Supplemental Digital Content

*Study cohort and sequence data*

A total of 24,550 HIV-1 subtype B partial *pol* gene sequences from the UK HIV Drug Resistance database were analysed. Detail of the database is given elsewhere [[1](#_ENREF_1)]. In brief, sequences were sampled from HIV positive individuals across the UK though routine genotypic resistance testing between 1997 and 2011. Sequences span the entire protease (PR; 297 nucleotides) and first 1,248 nucleotides of the reverse transcriptase (RT) of the virus. For patients with multiple sequences, only the first available sequence was included in the analysis. Sequences were linked to pseudo-anonymised clinical and demographic information including gender, risk group, ethnicity, age group, and treatment status.  Some DRMs were much more frequent than others, such as L90M and K103N, harboured by 0.9% (168/18,354) and 2.7% (490/18,354) of drug-naïve sequences, respectively.

*Phylogenetic reconstruction*

Sequences were aligned with ClustalX [[2](#_ENREF_2)] and manually edited with the software Se-Al version 2.0a11 ([*http://tree.bio.ed.ac.uk/software/seal/*](http://tree.bio.ed.ac.uk/software/seal/)). In the alignment, positions corresponding to drug-resistance mutations were removed in order to avoid biased topologies reflecting convergent evolution due to antiretroviral treatment. In order to assess the impact of phylogenetic uncertainty on our estimates, ten alternative phylogenies were reconstructed from the sequence alignment. First, 10,000 equivalently most parsimonious trees were generated using the software TNT [[3](#_ENREF_3)] version 1.1 (with the following parameters : tshrink !; rseed !; mult: wc 1000; bbreak: clu 40; mult 5 = hold 1 keep). A set of 100 trees were uniformly drawn from these parsimonious trees and used as starting topologies for maximum likelihood (ML) inference with FastTree v2.1.5 [[4](#_ENREF_4)] under the General Time Reversible model of nucleotide substitutions and varying evolutionary rates across sites (GTR+CAT; with the following parameters: -gamma -nt -gtr -nome -mllen -intree). The 10 trees with the highest likelihood values were selected and analysed independently. Mean estimates over the 10 tree topologies are reported.

*Phylotype identification*

A phylotype is a set of monophyletic sequences sharing a common trait (e.g. a particular DRM) in a phylogeny [[5](#_ENREF_5)]. We selected phylotypes as subsets of strains being stable across the 10 very large ML phylogenies (n = 24,550) in terms of membership and filiation. Such well characterised subsets enabled accurate estimations of the parameters of interest (origin of DRMs in naïve patients, reversion time, etc.). Both drug-naïve and mixed phylotypes were studied. First, an ancestral character state (presence/absence) reconstruction of all positions associated with drug resistance was performed along each of the 10 ML trees, using Fitch’s parsimony with the DELTRAN option [[6](#_ENREF_6)], as implemented in Phylotype version 7.0 (*http://phylotype.org/*). Fitch’s parsimony with ACCTRAN option yielded very similar results. From these, phylotypes were identified using the following criteria: (i) ≥3 sequences harbouring the same DRM (Phylotype option: Size≥3); (ii) a maximum intra-clade genetic distance of 4.0% (Diversity≤0.02), to avoid focusing on subtrees where missing data are (relatively) probable; and (iii) a basal branch support ≥90% (Support≥90%; branch support calculated by SH-like test, as implemented in FastTree) to ensure result stability across the 10 ML trees. Phylotypes for which the number of sub-clades without the shared DRM exceeded the number of sequences with the shared DRM were excluded (Size/Different≥1; Different is the number of sub-clades without the shared DRM at sub-clade root). Lastly, we used another criterion that ensures that both direct descendants of the most recent common ancestor (MRCA) present the DRM annotation (Persistence≥1). Resistant phylotypes are schematized in Figure 1A.

To assess the statistical significance of the phylogeny/DRM association found in the selected phylotypes, DRM annotations were reassigned at the phylogenies’ tips of the trees through 1,000 random permutations, simulating the null hypothesis that the distribution of DRMs in the tree is independent to the sequences relatedness. Phylotypes for which the phylogeny/DRM association didn’t reach statistical significance at the p = 0.01 (10/1,000) level were excluded from the analysis.

*Classification of phylotypes*

We distinguish three categories of resistant phylotypes: (i) phylotypes comprising treatment-naïve individuals only (called naïve phylotypes), (ii) phylotypes comprising both treatment-naïve and experienced individuals (called mixed phylotypes), and (iii) phylotypes comprising treatment-experienced individuals only (called experienced phylotypes).

Both naïve and experienced phylotypes are straightforward to identify from Phylotype program’s results, and to analyze. For mixed phylotypes, we developed a methodology to infer those that are topologically stable across the 10 alternative ML trees (see Section Inference of Phylotype Consensus).

Both naïve and mixed phylotypes are meant to reflect the transmission of DRMs among the individuals. Conversely, in experienced phylotypes, DRMs could be either transmitted or acquired independently among the individuals by selective pressure due to drugs.

An average of 84 naïve and mixed phylotypes were identified, while only 14 experienced phylotypes were discovered. Moreover, the number of sequences in the naïve and mixed phylotypes varied from 3 to 24 sequences. On the contrary, experienced phylotypes were small and comprised 3 to 4 sequences. These results show that (as expected) phylotypes mainly comprise naïve-to-naïve and experienced-to-naïve transmissions of DRMs, but encompass only a few number of experienced-to-experienced transmissions or independently acquired resistances.

*Assessment of phylotype stability*

Since phylotypes are topology-dependent, the size, composition and topology of a given phylotype may vary across the 10 alternative ML trees. The stability of selected phylotypes was assessed by calculating pairwise similarity values between two phylotypes (from tree) and (from tree ) as:

**,**

with being the set of sequences found in phylotype . The resulting similarity values range from 0 to 1, with 1 indicating absolute similarity between the two phylotypes (i.e. the phylotype from tree and the phylotype from tree contain the same strains and are likely to be stable across trees).

A matrix of similarity values was computed for each possible pair of ML trees. For a given pair, the percentage of matched phylotypes was calculated as:

,

where a pair of phylotypes and is considered as matching if . was averaged over all possible pairs of trees. The percentage of matched phylotypes was also averaged over all DRM positions. Using this approach, we were able to assess to which extent the same phylotypes were found in the 10 ML trees, as a stable phylotype is expected to be found in all trees.

*Inference of phylotype consensus*

To infer phylotypes that are topologically stable across trees, we computed phylotype consensuses between trees. This approach was used with mixed phylotypes only (see text). Considering a phylotype (from tree) and a phylotype (from tree ) that are meant to represent the same stable phylotype, we seek to infer a phylotype that is the consensus of the two formers. To compute phylotype consensuses, we developed the following methodology. First, for a given DRM, pairwise similarities between phylotypes of the 10 trees were calculated. For each DRM, an undirected graph was then built connecting phylotypes between the 10 trees. In the graph, two nodes (e.g. a phylotype from tree and a phylotype from tree ) are connected by an edge if their pairwise similarity satisfies . In the graph, connected components with ≥ 6 nodes were considered relevant for the next steps. These connected components are meant to reflect stable phylotypes. For each of these connected components, two supertrees were built from the set of phylotype topologies, each one coming from a different ML tree. A matrix representation with parsimony tree (MRPT) was built using the Phytools R package version 0.4-31 (http://cran.r-project.org/web/packages/phytools/index.html), and the maximum agreement subtree (MAST) was inferred using PhyloNet version 3.5.1 (PhyloNet: Phylogenetic Networks Toolkit). The MRPT was considered as the consensus among the phylotypes belonging to the same connected component in the graph. But the MRPT reconstruction was considered valid only if the tree size ratio:

**,**

with being the sequences of the maximum agreement subtree calculated from , and being the sequences of the maximum agreement subtree calculated from . In other words, was retained when a core (MAST) tree was included in all trees of the connected component, with at least 50% of sequences in .

*Quantification of naïve-to-naïve transmission of drug resistance*

Not all transmitted DRMs in a resistant phylotype have a treatment-naïve source. Since DRMs exclusively emerge through treatment-associated selective pressure in a treated individual, at least one treatment-naïve individual in the transmission chain has a treated source. We therefore assumed that the number of naïve-to-naïve transmissions of DRMs in a naïve phylotype only corresponded to the number of treatment-naïve sequences in the resistant phylotype minus one.

For mixed phylotypes, the proportion of naïve-to-naïve DRM transmission was quantified using a parsimony ancestral reconstruction framework specifically designed for the study. Ancestral reconstruction can only be done for phylotypes whose topologies are stable, and hence we used phylotype consensuses. We selected phylotype consensuses (MRPTs) that are mixed and used the following procedure. Ancestral treatment status was reconstructed using post-order traversal of the MRPT and 3 particular parsimony rules, which comply with our common understanding of resistance transmission: (i) if the two children nodes of a node are drug-naïve, then the node is considered as naïve, and we have one naive-to-naive transmission; (ii) if the two children nodes are drug-experienced, then the node is considered as experienced, and we don’t have any naïve-to-naïve transmission; and (iii) if one child node is naïve and the other child node is experienced, then the node is considered as experienced, and we don’t have any naïve-to-naïve transmission. Note that using these rules, the root of mixed phylotypes is (as expected) always experienced. From this reconstruction of treatment status, we were able to count the number of naïve-to-naïve transmissions in phylotype consensuses that are mixed. In these phylotypes, the number of naïve-to-naïve transmissions was calculated as the number of naïve sequences minus the number of edges that satisfies ).

For each DRM, the rate of naïve-to-naïve transmission over all phylotypes was calculated as:

### ,


### where the numerator is the number of sequences that possess a mutation resulting from naïve-to-naïve transmissions, and the denominator is the number of sequences from naïve patients in phylotypes (naïve or mixed).

*Rate of drug-resistance loss in the untreated population*

### In a phylotype, most recent common ancestors of sub-clades without the shared trait (i.e. a particular DRM) are called exceptions. Note that an exception does not belong to a phylotype but is connected to. In the resistant phylotypes that we selected, an exception represents the loss of a DRM in the population through reversion to wild type (Fig. 1A). For each phylotype, the population loss rate (PLR) of a mutation was calculated as:


### For a given DRM, the PLR was averaged over all phylotypes found for that mutation:

### ,


### with weight the number of phylotype sequences plus the number of exceptions for a phylotype . Only naïve phylotypes were used in these estimations. Mixed phylotypes resulted from consensus inference that lost exception information and thus could not be used for PLR estimation.


### *Time of persistence in the drug-naïve population*

### The persistence of DRMs in treatment-naïve resistant phylotypes in the HIV infected population was estimated using a molecular clock inference approach. Fast least-squares inference was used to date all nodes in the whole phylogeny, thus obtaining dates for the ancestral nodes of each of the phylotypes; to this purpose we used the Least-Squares Dating program (LSD) version 1.0 with temporal constraints (*http://www.atgc-montpellier.fr/LSD/*). LSD time estimates were compared to those obtained by Bayesian Markov chain Monte Carlo inference using BEAST v1.8.0 [[7](#_ENREF_7)], under the SRD06 model of nucleotide substitution [[8](#_ENREF_8)], an uncorrelated lognormal (UCLN) model of molecular evolutionary rate heterogeneity and a Bayesian skyline tree coalescent prior (other options: logNormalPrior mean=”0.0025”, stdev=”0.1”, offset=”0”; MCMC chainLength= ”50,000,000” generations, logEvery=”10,000” generations). Since BEAST inference is not scalable to large HIV *pol* gene phylogenies, phylotypes were pooled in groups of 150 to 200 sequences and divergence times were estimated as previously described [[9](#_ENREF_9)].

### For a given phylotype, we estimated the date of the most recent common ancestor (tMRCA) of all drug-naïve sequences in that phylotype. For each DRM, the date of emergence (in the naïve population) was defined as the most ancient tMRCA over all phylotypes found for that mutation. For a given phylotype, the time of persistence of a DRM was estimated as the difference between tMRCA and the sampling date of the most recent drug-naïve sequence in the phylotype. For each DRM, times of persistence were averaged over all phylotypes found for that mutation.

### *Evolution of drug-naïve, resistant viral reservoirs over time*

### We compared the phylotypes observed in the present study with the treatment-independent, drug-resistant transmission clusters reported by Hué *et al.* in 2009 [[10](#_ENREF_10)]. For this purpose, the 38 viral sequences representing naive-to-naive transmission of resistance in the original study of Hué *et al.* were aligned to the phylotypes’ naïve sequences from the present analysis. From the alignment of sequences, a maximum likelihood phylogeny was inferred using the program FastTree v2.1.5 [[4](#_ENREF_4)] under the General Time Reversible model of nucleotide substitutions and varying evolution rates across sites (GTR+CAT). Transmission clusters from the 2009 study were retrieved from the topology and scanned for additional sequences from the present study.

### References

### 1. Committee UKCHCS. The creation of a large UK-based multicentre cohort of HIV-infected individuals: The UK Collaborative HIV Cohort (UK CHIC) Study. *HIV Med* 2004,5:115-124.

### 2. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997,25:4876-4882.

### 3. Goloboff PA, Farris J, Nixon K. T.N.T.: tree analysis using new technology. *Available from:* <http://www.zmuc.dk/public/phylogeny/TNT>2003.

### 4. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010,5:e9490.

### 5. Chevenet F, Jung M, Peeters M, de Oliveira T, Gascuel O. Searching for virus phylotypes. *Bioinformatics* 2013,29:561-570.

### 6. Fitch WM. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology* 1971,20:406–416.

### 7. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007,7:214.

### 8. Shapiro B, Rambaut A, Drummond AJ. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol* 2006,23:7-9.

### 9. Hue S, Brown AE, Ragonnet-Cronin M, Lycett SJ, Dunn DT, Fearnhill E*, et al.* Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. *AIDS* 2014,28:1967-1975.

### 10. Hue S, Gifford RJ, Dunn D, Fernhill E, Pillay D, Resistance UKCGoHD. Demonstration of sustained drug-resistant human immunodeficiency virus type 1 lineages circulating among treatment-naive individuals. *J Virol* 2009,83:2645-2654.

**Supplementary Figure 1.** (A) Comparison of the mean time of persistence of resistant phylotypes estimated by fast least-squares inference (LSD; x axis) and Bayesian MCMC inference (BEAST; y axis), per resistance mutation. (B) Comparison of mean time of emergence of the resistant phylotypes in the naïve population estimated by LSD (x axis) and BEAST (y axis), per resistance mutation. Resistance to protease inhibitors (PIs), nucleoside analog reverse transcriptase inhibitors (nRTIs) and non-nucleoside analog reverse transcriptase inhibitors (nnRTIs) is indicated in black, grey and white respectively.

B

Mean maximum persistence in LSD (years)

Mean maximum persistence in BEAST (years)

A

V108I

K219E

D67N

K219Q

M46L

Y181C

Y188L

L210W

T215S

K103N

V82L

V82A

T215E

T215V

T215C

T215D

M41L

Oldest tMRCA in LSD (calendar years)

Oldest tMRCA in BEAST (calendar years)

T215D

T215S

M41L

K219E

L210W

K103N

V82L

Y188L

M46L

T215C

D67N

T215E

T215V

Y181C

V108I

K219E

V82A

PI

nRTI

nnRTI

PI

nRTI

nnRTI

**Error! Reference source not found.**