Supplementary data:

HIV-1 p24Gag adaptation to modern and archaic HLA-allele frequency differences in ethnic groups contributes to viral subtype diversification

Nicolaas C. Kist, Ben Lambert, Samuel Campbell, Aris Katzourakis, Daniel Lunn, Philippe Lemey, Astrid K.N. Iversen

Overview:

1) Supporting figures 1 to 12:

Fig. S1: Subtype-specific differences between the two conserved, clinically important HIV Gag Regions in HIV-B (B) and HIV-C (C)

Fig. S2: Outline of HLA-associated single-epitope and multiple-epitope selective pressures.

Fig. S3: HIV-1 subtype distribution in Sub-Saharan Africa

Fig. S4: The patients' haplotype score in PC1 and PC2

Fig. S5: HIV-C, HIV-B and selected HLA frequencies in Southern Africa, the US, and Europe

Fig. S6: HIV-B phylogeny

Fig. S7: HIV-C phylogeny

Fig. S8: MCA of each patient's HIV-1 subtype-specific amino acids

Fig. S9: Modeling of the evolution of HIV-C SSAA using HIV-B-related HLA ORs

Fig. S10: HLA diversity in mixed African American and US Caucasian populations

Fig. S11: Proportion of African Americans in the US HIV-infected population over time

Fig. S12: Phylogeny of the American HIV-B epidemic combined with demographic data

2) Supporting tables 1 to 8:

Table S1: African populations with HLA A, B, and C data

Table S2: The number of p24 Gag sequences per country and the number of HIV-1 subtypes, CRFs and URFs per country and the country code abbreviation key

Table S3: Complete key to the African ethnicities in **Fig. 3C**

Table S4: Complete key over HIV-1 subtypes shown in **Fig. 3D**

Table S5: Complete key over African languages shown in **Fig. 3E**

Table S6: HLA class I allele frequencies in worldwide populations

Table S7: Overview of the amino acid variation at SSP in HIV-1 subtype consensus sequences

Table S8: LANL patient IDs of patients with imputed four digit HLAs

Figure S1

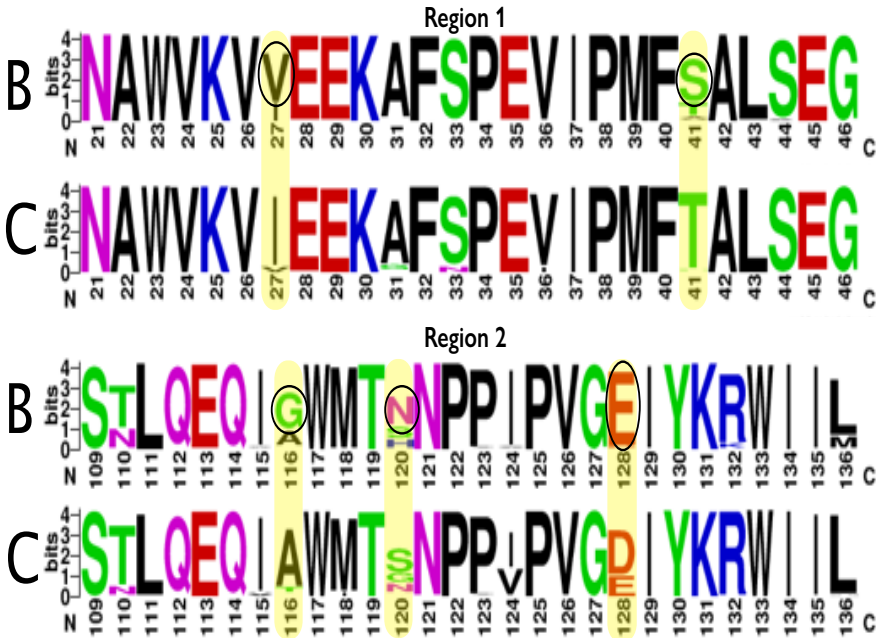

Fig. S1: Subtype-specific differences between the two conserved, clinically-important HIV Gag Regions in HIV-B (B) and HIV-C (C)

The subtype-specific positions (SSP) 27, 41, 116, 120, and 128 are highlighted in yellow, and the letters representing site-specific amino acids are sized according to their associated frequencies. HIV-B position 128 and HIV-C positions 27, 41, and 116 are almost always occupied by the consensus amino acids and cannot be used in the analysis. The HIV-B consensus amino acids are circled. This figure was generated using a single sequence from each HLA annotated patient in the HIV database (Foley et al. 2018).

Figure S2

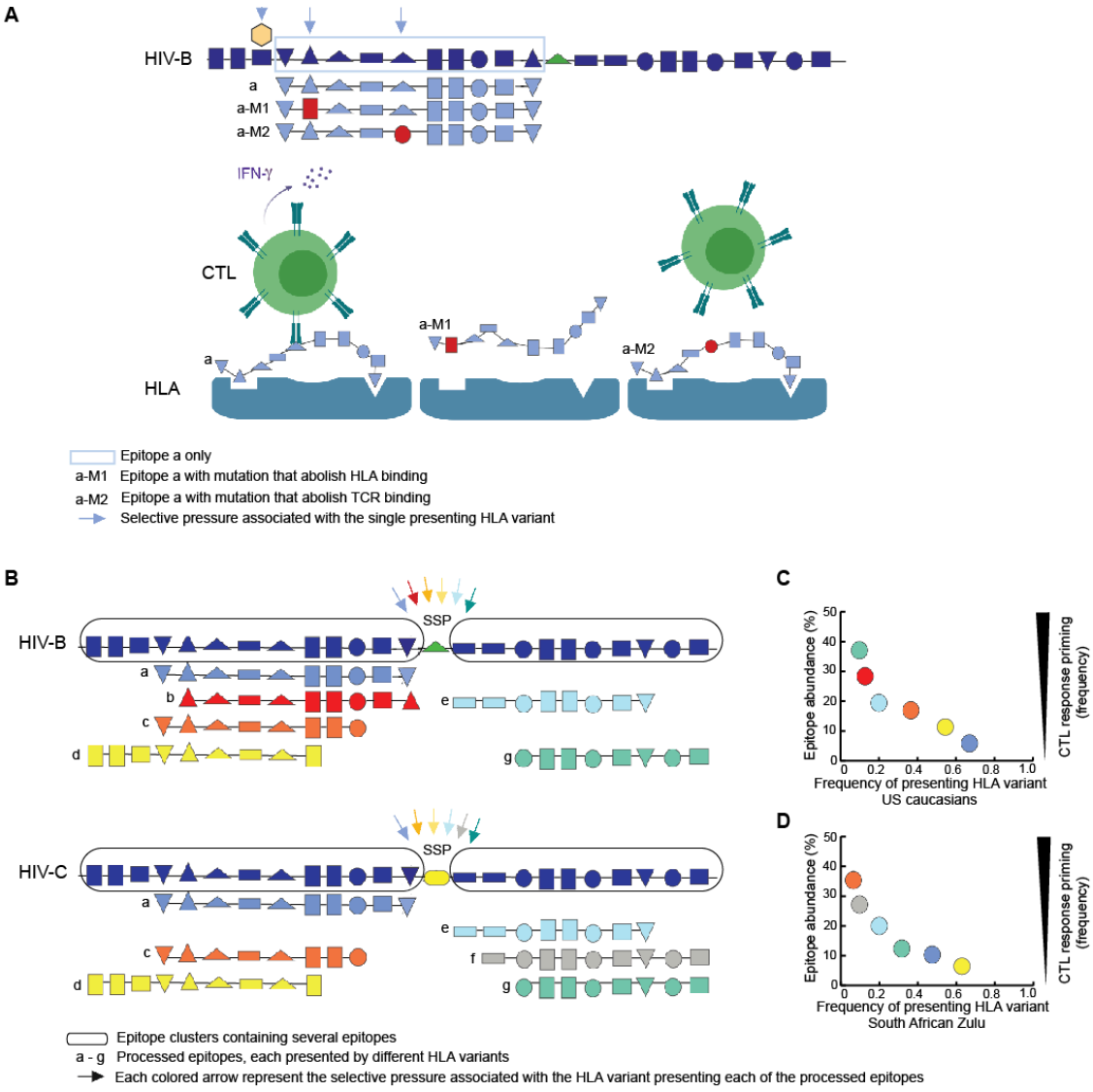

Fig. S2. Outline of HLA-associated single-epitope and multiple-epitope selective pressures.

(A). Three common forms of HLA-associated single epitope selective pressures. CTL responses targeting an epitope might select for an upstream mutation (indicated by the orange polygon) that disrupts ERAP trimming (as in (Draenert et al. 2004)) or intra-epitope mutations that abrogates epitope binding to the presenting HLA molecule or T cell receptor (TCR) recognition of the epitope (reviewed in (Goulder and Watkins 2008; Goulder and Walker 2012)). Arrows signify different forms of CTL selective pressures, geometric symbols indicate amino acids, and the green triangle represents an HIV-B subtype-specific amino acid (as in (**B**)).

(B). Schematic representation of a conserved p24Gag region in HIV-B and HIV-C with a subtype-specific amino acid position (SSP; HIV-B, green triangle, HIV-C, yellow ellipse) and down- and upstream epitope clusters; epitopes a-g indicate epitopes processed by intra-cellular proteasomes. When a patient presents any of these epitopes, additional HLA-associated selective pressures might select for the intra-epitope CTL-escape mutations described in **Figure S2A**. The nature of the amino acid in the subtype-specific position controls intra-cellular proteasomal production of the surrounding epitope-clusters (Tenzer et al. 2014), which each can contain 20-50 epitopes presented by approximately as many HLA variants (Foley et al. 2018). The outlined structure consisting of partly overlapping epitope sequences in clusters in hydrophobic regions that are separated by a subtype-specific position is also common in other HIV-1 proteins (Foley et al. 2018; Lucchiari-Hartz et al. 2003).

(C), (D). Schematic outline of the outcome of the HLA-associated selective pressure on the subtype-specific positions (experimentally demonstrated in (Tenzer et al. 2014)). The combined HLA-associated selective pressure on the SSP in an HIV-infected population results in an inverse relationship between the abundance of a processed epitope and the frequency of the presenting HLA allele in the population in which the virus circulates (Tenzer et al. 2014). In this hypothetical example, the grey epitope is not produced when the proteasome processes HIV-B because the HLA variants that can present the grey epitope is very common in US Caucasians. In contrast, the red epitope is not produced when the proteasome digests HIV-C because the restricting HLA variant is very common in the South African Zulu population. The likelihood of CTL priming will increase with the amount of presented epitope on the infected cell surface (Faroudi et al. 2003; Tenzer et al. 2009).

Figure S3

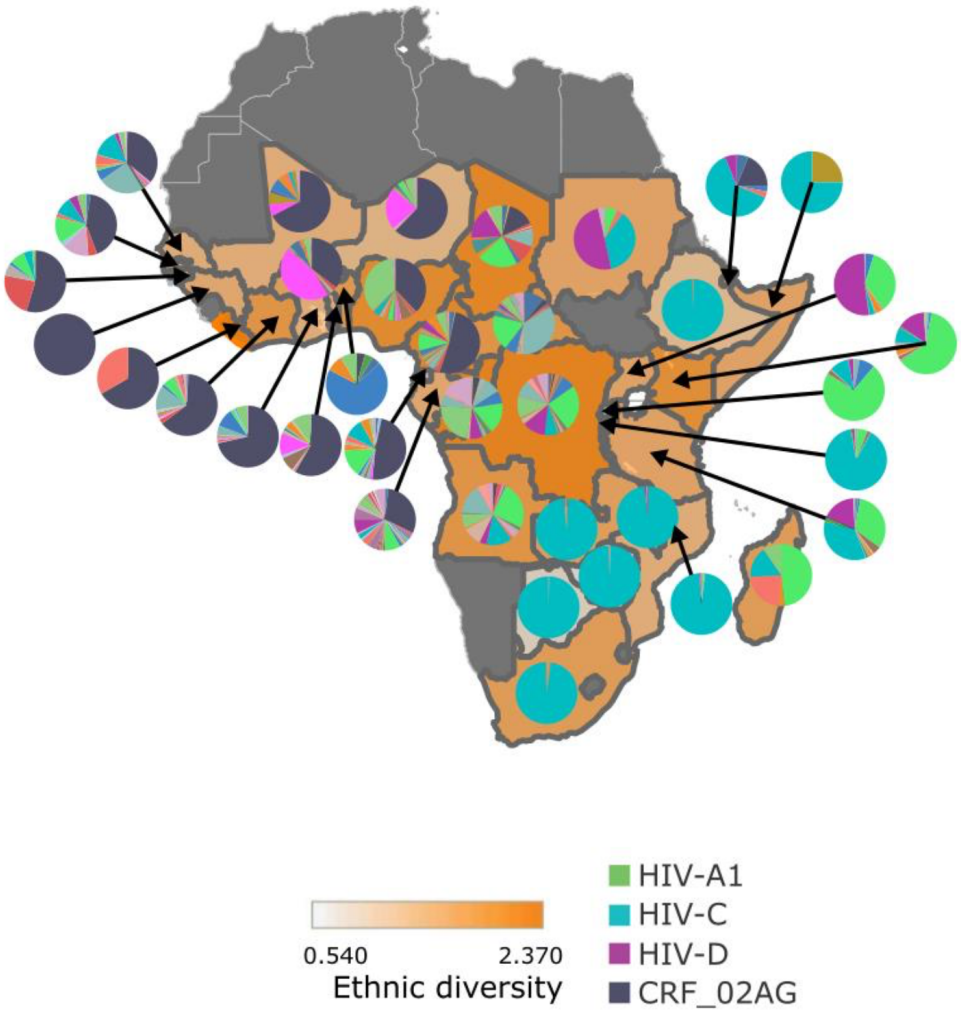

Fig. S3: HIV-1 subtype and CRF distribution in Sub-Saharan Africa

This figure is similar to **Fig. 3d**, but the pie charts have not been scaled according to the number of sequences from each country. The map showing the ethnic diversity within Sub-Saharan Africa (orange shaded background) is overlaid with pie charts demonstrating HIV-1 subtype diversity within each country. Missing countries (dark grey) lacked either HIV-1 or ethnic fractionalization data; complete subtype key can be found in **Table S4**.

Figure S4

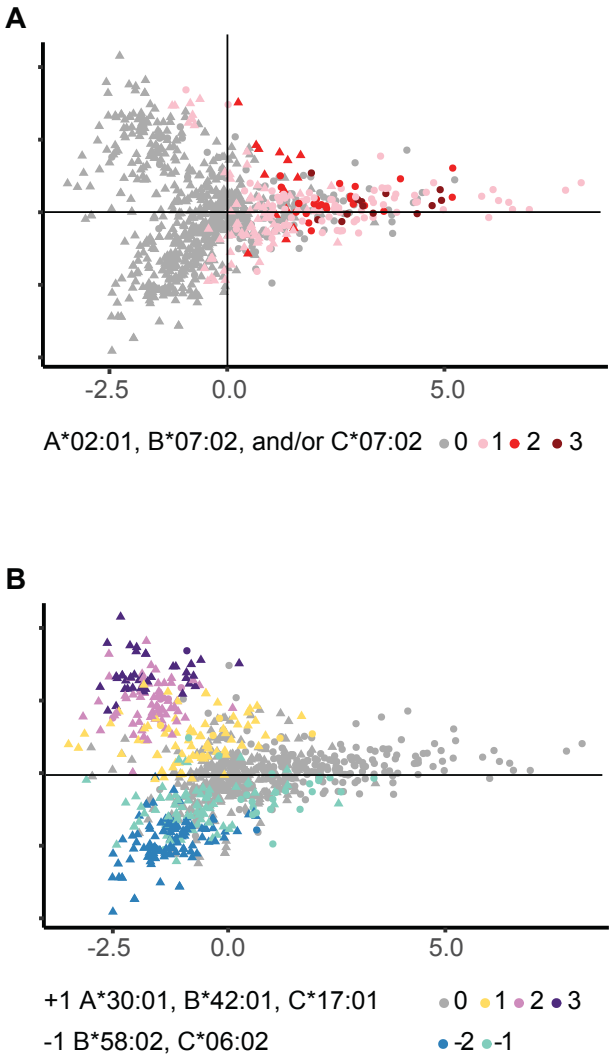

Fig. S4: The patients' haplotype score in PC1 and PC2

(A). The first two PCs explained 14.5% of the variance (PC1 = 7.7%, PC2 = 6.8%). The haplotype score is not stratified for Caucasian HLA variants as the HLA A*02:01 contributes most to the variance. HLA B*07:01 and HLA C*07:01 are in LD in Caucasians (a haplotype inherited from Neanderthals (Abi-Rached et al. 2011)), but not in Africans (Gonzalez-Galarza et al. 2015).

(B). The patient's haplotype score is incremented by one for each allele from haplotype HLA A*30:01-B*42:01-C*17:01 and reduced by one for each allele from haplotype HLA B*58:02-C*06:02 showing near-complete stratification of PC2 due to these haplotypes.

Figure S5

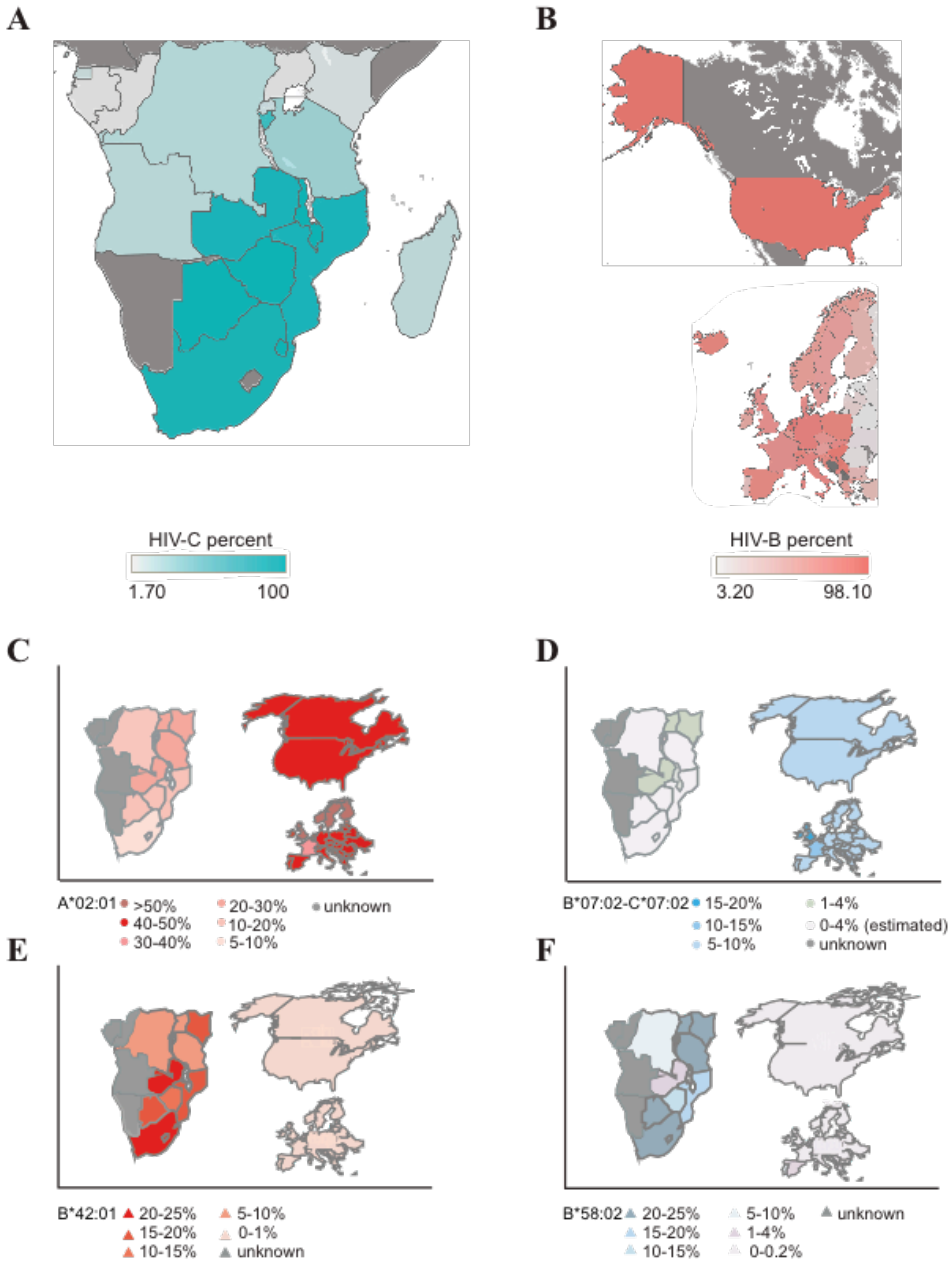

Fig. S5: HIV-C, HIV-B and selected HLA frequencies in Southern Africa, the US, and Europe

(A). The percentage of HIV-C in countries in Southern African. Dark grey signifies either no data or that the region is outside the geographical region of interest.

(B). The percentage of HIV-B in the US and Europe. The percentage of non-HIV-B sequences in Europe is higher than in the US due to a higher proportion of non-B HIV-1-infected immigrants and refugees primarily from Africa. Furthermore, the epidemic in some European countries was in part, or mostly, founded by non-B subtypes; for example, the percentage of HIV-B is low in Russia and former Soviet Union countries due to the introduction of HIV-A from the DRC and HIV-G in Portugal because of the introduction of HIV-G from Cape Verde, a former Portuguese colony (Beloukas et al. 2016; Diez-Fuertes et al. 2015).

(C-F). The HLA frequencies of key HLA variants in PC1 and PC2. Note the low frequency of HLA-B*58:02 in Zambia where the epidemic has lasted longer than in South Africa, and where HLA B*42:01, but not HLA B*58:02, is associated with facilitated HIV-1 transmission. The HLA B*07:02-C*07:02 haplotype was acquired from Neanderthals and is found primarily in Eurasians (Abi-Rached et al. 2011). HLA B*07:02 and HLA C*07:02 can be found individually in other combinations in some African ethnic groups (Gonzalez-Galarza et al. 2015).

Figure S6

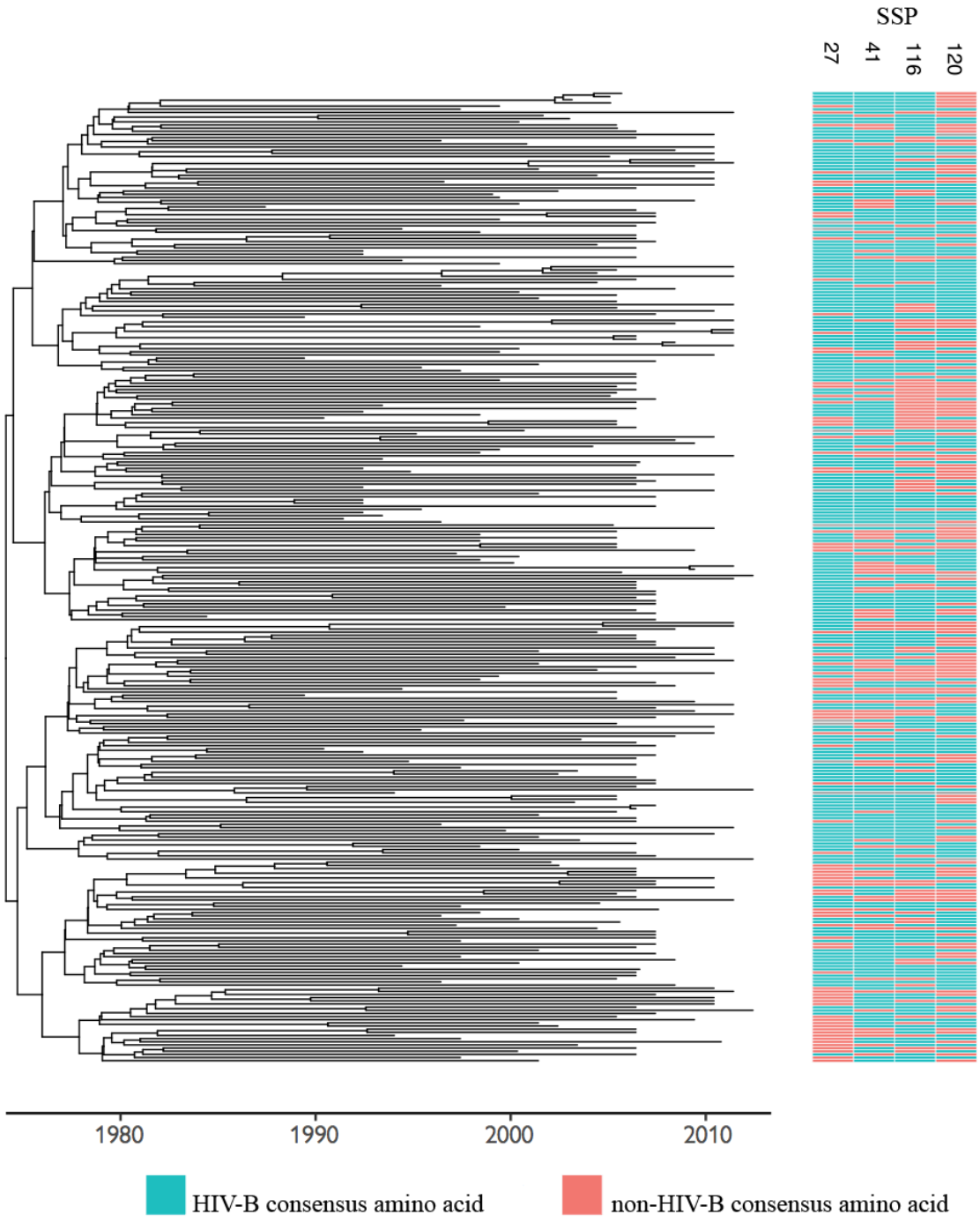

Fig. S6: HIV-B phylogeny

The HIV-B BEAST Maximum Clade Credibility phylogeny annotated with the subtype-specific positions (SSP) 27, 41, 116, and 120 (see **Fig. S1**). Every leaf-node on the phylogeny represents a patient, and the four lines next to the leaf node show whether the majority rule amino acid in each

of the four eligible SSPs was identical to the HIV-B consensus amino acid. The amino acid patterns shown here follow the phylogeny (i.e., continuous patches of red are more vertical than horizontal, and are therefore shared between closely related sequences rather than shared between multiple positions on the same patient), and demonstrate the necessity of using a multiple response random effect model where each amino acid is allowed to evolve independently. If the color pattern had been more horizontal, that would have meant a single patient's subtype-specific amino acids tended to change as a block, in which case a single patient parameter (random effect model) would have been more appropriate than the multiple random effect phylogenetic model used in this study.

Figure S7

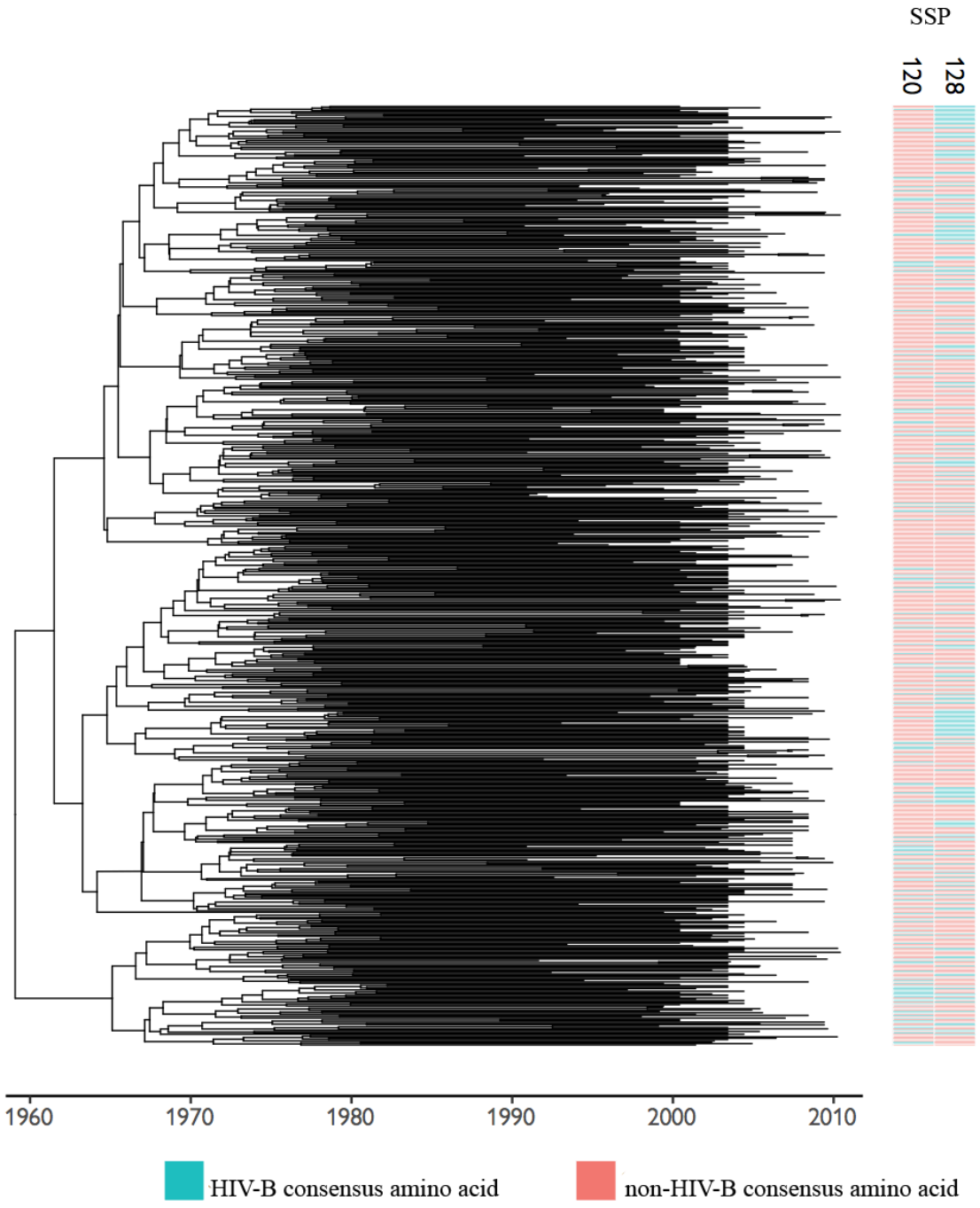

Fig. S7: HIV-C phylogeny

The HIV-C BEAST Maximum Clade Credibility phylogeny annotated with the subtype-specific positions (SSP) 120 and 128. Patients, where the majority rule amino acid is the same as the HIV-B (see fig. S1), are colored blue; all other amino acids are colored red.

Figure S8

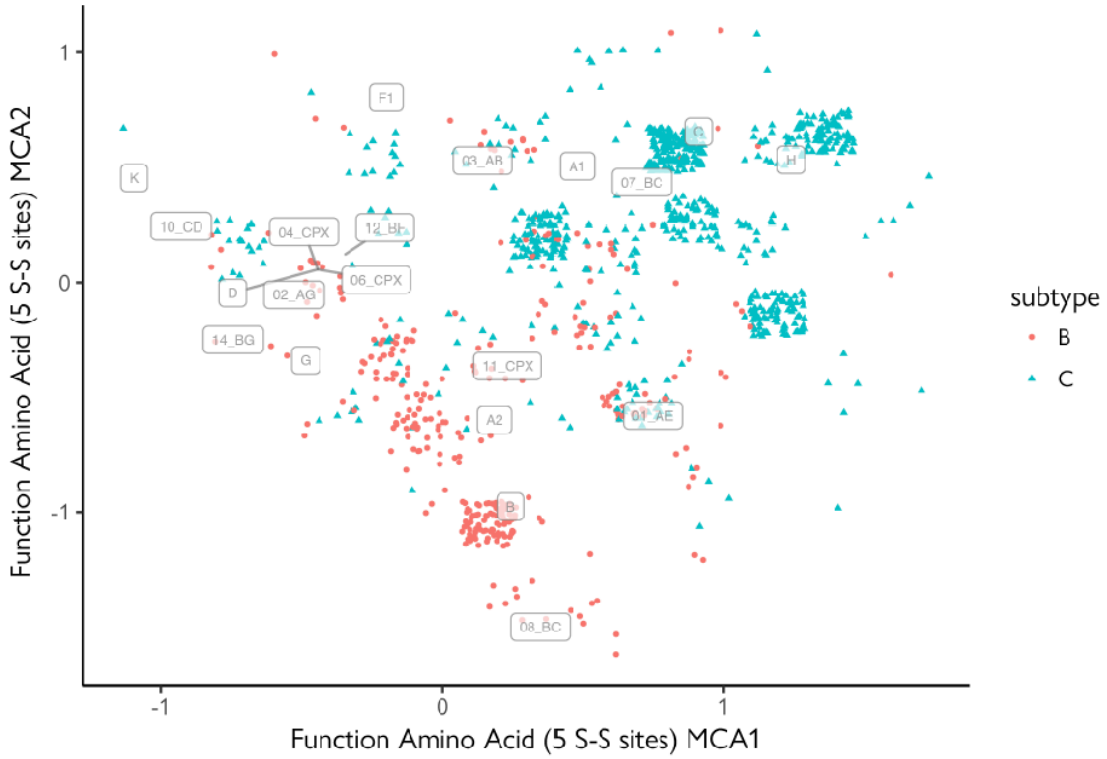

Fig. S8: MCA of each patient's HIV-1 subtype-specific amino acids

The MCA converts each patient's HIV subtype-specific position amino acid profile into two dimensions. This MCA was trained using the subtype consensus sequences (labels), the MCA was then used to project the patient's HIV subtype-specific amino acids onto the two-dimensional space. Points were jittered to prevent them obscuring one another. While many patients' subtype-specific amino acids are identical to the subtype consensus (see large clusters around the HIV-B and HIV-C labeled consensus sequences), other HIV-B and HIV-C sequences are identical in the five Gag subtype-specific sites studied here (e.g., around the labeled HIV-1 circulating recombinant (CRF) 01_AE consensus).

Figure S9

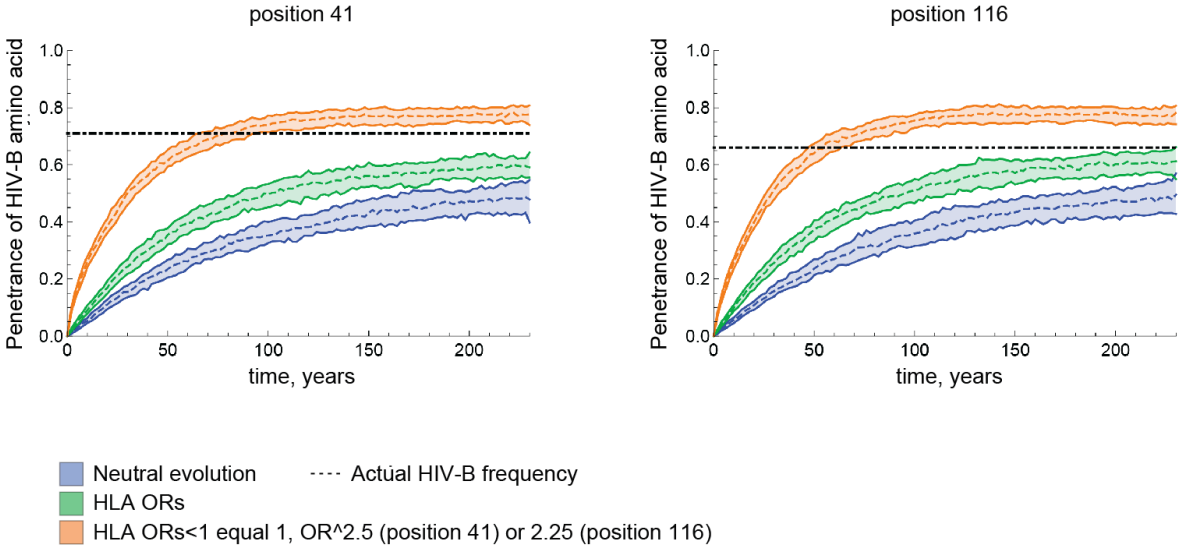

Fig. S9: Modeling of the evolution of HIV-C SSAA using HIV-B-related HLA ORs

We used an agent-based model (**1.3**) to estimate the effect of neutral and HLA-mediated selection pressures, respectively, on the evolution of p24Gag positions 41, and 116 on a fictive HIV-1 with HIV-C-like subtype-specific amino acids (SSAAs). For position 41, raising the odds to 2.9 or 3.0 made little difference. These results suggest that other, unidentified, HLA variants might influence the evolution of these positions and/or that structural co-evolutionary factors might be in play.

Figure S10

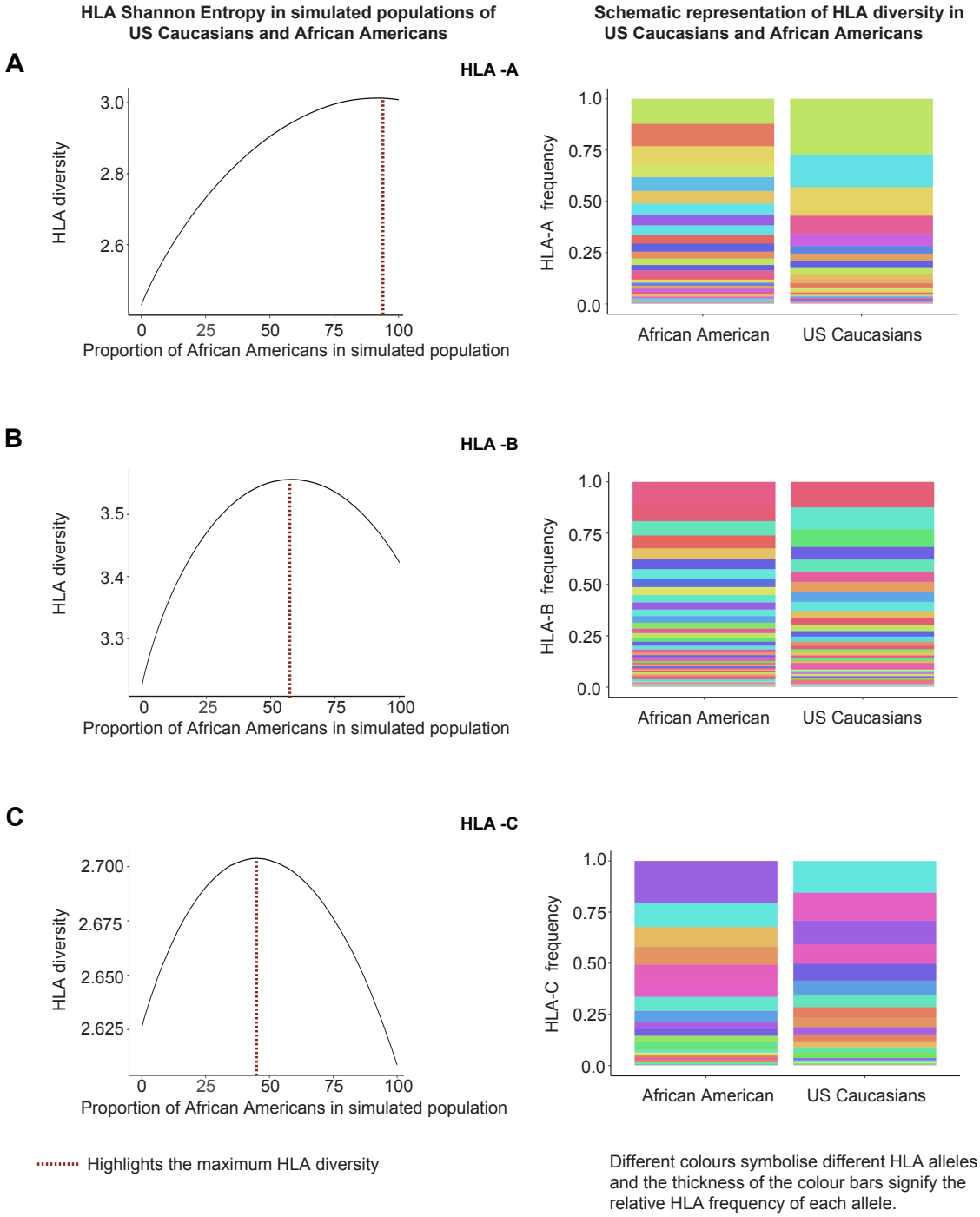

Fig. S10: Simulated mixtures of African Americans and US Caucasians (left) and HLA diversity in African American and US Caucasian populations (right)

(A). HLA A diversity in simulated populations with different proportions of African Americans and US Caucasians, and the HLA A variant frequencies in African Americans and US Caucasians (modified from (Gragert et al. 2013)).

(B, C). As in **A** for **B** (HLA B), and **C** (HLA C).

Figure S11

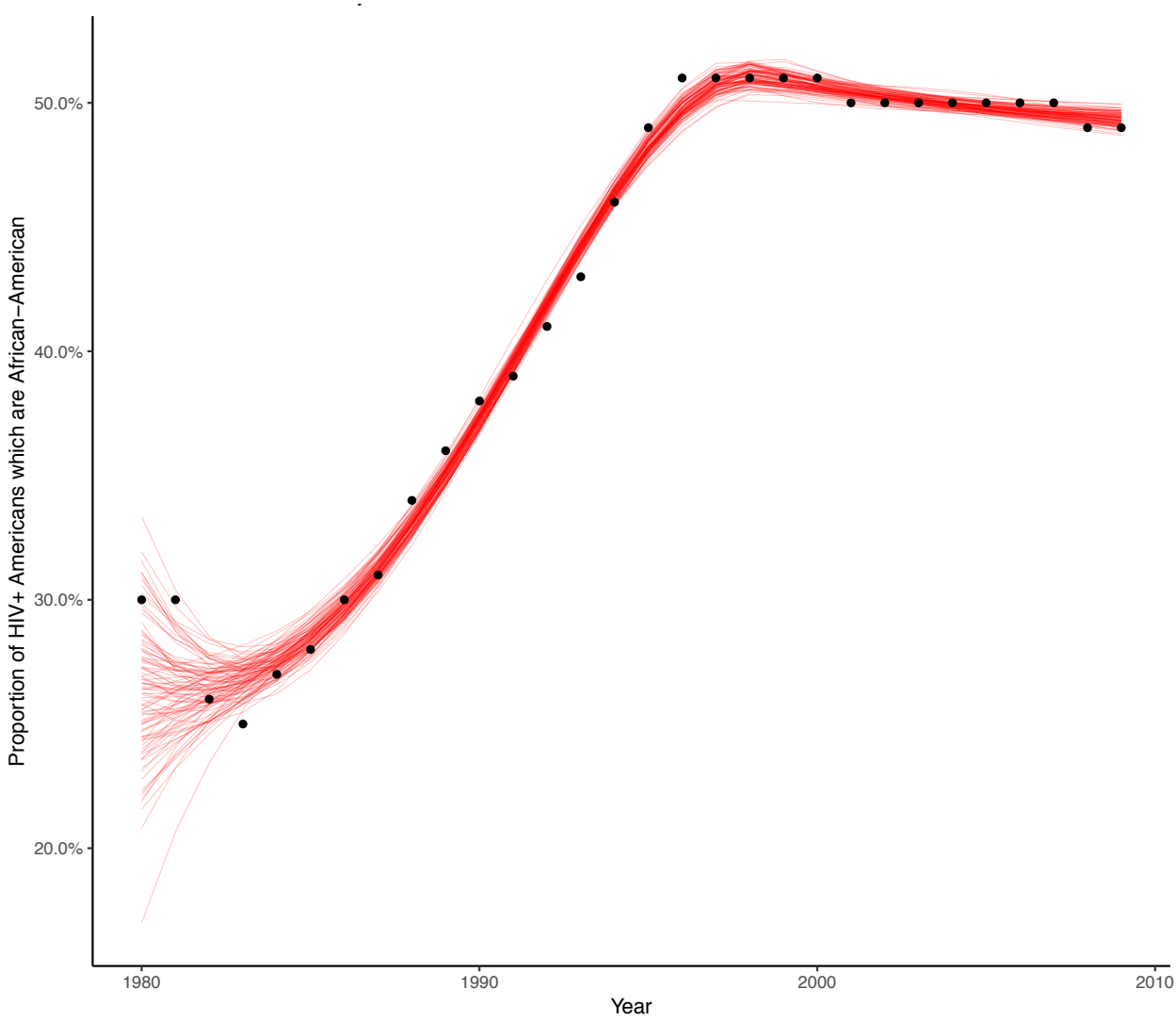

Fig. S11: **Proportion of African Americans in the US HIV-infected population over time**

The ethnic demographic proportion of African Americans in the US HIV-infected population (data from (Hall et al. 2008) and (Prejean et al. 2011)) as calculated in the simple population model with the proportion of African Americans shown as points. Samples drawn from the posterior distribution of the change-point model are shown as lines. Note the increased uncertainty around 1980 due to data limitations.

Figure S12

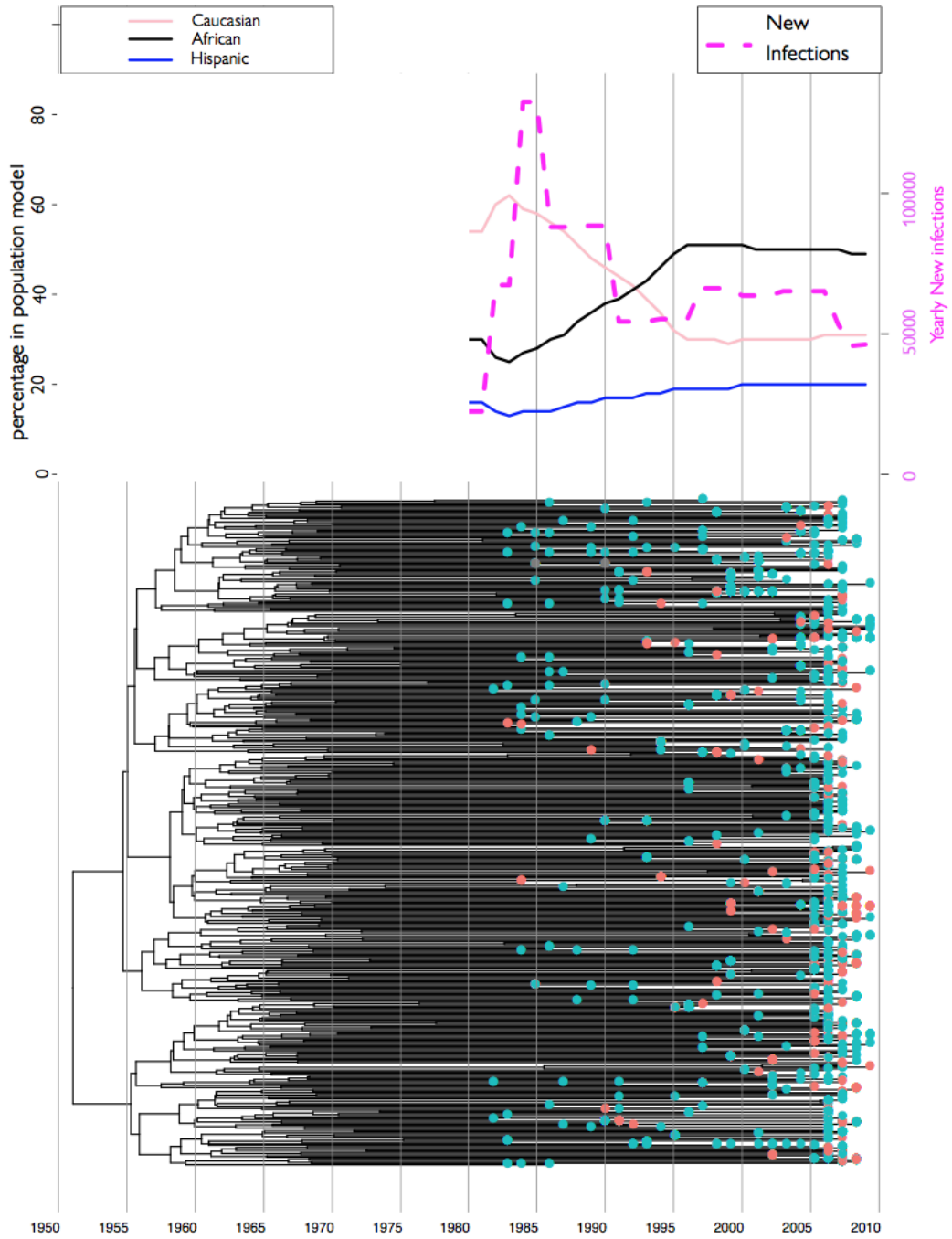

Fig. S12: Phylogeny of the American HIV-B epidemic combined with demographic data

The Maximum Clade Credibility tree, annotated with individual sequence data for position 27 (bottom, blue if a sequence is identical to the HIV-B consensus amino acid; red otherwise). In this case all non-B consensus amino acids were similar to the amino acid found in the HIV-C

consensus sequence due to the limited variation possible at this position (**Table S8**). The phylogeny was generated using one sequence per patient, but position 27 amino acid data from multiple sequences sampled from the same patient have been superimposed onto the phylogeny. These annotations illustrate the gradual increase of non-HIV-B consensus amino acids on position 27 as the African-American proportion of HIV-infected individuals in the United States increases (black line, top figure). No demographic data are available before 1980.

Table S1: African populations with HLA A, B, and C data

Table showing the only African populations with publicly available 'gold standard' HLA data (i.e., all HLA types were reported as 4-digits HLA variants)(Gonzalez-Galarza et al. 2015).

Number	Population or ethno-linguistic group	Country	Number of people in cohort
1	Bandiagara	Mali	138
2	Ga-Adangbe	Ghana	131
3	Kenya* (likely Masai)	Kenya	144
4	Luo	Kenya	265
5	Nandi	Kenya	240
6	South Africa Black**	South Africa (Free State)	200
7	Zulu	South Africa (KwaZulu Natal)	606
8	Kampala (pop 2)***	Uganda	175

* The population named 'Kenya' has no ethnic group specifications but the location coordinates (obtained from the database (Gonzalez-Galarza et al. 2015)) are situated within the Masai ethnic group's territory, and the population is described as 'rural.' These coordinates differ from those of the Luo and Nandi ethnic groups, and for this reason, we assumed all three groups were distinct from each other. The data have been presented in a workshop but have not been published. The Luo people in western Kenya are closely related to Luo people in Northern Uganda, and Northern Tanzania, and are part of a larger group of ethno-linguistically-related Luo people in South Sudan and South-western Ethiopia. Their language (Dholuo) belongs to the Western Nilotic branch of the Nilo-Saharan language family. The Nandi people live in Nandi Country (previously Rift Valley Province of Kenya) and speak Nandi, which belongs to the Southern Nilotic language group. The Masai people live in Northern, Central, and Southern Kenya and in Northern Tanzania. Their language (Maasai or Maa) belongs to the Eastern Nilotic language group.

** The location coordinates of the South Africa Black HLA data is situated within the Free State province of South Africa, a region where most people speak Sotho and belong to the Sotho-

Tswana people, however, in the paper by Paximadis, M et al., the HLA data is reported to represent “a cross-section of the black and Caucasian subgroups of the South African population (Paximadis et al. 2012).” The South Africa Black and Caucasian HLA data are recorded separately in the database. The number of South Africa Black ethnic groups is ~10, but the majority of the Zulu population lives in the KwaZulu Natal and Gauteng provinces (http://www.sahistory.org.za/).

*** The HLA data from Uganda does not belong to a distinct population; “All the individuals were Black Africans from Kampala, but no other ethnic affiliation was recorded” (from (Kijak et al. 2009)).”

Table S2: The number of p24 Gag sequences per country and the number of HIV-1 subtypes, CRFs and URFs per country and the country code abbreviation key

ISO alpha-2 two-letter Country code	Country name	p24 sequences / country
BW	Botswana	51
CD	Democratic Republic of the Congo	22
CI	Ivory Coast	6
CM	Cameroon	318
ET	Ethiopia	35
GA	Gabon	6
GH	Ghana	13
KE	Kenya	633
NE	Niger	72
NG	Nigeria	19
SN	Senegal	27
TD	Chad	18
TZ	United Republic of Tanzania	38
ZA	South Africa	935
ZM	Zambia	12

ISO alpha-2 two-letter Country code	Country name	HIV-1 subtypes, CRFs, and URFs (group M) / country
AO	Angola	21
BF	Burkina Faso	13
BI	Burundi	8
BJ	Benin	6
BW	Botswana	7
CD	Democratic Republic of the Congo	46
CF	Central African Republic	30
CG	Congo	20
CI	Ivory Coast	22
CM	Cameroon	78
DJ	Djibouti	6
ET	Ethiopia	7
GA	Gabon	33
GH	Ghana	18
GM	Republic of the Gambia	10
GN	Guinea	1

GQ	Equatorial Guinea	16
GW	Guinea Bissau	7
KE	Kenya	35
LR	Liberia	2
MG	Madagascar	5
ML	Mali	10
MW	Malawi	8
MZ	Mozambique	8
NE	Niger	13
NG	Nigeria	23
RW	Rwanda	10
SD	Sudan	5
SN	Senegal	35
SO	Somalia	2
TD	Chad	19
TG	Togo	21
TZ	United Republic of Tanzania	19
UG	Uganda	33
ZA	South Africa	30
ZM	Zambia	14
ZW	Zimbabwe	7

Data were obtained from the HIV data base (Foley et al. 2018).

Table S3: Complete key to the African ethnicities in Fig. 2C

Table S4: Complete key over HIV-1 subtypes shown in Fig. 2D

HIV-B and HIV-C are indicated by stars.

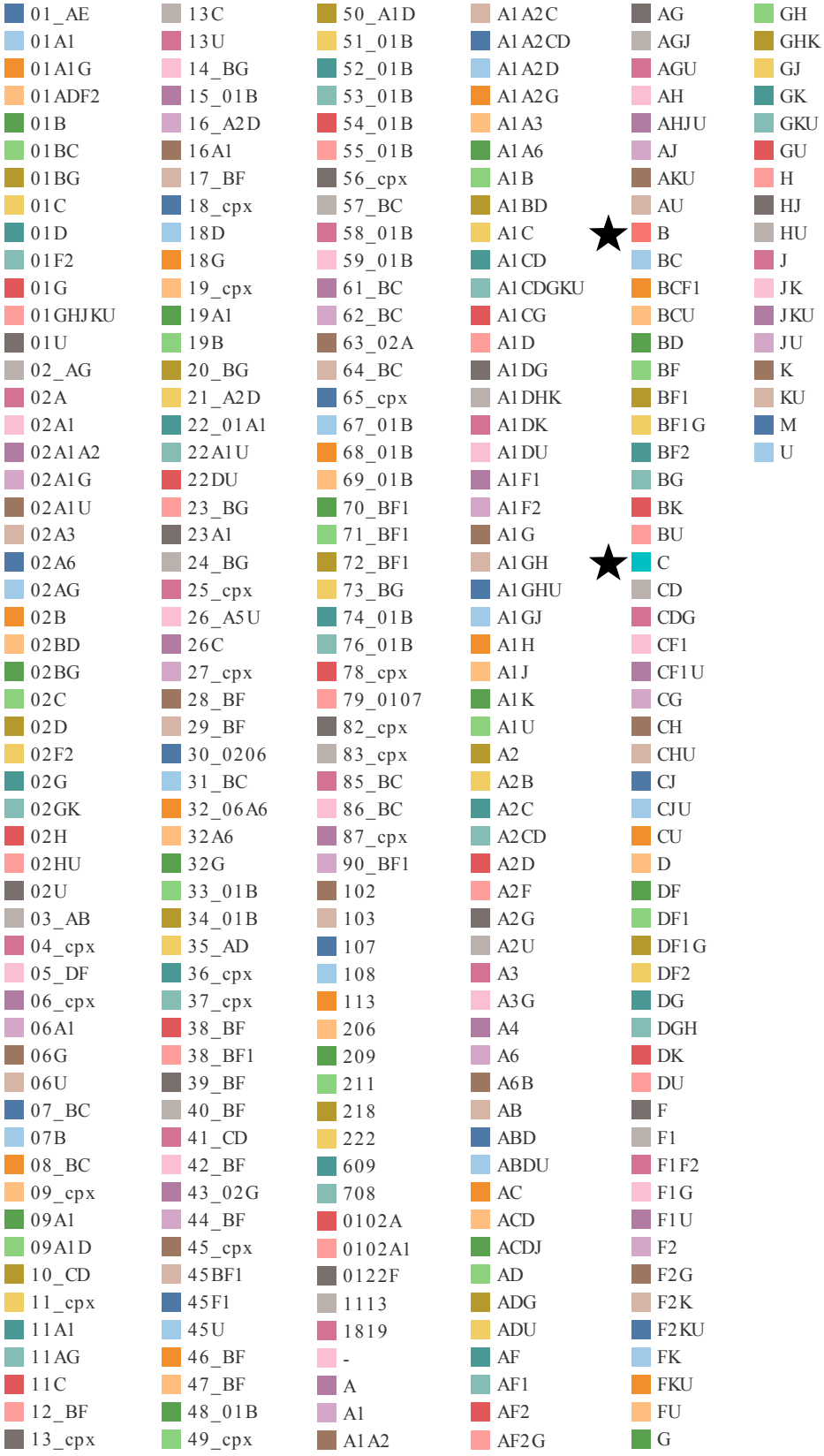

Table S5: Complete key over African languages shown in Fig. 2E

Bantu languages are indicated by stars.

Table S6: HLA class I allele frequencies in worldwide populations

	ZA Blacks	ZA Zulu	US African	US Caribbean	Thailand	US Chinese	US Caucasian	UK central
A*02:01	0.083	0	0.123	0.111	0.255	0.0946	0.2755	0.2697
B*07:02	0.046	0.07	0.073	0.073	0.038	0.0079	0.1306	0.1391
C*07:02	0.058	0.08	0.073	0.067	0	0.1945	0.1415	0.155
A*30:01	0.101	0.1	0.068	0.072	0.013	0.0274	0.013	0.011
B*42:01	0.089	0.11	0.053	0.053	0	0.0001	0.0003	0.0019
C*17:01	0.111	0.05	0.068	0.068	0	0.0005	0.0088	0.004
B*58:02	0.094	0.11	0.042	0.032		0	0.0001	0
C*06:02	0.149	0.15	0.087	0.072	0	0.0447	0.0932	0.11
Population	ZA Blacks	ZA Zulu	US African	US Caribbean	Thailand	US Chinese	US Caucasian	UK central
Sum of HLA variants common in Caucasian	0.187	0.15	0.269	0.251	0.293	0.297	0.5476	0.5638
Sum of HLA variants common in Africans	0.544	0.52	0.318	0.297	0.013	0.0727	0.1154	0.1269

The individual HLA alleles frequencies of HLA variants that contributed most to the first two principal components (PCs) were also found in common African and Caucasian haplotypes. Here we show the individual frequencies of each HLA variant and the sum of the variants that can form haplotypes in Africans and Caucasians. The data derive from (Gonzalez-Galarza et al. 2015), and the South Africa (ZA) Zulu data derive from the Females Rising through Education, Support, and Health (FRESH) project.

Table S7: Overview of the amino acid variation at SSP in HIV-1 subtype consensus sequences

HIV subtype consensus	Position 27 amino acid	Position 41	Position 116	Position 120	Position 128
HIV-A1	I	S	G	G	D
HIV-A2	V	T	G	S	E
HIV-B	V	S	G	N	E
HIV-C	I	T	A	S	D
HIV-D	I	S	G	S	E
HIV-F1*	I	S	Q	S	D
HIV-G	V	S	R	S	E
HIV-H*	V	S	A	G	D
HIV-K*	I	S	T	S	E

(From www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html, (Foley et al. 2018))

*HIV-F1 predominates in South America, especially in Brazil, Uruguay, and Argentina, HIV-H is found in Central Africa, Eastern Europe, and Central Asia, and HIV-K is found in Central Africa and Pakistan (Khan et al. 2018) at low frequency.

The subtype distribution of HIV-F1, HIV-H and HIV-K outside of Africa – and the presence of, e.g., HIV-C in India and China - is due to founder effects. The evolution of these subtypes over time in non-African countries is unknown due to limited sampling and the brevity of the ongoing HIV epidemic (Buonaguro et al. 2007; Hemelaar et al. 2011; Osmanov et al. 2002). Overall, HIV-F (0.59%), HIV-H (0.17%), HIV-J (0.14%), and HIV-K (0.04%) together cause fewer than 1% of HIV-1 infections worldwide (Buonaguro et al. 2007; Hemelaar et al. 2011). No consensus sequences for HIV-J p24Gag are available from the Los Alamos HIV sequence database (Foley et al. 2018).

Table S8: LANL HIV database patient IDs of patients with imputed four digit HLAs

B*27 → B*27:05:

402, 3393, 3394, 10542, 11222, 13225, 15792, 19252, 19542, 19878, 22972, 22973, 22974, 22975, 22982, 22984, 22986, 23077, 27220, 34369, 36112, 36113, 49996

B*35 → B*35:01:

24029, 30977, 35886, 36113, 49994, 51971

B*57 → B*57:01:

10487, 10488, 10489, 10490, 10491, 10492, 10493, 10494, 10495, 10496, 10497, 10498, 10499, 10500, 10501, 10502, 10503, 10504, 10505, 10506, 10507, 10508, 10509, 10510, 11222, 13224, 15396, 15397, 19539

References

Abi-Rached, L., et al. (2011), 'The shaping of modern human immune systems by multiregional admixture with archaic humans', *Science,* 334 (6052), 89-94.

Beloukas, A., et al. (2016), 'Molecular epidemiology of HIV-1 infection in Europe: An overview', *Infect Genet Evol,* 46, 180-89.

Buonaguro, L., Tornesello, M. L., and Buonaguro, F. M. (2007), 'Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications', *J Virol,* 81 (19), 10209-19.

Diez-Fuertes, F., Cabello, M., and Thomson, M. M. (2015), 'Bayesian phylogeographic analyses clarify the origin of the HIV-1 subtype A variant circulating in former Soviet Union's countries', *Infect Genet Evol,* 33, 197-205.

Draenert, R., et al. (2004), 'Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection', *J Exp Med,* 199 (7), 905-15.

Faroudi, M., et al. (2003), 'Lytic versus stimulatory synapse in cytotoxic T lymphocyte/target cell interaction: manifestation of a dual activation threshold', *Proc Natl Acad Sci U S A,* 100 (24), 14145-50.

Foley, Brian, et al. (2018), *HIV Sequence Compendium 2018* (Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, New Mexico 87545 U.S.A.LA-UR-18-25673).

Gonzalez-Galarza, F. F., et al. (2015), 'Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations', *Nucleic Acids Res,* 43 (Database issue), D784-8.

Goulder, P. J. and Watkins, D. I. (2008), 'Impact of MHC class I diversity on immune control of immunodeficiency virus replication', *Nat Rev Immunol,* 8 (8), 619-30.

Goulder, P. J. and Walker, B. D. (2012), 'HIV and HLA class I: an evolving relationship', *Immunity,* 37 (3), 426-40.

Gragert, L., et al. (2013), 'Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry', *Hum Immunol,* 74 (10), 1313-20.

Hall, H. I., et al. (2008), 'Estimation of HIV incidence in the United States', *JAMA,* 300 (5), 520-9.

Hemelaar, J., et al. (2011), 'Global trends in molecular epidemiology of HIV-1 during 2000-2007', *AIDS,* 25 (5), 679-89.

Khan, S., et al. (2018), 'HIV-1 genetic diversity, geographical linkages and antiretroviral drug resistance among individuals from Pakistan', *Arch Virol,* 163 (1), 33-40.

Kijak, G. H., et al. (2009), 'HLA class I allele and haplotype diversity in Ugandans supports the presence of a major east African genetic cluster', *Tissue Antigens,* 73 (3), 262-9.

Lucchiari-Hartz, M., et al. (2003), 'Differential proteasomal processing of hydrophobic and hydrophilic protein regions: contribution to cytotoxic T lymphocyte epitope clustering in HIV-1-Nef', *Proc Natl Acad Sci U S A,* 100 (13), 7755-60.

Osmanov, S., et al. (2002), 'Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000', *J Acquir Immune Defic Syndr,* 29 (2), 184-90.

Paximadis, M., et al. (2012), 'Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population', *Hum Immunol,* 73 (1), 80-92.

Prejean, J., et al. (2011), 'Estimated HIV incidence in the United States, 2006-2009', *PLoS One,* 6 (8), e17502.

Tenzer, S., et al. (2014), 'HIV-1 adaptation to antigen processing results in population-level immune evasion and affects subtype diversification', *Cell Rep,* 7 (2), 448-63.
Tenzer, S., et al. (2009), 'Antigen processing influences HIV-specific cytotoxic T lymphocyte immunodominance', *Nat Immunol,* 10 (6), 636-46.