

## 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**

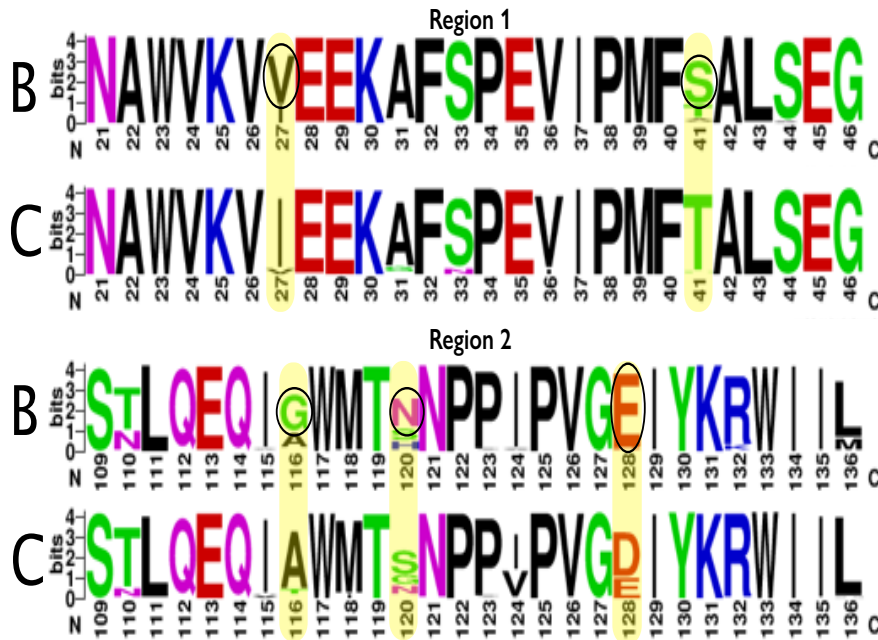
**Table S5:** Complete key over African languages shown in **Fig. 3E**

**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

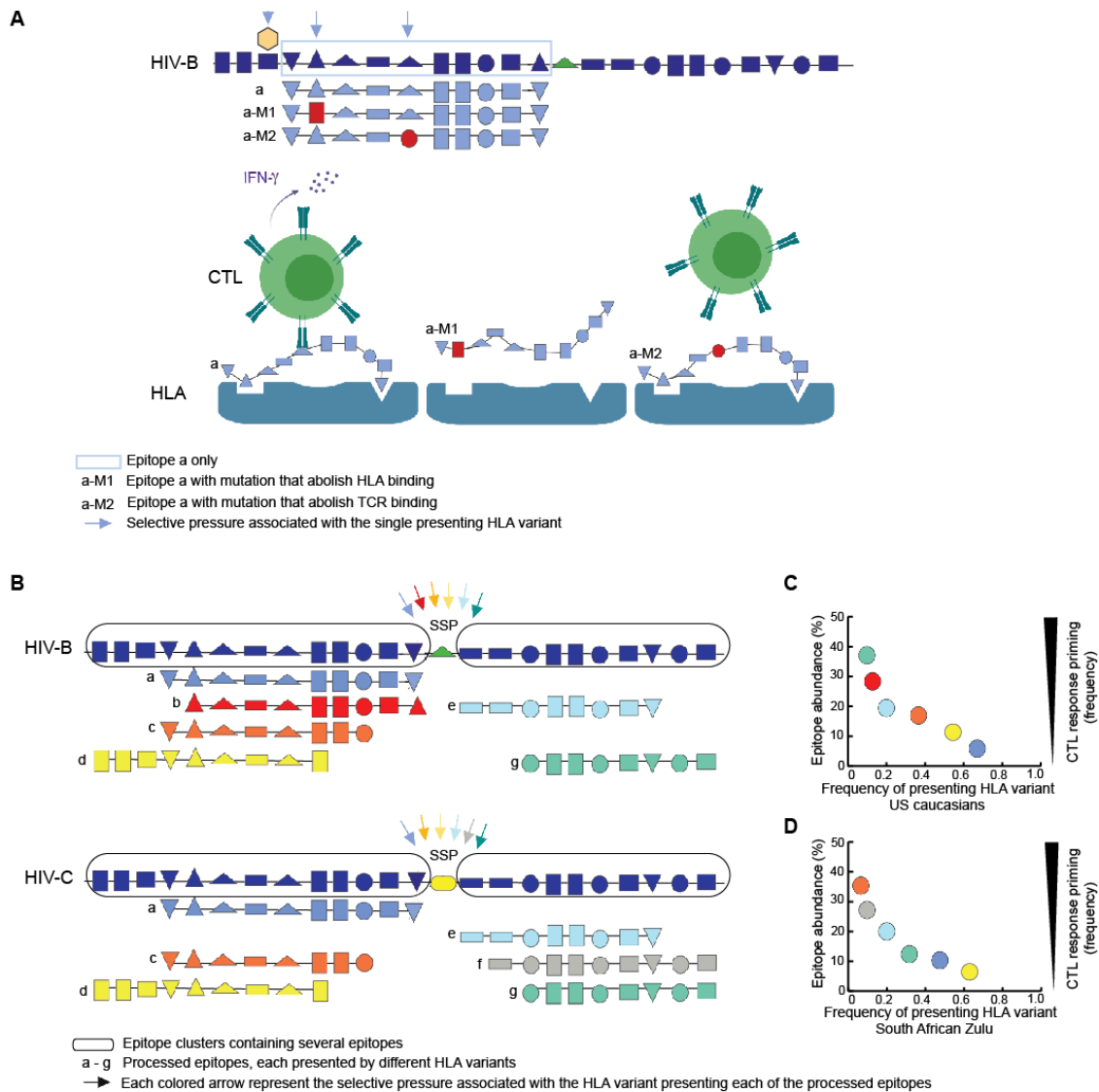
**Table S8:** LANL patient IDs of patients with imputed four digit HLAs

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).

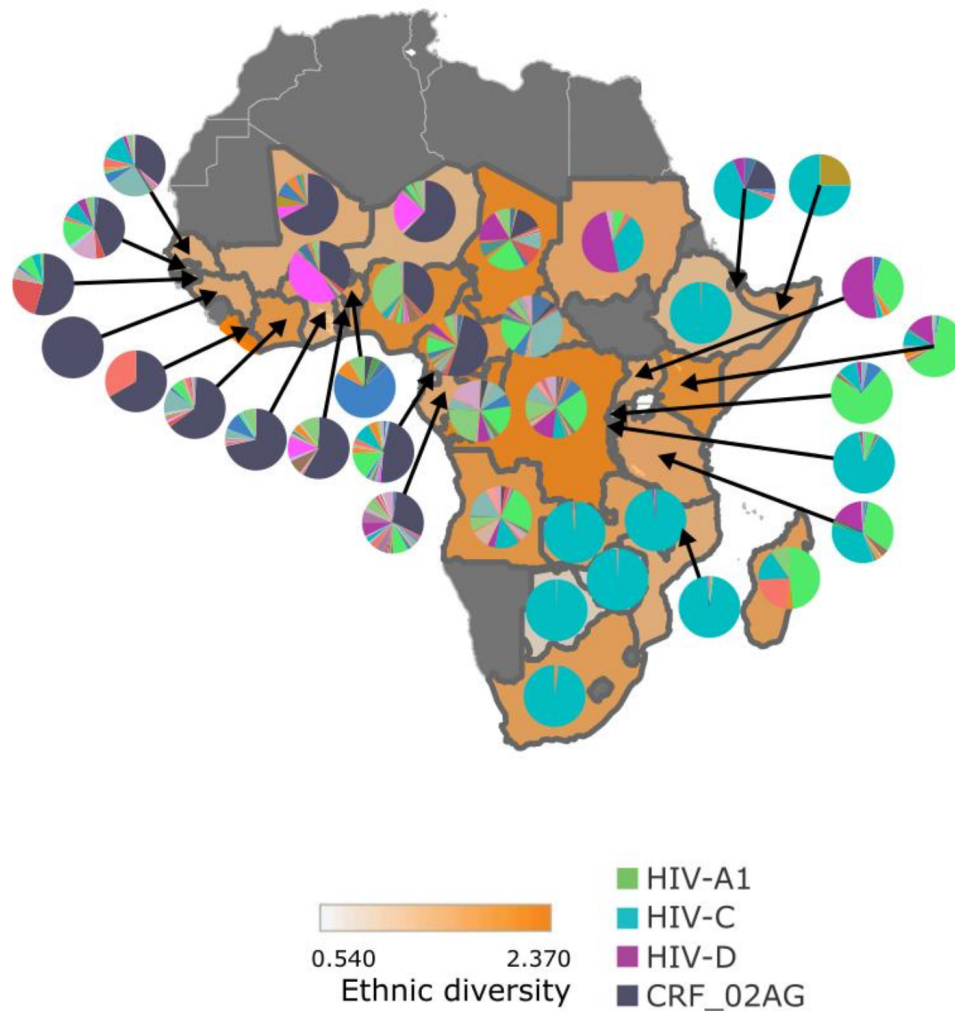
**Figure S2****Fig. S2. Outline of HLA-associated single-epitope and multiple-epitope selective pressures.**

(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).

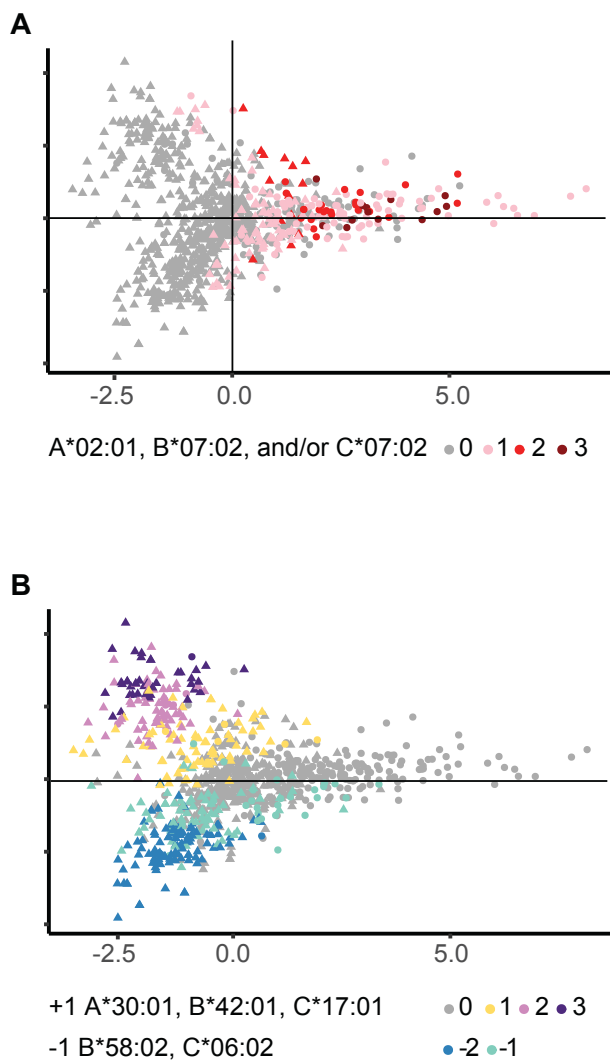
**(C), (D).** Schematic outline of the outcome of the HLA-associated selective pressure on the subtype-specific positions (experimentally demonstrated in (Tenzer et al. 2014)). The combined HLA-associated selective pressure on the SSP in an HIV-infected population results in an inverse relationship between 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**.

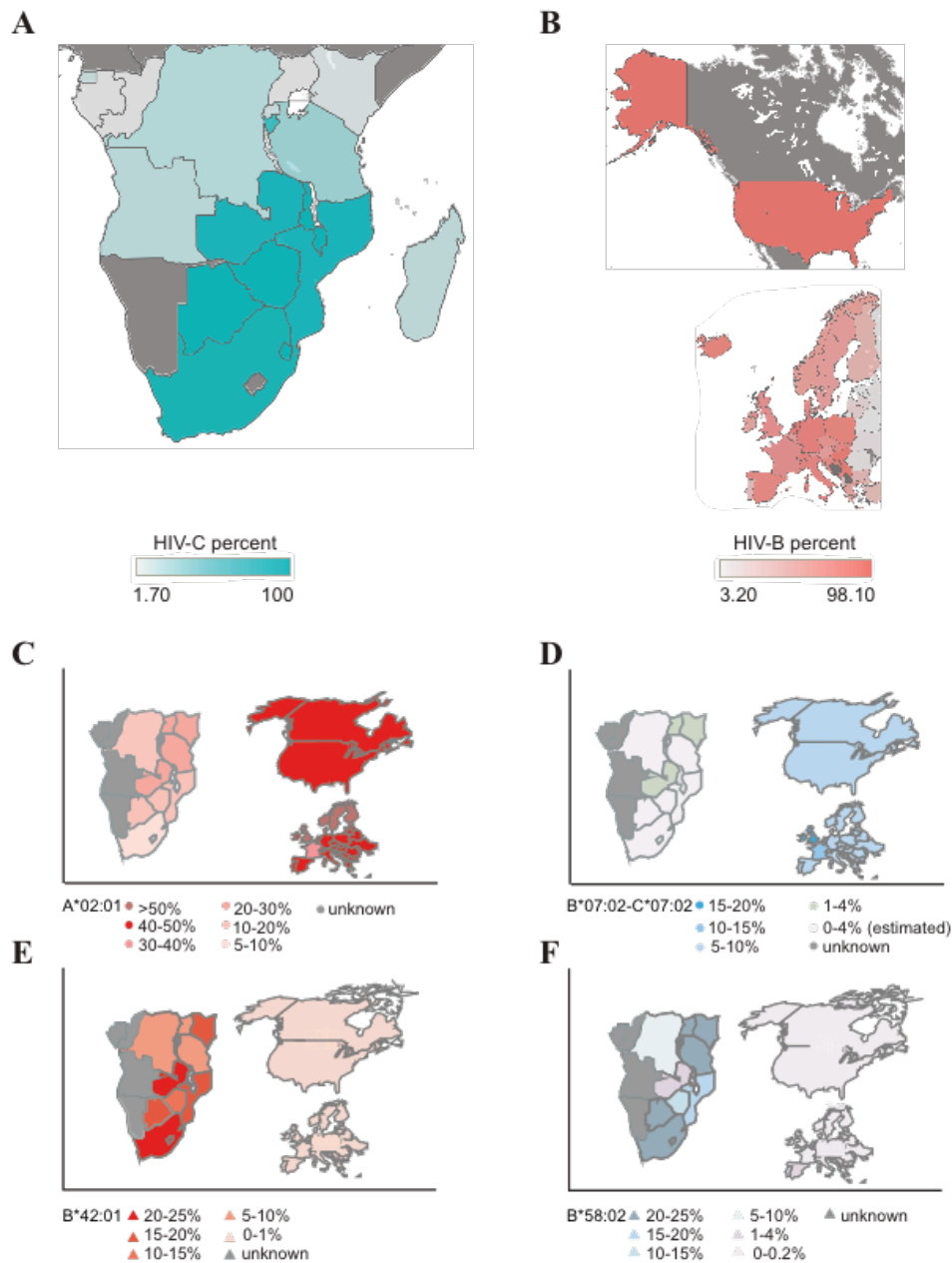
**Figure S4****Fig. S4: The patients' haplotype score in PC1 and PC2**

(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**



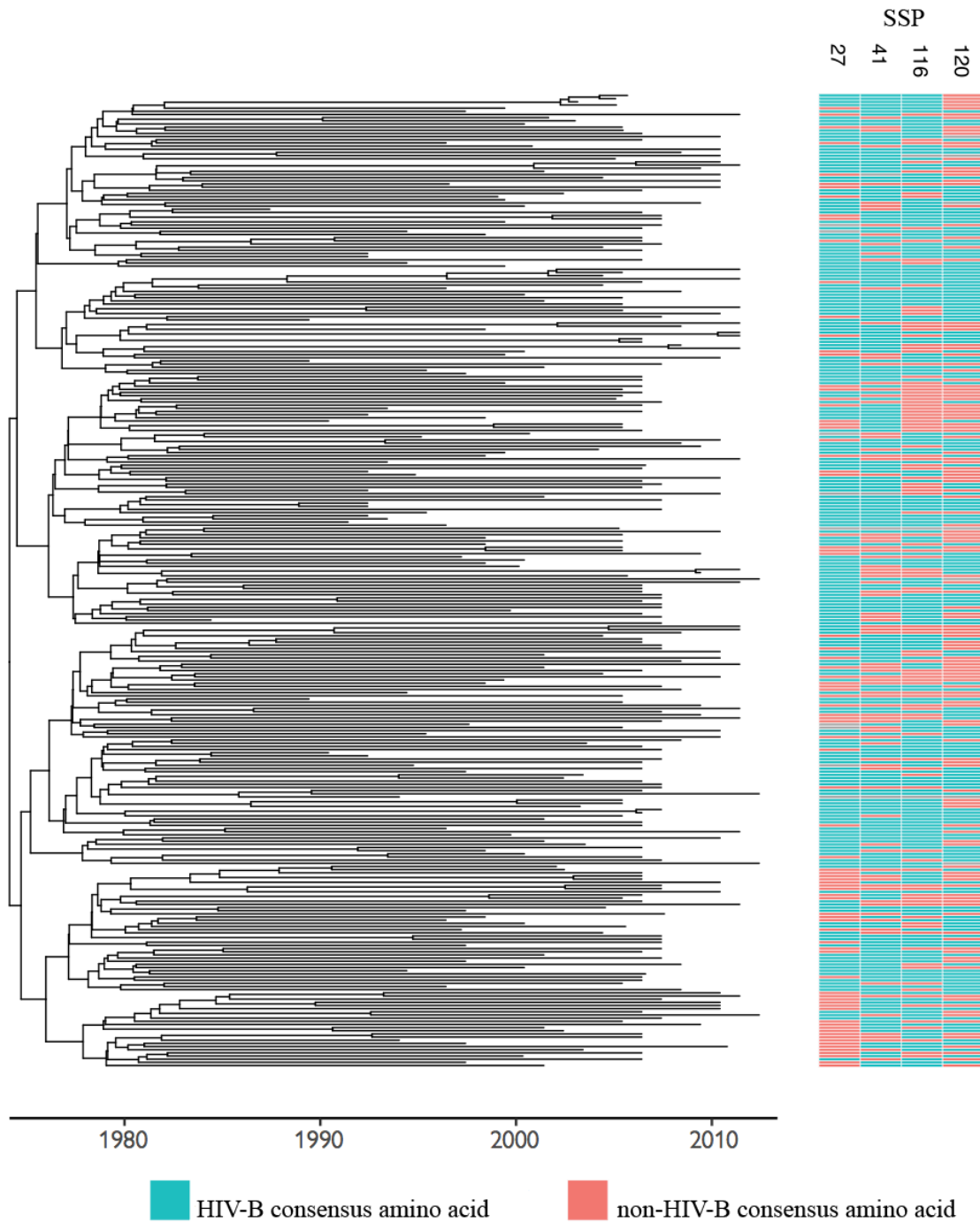
**Fig. S5: HIV-C, HIV-B and selected HLA frequencies in Southern Africa, the US, and Europe**

(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

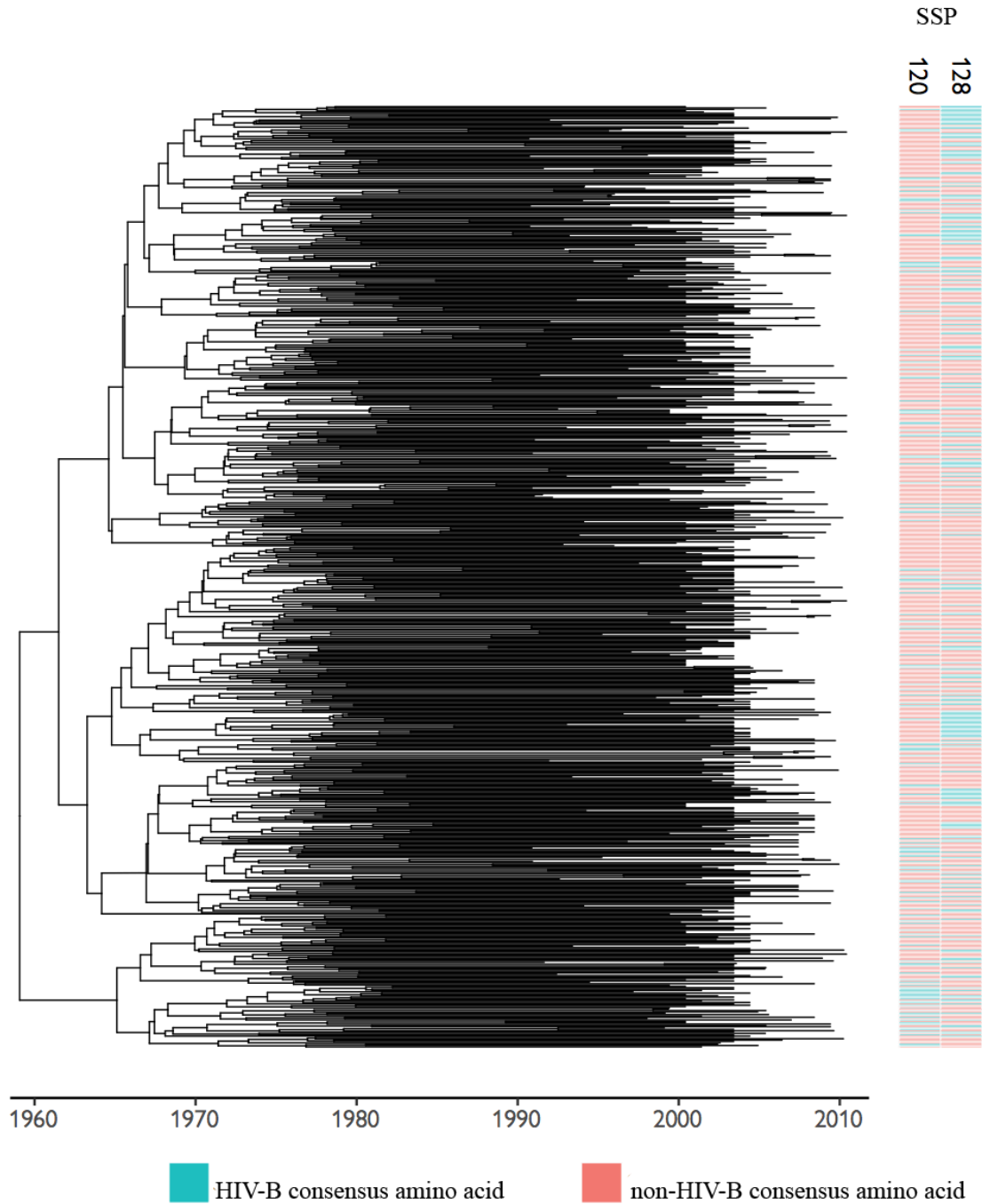


**Fig. S6: HIV-B phylogeny**

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.

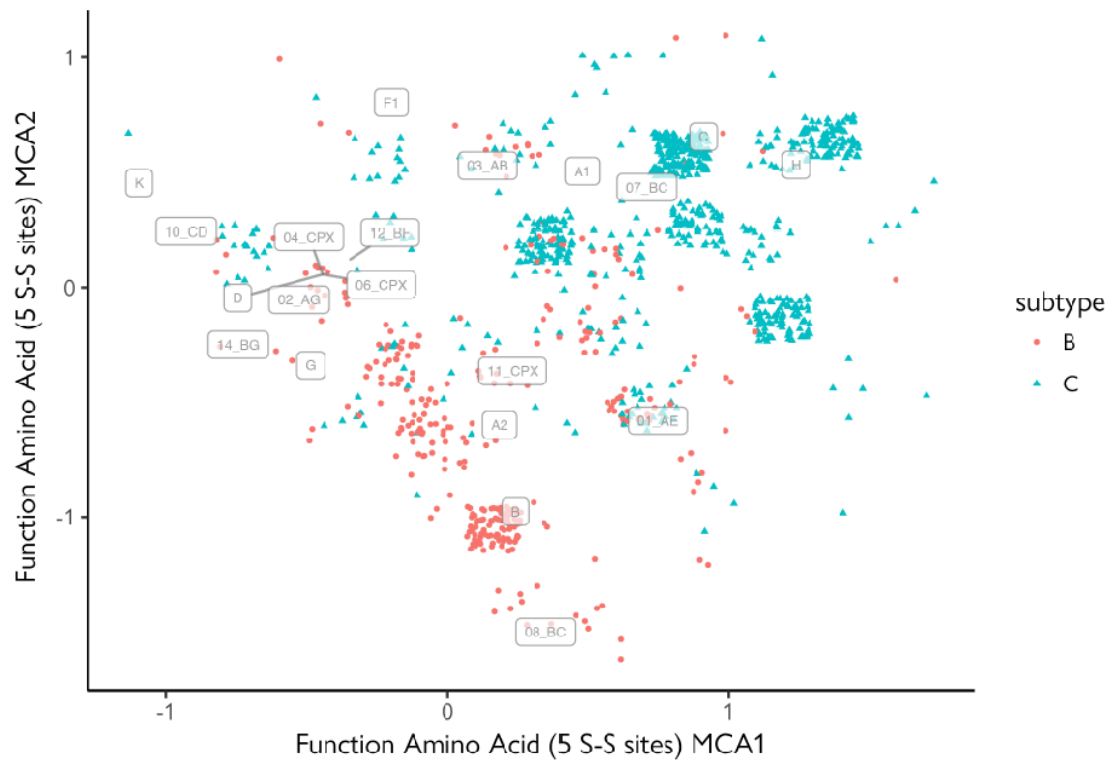
Figure S7



**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.

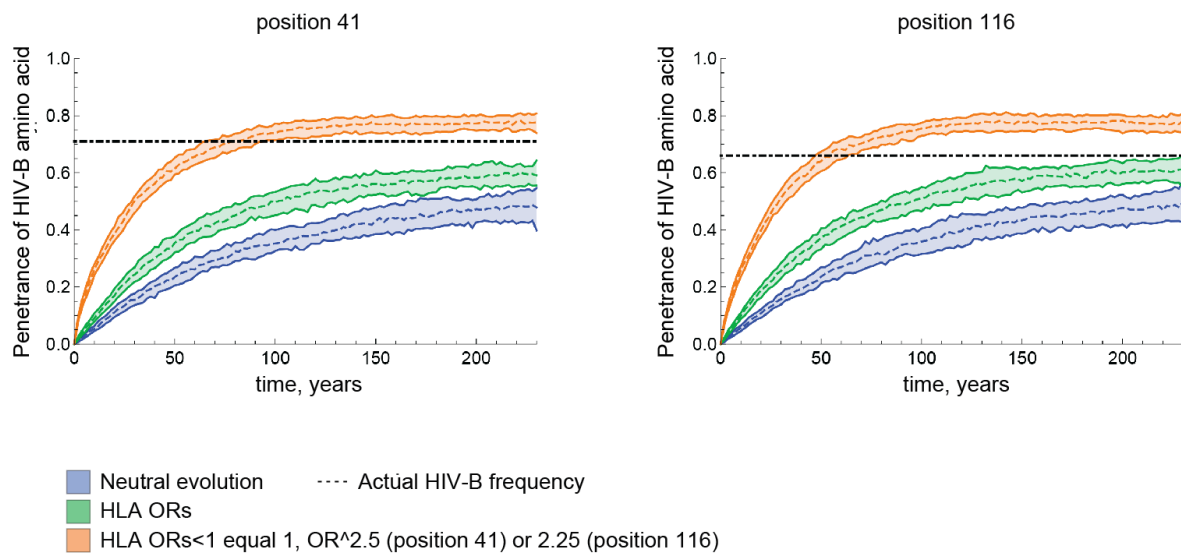
**Figure S8**



**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).

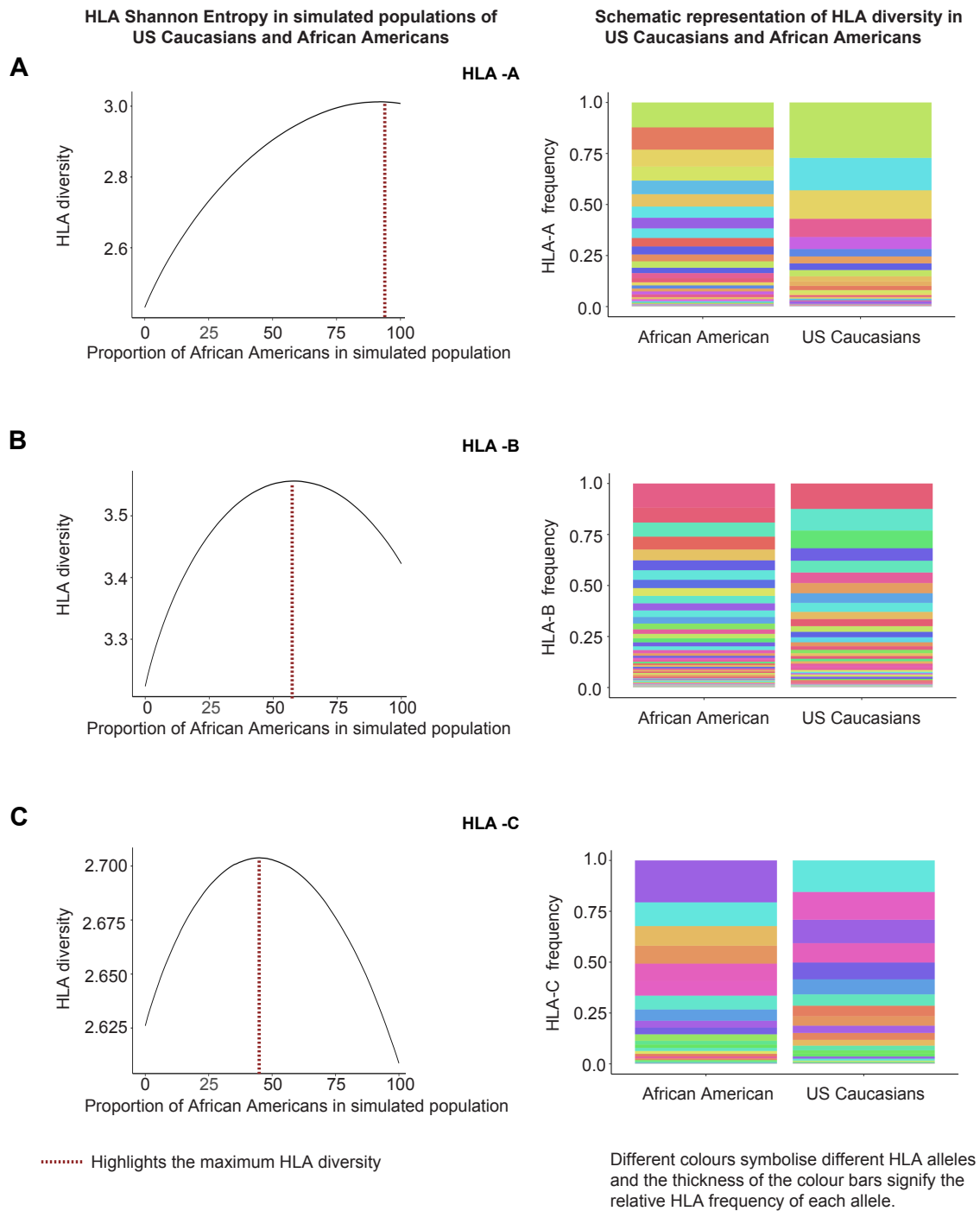
**Figure S9**



**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**



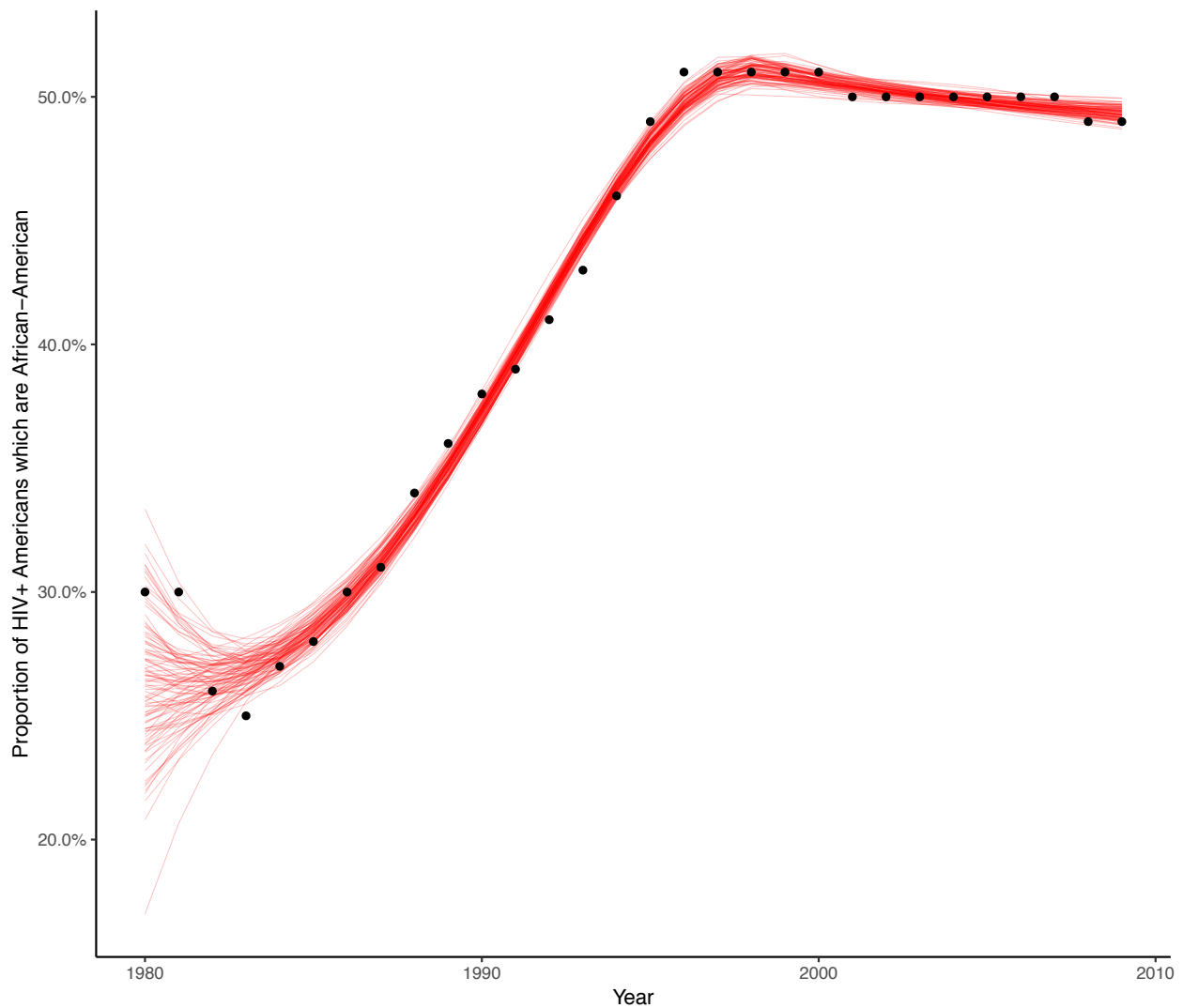
**Fig. S10: Simulated mixtures of African Americans and US Caucasians (left) and HLA diversity in African American and US Caucasian populations (right)**



**(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).

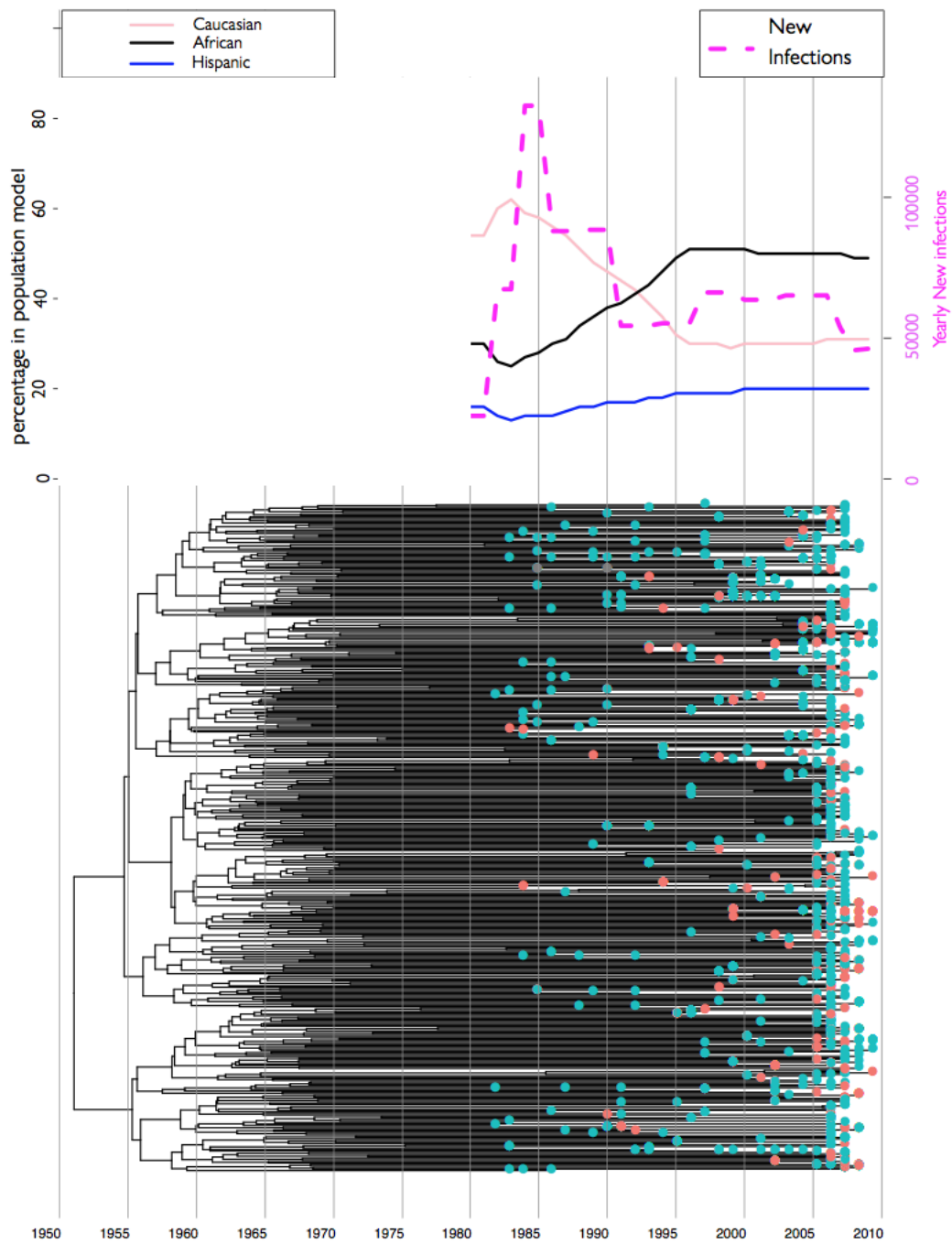
**Figure S11**



**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**

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

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

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

ABABDAH	BALLO	BOGHOM	DAN	FONGORO	HADIYYA
ABBALA	BAILUNDU	BOGURU	DAN-GEH	FOODO	HAM
ABE	BAINUK	BOKALA	DANAKIL	FUFULDE	HAMAMA
ABIDJI	BAKA	BOKIBA	DANGYO	FULANI	HAMBA
ABINSI	BAKO	BOKKOS	DARI	FULANKRIABE	HAMYAN
ABO	BAKPINKA	BOKO BUSA	DASS	FULBE	HANDA
ABRON	BAKUNDU	BOKONGO	DAY	FULLIWA	HANGAZA
ACPA	BAKWE	BOKYI	DCRIKU	FULIRU	HANYA
ACQU	BAKWELE	BOLENDO	DEG	FUNGWA	HARARGE OROMO
ADARAWA	BALANTA	BOLENGE	DEGEMA	FUNGWE	HARARI
ADJUKRU	BALONG	BOLGO	DEKAKIRE	FUR	HAVU
AFAR	BAMBARA	BOLOKI	DELIM	FURU	HAWIYA
AFO	BAMILEKE	BOLOMBOKI	DEMBO	FYAM	HAYA
AGAR	BAMUN	BOLON	DENYA	FYER	HEHE
AGATU	BANDA	BOLONGO	DEY	GA'ANDA	HEMBA
AGOI	BANDI	BOM	DIA	GAAJAK	HEMBA, BANGU-BANGU
AGUA	BANGANTU	BOMASA	DIDA	GAAJOK	HEMBA, BUYU
AHANTA	BANGBA	BOMU	DIDINGA	GAALIN	HENGA
AIKE	BANGU-BANGU	BONDEI	DIGO	GAAM	HERERO
AIT WARAIN	BANGWINJI	BONDJO	DIGODIA	GADAMES	HIMA
AJA	BANNA	BONDO	DII	GADE	HINGA
AKAN	BARA	BONDONGO	DIJIM	GAFSA	HOLO-HOLO
AKANIKI	BARABRA	BONGI	DIKIDIKI	GAGU	HOLHOLO
AKARE	BARAMBO	BONGILI	DINKA	GALA	HOLU
AKEBON	BARUA	BONGO	DIO	GALAMBU	HOMBO
AKOKO-EDO	BARAIN	BOR	DIOBO	GANAGA	HOROM
AKOOSE	BARII	BORORO	DIZI	GANDA	HUBA
AKPA	BARIBA	BOSOKU	DJAGADA	GANJA	HUM
AKPA-YACHE	BARWE	BOZO	DJEBALA	GARREH-AJURAN	HUNDE
AKPES	BASA	BUA	DJEM	GAYA	HUNGANO
AKPINI	BASAA	BUSIA	DJENNEKE	GBAGA	HUNGO
AKPOSO	BASHAR	BUDAMONO	DJERBA	GBAGYI	HUNGWORO
AKYE	BASONGO-MENO	BUDDU	DODOTH	GBANZIRI	HUTU
ALAGO	BASSA	BUDJA	DOE	GBARI	IBAJI
ALEGE	BASSOSSI	BUDUKWA	DOGON	GBAYA	IBIBIO
ALGERIAN	BATA	BUDUMA	DOKA	GBERI	IBINO
ALUR	BATAHIN	BUILSA	DOKO	GBII	ICEN
AMARAR	BATI	BUKA	DOKO-UYANGA	GBINDA	ICHEVE
AMBA	BAULE	BULAANG	DONDO	GBIRI	IDOMA
AMBO	BAUSHI	BULALA	DONGO	GBOMA	IDON
AMBOIM	BAYGO	BULLOM	DRAWA	GEJI	IFUMBA
ANAANG	BEBIL	BUMA	DUAISH	GENYA	IGALA
ANFILLO	BEDAWI	BUMAJI	DUALA	GERAWA	IGBO
ANGAS	BEEMBE	BUMALI	DUI-MENIA	GERI	IGEDE
ANGERE	BEKWARRA	BUNGU	DUKA	GHANGALA	IGUTA
ANIA	BELANDA	BURA-PABIR	DULBU	GHULFAN	IJAW
ANO	BELANDA VIRI	BURAK	DULI	GIBE	IJE
ANTANDROY	BELLE	BURJII	DUMBULE	GIDAR	IKA
ANTANUSI	BEMBA	BURJU	DURUMA	GIWUO	IKASSA
ANTEMURU	BEMBE	BURUN	DUJPA	GIMIRO	IKENGA
ANTESAKA	BENA	BURUNGE	DUIWAI	GIMMA	IKIZU
ANUAK	BENDE	BUSHOONG	DYALONKE	GIMINIME	IKOMA
ANUFO	BENGA	BUYU	DYE	GIMR	IKONGO
ANYI	BENI GIL	BWAALA	DYULA	GINGO	IKPESHI
AODJILA	BENI MENASSER	BWARI	DZA	GISU	IKULU
AOUK	BENI MZAB	BWENDE	DZING	GIZIGA SOUTH	ILA
APAKIBETI	BENI-AMER	BWILE	EBANGA	GOBU	ILA-TONGA
ARAD	BERABISH	BWISI	EBIRA	GODIE	ILEBO
ARGOBA	BERANDJOKO	BYEP	EDO	GOEMAI	ILWANA
ARIAAL	BERGDAMA	CARA	EFUTOP	GOGO	IMBANGALA
ARUM	BERIBERI	CHAAMBA	EGGON	GOKANA	IMOMA
ARUSI	BEROM	CHAGGA	EGYPTIAN	GOLA	IMRAGEN
ASANFWE	BERTA	CHAKALI	EJAGHAM	GOMA	INJOLO
ATEN	BERTI	CHAM	EJAMAT	GOMANI'S NGUNI	IPANGA
ATSAM	BESOMBO	CHAMBA	EKIT	GONDO	IPULO
ATTA	BETE	CHAOUJA	ELEKU	GONJA	IRANGI
ATUOT	BETE-BENDE	CHE	ENDO	GOROWA	IRAQW
AULYEHAN	BETI	CHISHINGA	ENYELLE	GOUIN	IRIGWE
AUSHI	BETSILEO	CHIWERE'S NGUNI	EOTILE	GOWA	ISAAQ
AUYUKAWA	BETSIMISARAKA	CHOKWE	EPEITA	GREBO	ISANZU
AVIKAM	BEYRU	CHOKWE, LELE	ESAN	GUA	ISHA
AVOKAYA-AJIGU	BEZANUZANU	CHOKWE, LUNDA	ESHIRA	GUDE	ISOKO
AVOKAYA-OJILA	BIAFADA	CHUPI	ESIMBI	GUDU	ISSA
AWAK	BIALI	CHUKMA	ETON	GUENZE	ISUWU
AWIYA	BIDEYAT	CINGLO	ETULO	GULA	ITSEKIRI
AWUTO	BIDIANKAMBA	CPALA	EWE	GULA KARA OF SUDAN	ITU
AYIGHA	BILE	CISANJI	EWONDO	GURAGE	IVBIOSAKON
BA	BILLEN	QITATA	FALI	GURENSI	IWA
BAALI	BIMBOBA	QVULVA	FANYAN	GURMA	IWELLEMEDEN
BABALE	BINDI	COMO KARIM	FEZARA	GURMAMA	IYEMBE
BABANKI	BINGA	CWAMHI-WURI	FEZZAN	GWE	IZDRA
BABILE	BIRGIT	DAAROOD	FIA	GWANDARA	IZI
BABOLE	BIRIFOR	DABA	FIGIG	GWENO	JA' ALIYYIN GAALIN
BADA	BIRKED	DADIYA	FIHERENA	GWOMU	JAMALA
BADE	BISA	DAGAARA	FILALA	GYEM	JANJERO
BAGA	BOBANGI	DAHASU	FIPA	HA	JANJI
BAGA MANDORI	BOBILI	DAJU	FODONON	HADENDOWA	JARA
BAGGARA	BOBO	DAKAKARI	FON-GBE		JARAWA
BAGHAP	BOBO-OULE	DAKHLA	FONGE		JARSO
BAGIRMI	BODO	DAMA			JEKANG
BAHARIYA	BOFI				JERA
BAI					JERID

JIBU	KERARISH	KWAFI	LUNDA, CHOKWE	MBEKO	NALU
JIDA	KERWE	KWAKUM	LUNDA, KETE	MBELIME	NAMBA
JIE	KETE	KWAME	LUNDA, NDEMBU	MBELO	NAMJI
JJI	KETE EAST	KWANDI	LUNGU	MBEMBE	NANCERE
JMI	KETE LULUA	KWANGARI	LUNTU	MBEMBRE	NANDE
JMINI	KETE NORTH	KWANGWA	LURI	MBENE	NANDI
JIRU	KETE SOUTH	KWANJA	LUSHANGI	MBERE	NANDU-TARI
JITA	KEYO	KWANKA	LUUNDA	MBESA	NANDYA
JJU	KHARGA	KWAYA	LUYIA	MBINSA	NATENI
JONGA	KHUMALO	KWERE	LWALU	MBO	NATIORO
JOOLA	KIBET	KWESE	LWENA	M'BOKO	NAWDM
JOPADHOLA	KIGA	LAALI	LWER	MBOLE	NBWERA
JUKUN	KIKUYU	LAWWOR	LWIMBI	MBOMOTABA	NDAAKA
JUKUN KONA	KIM	LADO	LYANGALILE	MBONGO	NDALI
JUKUN WASE	KIMBU	LAGBA	LYELE	MBOSHI	NDAM
JUKUN WURKUM	KIMBUNDU	LAGUAT	MA	MBUGWE	NDAMBA
JUNGO	KINGA	LAGWAN	MAAKA	MBUKUSHU	NDAU
JUR	KIONG	LAKA	MAASAI	MBULI	NDEBELE
JUR MODO	KIPSIKIS	LALA	MAAZA	MBULU	NDEMBU
JWIRA-PEPESA	KIR-BALAR	LALA ROBA	MABA	MBUM	NDEMLI
KAAMBA	KISAMA	LALIA	MABENDI	MBUNDA	NDENDEULE
KABA	KISSI	LAMA	MABEZA	MBUNGA	NDENGEREKO
KABA DEMI	KITA	LAMBA	MABIHA	MBUUN	NDENGESE
KABA NAA	KIYAK	LAMBAYA	MABINZA	MBWELA	NDIBU
KABA SO	KLAOH	LAMJA	MACHINGA	MBWELA KOLWE	NDO
KABONGA	KO	LANDUMA	MADA	MBWERA	NDOGO
KABRE	KOBIJANA	LANGA	MADI	MDUNDULU	NDOLO
KABYE	KOFFA	LANGO	MAGHARBA	MEDJE	NDOOOLA
KABYLE	KOFYAR	LARU	MAGUZAWA	MEDOGO	NDULU
KADARA	KOHUMONO	LEELAU	MAHAFALY	MENDE	NDUNGA
KADARU	KOKE	LEGA	MAHONGWE	MENING	NDZABI
KAFFA	KOKO	LEGBO	MAK	MERINA	NDZUNZA
KAGOMA	KOL BIKELE	LEKA	MAKAA	MERU	NEFUSA
KAGORO	KOLBILA	LELE	MAKERE	MESME	NEYO
KAGURU	KOM	LEMBWE	MAKOMA	MFINU	NGABA
KAJAKSE	KOMA	LEMORO	MAKONDE	MFUNU	NGALA
KAKO	KOME	LENDU	MAKUA	MIDOB	NGALANGI
KAKONDA	KONGO DINGA	LENGI	MAKWA	MIGAAMA	NGAM
KALAMSE	KONJI	LENGOLA	MALINKE	MILTU	NGAM GIR BOR
KALANGA	KONJO	LENJE	MALWAL	MIMI	NGAMBAI
KALIKO	KONKOMBA	LESE	MAMBAI	MINIANKA	NGAMO
KALUNDWE	KONO	LIA	MAMBILA	MINUNGO	NGANDA
KAM	KONONGO	LIBOLO	MAMBWE	MINUNGU	NGANDO
KAMANGA	KONSO	LIGBI	MAMPRUSI	MITSOGHO	NGANDU
KAMANTAN	KONYANKE	LIGI	MAMVU	MITTU	NGANDYERA
KAMBA	KOONI	LJILI	MANALA	MITUKU	NGANGELA
KAMBARI	KORO	LKO	MANDA	MIYA	NGARE
KAMBATA	KOROP	LIKWALA	MANDARI	MOBA	NGASA SHAKA
KAMIR	KOTA	LIMA	MANDINKA	MOBANGO	NGATA
KAMKAM	KOTE	LIMO	MANDYAK	MOBENGE	NGBAGA
KAMO	KOTOKO	LO	MANGAS	MOKOLE	NGBAKA
KAMUKU	KOTOPO	LOBI	MANGAYAT	MOLO	NGBANDI
KAMWE	KOUYA	LOBI LORHON	MANGBELE	MONDOMBE	NGEENDE
KANAKURU	KPELLE	LOGO	MANGBETU	MONDZOMBO	NGENGE
KANDAWIRE	KPESSI	LOGORIK	MANGBUTU	MONGO	NGENGELE
KANDERMA	KRAN, NGERE	LOI	MANINKA	MONO	NGENZA
KANEMBU	KRESH	LOKALO	MANKANYA	MONTOL	NGGWAHYI
KANGO	KRIM	LOKELE	MANO	MOROCCAN	NGHWELE
KANTANA	KUANG	LOKO	MANYANGA	MORU	NGINDO
KANU	KUDU	LOKORO	MANYIKA	MOSSI	NGIZIM
KANUFI	KUFRA	LOKOYA	MANZA	MOTEMBO	NGOLA
KANURI	KUKA	LOMA	MAO SOUTH	MOYE	NGOMBE
KANYOK	KUKELE	LOMBI	MAPOPOI	MPAMA	NGONGI
KAONDE	KUKU	LOMOTWA	MARARIT	MPANGU	NGONGO
KAONDE, SANGA	KUKUYA	LOMWI	MARBA	MPE	NGOUNGWONI
KARA	KULANGO	LONDO	MARFA	MPEZENI'S NGUNI	NGUL
KARABORO	KULERE	LONGARIM	MARGHI	MPONGMPONG	NGULU
KARANGA	KULUNG	LONGTO	MARGHI SOUTH	MPOTO	NGUMBA
KARE	KUMU	LOPA	MARKA	MPUT	NGUMBO
KAREKARE	KUNAMA	LOSAKANYI	MASAKIN	MPUTU	NGUNDI
KASEM	KUNDA	LOTSU-PIRI	MASALIT	MPYEMO	NGUNGULU
KASENA	KUNDU	LOVALE	MASANA	MSER	NGUNI
KASONGI	KUNTA	LOVEDU	MASHASHA	MUBI	NGURIMI
KASONGO	KUNUZ	LOZI	MASHI	MUCHIKONGO	NGWABA
KASONKE	KUNYI	LUANO	MASMAJE	MUKULU	NHAM
KATAB	KUPTO	LUBA	MASONGO	MUMUYE	NIABWA
KATYA	KURAMA	LUBA EASTERN	MASSA	MUN	NIELLM
KAWENDI	KURANKO	LUBA KASAYI	MASSALAT	MUNDANG	NIMBARI
KAYAMBA	KURFEI	LUBA SHANKADI	MATENGO	MUNDU	NINZAM
KAYLA	KURI	LUBA UPEMBA	MATUMBI	MUNGA	NJAJGULGULE
KEDERU	KURUMFE	LUBA, LUNDA	MAU	MURLE	NJAMUS
KEENGE	KURYA	LUBA, NDEMBU	MAURI	MUSGU	NKANGALA
KEL	KUSAL	LUBA, NDEMBU, LUNDA	MAYOGO	MUSSEY	NKANSI
KEL AHAGGAR	KUSHI	LUBILA	MBA	MUSSUMBE	NKANU
KEL AYR ASBEN	KUSU	LUCHAZI	MBAAMBA	MVELE	NKHUMBI
KEL GRES	KUTEP	LUGBARE	MBAATI	MWAGHAVUL	NKOLE
KEL IFFORA	KUTIN	LUGURU	MBAGANI	MWENYI	NKOMI
KELE	KUTU	LULA	MBAI BEDIONDO	MYANG	NKOYA
KENGA	KUTURMI	LULUBA	MBALA	MYENE	NKUKOLI
KENYANG	KUYERI	LULUWA	MBANGWE	NABAN	NKUTSHU
KENYI	KUYU	LUMBU	MBANZA	NAFANA	NKWE
KERA	KWAAMI	LUNDA	MBATA	NAGUMI	NOBIIN

NOLE	PONGWE	SISAALA	TIV	WOLOF
NOWOLO	POTO	SISYA	TOBANGA	WOM
NSAMBA	PUGULI	SIWA	TOGBO	WONGO
NSAFO	PUKU	SIZAKI	TOKA	WOYO
NSENGA	PUNU	SO	TOMA	XHOSA
NSONGO	PYAANG	SOKORO	TONGA	YAA
NTANDU	QUILENGUE	SOLA	TONGA-INHAMBANE	Yaelima
NTCHAM	RASHAD	SOLI	TONGWE	YAGOUTE
NTOMBA	REGEIBAT	SOLONGO	TOPOKE	YAHl
NUBA	RENDILL	SOLU	TOPOSA	YAKA
NUER	RIF	SOMRAI	TORAM	YAKA, SUKU
NUMANA	RIYAH	SOMYEWE	TORO	YAKOMA
NUNA	RONGA	SONGHAI	TOROBE	YALIWASA
NUNGU	RUARHA	SONGO	TOTELA	YAMAIE
NUNU	RUBASA	SONGOLA	TOUGOURT	YAMANDUNDU
NUNUMA	RUFUJl	SONGYE	TOUYO	YAMONGO
NUPE	RUNGA	SONINKE	TOW	YANA
NWENSHI	RUNYANKOLE	SONJO	TRARZA	YANZI
NWERA	RURI	SOONDE	TRIBOUe	YAO
NYAKYUSA	RUSHA	SOSSO	TRIPOLITANIAN	YARSE
NYALA	RUUND	SOTHO	TSAAM	YASA
NYALI	RUWENG	SOUTHERN BANGANTU	TSHIAKA	YASAMA
NYAMWANGA	SAAB	SUAFA	TSIENIMBALALA	YEKE
NYAMWEZI	SAADI	SUBA	TSIMIHETY	YELA
NYANEKA-HUMBE	SABA	SUBIYA	TSONG	YESKWA
NYANGA	SAFWA	SUGA	TSONGA	YEYE
NYANGATOM	SAGALA	SUK	TUAT	YIWOM
NYANGBARA	SAGARA	SUKU	TUKEN	YOKO
NYANJA	SAHEL	SUKUMA	TUKULOR	YOMBE
NYARAFOLO	SAHO	SUKWA	TULA	YORUBA
NYATURU	SAKA	SUMA	TULAMA	YOWA
NYEMBA	SAKALAVA	SUMBWA	TUMBUKA	YUKUTARE
NYENGO	SAKATA	SUNDI	TUMBWE	YUNGUR
NYIHA	SALA	SUNGOR	TUMTUM	ZAGHAWA
NYILAMBA	SALA MPASU	SURI	TUNDJUR	ZANAKI
NYIMANG	SAMBA	SURUBU	TUNGU	ZANDE
NYINDU	SAMBA LEKO	SUSU	TUNISIAN	ZANDE ABANDIA
NYONG	SAMBO	SWAKA	TUPURI	ZANDE AVUNGARA
NYORO	SAMBU	SWAZI	TURA	ZARAMO
NYULI	SAMBURU	SYEMU	TURKA	ZARI
NZAKARA	SAN	TABWA	TURKANA	ZARMA
NZANYI	SANGA	TAFIRE	TURUMBU	ZEEM
NZIMA	SANGO	TAGWANA	TUSIAN	ZEKARA
OBOLO	SANGU	TAHOu	TUTSI	ZELA
OBULKOM	SANUSI	TAJAKANT	TYEMBARA	ZENAGA
ODUT	SAPAUT	TAJUASO	TYENGA	ZIBAN
OKAK	SAR	TALE	UBANG	ZIGUA
OKO-ENI-OSAYEN	SARA	TALENSI	UHAMI-IYAYU	ZILMANU
OKPE	SARA GULA	TAMA	UJOGOMA	ZIMBA
OKPE-AKUKU	SARA GULAY	TAMAZIGHT	UKAAN	ZINZU
OKPELA	SARO	TAMBAHUAKA	UKPE	ZOMBO
OMBO	SARUA	TAMBAS	UKPET	ZULA
OMETO	SASARU	TAMBERAMA	UKWUANI	
OMONO	SAYA	TAMBO	ULED NAIL	
OOLl	SEBA	TAMEZRET	UMON	
OPA	SEEKU	TAMPOLENSE	UNGA	
OPAMERI	SEKE	TANALA	UREGU	
OPUJU	SENA	TANDA	URHOBO	
ORANA	SENGA	TANGA	UTUGWANG	
ORING	SENGELE	TANGALE	UZEKWE	
OROMO	SERE	TANGBAGO	VAGALA	
ORON	SERER	TANKARA	VAI	
OTANG	SESE	TAPSHIN	VEKHEE	
OTUHU	SH	TAROK	VENDA	
OUASSA	SHAGAWU	TASUMSA	VERE	
PAI	SHAIKIA	TATOG	VEZO	
PALLAKA	SHALL	TAWANA	VIDUNDA	
PAMBIA	SHAMBAA	TAZARAWA	VILI	
PANA	SHANJO	TCHWABO	VIN	
PANDE	SHATT	TEDA	VINZA	
PANI	SHEBELLE	TEFASI	VIRA	
PAPEL	SHERBRO	TEGE	VIVe	
PARE	SHI	TEITA	VUMA	
PARE, SAA	SHIKI	TEM	VUTE	
PATU	SHILA	TEMNE	WAAMA	
PAYE	SHILLUK	TENDA	WADA	
PENDE	SHINJI	TEPETH	WAJA	
PERE	SHIRAWA	TERA	WALA	
PERO	SHLUH	TERE	WALLAGA	
PEVE	SHONA	TESO	WAMBU	
PHOKA	SHOO	TETELA	WANDYA	
PIMBWE	SHUBI	TEVUNDRU	WARGLA	
PINDA	SHUKRIA	TIBA	WARJI	
PINDI	SHUWA	TIBEA	WASA	
PITI	SIDAMO	TIE	WASI ALAWA	
PIYA	SIDI	TIENE	WE	
PODJULU	SIHANAKA	TIGRAY	WENYA	
PODZO	SIMAA	TIGRINYA	WILE	
POGORO	SINASHA	TIKAR	WILLE	
POKOMO	SINGA	TIKUU	WINIAMA	
POMBO	SINYAR	TIO	WINJI	
POMO	SIRTICAN	TITU	WOLLO	



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

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

01_AE	13C	50_A1D	A1A2C	AG	GH
01A1	13U	51_01B	A1A2CD	AGJ	GHK
01A1G	14_BG	52_01B	A1A2D	AGU	GJ
01ADF2	15_01B	53_01B	A1A2G	AH	GK
01B	16_A2D	54_01B	A1A3	AHJU	GKU
01BC	16A1	55_01B	A1A6	AJ	GU
01BG	17_BF	56_cpx	A1B	AKU	H
01C	18_cpx	57_BC	A1BD	AU	HJ
01D	18D	58_01B	A1C	B	HU
01F2	18G	59_01B	A1CD	BC	J
01G	19_cpx	61_BC	A1CDGKU	BCF1	JK
01GHJKU	19A1	62_BC	A1CG	BCU	JKU
01U	19B	63_02A	A1D	BD	JU
02_AG	20_BG	64_BC	A1DG	BF	K
02A	21_A2D	65_cpx	A1DHK	BF1	KU
02A1	22_01A1	67_01B	A1DK	BF1G	M
02A1A2	22A1U	68_01B	A1DU	BF2	U
02A1G	22DU	69_01B	A1F1	BG	
02A1U	23_BG	70_BF1	A1F2	BK	
02A3	23A1	71_BF1	A1G	BU	
02A6	24_BG	72_BF1	A1GH	C	
02AG	25_cpx	73_BG	A1GHU	CD	
02B	26_A5U	74_01B	A1GJ	CDG	
02BD	26C	76_01B	A1H	CF1	
02BG	27_cpx	78_cpx	A1J	CF1U	
02C	28_BF	79_0107	A1K	CG	
02D	29_BF	82_cpx	A1U	CH	
02F2	30_0206	83_cpx	A2	CHU	
02G	31_BC	85_BC	A2B	CJ	
02GK	32_06A6	86_BC	A2C	CJU	
02H	32A6	87_cpx	A2CD	CU	
02HU	32G	90_BF1	A2D	D	
02U	33_01B	102	A2F	DF	
03_AB	34_01B	103	A2G	DF1	
04_cpx	35_AD	107	A2U	DF1G	
05_DF	36_cpx	108	A3	DF2	
06_cpx	37_cpx	113	A3G	DG	
06A1	38_BF	206	A4	DGH	
06G	38_BF1	209	A6	DK	
06U	39_BF	211	A6B	DU	
07_BC	40_BF	218	AB	F	
07B	41_CD	222	ABD	F1	
08_BC	42_BF	609	ABDU	F1F2	
09_cpx	43_02G	708	AC	F1G	
09A1	44_BF	0102A	ACD	F1U	
09A1D	45_cpx	0102A1	ACDJ	F2	
10_CD	45BF1	0122F	AD	F2G	
11_cpx	45F1	1113	ADG	F2K	
11A1	45U	1819	ADU	F2KU	
11AG	46_BF	-	AF	FK	
11C	47_BF	A	AF1	FKU	
12_BF	48_01B	A1	AF2	FU	
13_cpx	49_cpx	A1A2	AF2G	G	

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

Bantu languages are indicated by stars.

	Adamawa-Ubangian
	Adamawa-Ubangian / Chari-Nile
	Bantoid
★ 	Bantu
★ 	Bantu / Bantu
	Berber
	Chadic
	Chadic / Cushitic
	Chadic / Ffulde
	Chari-Nile
	Chari-Nile / Nilotic
	Cushitic
	Ffulde
	Ffulde / Adamawa-Ubangia
	Fur
	Gbaya
	Khoi: Nama, Bergdama
	Kordofanian
	Kru
	Kwa
	Maban
	Malagasy
	Miscellaneous / Unclassified
	Nilotic
	Nilotic / Bantoid
	Nilotic / Bantu
	Northern Mande
	Other
	Saharan
	Saharan / Cushitic
	Saharan / Nilotic
	San
	Sandawe
	Semitic: Arab, Bedouin
	Songhai
	Southern Mande
	Voltaic
	West Atlantic

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