SUPPLEMENTARY MATERIAL

Sequence Cleanup and Bioinformatics Analysis

The procedure of the bioinformatics pipeline consists of the following steps:

1. Baseline quality control: instrument quality (q) scores are used to remove reads that are too short (≤50 bp), have too many (≥8 bases with q≤15) low quality bases, and to mask all bases with low scores with 'N' characters.
2. Error correction: reads from step 1 are aligned to the corresponding amplicon reference sequence (HXB2) using a codon-level extension to the Smith-Waterman local alignment algorithm, which directly accounts for frameshift errors caused by homopolymer length miscalls – the most common error modality for the 454 platform. Pairwise alignments are merged into a global alignment of all reads, so that each base is mapped to a consistent system of genomic coordinates. Remaining sequencing errors are modeled as a mixture of multinomial distributions, and all individual single nucleotide variants, which cannot be reliably assigned to non-error multinomial components (posterior probability > 0.999, assumed error rate ≤0.5%) are reverted to position consensus.
3. Extraction of common and representative reads: because the median read length is >50% of the length of sequenced amplicons, we performed simple clustering of error-corrected reads, where two reads were merged in the same cluster if (a) they did not differ except in positions where one of the sequences had an error-induced 'gap'; (b) they matched at X or more nucleotide positions, where X is the maximum of 100 half of the median read length. Each cluster was represented by its consensus sequence, and by the number of reads assigned to the cluster.
4. Recombination screening: we screened all sets of representative reads from step 3 for evidence of recombination using GARD [1]. Note that because step 4 does not rely on the assumption that all reads are related by a single phylogeny, it will not be adversely affected by the presence of recombinants.

The pipeline is available at *https://github.com/veg/HIV-NGS*

Table Suppl.1. Summary of the Algorithm for Computing the Estimated Date of HIV-1 Infection. Adapted from [2].

Class	Definitions for stage of HIV-1 infection
A1.0	If there is a first positive RNA† and negative enzyme immunoassay (EIA) within 7 days of the first positive RNA, and no prior positive/indeterminate western blot (WB), then EDI = first positive RNA date − 11 days. (~Fiebig Stages I-II‡)
A2.0	If there is an indeterminate WB within 7 days of the first positive RNA, then EDI = first positive RNA date − 20 days. (~Fiebig Stages III-IV‡)
A3.0	If the last negative EIA or negative/indeterminate WB occurred ≤ 30 days before the first positive WB (with associated positive RNA), then EDI = midpoint of the positive WB date and the negative EIA or negative/indeterminate WB date (earlier of two) − 19 days. (~Fiebig Stage IV‡)
A3.1	If the first positive WB p31/32 band is absent, then EDI = first positive WB date − 89 days. (~Fiebig Stage V‡)
E1.0A	If there is a detuned EIA (dtEIA) consistent with infection of ~3 mo within 30 days of the first positive WB and CD4 count > 200 or CD4% > 14 within 30 days of the first positive WB, then EDI = first dtEIA date − 70 days§. (Fiebig VI‡)
E1.0B	If there is a dtEIA consistent with infection of ~3-6 mo within 30 days of the first positive WB and CD4 count > 200 or CD4% > 14), then EDI = first dtEIA date − 133 days§. (Fiebig VI‡)
E1.0C	If there is a dtEIA consistent with infection of ~6-12 mo within 30 days of the first positive WB and CD4 count > 200 or CD4% > 14, then EDI = first dtEIA date − 170 days§. (Fiebig VI‡)
E2.0	If there is a first positive WB and a negative EIA within 365 days participant enrollment (Day 0), then EDI = midpoint between the last negative EIA and Day 0. (Fiebig VI‡)

* Each rule applied sequentially until EDI criteria satisfied.
† Positive RNA was defined as a NAT/viral load exceeding the detectable level for a given assay.
‡ Indexed to algorithm from Fiebig *et al.* [3]. Note: no endpoint was defined for stage VI by Fiebig.
§ dtEIA threshold from Kothe *et a*l. [4].

Table Suppl.2. Next Generation Sequencing Characteristics

ID	Cpt	Time Points	Time from EDI (in days)	Nb of Haplotypes	Genome Coverage	Nb of Reads
S1	BP	1	20	17	1656	2533
	PBMC	1		16	1682.5	2205
	SP	1		26	2462	12527
	BP	2	48	15	1690	2663
	BP	3	104	5	2689	2802
	PBMC	3		7	815	1094
	SP	3		39	7332	8810
S2	BP	1	87	45	4774	5140
	SP	1		33	4639	5358
	SP	2	227	23	4687	4843
	BP	3	248	46	5562.5	5945
	BP	4	289	10	769	870
	SP	4		2	1367	1823
	PBMC	5	2096	20	1798	1875
S3	BP	1	84	27	3356.5	3458
	BP	2	98	19	1767.5	1919
	PBMC	2		20	1745	1963
	SP	2		33	3306.5	3726
	BP	3	154	23	1579	1731
	BP	4	182	15	873	1012
	SP	4		37	4023	4203
S4	BP	1	40	19	2223.5	2473
	PBMC	1		17	3096	3321
	SP	1		31	2287	2573
	BP	2	42	26	6674.5	6992
	BP	3	47	4	1683	4886
	SP	3		22	2748	3003
	BP	4	75	10	754	1230
	SP	4		10	879	1274
	BP	5	118	5	563.5	832
	PBMC	5		13	1293	1441
S5	BP	1	77	34	4184.5	4326
	PBMC	1		24	1807	1893
	SP	1		21	2480	2692
	BP	2	162	24	2326	3936
	SP	2		28	11863	13611
	BP	3	254	29	2787	2982
	SP	3		5	596	1272
S6	BP	1	169	50	3869.5	5342
	BP	2	177	31	1936	2041
	PBMC	2		12	1520	1925
	SP	2		17	1603	2032
	BP	3	190	26	1831	1978
	BP	4	288	42	3351	4180
	BP	5	457	12	686	824
	SP	5		34	6538.5	12309
	BP	6	514	15	1148	1313
	PBMC	7	762	15	850	948
median		4.5	218	21	1884	2553
mean		4.5	586	22	2711	3503

Cpt: Compartment; EDI: Estimated date of infection; Nb: Number; BP: Blood Plasma; SP: Seminal Plasma; PBMC: Peripheral Blood Mononuclear Cells

Table Suppl.3. Sites under positive and selective pressure by compartment

	Positive Selection			Purifying Selection		
Subject	Blood plasma	PBMC	Seminal plasma	Blood plasma	PBMC	Seminal plasma
S1	2.	3.	2.	4.	3.	4.
S2	6.	0.	5.	7.	3.	3.
S3	3.	4.	4.	4.	3.	4.
S4	3.	1.	2.	7.	5.	4.
S5	3.	2.	5.	5.	0.	3.
S6	4.	4.	4.	16.	4.	5.

Sites under selective pressure were assessed across compartment for each participants with the Fast Unconstrained Bayesian AppRoximation (FUBAR) program [5].

Table Suppl.4. Markov Jump Counts Summary

Subject	Markov Jump Count [95%HPD]	ME From BP [95%HPD], (%)	ME From BP to SP, No. (%)	ME From BP to PBMC, No. (%)
S1	52.9 [41-63]	43.4 [35-58], (82)	27.2 (62.6)	16.2 (37.4)
S2	22.8 [17-28]	13.0 [6-18], (57)	12.1 (93.1)	0.8 (6.9)
S3	29.0 [22-36]	27.1 [20-34], (93.4)	13.7 (50.6)	13.4 (49.4)
S4	64.8 [53-77]	60.5 [46-75], (93.3)	31.8 (52.6)	28.6 (47.3)
S5	43.7 [35-51]	41.4 [33-50], (94.7)	21.9 (52.9)	19.5 (47.1)
S6	23.6 [18-29]	22.4 [16-27], (94.9)	7.0 (31.2)	15.4 (68.8)
Mean	39.5	34.6 (85.9)	18.9 (57.2)	15.7 (42.8)

ME: Migration events; BP: Blood Plasma; SP: Seminal Plasma; 95%HPD: highest posterior density intervals at 95%.

Table Suppl.5 Cytokines quantification in seminal plasma at baseline

ID	IL-6	MCP-1/CCL2	RANTES/CCL5	IFN-gamma	IP-10/CXCL10	TNF-alpha
S1	43.6	2175.8	324.1	109.4	44532.4	2.1
S2	1.3	1577.9	508.4	321.6	615464.2	0.5
S3	32.2	605.6	292.4	3.2	34033.4	7.1
S4	7.4	2222.1	558.2	177.9	353660.1	4.7
S5	355.5	570.3	251.7	160.5	36247.3	37.1
S6	79.4	2601.4	677.4	107.8	148871.9	9.3

MCP-1: monocyte chemotactic protein; IL-6: interleukin; TNF- α: tumor necrosis factor; Interferon-γ, RANTES: regulated on activation normal T cell expressed and secreted; IP-10: Interferon-γ induced protein.

Table Suppl.6. Human Herpes Virus copies/ml in seminal plasma at baseline (Log_{10}).

ID	HSV-1	HSV-2	CMV	EBV	HHV6	HHV7
S1	0.00	0.00	0.00	0.00	3.52	0.00
S2	0.00	0.00	0.00	2.77	0.00	0.00
S3	0.00	3.34	5.37	4.34	0.00	0.00
S4	0.00	0.00	0.00	0.00	0.00	0.00
S5	3.00	0.00	5.74	0.00	0.00	0.00
S6	0.00	0.00	7.38	3.70	0.00	0.00

HSV-1: Herpes Simplex Virus 1; HSV-2: Herpes Simplex Virus 2; CMV: Cytomegalovirus; EBV: Ebstein Barr Virus; HHV-6: Human Herpes Virus 6; HHV-7: Human Herpes Virus 7.

Figure Suppl.1. Virological characteristics and sampling for the 6 individuals.

For each individual, top plots indicate the evolution of HIV RNA levels (in red) and semen (in blue) and CD4 T cell count (in black). Bottom plots indicate the timing of blood plasma (red), seminal plasma (blue) and PBMC (green) sampling.

FIGURE Suppl.1 Virological characteristics and sampling for the 6 individuals

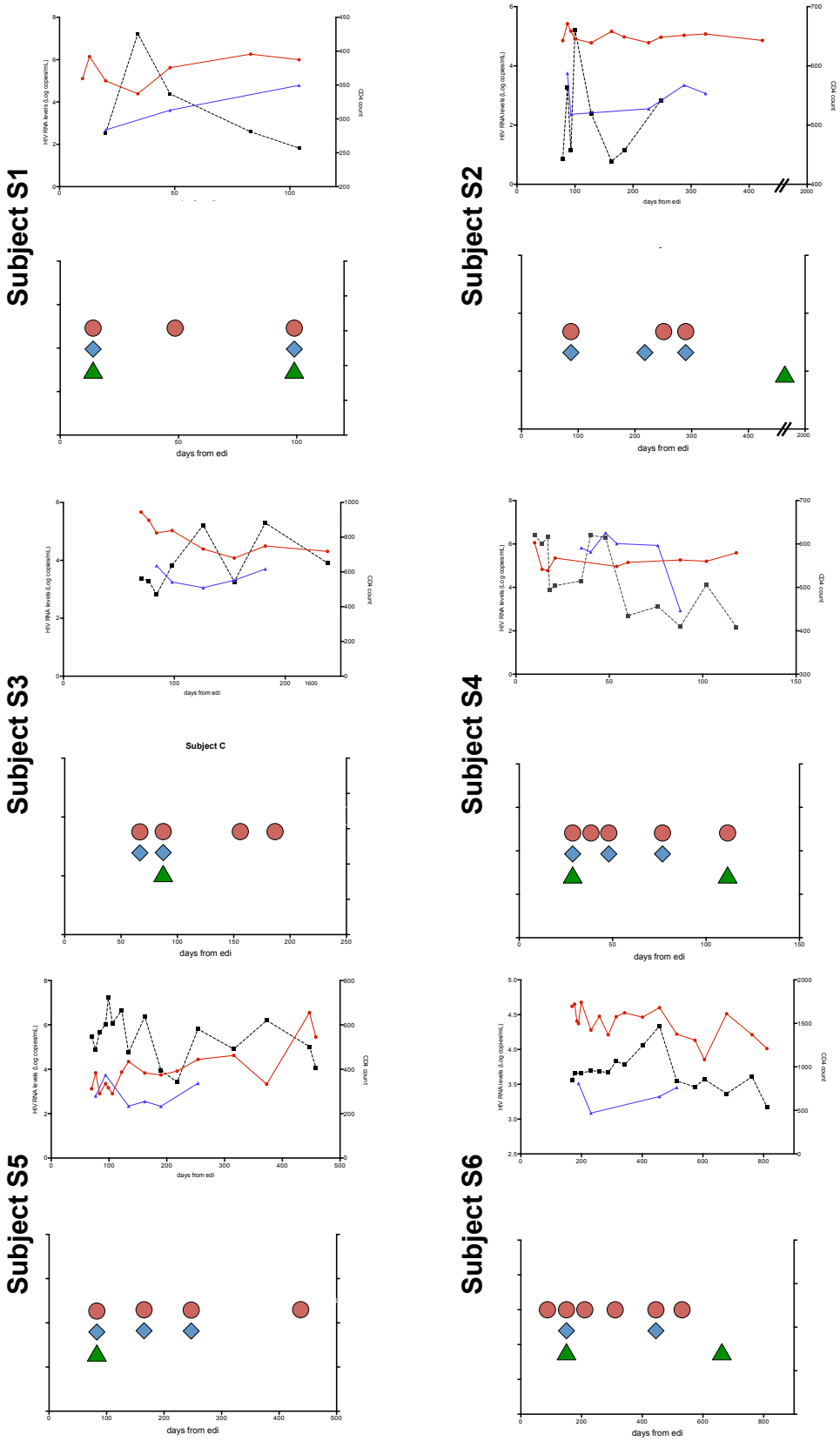
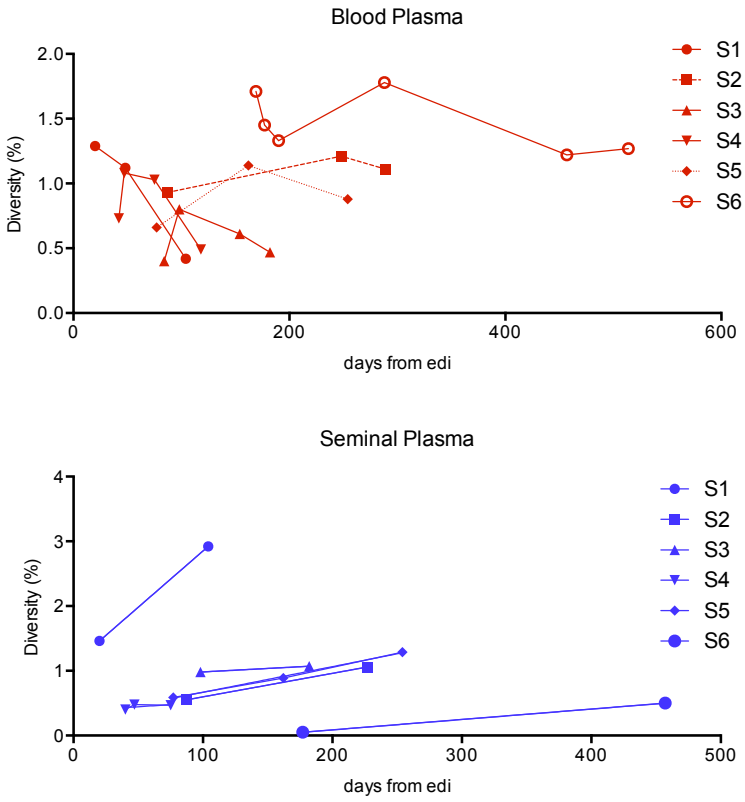


Figure Suppl.2. Env nucleotide diversity in blood plasma and seminal plasma derived viral populations. The mean of all pairwise Tamura-Nei 93 distances between reads with at least 100 overlapping base pairs was computed to measure the mean pairwise diversity [6] (2). Blood and seminal plasma samples are colored in red and blue respectively.

FIGURE Suppl.2. Env nucleotide diversity in blood plasma and seminal plasma derived viral populations.



REFERENCES

1. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: a genetic algorithm for recombination detection. *Bioinformatics (Oxford, England)* 2006,**22**:3096-3098.
2. Le T, Wright EJ, Smith DM, He W, Catano G, Okulicz JF*, et al.* Enhanced CD4+ T-cell recovery with earlier HIV-1 antiretroviral therapy. *The New England journal of medicine* 2013,**368**:218-230.
3. Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, Peddada L*, et al.* Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. *AIDS (London, England)* 2003,**17**:1871-1879.
4. Kothe D, Byers RH, Caudill SP, Satten GA, Janssen RS, Hannon WH*, et al.* Performance characteristics of a new less sensitive HIV-1 enzyme immunoassay for use in estimating HIV seroincidence. *Journal of acquired immune deficiency syndromes (1999)* 2003,**33**:625-634.
5. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL*, et al.* FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Molecular Biology and Evolution* 2013,**30**:1196-1205.
6. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution* 1992,**9**:678-687.