A de novo approach to inferring within-host fitness effects during untreated HIV-1 infection

Christopher J. R. Illingworth[1,2,3,*], Jayna Raghwani[4,5], David Serwadda[6], Nelson K. Sewankambo[6,7], Merlin L. Robb[8,9], Michael A. Eller[8,9], Andrew R. Redd[9,11], Thomas C. Quinn[10,11], Katrina A. Lythgoe[4,5,*]

1 Department of Genetics, University of Cambridge, Cambridge, United Kingdom
2 Department of Applied Mathematics and Theoretical Chemistry, University of Cambridge, Cambridge, United Kingdom
3 School of Chemical and Biological Sciences, Queen Mary University of London, London, United Kingdom
4 Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
5 Department of Zoology, Peter Medawar Building, University of Oxford, Oxford, United Kingdom
6 Rakai Health Sciences Program, Kalisizo, Uganda, School of Public Health, Makerere University, Kampala, Uganda
7 School of Medicine, Makerere University, College of Health Sciences, Kampala, Uganda
8 U.S. Military HIV Research Program, Walter Reed Army Institute of Research, Silver Spring, MD, United States of America
9 Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD, United States of America
10 Laboratory of Immunoregulation, Division of Intramural Research, NIAID, NIH, Baltimore MD, United States of America
11 Department of Medicine, Johns Hopkins School of Medicine, Johns Hopkins University, Baltimore, MD, United States of America

*Corresponding authors: chris.illingworth@gen.cam.ac.uk, katrina.lythgoe@bdi.ox.ac.uk

Abstract

In the absence of effective antiviral therapy, HIV-1 evolves in response to the within-host environment, of which the immune system is an important aspect. During the earliest stages of infection, this process of evolution is very rapid, driven by a small number of CTL escape mechanisms. As the infection progresses, immune escape variants evolve under reduced magnitudes of selection, while competition between an increasing number of polymorphic alleles (i.e., clonal interference) makes it difficult to quantify the magnitude of selection acting upon specific variant alleles. To tackle this complex problem, we developed a novel multi-locus

inference method to evaluate the role of selection during the chronic stage of within-host infection. We applied this method to targeted sequence data from the p24 and gp41 regions of HIV-1 collected from 34 patients with long-term untreated HIV-1 infection. We identify a broad distribution of beneficial fitness effects during infection, with a small number of variants evolving under strong selection and very many variants evolving under weaker selection. The uniquely large number of infections analysed granted a previously unparalleled statistical power to identify loci at which selection could be inferred to act with statistical confidence. Our model makes no prior assumptions about the nature of alleles under selection, such that any synonymous or non-synonymous variant may be inferred to evolve under selection. However, the majority of variants inferred with confidence to be under selection were non-synonymous in nature, and in nearly all cases were associated with either CTL escape in p24 or neutralising antibody escape in gp41. Sites inferred to be under selection in multiple hosts have high within-host and between-host diversity albeit not all sites with high between-host diversity were inferred to be under selection at the within-host level. Our identification of selection at sites associated with resistance to broadly neutralising antibodies (bNAbs) highlights the need to fully understand the role of selection in untreated individuals when designing bNAb based therapies.

Author Summary

During the within-host evolution of HIV-1, the diversity of the viral population increases, with many beneficial variants competing against each other. This competition, known as clonal interference, makes the identification of variants under positive selection a challenging task. We here apply a novel method for the inference of selection to targeted within-host sequence data describing changes in the p24 and gp41 genes during HIV-1 infection in 34 patients. Our method adopts a parsimonious approach, assigning selection to the smallest number of variants necessary to explain the evolution of the system. The large size of our dataset allows for the confident

identification of variants under selection, alleles at certain loci being repeatedly inferred as under selection within multiple individuals. While early CTL escape mutations have been identified to evolve under strong positive selection, we identify a distribution of beneficial fitness effects in which a large number of mutations are under weak selection. Variants that were confidently identified under selection were primarily found to be associated with either CTL escape in p24 or neutralising antibody escape in gp41, including sites associated with escape from broadly neutralising antibodies. We find that the most frequently selected loci have high diversity both within-host and at the between-host level.

Introduction

In the absence of effective antiretroviral therapy, HIV-1 evolves rapidly during infection. A key driver of evolution is the influence of the host immune system; cytotoxic CD8+ T-cells (CTLs) and neutralising antibodies (nAbs) impose selection on the virus, leading to the emergence of immune escape mutations[1]. However, other factors also influence viral evolution. For example, the host-specific nature of the immune response leads to the accumulation of mutations which are deleterious to the virus upon transmission to a new host. During the course of a new infection such variants are often lost, in particular where they occur at sites which in general are under strong purifying selection[2,3]. Selection may further act for protein or RNA secondary structure[4,5].

The complex nature of selection has led to multiple studies evaluating how the viral genotype may be both constrained and shaped during the course of evolution. These include the use of techniques for *in vitro* mutagenesis, and analyses of viral sequence data, evaluated at the level of population consensus or through deep sequencing exploring within-host variation at one or more time points during infection. For example, mutagenesis of HIV-1 proteins has allowed the

measurement *in vitro* of the effect of specific mutations [6]. The development of technologies for high-throughput mutagenesis has enabled such measurements to be made across very large sets of potential mutations[7–9]. Mathematical methods combining such results have been used to generate an overview of fitness costs and epistatic effects for the virus[10]. Measurements of this form provide a base-level estimation of the general fitness landscape of the virus, albeit that *in vitro* data may be limited in the extent to which it captures the behaviour of the virus within a human host.

Many years of study of HIV-1 have led to the collection of consensus genome sequence data for a large number of individual infections[11]. Such data have allowed techniques such as the fitting of maximum entropy models, which characterise the fitness costs of non-consensus variants in regions of the viral genome[12–14]. In this case, the fitness effects learned give a picture of the generic landscape upon which HIV-1 evolves, averaged over all within-host environments represented by the sequence database, albeit those host-specific environments may differ[15,16]. The extent of conservation at a particular genetic locus indicates the extent to which purifying selection acts upon the majority allele[17].

Short-read deep-sequencing data has provided valuable insights into how fitness effects shape the evolution of HIV-1. Studies can be broadly categorised into those that consider purifying selection, and those that consider positive selection. Purifying (or negative) selection represents the process by which deleterious variants are purged from a population. Over time the frequency of a variant under purifying selection evolves in a statistically predictable way towards an equilibrium state via mutation-selection balance[18]. Exploiting this fact, changes in allele frequency observed over time from across the course of untreated infections[2], or comparing single time points across multiple individuals, have been used to estimate the magnitude of selection and the mutation rate acting upon distinct regions of the genome[19][17].

117

118 Positive selection represents the process by which favourable variants are driven towards fixation,
119 and as with purifying selection population genetic methods can be adopted for the inference of
120 fitness effects. For example, a series of models have been developed for the inference of HIV-1
121 escape rates from CTL responses. Whereas earlier approaches to this problem considered viral
122 escape from a single CTL response[20–23], more recent studies have considered the multiple
123 immune responses that arise successively during infection [24–26]. Under such circumstances,
124 interference between beneficial viral mutations affects the population dynamics [27]. Therefore,
125 accounting for this clonal interference is critical if the role of selection is to be correctly inferred[28–
126 30].

127

128 Studies assessing fitness effects in within-host HIV-1 infection have often focused upon the
129 earliest stages of infection when strong selection on CTL escape mutations typically dominates
130 the viral population dynamics[24,30]; in this circumstance, we can model evolution as a
131 competition between a relatively small number of viral genotypes [24,31]. Later in infection, where
132 escape mutations are less strongly beneficial, and where synonymous diversity has had longer
133 to accumulate [2], the potential for hitchhiking and clonal interference is greater, such that variants
134 observed at high frequency are less certain to have evolved under positive selection.
135 Consequently, distinguishing selected from non-selected variants poses a substantial challenge.
136 To address this, we present a *de novo* approach for inferring selection from HIV-1 sequence data
137 in which any variant allele may, in theory, be detected as under selection. We adopt a
138 parsimonious approach, assigning selection to the smallest set of variants required to explain the
139 observed multi-locus sequence data under a likelihood model. Applied to targeted sequence data
140 from a substantial cohort of 34 untreated individuals living in Uganda[32], we determine how
141 selection drives viral evolution. In the presence of pervasive interference between alleles in
142 linkage disequilibrium with one another, our consideration of data from a large number of

individuals is fundamental in providing statistical confidence in the assignment of selection. Specifically, the repeated inference of selection at the same locus in independent cases of infection enhances the power of our study to elucidate how selection during individual infections shapes genetic diversity at the population level.

Results

We applied an evolutionary inference method to deep-sequencing data spanning multiple years of infection from 34 untreated individuals living in Rakai, Uganda, enabling us to infer positive selection acting on part of the gp41 region of *env* (324 base pairs) and the p24 region of *gag* (387 base pairs)[32].

Extent of selection

An initial application of our method found extensive evidence of positive selection in the viral genome, with 74 (out of 387) nucleotide sites in p24, and 81 (out of 324) nucleotide sites in gp41, being inferred to evolve under positive selection in at least one individual (S1 Table and S1 Fig). Our method explicitly accounts for linkage disequilibrium between alleles observed in the sequencing data[28,33]; however a potential exists for interactions between observed alleles, in the targeted sequence region, and non-observed alleles in flanking regions of the genome. Simulation studies showed that our approach performs well in multi-locus systems where all selected alleles are observed, but is prone to generating false positive inferences of selection where non-observed selected alleles are in linkage disequilibrium with observed alleles (S1 Text). Despite not always attributing selection to the correct locus, the distribution of the inferred strengths of selection, in cases where these could be inferred with confidence, was not statistically distinguishable from the ‘true’ distribution of strengths used to conduct the simulations (S2 Fig;

169 p=0.078, Kolmogorov-Smirnov test). Interpreted at face value, these results reflect the extent of
170 selection in and around the regions of the genome under analysis without necessarily correctly
171 attributing selection to the correct genetic variant.
172
173 *Strength and time of onset of selection*
174
175 Inferred variants generally evolved under weak selection. Taking our inferences at face value, we
176 identified selected variants for which the magnitude of selection could be quantified with some
177 degree of precision. The uncertainty in our estimates depends upon the extent to which data are
178 available; for example, when a variant emerges and fixes between two time points, only a lower
179 bound on the strength of selection can be inferred[34]. For this analysis, we generated confidence
180 intervals for each inferred magnitude of selection using a likelihood-based method, retaining only
181 variants for which the upper and lower bounds of this interval differed by no more than an order
182 of magnitude. Our model assumed that a selected variant evolved under neutrality until some
183 fixed point in time, following which it conferred a constant fitness advantage or disadvantage to
184 the virus. Distributions of fitness effects compiled from the inference showed that most of the
185 identified alleles under selection experienced very weak selective effects, with long tails of alleles
186 evolving under strong positive selection (Fig 1). In p24, around 65% of inferred selected variants
187 had a strength of selection of less than 5% per generation, while in gp41 approximately half of
188 variants were under this threshold. In our model, selection of 5% per generation would cause, in
189 the absence of interference effects, a change in allele frequency from 5% to 95% in a period of
190 just under eight months. Our approach likely underestimates the proportion of variants under
191 weak selection; very weak selection would not produce a change in the population sufficient to be
192 identified from the data. Variants which fixed in the population before the first sample was
193 collected also cannot be identified from our data, such that the extreme beneficial end of the
194 distribution may also be under-represented. Our result, of a preponderance of weakly beneficial

195 mutations during chronic infection, is robust to these two qualifications. Most selection acting upon
196 variants was inferred to kick in during the first year of infection and almost all within two years; we
197 note that the finite length of sequencing could restrict the inference of more lately selected
198 variation.
199
200 *Distribution of selected variants among individuals*
201
202 While the raw output of our inference method contains false positive inferences of selection, we
203 can use statistical analysis to identify specific sites in the genome containing variants under
204 selection. A unique aspect of this dataset is the large number of untreated individuals included in
205 the study. Where alleles at the same locus were inferred to be under selection in multiple
206 individuals, a statistical approach was used to infer loci at which, under conservative assumptions,
207 at least one of the inferred variants is genuinely under selection. Taking into account different
208 patterns for nonsynonymous and synonymous mutations, we estimated per-site false-positive
209 rates for nonsynonymous and synonymous mutations, which were subsequently used to identify
210 with statistical confidence sites that were under selection in at least one individual in our dataset.
211 Taking all 34 individuals into consideration, we calculated that in p24 we could be confident that
212 a site is under selection in at least one individual if mutations were inferred to be under selection
213 in at least five individuals, and/or if nonsynonymous mutations were inferred to be under selection
214 in two or more individuals, while in gp41 we could be confident if at least five mutations and/or
215 three nonsynonymous mutations were inferred to be under selection. Applying these criteria, we
216 identified 11 specific nucleotide sites, representing 10 amino acid (AA) positions, under selection
217 in p24, and likewise we identified 11 such sites, representing 8 AA positions, in gp41. All but four
218 mutations at these AA positions represented nonsynonymous changes (see Table 1).
219 Occasionally two different codons at the same AA position were found to be subject to selection

220 in a single individual, but in general repeated inferences of selection at an AA position occurred

221 in distinct individuals (S1 Table).

222

223 **Table 1:** Summary of amino acid positions containing sites under selection.

224

Region	Amino Acid Position[a]	Sensitivity[b]	Nonsynonymous[c]			Synonymous
			Reversion	Escape	Neither	
p24	215[35–38]	CTL	0	2	1	2
p24	219[38–41]	CTL	1	0	1	0
p24	223[38–41]	CTL	2	4	2	0
p24	228[40,41]	CTL	1	1	0	0
p24	242[41–46]	CTL	2	0	1	0
p24	252[37,41]	CTL	1	2	0	0
p24	286		1	1	0	0
p24	302[47]	CTL	2	0	0	0
p24	310[48–51]	CTL	1	1	0	0
p24	312[37,47,49, 52,53]	CTL	3	2	0	0
gp41	620[54–58]	NAb	3	3	4	1
gp41	624[54,55]	NAb	3	2	4	0
gp41	641[59]	HGF24	0	2	1	1
gp41	644[59]	HGF24	2	0	1	0
gp41	648[59]	HGF24	1	3	0	0
gp41	655[60]	NAb	1	3	0	0
gp41	674[57,61–66]	NAb	0	2	1	0
gp41	677[57,63,64]	NAb	1	3	0	0

225

226 [a]Amino Acid (AA) positions in *gag* (p24) or *env* (gp41) relative to the HXB2 reference genome

227 [b]Sensitivity to CTLs or NAbs using the Los Alamos HIV database (http://www.hiv.lanl.gov; see

228 Methods for further details). AA position 228 in *gag* is associated with compensatory mutations,

229 rather than affecting sensitivity to CTLs directly. Position 286 in *gag* occurs in an epitope position
230 targeted by a common HLA allele in the Ugandan population and so is likely associated with CTL
231 escape. Positions 644, 648 and 655 were not associated with sensitivity to NAbs, but are in an epitope
232 region recognised by NAb HGF25.
233 [c]A nonsynonymous change was classed as a reversion if the nucleotide changed towards the subtype-
234 specific population-level consensus, as an escape if the nucleotide changed away from the subtype-
235 specific population-level consensus, and as neither of these if there was an nucleotide change which
236 remained different to the subtype-specific population-level consensus.
237
238
239 Using the Los Alamos HIV database (http://www.hiv.lanl.gov), we determined for each of these
240 AA positions whether they have previously been associated with changes in CTL susceptibility or
241 NAbs (see Methods, Fig 2, S2 Table). In p24, nine of the 10 AA positions identified have
242 previously been associated with changes in CTL susceptibility and/or compensatory mutations.
243 The remaining codon (residue 286 in HXB2 *gag*) lies within CTL epitopes recognised by human
244 leukocyte antigen (HLA) alleles that are relatively common in the Ugandan population (A1101 and
245 B27[67]) and therefore mutations at this site probably affect sensitivity to CTLs, or possibly
246 compensatory mutations. In gp41, of the eight AA positions identified as being under selection,
247 five (residues 620, 624, 655, 674 and 677 of *env*) have previously been associated with NAbs,
248 including broadly neutralising antibodies (bNAbs). These are associated with the gp120/g41
249 interface and the membrane-proximal external region (MPER). The other three codons (residues
250 641, 644 and 648 of *env*) all lie within an epitope region that is targeted by the monoclonal
251 antibody HGF24, which has been shown to neutralise some viruses isolated from the African
252 continent[59]. Although we cannot rule out other sources of selection in *env*, particularly due to
253 CTLs, given the position of these codons on the genome, we believe selection due to humoral
254 immune pressure is more likely. As such, all of the sites we identified as under selection in p24
255 probably affect susceptibility to CTLs or associated compensatory mutations, whereas all of the
256 sites under selection in gp41 probably affect susceptibility to NAbs or associated compensatory
257 mutations.

258

259 *Direction of selection*

260

261 Since mutations that enable the virus to evade host immune responses are often costly in terms
262 of viral replication[68,69], sites harbouring escape mutations are likely to evolve towards the
263 population-level consensus when transmitted to a new host since the selective pressure of a
264 specific immune response is removed, but costs associated with the mutation remain [2,32,70–
265 73]. This hypothesis is supported by the observation that allelic substitutions during the course of
266 untreated infection occur towards population consensus much more frequently than expected by
267 chance [2,32]. If analysis is restricted to sites where substitutions are observed, this bias is
268 confined mainly to nonsynonymous substitutions[32], which is expected if immune escape
269 mutations are generally nonsynonymous.

270

271 For the nucleotide sites that were confidently inferred to be under selection, we classified selected
272 variants as escapes if they resulted in an AA change away from the subtype-specific population-
273 level consensus in Uganda during the period the individuals were being sampled (see methods
274 for full details). Conversely, inferred selected variants were classified as reversions if they resulted
275 in change towards the subtype-specific population-level consensus. Of the 32 selected
276 nonsynonymous mutations identified in p24, 13 were classified as escapes, 14 as reversions, and
277 five as neither. The similar number of escapes and reversions observed is expected if CTL
278 immune escape mutations are costly in non-HLA matched hosts, with escape in one individual
279 followed by reversion in the next. Of the 40 nonsynonymous mutations identified in gp41, 18 were
280 classified as escapes, 11 as reversions, and 11 as neither. The fact that over a quarter of AA
281 changes are towards population consensus suggests that in many cases antibody escape
282 mutations are deleterious in hosts without a matching antibody, but the stereotypical pattern of
283 adaptation in one individual followed by reversion in another at CTL immune epitopes is likely

284 more complex for antibody-escape evolution, where escape mutations are sometimes, but not
285 always costly in the absence of antibody response [74,75].
286
287 We also determined the direction of change for all variants inferred to be under selection, which
288 will likely include variants under direct selection and variants changing in frequency due to hitch-
289 hiking, and whether they represented synonymous or nonsynonymous changes (S3 Fig). We
290 found a clear correlation between the number of times that selection was inferred at an AA position
291 and the proportion of selected variants that were nonsynonymous (linear regression, p24
292 $p=0.009$, $r^2=0.85$; gp41 $p=0.007$, $r^2=0.67$). For gp41, a correlation was also found between
293 repeated selection at an AA position and a pattern of evolution towards the population consensus
294 ($p=0.011$, $r^2=0.63$); around half of inferred selection events were towards population consensus
295 at the AA position most frequently inferred to be under selection. Although for p24 the linear
296 regression did not reveal a significant trend ($p=0.400$, $r^2=0.18$), there is a distinction between AA
297 positions selected 2 or more times, which have a high probability of being towards subtype-
298 specific population consensus (44%), and AA positions selected once, which have only a small
299 probability of being towards the consensus (5%). Our interpretation is that AA positions
300 represented once in our analysis disproportionately represent mutations increasing in frequency
301 due to hitch-hiking, with these mutations tending to be synonymous and away from population
302 level consensus. AA positions represented multiple times, on the other hand, are more likely to
303 represent immune escapes and reversions, and therefore tend to be nonsynonymous but with
304 only around half of mutations away from consensus.
305
306 *Comparing codon diversity at the within-host and population scale*
307
308 Our data showed a strong relationship between within- and between-host sequence diversities,
309 where the diversity of codons was measured at each AA position. Within-host diversity was

310 measured approximately three years after seroconversion for each of the 34 individuals, and the
311 mean calculated. Diversity at the population scale was calculated as the mean of the diversities
312 for each of the subtypes A, D, and C, using virus sequences from a large number of individuals
313 living in Uganda around the same time as the 34 individuals in our study (see Methods).
314 Consistent with previous studies [2], we identified a strong relationship between measurements
315 of sequence diversity calculated at the within-host and population scales (Fig 3), with all AA
316 positions found to be highly diverse at the within-host level also highly diverse at the population
317 level. Moreover, all but one of the AA positions containing nucleotide sites that we are confident
318 are under selection are also diverse at the population level (the exception being residue 302 in
319 *gag*). Given that most changes at these sites probably reflect escape from host immune
320 responses, compensatory mutations, or reversions of these escapes in subsequent individuals,
321 diversity at the population scale at these AA positions is likely maintained by the differing selection
322 pressures faced by variants in different hosts due to different immunological backgrounds.
323
324 Not all AA positions found to be highly diverse at the population level were detected as being
325 under selection, with a high degree of confidence, at the within-host scale. This observation could
326 arise if codons at these positions are under selection in some individuals in the population, but
327 we failed to observe it with confidence, either because we sampled too few individuals, selection
328 was too weak to be detected, and/or selection drove fixation events before the first sampling time
329 point. In addition, at unconstrained sites experiencing little or no selection, diversity might
330 gradually accumulate at the population scale due to drift, exacerbated by small bottleneck sizes
331 at transmission. This could explain the high levels of population diversity at positions 235 in *gag*
332 and 609 in *env*; almost all of the diversity observed at these codon positions is due to the presence
333 of synonymous variants.
334
335 **Discussion**

336

337 We developed a novel inference framework to infer the extent of selection acting upon variants
338 which drive the evolution of within-host HIV-1 populations, considering data from the p24 region
339 of *gag*, and gp41 of *env*, from 34 longitudinally sampled untreated individuals. A frequent
340 assumption is that beneficial mutations will rapidly spread within individuals once they occur[19].
341 For example, it is well established that CTL-escape mutations accumulate and spread rapidly
342 during acute infection [20–26], albeit that the rate of allele fixation decreases considerably during
343 chronic infection[23,26]. However, estimating the extent and strength of positive selection during
344 infection more generally is challenging due to genetic linkage among variant alleles, which makes
345 differentiating between selected variants and variants that are increasing in frequency due to
346 linkage with a selected variant (hitch-hiking) difficult. Our *de novo* approach incorporates genetic
347 linkage and recombination. Furthermore, it is generally assumed that during untreated infection
348 selected variants are associated with immune escape; our approach is agnostic with regards to
349 phenotypic data, potentially allowing any polymorphic site in the genome to be identified as under
350 positive selection.

351

352 Our results indicate a pattern of weak and slow selective sweeps characterising evolution during
353 chronic HIV-1 infection, with stronger faster selective sweeps being relatively rare. We note that
354 where 'weak' selection was inferred, this was still on a scale outweighing the effects of genetic
355 drift. Studies of the effective population size of HIV-1 have indicated a value in excess of 10^5 [76];
356 given such a value, selection of the order of 5% per generation is comfortably within a realm
357 whereby the influence of selection dominates that of genetic drift[77]. An important caveat is that
358 the first sampling time point for each individual in our analysis is estimated to be between 150
359 and 425 days since seroconversion, and therefore we will not detect variants that were under
360 strong selection and rapidly reached fixation before the first sampling time point. Furthermore, the
361 magnitude of the most strongly selected variants could not always be quantified; where fixation

362 occurs entirely in the interval between two consecutive time points, no upper bound on the
363 magnitude of selection could be fixed.
364
365 A unique aspect of our study is the large number of individuals for which we have data.
366 Comparisons among individuals revealed AA positions which were inferred to be under selection
367 in multiple (up to ten) individuals, with selected changes at these sites likely reflecting the gain or
368 loss of CTL-escape mutations in p24, antibody-escape mutations in gp41, and escape-related
369 compensatory mutations. Moreover, these sites were also found to be highly diverse at the
370 population level. This is consistent with a pattern where a minority of codons are repeatedly under
371 selection in multiple individuals, likely representing adaptation to the immunological background,
372 which revert upon transmission to subsequent individuals; a pattern which has been referred to
373 as “adapt and revert”[78]. Although we do not know the HLA-type of the infected individuals in our
374 study, the number of putative CTL escapes and reversions is consistent with the frequency of
375 different HLA alleles in the Ugandan population.
376
377 Perhaps less expected in our analysis was the identification of AA positions that are associated
378 with neutralising antibodies and which were found to be under selection in a large number of
379 individuals; for one site selection was inferred in nearly a third of individuals, with another inferred
380 in a quarter of individuals. The implication is that the same epitopes are frequently targeted by
381 antibodies in different individuals, and with similar means of viral escape. Moreover, since around
382 a quarter of changes at these sites are towards the subtype-specific population level consensus,
383 many may well represent the reversion of costly antibody-escape mutations from previous
384 individuals, supporting the observation that some but not all antibody-escape mutations are costly
385 [69,74,79–83]. These patterns can help explain why resistance to antibodies has increased over
386 the course of the epidemic [84–88], but also highlights that viral evolution at the population level

387 in response to bNAb-based interventions is likely to be complex, involving evolutionary responses
388 to both naturally and therapeutically induced immune responses.
389
390 Even though our framework explicitly accounts for linkage disequilibrium between observed
391 variants, it is still vulnerable to false positive inferences of selection due to linkage disequilibrium
392 with unobserved variants flanking the genetic regions we analysed. Although simulated data
393 suggested that the overall distribution of fitness effects was robust to this vulnerability, our study
394 should serve as a cautionary note; where multiple alleles evolve in linkage disequilibrium, care is
395 needed in identifying selection with any particular allele. The large number of individuals included
396 in our study enabled us to partly circumvent this problem by only assigning confidence that any
397 particular nucleotide site is under selection if it is inferred to be under selection in multiple
398 individuals. Indeed, evidence for the validity of our method is provided by the repeated
399 observation of variants under weak selection across multiple individual infections, with these
400 changes making biological sense under the “adapt and revert” hypothesis. Our results emphasise
401 the role of immune escape in driving evolution during chronic infection, shaping patterns of
402 diversity at the population level, and provides new insights that could be useful in the development
403 of immune-based interventions, particularly in the context of viruses circulating in Africa.
404
405 **Methods**
406
407 In order to evaluate selection within a host, we employed a likelihood-based inference framework
408 to infer the most parsimonious explanation of the sequence data in terms of a model of selection
409 acting for specific nucleotides in the viral population. Some of the mathematical aspects of this
410 framework have previously been applied in studies of the within-host evolution of the influenza
411 virus[33,89], albeit that the model used in this case is specific to HIV-1. Our model explicitly

accounts for linkage disequilibrium between alleles and builds upon earlier approaches for inferring selection in cases where linkage is of importance for evolution [24,28,90,91].

Sequencing data

For our evolutionary analysis we used previously generated deep-sequence data from 34 longitudinally sampled individuals participating in the Rakai Community Cohort study and co-enrolled in the Molecular Epidemiology Research (MER) seroconverter study. Targeted short-read deep-sequence data from the p24 region of gag (390 bp; HXB2 reference genome positions 1429-1816) and the gp41 region of env (324 bp; HXB2 7941-8264) had been sequenced using the 454 sequencing platform (Roche, Branford, CT). All individuals were untreated, with a first sampling time point around one year since seroconversion, and typically 3 or 4 subsequent time points spanning between two and seven years of infection (see Table 1 in Raghwani et al 2019). Aligned sequences can be found at https://github.com/katrinalythgoe/ RakaiHIV. Further details on the individuals and sequencing methods used have been given elsewhere[32,92].

Calling of variants from sequence data

As a preliminary step in the evaluation of sequence data, an estimate was made of the extent of noise in the data. While short-read sequencing allows very high read depths to be achieved, the incomplete sampling and inaccurate sequencing of a viral population imply that the extent of information in a viral dataset is often less than that assumed by a binomial or multinomial model[93]. While sequencing noise may be complex in nature, one approach to accounting for noise within an analytical likelihood framework is to employ a Dirichlet multinomial likelihood function with parameters equal to a constant term, C, multiplied by the underlying genotype

frequencies[33]; the constant defining a generalised overdispersion in the data relative to a multinomial model. In the case of within-host influenza evolution a conservative estimate of this parameter may be generated by fitting a model of neutral evolution to a subset of genotype frequencies which are believed not to change substantially during the course of infection[33]; the inferred variance of the model, accounting for the read depth of sampling, provides an estimate of the extent of noise in the data.

HIV-1 infection occurs over a substantially longer period than does influenza infection, requiring a new strategy for the estimation of noise; we here exploited effects caused by genetic hitch-hiking[94]. If two alleles appear uniquely upon a shared genetic background, they will initially share an identical allele frequency. Over time the allele frequencies will change in a very similar manner, differences arising over time as a result of recombination between distinct haplotypes. We used putatively hitch-hiking trajectories to derive an estimate of noise in the sequence dataset. Loci were identified within the data from each patient at which a minor allele frequency of at least 10% was observed for at least two points in time. In a given patient we denote the trajectory

$$(i, a) = \{\tilde{q}_i^a(t_k)\}_{k=1,\ldots,K}$$

comprising the observed frequencies of allele *a* at locus *i* across all recorded times t_k. For each pair of identified loci (*i, j)*, we found the alleles *a*, b** in {A, C, G, T} minimising the statistic

$$d_{ij}^{a^*b^*} = \min_{a,b} \left\{ \frac{1}{K} \sum_k |\tilde{q}_i^a(t_k) - \tilde{q}_j^b(t_k)| \right\}$$

Initially, pairs of trajectories (i,a*) and (j,b*) were denoted as being ‘similar’ if

$$d_{ij}^{a^*b^*} < 10\%.$$

A set of further heuristic steps was then applied to refine these sets of trajectories. On timescales close to those over which the data for this study was measured, recombination in HIV-1 has been noted as being of importance over genetic distances greater than 100 nucleotides [2]; here a distance cutoff of less than or equal to 50 nucleotides was imposed between trajectories.

Next, pairs of variants were required to have similar frequencies at the first time of observation, requiring that

$$|\tilde{q}_i^{a^*}(t_1) - \tilde{q}_j^{b^*}(t_1)| < 5\%.$$

Further, so as to remove pairs of trajectories for which only one was polymorphic at a given time, it was required that the maximum ratio between minor allele frequencies did not exceed 10^3 for the duration of the trajectories. Remaining sets of pairs identified in each patient were clustered into sets via an iterative process; beginning with an initial pair of trajectories (i_1, a_1) and (i_2, a_2), a trajectory (i_m, a_m) was added to the set if, following filtering, (i_m, a_m) was similar to a trajectory (i_n, a_n) already in the set. Sets of trajectories generated by the above process are shown in S4 Fig.

Having identified sets of trajectories, an inference process was used to evaluate the extent of noise in the data. Conservatively, the 'true' allele frequencies of each set were calculated as a simple mean of the observations, thereby assuming that all differences in frequencies result from 'noise' in the sequencing process. Given a set of trajectories *S* from a single viral population, where the allele frequency $\tilde{q}_i^a(t_k)$ was calculated as $n_i^a(t_k)/N_i(t_k)$, the frequency of observations reporting the given allele, the inferred frequency at time t_k was calculated as

$$q^S(t_k) = \frac{\sum_m n_{i_m}^{a_m}(t_k)}{\sum_m N_{i_m}(t_k)}$$

A Dirichlet multinomial model was then parameterised across all trajectory sets, finding the value of *C* satisfying

$$\min_C \left\{ \sum_S \sum_m \sum_k \mathcal{L}(N_{i_m}(t_k), C, q^S(t_k), n_{i_m}^{a_m}) \right\}$$

where

$$\mathcal{L}(N, C, q, n) = \log \frac{\Gamma(N+1)}{\prod_a \Gamma(n^a+1)} \frac{\Gamma\left(\sum_a cq^a\right)}{\Gamma\left(\sum_a n^a + Cq^a\right)} \prod_a \frac{\Gamma(n^a + Cq^a)}{\Gamma(Cq^a)}$$

The derived value of *C* provides a proxy measurement of the extent of noise in the data and was used in further likelihood calculations; we inferred the value C = 100.359. We note that patterns of noise in genome sequence data may be substantially more complex than represented by our model; our likelihood, combined with the BIC model selection framework, provides a simple yet analytically tractable approach for the inference of selection parameters from real genome sequence data.

Calling of variants from sequence data

Single-locus variants were identified in the data using the SAMFIRE software package[95]. Variants with a minimum allele frequency of at least 1% for at least one time point in the course

511 of infection were identified. Considering frequency data from each variant allele, 'potentially non-
512 neutral' loci [91] were identified, corresponding to positions in the genome at which significant
513 changes in allele frequency were identified.
514
515 Variants were then filtered, identifying 'potentially non-neutral' loci [91] at which significant
516 changes in allele frequency were observed over time. In this process, where $q^1(t)$ denotes the
517 frequency of the variant allele at locus *i* at time *t*, deterministic models were fitted to the single-
518 locus trajectories, using the equation
519

520 $$q_i^1(t_{k+1}) = \frac{q_i^1(t_k)e^{\sigma_k(t_{k+1}-t_k)}}{1 - q_i^1(t_k) + q_i^1(t_k)e^{\sigma_k(t_{k+1}-t_k)}}$$

521
522 for neutral (σ_k=0), constant (σ_k=σ), and time dependent models of selection, retaining trajectories
523 for which the constant or time-dependent models of selection outperformed the neutral model
524 when evaluated using the Bayesian Information Criterion (BIC) to account for the increased
525 complexity of the models including selection [96].
526
527 Targeted sequence data were processed into representative haplotypes, expressing each read
528 in terms of the alleles present at the set of potentially non-neutral loci identified for that gene and
529 individual. The evolution of the system was subsequently assessed in terms of the changes in
530 frequencies of these haplotypes.
531
532 *Evolutionary model*
533
534 Our evolutionary model considered the effect of mutation, selection, and recombination upon the
535 population. Within the model the frequency of the haplotype *a* at generation t_k was specified by

the frequency $q_a(t_k)$, frequencies changing according to the three evolutionary processes. The model system was propagated within the space of observed haplotypes using a Wright-Fisher approach of discrete generations, with successive steps of mutation, recombination and selection.

Mutation was approximated as occurring between haplotypes that differ by a single nucleotide with rate μ per generation. Recombination was approximated as occurring in a pairwise manner between haplotypes with rate ρ per base per generation. That is, if a recombination event occurring between the loci indexed i and i+1, and involving the haplotypes a and b, were to produce the haplotype c, then in our model the new haplotype was produced at rate $\Delta_{i,i+1}\rho q_a(t_k)q_b(t_k)$ where $\Delta_{i,i+1}$ is the sequence distance between loci i and i+1.

A time-dependent model of selection acting upon haplotypes was applied, simulating the changing selection acting upon HIV during an infection. The time-dependent fitness w_a of a haplotype was modelled via a hierarchical model of single-locus terms

$$w_a(t_k) = \exp\left(\sum_i \sigma_i I_{i,K}\right)$$

where the parameter σ_i denotes selection acting for or against all haplotypes with a variant locus at locus *i*, and the parameter $I_{\{i,K\}} = 0$ for all $k \leq K$, and $I_{\{i,K\}} = 1$ for all $k > K$.

Selection then modifies the frequency of each haplotype according to the equation

$$q_a(t_k + 1) = \frac{w_a(t_k)q_a(t_k)}{\sum_b w_b(t_k)q_b(t_k)}$$

559

560 On the basis of previous studies[19,97–100], parameters for mutation and recombination were
561 set at μ=3 x 10^{-5} per generation and ρ=10^{-5} per base per generation, with a generation time of two
562 days[101]. Parameters σ_i and $I_{\{i,K\}}$, and initial frequencies $q^a(t_0)$, were learnt from the data
563 according to a hierarchical model framework. An initial, neutral model contained no parameters
564 σ_i. Next, single-locus selection models, with a single pair of selection parameters (σ_i, $I_{\{i,K\}}$), were
565 considered; in each model the optimal values for these parameters were identified. By means of
566 an iterative process, more complex models of selection were considered to the point of
567 discovering a model for which neither adding an additional pair of selection parameters, or
568 removing an existing pair of parameters, improved the model. As a conservative step, a model
569 with an additional pair of selection parameters was required to significantly outperform a simpler
570 model to be accepted, this being denoted by an improvement of 10 units of BIC upon the
571 optimisation of model parameters.

572

573 The above model is an approximation in so far as the evolutionary process is evaluated only over
574 the set of viral haplotypes which are observed in the sequencing data; the potential spread of the
575 population into other haplotypes is not modelled. This approximation is necessary to make the
576 calculation feasible; without such a constraint the exponentially large number of potential viral
577 sequences would not allow model fitting to be performed.

578

579 *Estimates of uncertainty in selection coefficients*

580

581 Error bars were calculated for each parameter σ_i inferred in the maximum likelihood calculation.
582 Supposing the maximum log likelihood for a given inferred system to be equal to some value L,
583 error bars were generated via a constrained exploration of the model space, in which a change
584 in model parameters was accepted if the resulting likelihood was not greater than L-2, and for

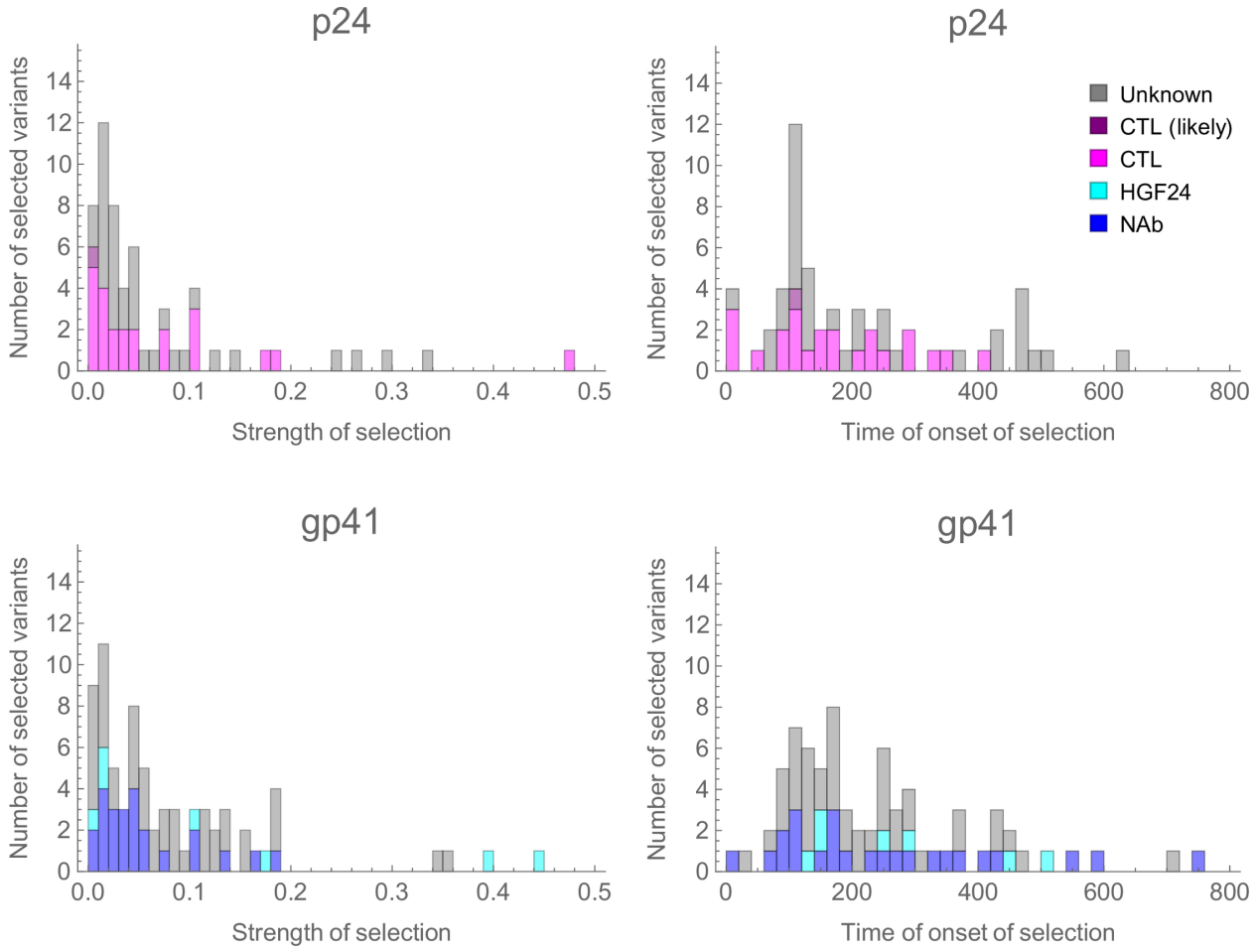
585 which changes to the parameter of concern, σ_i, were constrained so that this parameter could
586 only change in a specific direction; forcing this parameter to increase generated an estimate, after
587 repeated iteration, for the upper error bar of this parameter, while forcing this parameter to
588 decrease generated an estimate of the lower error bar of this parameter.
589
590 *Reporting selection coefficients*
591
592 In our study we report the respective fitness advantage conferred by each beneficial mutation as
593 a percentage per generation. This statistic, s, is calculated as
594
595 $$s = e^{\sigma} - 1$$
596
597 *Direction of Selection*
598
599 For each of the selected mutations identified in our analysis, we determined whether evolution
600 was towards or away from the subtype-specific population consensus. The subtype of each gene
601 region for each individual and the population consensus for subtypes A, D and C in Uganda during
602 a similar period to which the individuals were sampled was previously determined[32].
603
604 *Amino Acid position association with immune escape*
605
606 Using the Los Alamos HIV database (http://www.hiv.lanl.gov), we determined for each of the AA
607 positions in the p24 and gp41 regions that we analysed whether they have previously been
608 associated with changes in CTL susceptibility or neutralising antibodies. For CTL susceptibility
609 we used the Epitope Variant and Escape Mutation Database CTL variant search tool
610 (https://www.hiv.lanl.gov/content/immunology/variants/variant_search.html?db=ctl). AA positions

611 were marked as being associated with susceptibility to CTLs if susceptible and/or resistant
612 variants were returned in the search tool. We excluded codons inferred to be susceptible to CTLs
613 if the only evidence was high levels of diversity at the population level. For neutralising antibodies
614 we used the Neutralizing Antibody Contacts and Features search tool
615 (https://www.hiv.lanl.gov/components/sequence/HIV/featuredb/search/env_ab_search_pub.com
616 p), marking codons as associated with susceptibility if variants were shown or predicted to affect
617 neutralisation by antibodies, or binding to neutralising antibodies.
618
619 *Codon Diversity*
620
621 We used the diversity statistic π to measure mean intra- and inter-host diversity at the codon
622 level. Given an allele *a* with read depth *N*, where n_i of each of the codons $i \in \{A,C,G,T\}$ have been
623 observed, we define
624
625 $$D_a = \frac{N(N-1) - \sum_i n_i(n_i - 1)}{N(N-1)}.$$
626
627 Then the diversity statistic π is defined by
628
629 $$\pi = \sum_{a=1}^{L} D_a / L$$
630
631 where the region of the genome under consideration is of length *L*.
632

633 This diversity statistic has been shown to be less prone than some other metrics to biases at the
634 intra-host scale[102], and using codons (nucleotide triplet motifs) rather than amino acids or single
635 nucleotides means our measure incorporates synonymous and nonsynonymous diversity whilst
636 still enabling comparison with information on sensitivity to host immune responses, which is
637 typically given at the amino acid level. To determine the mean intra-host diversity at a given codon
638 position, we calculated π for each individual and then took the mean for all 34 individuals. To
639 determine the mean inter-host diversity at a given codon position we calculated π for each
640 subtype, and then took the mean for all three subtypes (A, C and D).
641
642

643 **Figures**

644



645

646

647 Fig 1: Histograms of inferred strengths of selection and times in days of onset of selection across
648 34 individuals. The maximum likelihood estimates are shown in each case, for the subset of the
649 data where the upper and lower bound confidence intervals on the strength of selection are within
650 an order of magnitude of each other. Coloured bars indicate variants at nucleotide sites we are
651 confident are under selection in at least one individual and are coloured according to whether
652 they are within AA positions associated with differences in susceptibility to CTLs (pink) or
653 susceptibility to NAbs (blue). The purple bar represents a probable CTL escape mutation, while
654 the cyan bars represent selected variants within an epitope position targeted by the neutralising

655 antibody HGF24. Where a codon is implicated in multiple responses, for clarity they are coloured

656 in order of preference NAb, HGF24, CTL (see S2 Table).

657

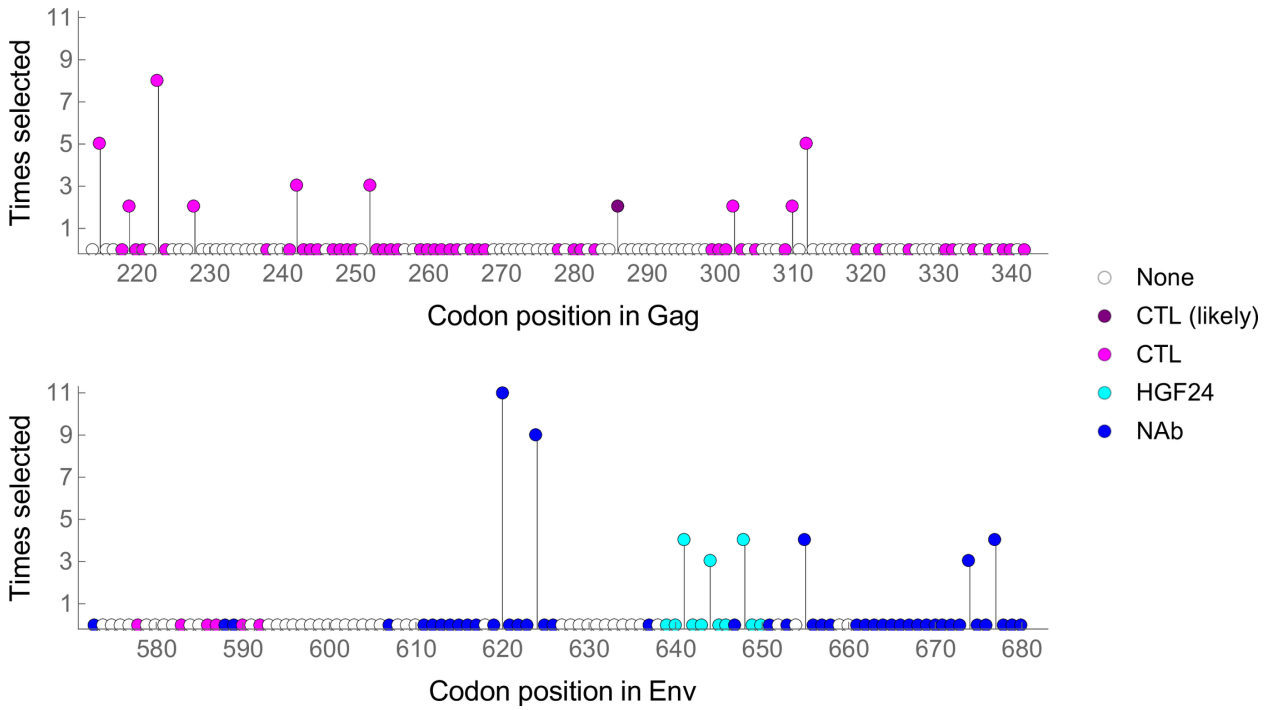


Fig 2: Amino Acid (AA) positions containing nucleotide sites that with confidence were inferred to be under selection in at least one individual. Codon positions are in relation to the HXB2 reference sequence. and the vertical y-axis gives the number of times that codon was inferred to be under selection across 34 individuals. Occasionally the same AA position was inferred to be under selection twice in the same individual. Pink: AA positions associated with changes in susceptibility to CTLs or compensatory mutations; Purple: AA probably affects susceptibility to CTLs; Blue: AA positions associated with susceptibility to NAbs; Cyan: AA positions in an epitope position targeted by the neutralising antibody HGF24. Where an AA position is implicated in multiple responses, for clarity they are coloured in order of preference NAb, HGF24, CTL.

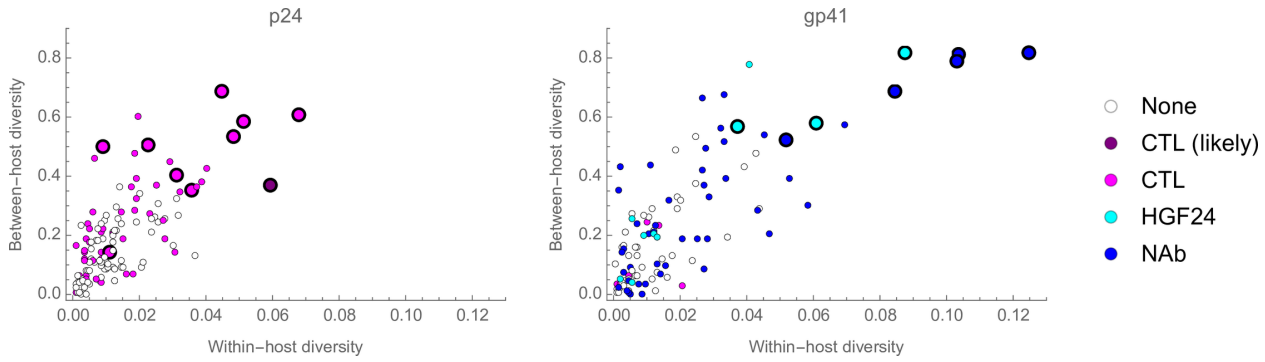


Fig 3: Within- and between-host codon diversity. For every AA position in our region of analysis we determined the within- and between-host codon diversity. Large markers denote AA positions in which we are confident selection is occurring in at least one individual. Markers are coloured if they are associated with changes in sensitivity to CTLs or compensatory mutations (pink) or NAbs (blue). In addition, AA postions in an epitope targeted by the neutralising antibody HGF24 are coloured in cyan. Where an AA position is implicated in multiple responses, for clarity they are coloured in order of preference NAb, HGF24, CTL.

679 **Supporting Figures and Tables**

680

681 S1 Fig: Codon positions containing nucleotide sites that are inferred to be under selection. This
682 includes codons that are genuinely under selection and those that are increasing in frequency
683 due to hitchhiking. Codon positions are in relation to the HXB2 reference sequence, and the y-
684 axis gives the number of times that codon is inferred to be under selection across 34 individuals.
685 Occasionally the same codon is inferred to be under selection twice in the same individual. Pink:
686 codons associated with changes in susceptibility to CTLs; Purple: codon probably affects
687 susceptibility to CTLs; Blue: codons associated with susceptibility to NAbs; Cyan: codons in an
688 epitope position targeted by the neutralising antibody HGF24. Where a codon is implicated in
689 multiple responses, for clarity they are coloured in order of preference NAb, HGF24, CTL.

690

691 S2 Fig: Distributions of input and inferred magnitudes of selection for simulated data in which the
692 observed data described a fraction of the simulated region of the virus. Data are shown for
693 variants at which the magnitude of selection could be inferred with confidence.

694

695 S3 Fig: Proportion of mutations inferred to be under selection that are towards population level
696 consenus or are nonsynonymous. This includes codons that are genuinely under selection and
697 those that are increasing in frequency due to hitchhiking. In all cases mutations are grouped
698 according to the number of times the codon in which they appear is inferred to be under selection
699 across the 34 individuals (x-axis). Top row: the number of codons in each group. Middle row: the
700 proportion of mutations in each group that are towards population level consensus. Bottom row:
701 the proportion of mutations in each group that are nonsynonymous.

702

703 S4 Fig: Sets of nucleotide trajectories that were identified as putatively hitchhiking. These
704 trajectories were used to create a conservative estimate of the extent of noise in the sequencing
705 data.
706
707 S5 Fig: True and inferred magnitudes and timings of selection for simulated data. Confidence
708 intervals for the inferred selection coefficients are shown, calculated using the method described
709 in the main text. The red dashed line indicates agreement between the true and inferred
710 parameters.
711 S6 Fig: Observed (solid lines) and inferred (dashed lines) haplotype frequencies for simulated
712 data in which all loci under selection were observed. In some cases the lines cannot be
713 distinguished from one another.
714
715 S7 Fig: Observed (solid lines) and inferred (dashed lines) haplotype frequencies for simulated
716 data in which only data from within a fraction of a simulated region was observed. In some cases
717 the lines cannot be distinguished from one another.
718
719 S1 Table: Summary of results for all sites inferred to be under selection
720
721 S2 Table: Characteristics of all codons analysed: Sensitivity to host immunity, within- and
722 between-host diversity, and the number of times the codon was inferred under selection
723
724 S3 Table: True and inferred selection parameters for a case in which sequence data describes
725 all selected alleles within a system.
726

727 S4 Table: True and inferred selection parameters for a case in which sequence data partially
728 describes the selected alleles within a system. Eleven out of 24 alleles were described in the data
729 used for the inference calculation.
730
731 S1 Text: Calculations performed on simulated data, where targeted reads describe the complete
732 evolution of the system, and where targeted reads do not describe the complete evolution of the
733 system.
734
735 **References**

736 1. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV
737 evolution. Nat Rev Genet. 2004;5: 52–61.

738 2. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of
739 intrapatient HIV-1 evolution. Elife. 2015;4. doi:10.7554/eLife.11282

740 3. Lythgoe KA, Gardner A, Pybus OG, Grove J. Short-Sighted Virus Evolution and a Germline
741 Hypothesis for Chronic Viral Infections. Trends Microbiol. 2017;25: 336–348.

742 4. Keating CP, Hill MK, Hawkes DJ, Smyth RP, Isel C, Le S-Y, et al. The A-rich RNA
743 sequences of HIV-1 pol are important for the synthesis of viral cDNA. Nucleic Acids Res.
744 2009;37: 945–956.

745 5. Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A. Mapping of positive selection sites in the
746 HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology.
747 2011;8: 87.

748 6. Parera M, Fernàndez G, Clotet B, Martínez MA. HIV-1 protease catalytic efficiency effects
749 caused by random single amino acid substitutions. Mol Biol Evol. 2007;24: 382–387.

750 7. Rihn SJ, Hughes J, Wilson SJ, Bieniasz PD. Uneven Genetic Robustness of HIV-1
751 Integrase. J Virol. 2014;89: 552–567.

752 8. Haddox HK, Dingens AS, Bloom JD. Experimental Estimation of the Effects of All Amino-
753 Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. PLoS Pathog.
754 2016;12: e1006114.

755 9. Haddox HK, Dingens AS, Hilton SK, Overbaugh J, Bloom JD. Mapping mutational effects
756 along the evolutionary landscape of HIV envelope. Elife. 2018;7. doi:10.7554/eLife.34420

757 10. Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, Whitcomb JM, et al. A systems
758 analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nat Genet.
759 2011;43: 487–489.

760 11. Rhee S-Y, Gonzales MJ, Kanor R, Betts BJ, Ravela J, Shafer RW. Human
761 immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic
762 Acids Res. 2003;31: 298–303.

763 12. Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. Translating
764 HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational
765 immunogen design. Immunity. 2013;38: 606–617.

766 13. Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. Relative
767 rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nat
768 Commun. 2016;7: 11660.

769 14. Louie RHY, Kaczorowski KJ, Barton JP, Chakraborty AK, McKay MR. Fitness landscape of
770 the human immunodeficiency virus envelope protein that is targeted by antibodies. Proc
771 Natl Acad Sci U S A. 2018;115: E564–E573.

772 15. Kiepiela P, Leslie AJ, Honeyborne I, Ramduth D, Thobakgale C, Chetty S, et al. Dominant
773 influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature.
774 2004;432: 769–775.

775 16. Matthews PC, Leslie AJ, Katzourakis A, Crawford H, Payne R, Prendergast A, et al. HLA
776 footprints on human immunodeficiency virus type 1 are associated with interclade
777 polymorphisms and intraclade phylogenetic clustering. J Virol. 2009;83: 4605–4615.

778 17. Theys K, Feder AF, Gelbart M, Hartl M, Stern A, Pennings PS. Within-patient mutation
779 frequencies reveal fitness costs of CpG dinucleotides and drastic amino acid changes in
780 HIV. PLoS Genet. 2018;14: e1007420.

781 18. Haldane JBS. The Effect of Variation of Fitness. Am Nat. 1937;71: 337–349.

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

784 20. Fernandez CS, Stratov I, De Rose R, Walsh K, Dale CJ, Smith MZ, et al. Rapid viral
785 escape at an immunodominant simian-human immunodeficiency virus cytotoxic T-
786 lymphocyte epitope exacts a dramatic fitness cost. J Virol. 2005;79: 5721–5731.

787 21. Asquith B, Edwards CTT, Lipsitch M, McLean AR. Inefficient Cytotoxic T Lymphocyte–
788 Mediated Killing of HIV-1–Infected Cells In Vivo. PLoS Biol. 2006;4: e90.

789 22. Ganusov VV, De Boer RJ. Estimating Costs and Benefits of CTL Escape Mutations in
790 SIV/HIV Infection. PLoS Comput Biol. 2006;2: e24.

791 23. Ganusov VV, Goonetilleke N, Liu MKP, Ferrari G, Shaw GM, McMichael AJ, et al. Fitness
792 costs and diversity of the cytotoxic T lymphocyte (CTL) response determine the rate of CTL
793 escape during acute and chronic phases of HIV infection. J Virol. 2011;85: 10518–10528.

794 24. Kessinger TA, Perelson AS, Neher RA. Inferring HIV Escape Rates from Multi-Locus
795 Genotype Data. Front Immunol. 2013;4. doi:10.3389/fimmu.2013.00252

796 25. van Deutekom HWM, Wijnker G, de Boer RJ. The rate of immune escape vanishes when
797 multiple immune responses control an HIV infection. J Immunol. 2013;191: 3277–3286.

798 26. Leviyang S, Ganusov VV. Broad CTL Response in Early HIV Infection Drives Multiple
799 Concurrent CTL Escapes. PLoS Comput Biol. 2015;11: e1004492.

800 27. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res.
801 1966;8: 269.

802 28. Illingworth CJR, Mustonen V. Distinguishing driver and passenger mutations in an
803 evolutionary history categorized by interference. Genetics. 2011;189: 989–1000.

804 29. Garcia V, Regoes RR. The Effect of Interference on the CD8 T Cell Escape Rates in HIV.
805 Front Immunol. 2015;5. doi:10.3389/fimmu.2014.00661

806 30. Garcia V, Feldman MW, Regoes RR. Investigating the Consequences of Interference
807 between Multiple CD8+ T Cell Escape Mutations in Early HIV Infection. PLoS Comput Biol.
808 2016;12: e1004721.

809 31. Yang Y, Ganusov VV. Kinetics of HIV-Specific CTL Responses Plays a Minimal Role in
810 Determining HIV Escape Dynamics [Internet]. Frontiers in Immunology. 2018.
811 doi:10.3389/fimmu.2018.00140

812 32. Raghwani J, Redd AD, Longosz AF, Wu C-H, Serwadda D, Martens C, et al. Evolution of
813 HIV-1 within untreated individuals and at the population scale in Uganda. PLoS Pathog.
814 2018;14: e1007167.

815 33. Illingworth CJR. Fitness Inference from Short-Read Data: Within-Host Evolution of a
816 Reassortant H5N1 Influenza Virus. Mol Biol Evol. 2015;32: 3012–3026.

817 34. Asquith B, McLean AR. In vivo CD8 T cell control of immunodeficiency virus infection in
818 humans and macaques [Internet]. Proceedings of the National Academy of Sciences. 2007.
819 pp. 6365–6370. doi:10.1073/pnas.0700666104

820 35. Yang OO, Daar ES, Jamieson BD, Balamurugan A, Smith DM, Pitt JA, et al. Human
821 immunodeficiency virus type 1 clade B superinfection: evidence for differential immune
822 containment of distinct clade B strains. J Virol. 2005;79: 860–868.

823 36. Watanabe K, Murakoshi H, Tamura Y, Koyanagi M, Chikata T, Gatanaga H, et al.
824 Identification of cross-clade CTL epitopes in HIV-1 clade A/E-infected individuals by using
825 the clade B overlapping peptides. Microbes Infect. 2013;15: 874–886.

826 37. Buckheit RW 3rd, Allen TG, Alme A, Salgado M, O'Connell KA, Huculak S, et al. Host
827 factors dictate control of viral replication in two HIV-1 controller/chronic progressor
828 transmission pairs. Nat Commun. 2012;3: 716.

829 38. Sipsas NV, Kalams SA, Trocha A, He S, Blattner WA, Walker BD, et al. Identification of
830 type-specific cytotoxic T lymphocyte responses to homologous viral proteins in laboratory
831 workers accidentally infected with HIV-1. J Clin Invest. 1997;99: 752–762.

832 39. Hoof I, Pérez CL, Buggert M, Gustafsson RKL, Nielsen M, Lund O, et al. Interdisciplinary
833 analysis of HIV-specific CD8+ T cell responses against variant epitopes reveals restricted
834 TCR promiscuity. J Immunol. 2010;184: 5383–5391.

835 40. Tang Y, Huang S, Dunkley-Thompson J, Steel-Duncan JC, Ryland EG, St John MA, et al.
836 Correlates of spontaneous viral control among long-term survivors of perinatal HIV-1

837 infection expressing human leukocyte antigen-B57. AIDS. 2010;24: 1425–1435.

838 41. Miura T, Brockman MA, Schneidewind A, Lobritz M, Pereyra F, Rathod A, et al. HLA-
839 B57/B*5801 human immunodeficiency virus type 1 elite controllers select for rare gag
840 variants associated with reduced viral replication capacity and strong cytotoxic T-
841 lymphocyte [corrected] recognition. J Virol. 2009;83: 2743–2755.

842 42. Chopera DR, Mlotshwa M, Woodman Z, Mlisana K, de Assis Rosa D, Martin DP, et al.
843 Virological and Immunological Factors Associated with HIV-1 Differential Disease
844 Progression in HLA-B*58:01-Positive Individuals [Internet]. Journal of Virology. 2011. pp.
845 7070–7080. doi:10.1128/jvi.02543-10

846 43. Koup RA, Roederer M, Lamoreaux L, Fischer J, Novik L, Nason MC, et al. Priming
847 immunization with DNA augments immunogenicity of recombinant adenoviral vectors for
848 both HIV-1 specific antibody and T-cell responses. PLoS One. 2010;5: e9015.

849 44. Feeney ME, Tang Y, Pfafferott K, Roosevelt KA, Draenert R, Trocha A, et al. HIV-1 viral
850 escape in infancy followed by emergence of a variant-specific CTL response. J Immunol.
851 2005;174: 7524–7530.

852 45. McKinnon LR, Capina R, Peters H, Mendoza M, Kimani J, Wachihi C, et al. Clade-specific
853 evolution mediated by HLA-B*57/5801 in human immunodeficiency virus type 1 clade A1
854 p24. J Virol. 2009;83: 12636–12642.

855 46. Payne RP, Branch S, Kløverpris H, Matthews PC, Koofhethile CK, Strong T, et al.
856 Differential escape patterns within the dominant HLA-B*57:03-restricted HIV Gag epitope
857 reflect distinct clade-specific functional constraints. J Virol. 2014;88: 4668–4678.

858 47. Du VY, Bansal A, Carlson J, Salazar-Gonzalez JF, Salazar MG, Ladell K, et al. HIV-1-
859 Specific CD8 T Cells Exhibit Limited Cross-Reactivity during Acute Infection. J Immunol.
860 2016;196: 3276–3286.

861 48. Vollbrecht T, Brackmann H, Henrich N, Roeling J, Seybold U, Bogner JR, et al. Impact of
862 changes in antigen level on CD38/PD-1 co-expression on HIV-specific CD8 T cells in
863 chronic, untreated HIV-1 infection. J Med Virol. 2010;82: 358–370.

864 49. Zhao S, Zhai S, Zhuang Y, Wang S, Huang D, Kang W, et al. Inter-clade cross-reactivity of
865 HIV-1-specific T cell responses in human immunodeficiency virus type 1 infection in China.
866 Curr HIV Res. 2007;5: 251–259.

867 50. Malhotra U, Nolin J, Mullins JI, McElrath MJ. Comprehensive epitope analysis of cross-
868 clade Gag-specific T-cell responses in individuals with early HIV-1 infection in the US
869 epidemic. Vaccine. 2007;25: 381–390.

870 51. Kaul R, Dong T, Plummer FA, Kimani J, Rostron T, Kiama P, et al. CD8(+) lymphocytes
871 respond to different HIV epitopes in seronegative and infected subjects. J Clin Invest.
872 2001;107: 1303–1310.

873 52. Aidoo M, Sawadogo S, Bile EC, Yang C, Nkengasong JN, McNicholl JM. Viral, HLA and T
874 cell elements in cross-reactive immune responses to HIV-1 subtype A, CRF01_AE and
875 CRF02_AG vaccine sequence in Ivorian blood donors [Internet]. Vaccine. 2008. pp. 4830–
876 4839. doi:10.1016/j.vaccine.2008.06.097

877 53. Migueles SA, Laborico AC, Imamichi H, Shupert WL, Royce C, McLaughlin M, et al. The
878 Differential Ability of HLA B*5701 Long-Term Nonprogressors and Progressors To Restrict
879 Human Immunodeficiency Virus Replication Is Not Caused by Loss of Recognition of
880 Autologous Viral gag Sequences [Internet]. Journal of Virology. 2003. pp. 6889–6898.
881 doi:10.1128/jvi.77.12.6889-6898.2003

882 54. Blattner C, Lee JH, Sliepen K, Derking R, Falkowska E, de la Peña AT, et al. Structural
883 delineation of a quaternary, cleavage-dependent epitope at the gp41-gp120 interface on
884 intact HIV-1 Env trimers. Immunity. 2014;40: 669–680.

885 55. Falkowska E, Le KM, Ramos A, Doores KJ, Lee JH, Blattner C, et al. Broadly neutralizing
886 HIV antibodies define a glycan-dependent epitope on the prefusion conformation of gp41
887 on cleaved envelope trimers. Immunity. 2014;40: 657–668.

888 56. Hraber P, Korber B, Wagh K, Giorgi EE, Bhattacharya T, Gnanakaran S, et al. Longitudinal
889 Antigenic Sequences and Sites from Intra-Host Evolution (LASSIE) Identifies Immune-
890 Selected HIV Variants. Viruses. 2015;7: 5443–5475.

891 57. Bricault CA, Yusim K, Seaman MS, Yoon H, Theiler J, Giorgi EE, et al. HIV-1 Neutralizing
892 Antibody Signatures and Application to Epitope-Targeted Vaccine Design. Cell Host
893 Microbe. 2019;26: 296.

894 58. Huang J, Kang BH, Pancera M, Lee JH, Tong T, Feng Y, et al. Broad and potent HIV-1
895 neutralization by a human antibody that binds the gp41-gp120 interface. Nature. 2014;515:
896 138–142.

897 59. Corti D, Langedijk JPM, Hinz A, Seaman MS, Vanzetta F, Fernandez-Rodriguez BM, et al.
898 Analysis of Memory B Cell Responses and Isolation of Novel Monoclonal Antibodies with
899 Neutralizing Breadth from HIV-1-Infected Individuals [Internet]. PLoS ONE. 2010. p. e8805.
900 doi:10.1371/journal.pone.0008805

901 60. Gnanakaran S, Daniels MG, Bhattacharya T, Lapedes AS, Sethi A, Li M, et al. Genetic
902 signatures in the envelope glycoproteins of HIV-1 that associate with broadly neutralizing
903 antibodies. PLoS Comput Biol. 2010;6: e1000955.

904 61. Haim H, Strack B, Kassa A, Madani N, Wang L, Courter JR, et al. Contribution of intrinsic
905 reactivity of the HIV-1 envelope glycoproteins to CD4-independent infection and global
906 inhibitor sensitivity. PLoS Pathog. 2011;7: e1002101.

907 62. Brunel FM, Zwick MB, Cardoso RMF, Nelson JD, Wilson IA, Burton DR, et al. Structure-
908 function analysis of the epitope for 4E10, a broadly neutralizing human immunodeficiency
909 virus type 1 antibody. J Virol. 2006;80: 1680–1687.

910 63. Nelson JD, Brunel FM, Jensen R, Crooks ET, Cardoso RMF, Wang M, et al. An affinity-
911 enhanced neutralizing antibody against the membrane-proximal external region of human
912 immunodeficiency virus type 1 gp41 recognizes an epitope between those of 2F5 and
913 4E10. J Virol. 2007;81: 4033–4043.

914 64. Chuang G-Y, -Y. Chuang G, Acharya P, Schmidt SD, Yang Y, Louder MK, et al. Residue-
915 Level Prediction of HIV-1 Antibody Epitopes Based on Neutralization of Diverse Viral
916 Strains [Internet]. Journal of Virology. 2013. pp. 10047–10058. doi:10.1128/jvi.00984-13

917 65. Montero M, Gulzar N, Klaric K-A, Donald JE, Lepik C, Wu S, et al. Neutralizing epitopes in
918 the membrane-proximal external region of HIV-1 gp41 are influenced by the
919 transmembrane domain and the plasma membrane. J Virol. 2012;86: 2930–2941.

920 66. Williams KL, Cortez V, Dingens AS, Gach JS, Rainwater S, Weis JF, et al. HIV-specific
921 CD4-induced Antibodies Mediate Broad and Potent Antibody-dependent Cellular
922 Cytotoxicity Activity and are Commonly Detected in Plasma from HIV-infected Humans
923 [Internet]. EBioMedicine. 2015. pp. 1464–1477. doi:10.1016/j.ebiom.2015.09.001

924 67. Cao K, Moormann AM, Lyke KE, Masaberg C, Sumba OP, Doumbo OK, et al.
925 Differentiation between African populations is evidenced by the diversity of alleles and
926 haplotypes of HLA class I loci [Internet]. Tissue Antigens. 2004. pp. 293–325.
927 doi:10.1111/j.0001-2815.2004.00192.x

928 68. Walker B, McMichael A. The T-Cell Response to HIV [Internet]. Cold Spring Harbor
929 Perspectives in Medicine. 2012. pp. a007054–a007054. doi:10.1101/cshperspect.a007054

930 69. Bar KJ, Tsao C-Y, Iyer SS, Decker JM, Yang Y, Bonsignori M, et al. Early Low-Titer
931 Neutralizing Antibodies Impede HIV-1 Replication and Select for Virus Escape [Internet].
932 PLoS Pathogens. 2012. p. e1002721. doi:10.1371/journal.ppat.1002721

933 70. Herbeck JT, Nickle DC, Learn GH, Gottlieb GS, Curlin ME, Heath L, et al. Human
934 Immunodeficiency Virus Type 1 env Evolves toward Ancestral States upon Transmission to
935 a New Host [Internet]. Journal of Virology. 2006. pp. 1637–1644. doi:10.1128/jvi.80.4.1637-
936 1644.2006

937 71. Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, et al. HIV evolution:
938 CTL escape mutation and reversion after transmission [Internet]. Nature Medicine. 2004.
939 pp. 282–289. doi:10.1038/nm992

940 72. Friedrich TC, Dodds EJ, Yant LJ, Vojnov L, Rudersdorf R, Cullen C, et al. Reversion of CTL
941 escape–variant immunodeficiency viruses in vivo [Internet]. Nature Medicine. 2004. pp.
942 275–281. doi:10.1038/nm998

943 73. Delport W, Scheffler K, Seoighe C. Frequent Toggling between Alternative Amino Acids Is
944 Driven by Selection in HIV-1 [Internet]. PLoS Pathogens. 2008. p. e1000242.
945 doi:10.1371/journal.ppat.1000242

946 74. Frost SDW, Wrin T, Smith DM, Pond SLK, Liu Y, Paxinos E, et al. Neutralizing antibody
947 responses drive the evolution of human immunodeficiency virus type 1 envelope during
948 recent HIV infection [Internet]. Proceedings of the National Academy of Sciences. 2005. pp.
949 18514–18519. doi:10.1073/pnas.0504658102

950 75. Richman DD, Wrin T, Little SJ, Petropoulos CJ. Rapid evolution of the neutralizing antibody
951 response to HIV type 1 infection [Internet]. Proceedings of the National Academy of
952 Sciences. 2003. pp. 4144–4149. doi:10.1073/pnas.0630530100

953 76. Pennings PS, Kryazhimskiy S, Wakeley J. Loss and recovery of genetic diversity in
954 adapting populations of HIV. PLoS Genet. 2014;10: e1004000.

955 77. Rouzine IM, Rodrigo A, Coffin JM. Transition between stochastic evolution and
956 deterministic evolution in the presence of selection: general theory and application to

957 virology. Microbiol Mol Biol Rev. 2001;65: 151–185.

958 78. Lythgoe KA, Fraser C. New insights into the evolutionary rate of HIV-1 at the within-host
959 and epidemiological levels [Internet]. Proceedings of the Royal Society B: Biological
960 Sciences. 2012. pp. 3367–3375. doi:10.1098/rspb.2012.0595

961 79. Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, et al. Antibody neutralization and
962 escape by HIV-1. Nature. 2003;422: 307–312.

963 80. van Gils MJ, Bunnik EM, Burger JA, Jacob Y, Schweighardt B, Wrin T, et al. Rapid escape
964 from preserved cross-reactive neutralizing humoral immunity without loss of viral fitness in
965 HIV-1-infected progressors and long-term nonprogressors. J Virol. 2010;84: 3576–3585.

966 81. Lynch RM, Rong R, Boliar S, Sethi A, Li B, Mulenga J, et al. The B Cell Response Is
967 Redundant and Highly Focused on V1V2 during Early Subtype C Infection in a Zambian
968 Seroconverter [Internet]. Journal of Virology. 2011. pp. 905–915. doi:10.1128/jvi.02006-10

969 82. Moore PL, Ranchobe N, Lambson BE, Gray ES, Cave E, Abrahams M-R, et al. Limited
970 neutralizing antibody specificities drive neutralization escape in early HIV-1 subtype C
971 infection. PLoS Pathog. 2009;5: e1000598.

972 83. Lynch RM, Wong P, Tran L, O'Dell S, Nason MC, Li Y, et al. HIV-1 fitness cost associated
973 with escape from the VRC01 class of CD4 binding site neutralizing antibodies. J Virol.
974 2015;89: 4201–4213.

975 84. Bunnik EM, Euler Z, Welkers MRA, Boeser-Nunnink BDM, Grijsen ML, Prins JM, et al.
976 Adaptation of HIV-1 envelope gp120 to humoral immunity at a population level. Nat Med.
977 2010;16: 995–997.

978 85. Bouvin-Pley M, Morgand M, Moreau A, Jestin P, Simonnet C, Tran L, et al. Evidence for a
979 continuous drift of the HIV-1 species towards higher resistance to neutralizing antibodies
980 over the course of the epidemic. PLoS Pathog. 2013;9: e1003477.

981 86. Bouvin-Pley M, Morgand M, Meyer L, Goujard C, Moreau A, Mouquet H, et al. Drift of the
982 HIV-1 Envelope Glycoprotein gp120 toward Increased Neutralization Resistance over the
983 Course of the Epidemic: a Comprehensive Study Using the Most Potent and Broadly
984 Neutralizing Monoclonal Antibodies [Internet]. Journal of Virology. 2014. pp. 13910–13917.
985 doi:10.1128/jvi.02083-14

986 87. Rademeyer C, Korber B, Seaman MS, Giorgi EE, Thebus R, Robles A, et al. Features of
987 Recently Transmitted HIV-1 Clade C Viruses that Impact Antibody Recognition:
988 Implications for Active and Passive Immunization. PLoS Pathog. 2016;12: e1005742.

989 88. Stefic K, Bouvin-Pley M, Essat A, Visdeloup C, Moreau A, Goujard C, et al. Sensitivity to
990 Broadly Neutralizing Antibodies of Recently Transmitted HIV-1 Clade CRF02_AG Viruses
991 with a Focus on Evolution over Time [Internet]. Journal of Virology. 2018.
992 doi:10.1128/jvi.01492-18

993 89. Sobel Leonard A, McClain MT, Smith GJD, Wentworth DE, Halpin RA, Lin X, et al. The
994 effective rate of influenza reassortment is limited during human infection. PLoS Pathog.
995 2017;13: e1006203.

90. Illingworth CJR, Mustonen V. Components of selection in the evolution of the influenza virus: linkage effects beat inherent selection. PLoS Pathog. 2012;8: e1003091.

91. Illingworth CJR, Fischer A, Mustonen V. Identifying selection in the within-host evolution of influenza using viral sequence data. PLoS Comput Biol. 2014;10: e1003755.

92. Redd AD, Mullis CE, Serwadda D, Kong X, Martens C, Ricklefs SM, et al. The Rates of HIV Superinfection and Primary HIV Incidence in a General Population in Rakai, Uganda [Internet]. The Journal of Infectious Diseases. 2012. pp. 267–274. doi:10.1093/infdis/jis325

93. Illingworth CJR, Roy S, Beale MA, Tutill H, Williams R, Breuer J. On the effective depth of viral sequence data. Virus Evol. 2017;3: vex030.

94. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene [Internet]. Genetical Research. 1974. p. 23. doi:10.1017/s0016672300014634

95. Illingworth CJR. SAMFIRE: multi-locus variant calling for time-resolved sequence data. Bioinformatics. 2016;32: 2208–2209.

96. Schwarz G. Estimating the Dimension of a Model. Ann Stat. 1978;6: 461–464.

97. Mansky LM. The mutation rate of human immunodeficiency virus type 1 is influenced by the vpr gene. Virology. 1996;222: 391–400.

98. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral Mutation Rates. J Virol. 2010;84: 9733–9748.

99. Neher RA, Leitner T. Recombination rate and selection strength in HIV intra-patient evolution. PLoS Comput Biol. 2010;6: e1000660.

100. Batorsky R, Kearney MF, Palmer SE, Maldarelli F, Rouzine IM, Coffin JM. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. Proc Natl Acad Sci U S A. 2011;108: 5661–5666.

101. Markowitz M, Louie M, Hurley A, Sun E, Di Mascio M, Perelson AS, et al. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. J Virol. 2003;77: 5037–5038.

102. Zhao L, Illingworth CJR. Measurements of intrahost viral diversity require an unbiased diversity metric. Virus Evol. 2019;5: vey041.