JP, Hillis DM, Gulati A, Ho C, Molineux IJ. 1997.

430 Exceptional convergent evolution in a virus. Genetics 147:1497–1507.

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

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

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

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

441 Cuevas JM, Domingo-Calap P, Sanjuán R. 2012. The fitness effects of synonymous mutations in DNA
442 and RNA viruses. Mol. Biol. Evol. 29:17–20.

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

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

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

450 Garcia-Knight MA, Slyker J, Payne BL, Pond SLK, de Silva TI, Chohan B, Khasimwa B, Mbori-
451 Ngacha D, John-Stewart G, Rowland-Jones SL, et al. 2016. Viral Evolution and Cytotoxic T Cell
452 Restricted Selection in Acute Infant HIV-1 Infection. Sci Rep 6:29536.

453 Giorgi EE, Korber BT, Perelson AS, Bhattacharya T. 2013. Modeling sequence evolution in HIV-1
454 infection with recombination. J. Theor. Biol. 329:82–93.

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

458 Gounder K, Padayachi N, Mann JK, Radebe M, Mokgoro M, van der Stok M, Mkhize L, Mncube Z,
459 Jaggernath M, Reddy T, et al. 2015. High frequency of transmitted HIV-1 Gag HLA class I-driven
460 immune escape variants but minimal immune selection over the first year of clade C
461 infection.Gray CM, editor. PLoS ONE 10:e0119886.

462 Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf
463 CM, Ryan EM, Gnerre S, et al. 2012. Whole Genome Deep Sequencing of HIV-1 Reveals the
464 Impact of Early Minor Variants Upon Immune Recognition During Acute Infection.Walker CM,
465 editor. PLoS Pathog. 8:e1002529.

466 Herbeck JT, Rolland M, Liu Y, McLaughlin S, McNevin J, Zhao H, Wong K, Stoddard JN, Raugi D,
467 Sorensen S, et al. 2011. Demographic processes affect HIV-1 evolution in primary infection
468 before the onset of selective processes. J. Virol. 85:7523–7534.

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

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

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

477 Kosakovsky Pond SL, Poon AFY, Zárate S, Smith DM, Little SJ, Pillai SK, Ellis RJ, Wong JK, Leigh
478 Brown AJ, Richman DD, et al. 2008. Estimating selection pressures on HIV-1 using phylogenetic
479 likelihood models.Gezmu M, editor. Stat Med 27:4779–4789.

480 Lee HY, Giorgi EE, Keele BF, Gaschen B, Athreya GS, Salazar-Gonzalez JF, Pham KT, Goepfert PA,
481 Kilby JM, Saag MS, et al. 2009. Modeling sequence evolution in acute HIV-1 infection. J. Theor.
482 Biol. 261:341–360.

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

485 Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, Tang Y, Holmes EC, Allen T,
486 Prado JG, et al. 2004. HIV evolution: CTL escape mutation and reversion after transmission. Nat.
487 Med. 10:282–289.

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

491 Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, Fire AZ, Roskin KM, Schramm CA, Zhang
492 Z, et al. 2013. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature
493 496:469–476.

494 Lindsey HA, Gallie J, Taylor S, Kerr B. 2013. Evolutionary rescue from extinction is contingent on a
495 lower rate of environmental change. Nature 494:463–467.

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

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

502 O'Rourke SM, Schweighardt B, Phung P, Mesa KA, Vollrath AL, Tatsuno GP, To B, Sinangil F,
503 Limoli K, Wrin T, et al. 2012. Sequences in glycoprotein gp41, the CD4 binding site, and the V2
504 domain regulate sensitivity and resistance of HIV-1 to broadly neutralizing antibodies. J. Virol.
505 86:12105–12114.

506 Park SY, Love TMT, Perelson AS, Mack WJ, Lee HY. 2016. Molecular clock of HIV-1 envelope
507 genes under early immune selection. Retrovirology 13:38.

508 Poon AFY, Swenson LC, Dong WWY, Deng W, Kosakovsky Pond SL, Brumme ZL, Mullins JI,
509 Richman DD, Harrigan PR, Frost SDW. 2010. Phylogenetic analysis of population-based and
510 deep sequencing data to identify coevolving sites in the nef gene of HIV-1. Mol. Biol. Evol.
511 27:819–832.

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

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

517 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M,
518 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.

Stayton CT. 2015. What does convergent evolution mean? The interpretation of convergence and its implications in the search for limits to evolution. Interface Focus 5:20150039.

Team RC. R: A Language and Environment for Statistical Computing. Available from: https://www.R-project.org/

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.

Wickham H. 2016. ggplot2. Springer

Wilke CO. 2016. Streamlined Plot Theme and Plot Annotations for "ggplot2" [R package cowplot version 0.8.0]. Available from: https://CRAN.R-project.org/package=cowplot

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.

Xue KS, Stevens-Ayers T, Campbell AP, Englund JA, Pergam SA, Boeckh M, Bloom JD. 2017. Parallel evolution of influenza across multiple spatiotemporal scales. Elife 6:46.

Yoshida I, Sugiura W, Shibata J, Ren F, Yang Z, Tanaka H. 2011. Change of Positive Selection Pressure on HIV-1 Envelope Gene Inferred by Early and Recent Samples.Liberles D, editor. PLoS ONE 6:e18630.

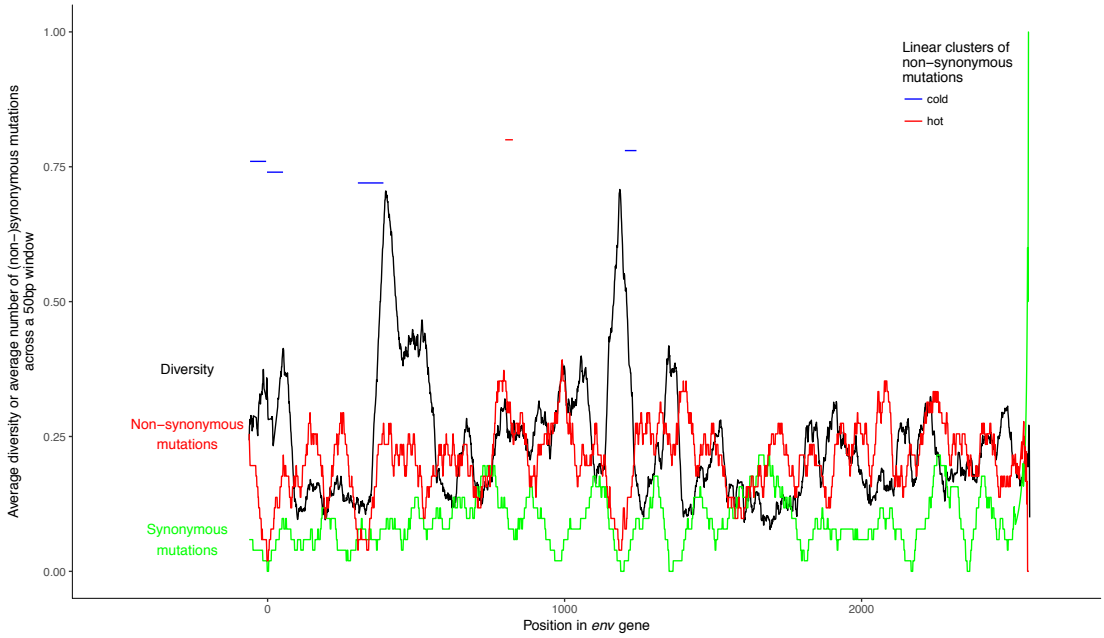
Zanini F, Neher RA. 2013. Quantifying selection against synonymous mutations in HIV-1 env evolution. J. Virol. 87:11843–11850.

Zhang Z, Townsend JP. 2009. Maximum-likelihood model averaging to profile clustering of site types across discrete linear sequences.Li W-H, editor. PLoS Comput Biol 5:e1000421.

Zhao Z-M, Campbell MC, Li N, Lee DSW, Zhang Z, Townsend JP. 2017. Detection of regional variation in selection intensity within protein-coding genes using DNA sequence polymorphism and divergence. Mol. Biol. Evol.

## Supporting Information

**Supplementary Data 1. Position of all identified mutations.** This file contains detailed information about the identity of the observed mutations. It provides the position in the HXB2 genome, the amino acid change they cause in the different genetic backgrounds and the number of HIV-1 populations the mutations occurs in.




557

558 **Supplementary Figure 1. Correlation between the presence of nonsynonymous and synonymous**
559 **mutations with diversity across the *env* gene. Linear clusters were computed with MACML**
560 **(Zhang and Townsend 2009).**

561

562 **Supplementary Program. With this program one can redo the analyses and simulations of the**
563 **manuscript.**