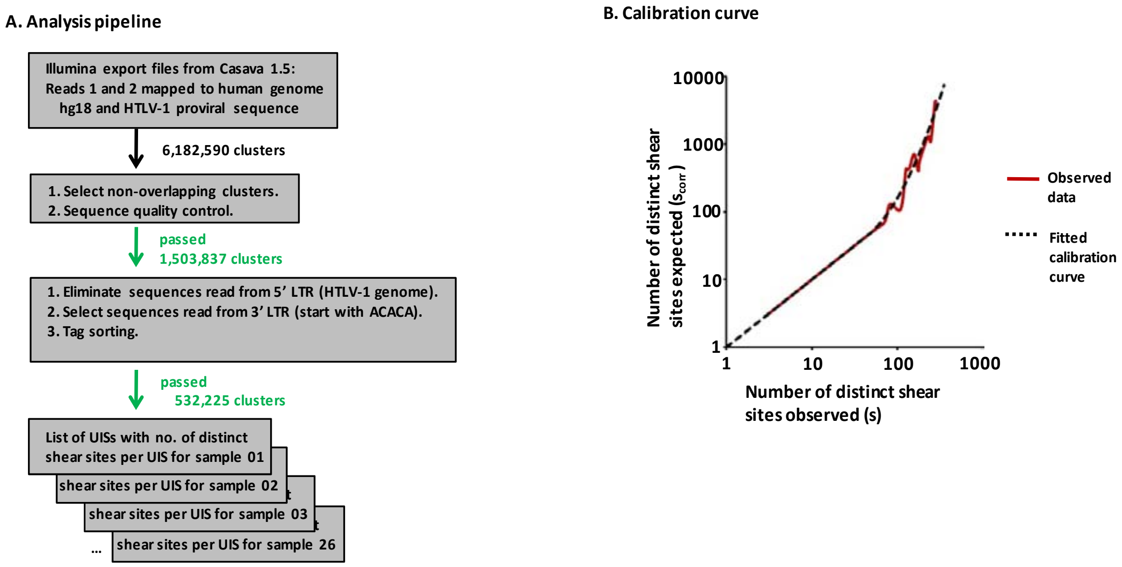
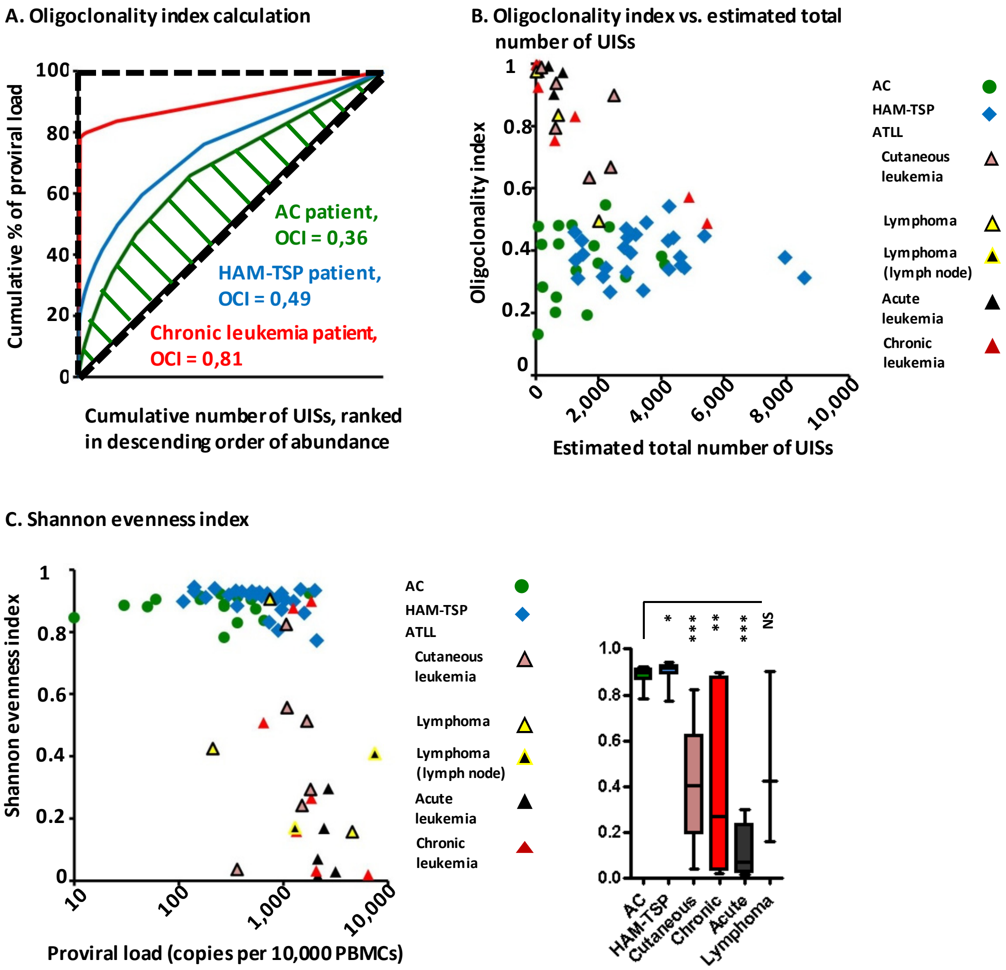

A. Analysis pipeline
Illumina export files from Casava 1.5: Reads 1 and 2 mapped to human genome hg18 and HTLV-1 proviral sequence
6,182,590 clusters
1. Select non-overlapping clusters.
2. Sequence quality control.
passed
1,503,837 clusters
1. Eliminate sequences read from 5' LTR (HTLV-1 genome).
2. Select sequences read from 3' LTR (start with ACACA).
3. Tag sorting.
passed
532,225 clusters
List of UISs with no. of distinct shear sites per UIS for sample 01
shear sites per UIS for sample 02
shear sites per UIS for sample 03
...
shear sites per UIS for sample 26
B. Calibration curve
10000
1000
100
10
1
Number of distinct shear sites expected (s_{corr})
1
10
100
1000
Number of distinct shear sites observed (s)
Observed data
Fitted calibration curve

Figure S1

(A) A cluster is the result of amplification of a single molecule during the solid phase PCR of a library on the Illumina flow cell. The numbers of clusters given here are from a representative experiment. (B) To generate the calibration curve, three dilutions of genomic DNA from an HTLV-1 infected individual (10µg, 1µg and 0,1µg) were analyzed, each in duplicate.



A. Oligoclonality index calculation

B. Oligoclonality index vs. estimated total number of UISs

C. Shannon evenness index

Figure S2

(A) For each individual, the curve shows the rise in the cumulative percentage of the proviral load as insertion sites are added in descending order of abundance. The oligoclonality index is defined by the area between these curves and the diagonal line (e.g. hatched area for AC) as a fraction of the triangle indicated by the dotted black line. The larger the area, the less uniform the distribution of insertion site abundance. A patient with a perfect monoclonal distribution (only one clone) will have an oligoclonality index of 1. A patient with a perfect polyclonal distribution (each clone has equal abundance) will have an oligoclonality index of 0. (B) The oligoclonality index did not correlate with the estimated total number of UISs, either in ACs or in patients with HAM-TSP. **C.** Shannon evenness index distinguished patients with ATLL from subjects with non-malignant HTLV-1 infection (Mann-Whitney, $p<0.0001$). Mean CV of Shannon evenness = 1,5% (triplicate analysis of 11 samples).

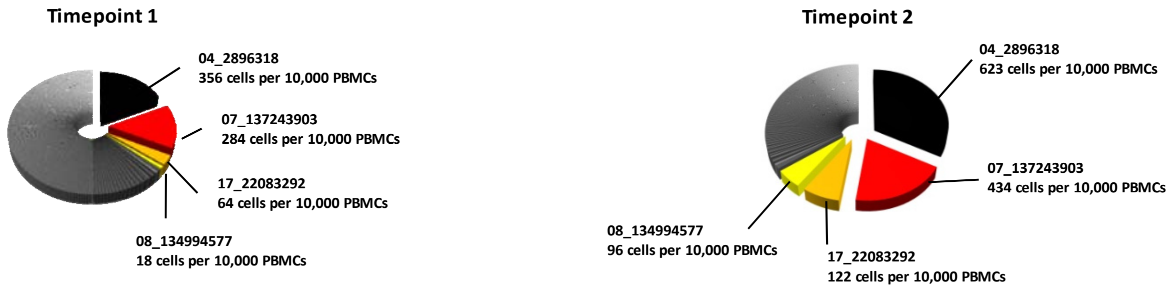


Figure S3. HTLV-1 clonality in the blood of patient TBK at timepoints 1 and 2. UIS position and abundance are given for the 4 largest clones

The coordinate 04_2896318 denotes insertion of the provirus in chromosome 4, nucleotide position 2896318. The data show that the increase in the oligoclonality index in this patient was due to the expansion of clones that were already abundant at timepoint 1.

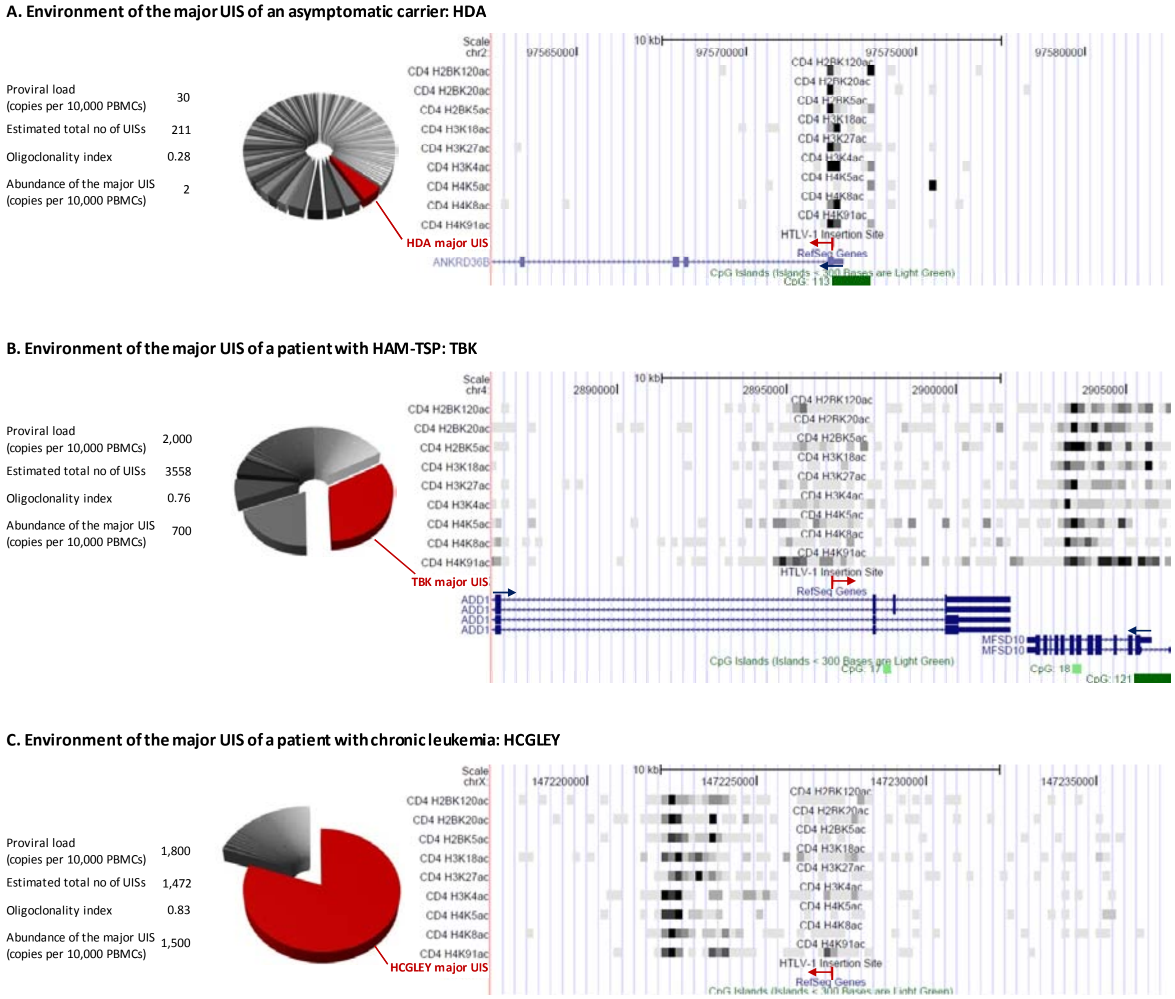


Figure S4. Genetic and epigenetic environment of the major UIS in 3 infected individuals

The clonal structure of the UIS population in each individual is illustrated by the pie charts. The red slice represents the major UIS. UCSC Genome Browser tracks were used to illustrate the environment within +/- 10kb of the insertion site (vertical red line with horizontal red arrow showing the provirus and its orientation). Different histone acetylation marks are represented together with a track for RefSeq genes (blue lines; the blue arrow gives the orientation of the gene) and for CpG islands (green boxes).

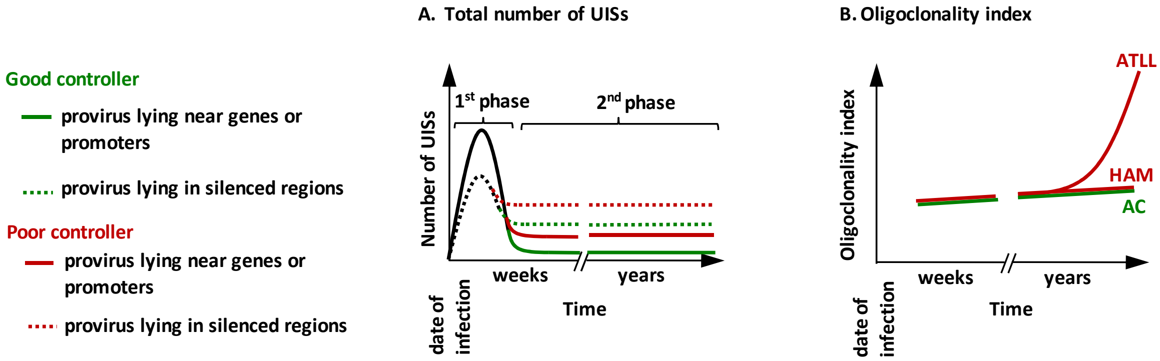


Figure S5. Progression of HTLV-1 replication and oligoclonality in hosts who differ in the efficiency of the anti-HTLV-1 immune response: a hypothetical scheme

(A) In the first phase of infection, before the immune response, the virus disseminates in the host by cell-to-cell transmission, generating a large number of distinct UISs. Proviruses integrated near host genes and promoters (solid black line) outnumber proviruses integrated in transcriptionally silent DNA (dotted black line). In the second phase, the host immune response negatively selects T-cells that express viral proteins, resulting in an inversion of the ratio of proviruses inserted in active transcribed genomic regions (solid lines) to proviruses integrated in silenced DNA (dotted lines). A good HTLV-1 controller (green curves), i.e. an individual able to mount an efficient anti-HTLV-1 immune response, limits the total number of UISs more effectively than a poor controller (red curves). In the chronic phase of infection, the less efficient immune (CTL) response in a poor controller allows provirus-expressing T-cells to proliferate faster than in the good controller. (B) The oligoclonality index increases continuously during the chronic phase; in a minority of individuals, one or more infected T-cell clones undergoes malignant transformation, resulting in a sharp rise in the oligoclonality index and culminating in ATLL.

Flow cell no.	Patient code	Disease status	Sample type	Age (years)	PVL (copy number per 10,000 PBMCs)	CD4 count	percentage of CD4 in lymphocytes	Number of reads (human sequences only)	Number of UISs observed	Estimated total number of UISs in host	OCI
0	TBP	HAM/TSP	PBMCs	69	350	760	44%	36279	3613	8569	0.31
0	TBU	HAM/TSP	PBMCs	65	110	1000	38%	70258	749	1245	0.46
0	TCQ	HAM/TSP	PBMCs	68	500	not available	not available	44629	2156	4743	0.35
0	HBX	AC	PBMCs	54	270	not available	not available	42889	644	1153	0.48
0	HDG	AC	PBMCs	24	270	1330	45%	90212	1048	2230	0.55
0	HFG	AC	PBMCs	58	160	1180	50%	44436	413	724	0.42
0	HCD	AC	PBMCs	53	490	470	50%	5181	179	633	0.20
0	HFE	AC	PBMCs	66	390	620	27%	98809	802	2199	0.92
1	HAY	AC	PBMCs	59	400	not available	not available	20957	993	2258	0.47
1	HT/UV1	UV	PBMCs	72	130	not available	not available	17456	699	1373	0.38
1	TAA	HAM/TSP	PBMCs	64	350	not available	not available	27947	1229	2768	0.44
1	TAL	HAM/TSP	PBMCs	63	930	not available	not available	16061	1564	3030	0.40
1	TAW	HAM/TSP	PBMCs	73	160	not available	not available	19120	571	1184	0.38
1	TAY	HAM/TSP	PBMCs	59	180	not available	not available	23944	832	1661	0.40
1	TBC	HAM/TSP	PBMCs	59	2210	not available	not available	20586	3684	8163	0.39
1	TBG	HAM/TSP	PBMCs	64	2300	not available	not available	16767	2091	3908	0.42
1	TBK	HAM/TSP	PBMCs	75	1990	not available	not available	25776	1159	3485	0.46
1	TBW	HAM/TSP	PBMCs	60	940	not available	not available	37206	1600	3323	0.48
1	TW	HAM/TSP	PBMCs	34	430	not available	not available	22054	1242	2562	0.43
2	HAY	AC	PBMCs	59	400	not available	not available	19714	1048	2443	0.48
2	HT/UV1	UV	PBMCs	72	130	not available	not available	11152	622	1255	0.36
2	TAA	HAM/TSP	PBMCs	64	350	not available	not available	23493	1157	2877	0.43
2	TAL	HAM/TSP	PBMCs	63	930	not available	not available	15628	1580	2748	0.42
2	TAW	HAM/TSP	PBMCs	73	160	not available	not available	19644	649	1363	0.37
2	TAY	HAM/TSP	PBMCs	59	180	not available	not available	16498	691	1272	0.38
2	TBC	HAM/TSP	PBMCs	59	2210	not available	not available	14994	3261	7414	0,37
2	TBG	HAM/TSP	PBMCs	64	2300	not available	not available	18897	2445	4369	0.45
2	TBK	HAM/TSP	PBMCs	75	1990	not available	not available	61868	1951	4772	0.57
2	TBW	HAM/TSP	PBMCs	60	940	not available	not available	28972	1652	3734	0.48
2	TW	HAM/TSP	PBMCs	34	430	not available	not available	24793	1371	3148	0.44
3	HAY	AC	PBMCs	59	400	not available	not available	21427	937	2306	0.48
3	HT/UV1	UV	PBMCs	72	130	not available	not available	17149	559	1181	0.34
3	TAA	HAM/TSP	PBMCs	64	350	not available	not available	34798	1393	3136	0.46
3	TAL	HAM/TSP	PBMCs	63	930	not available	not available	17518	1579	2692	0.40
3	TAW	HAM/TSP	PBMCs	73	160	not available	not available	23018	631	1272	0.36
3	TAY	HAM/TSP	PBMCs	59	180	not available	not available	27499	802	1584	0.39
3	TBC	HAM/TSP	PBMCs	59	2210	not available	not available	23558	3696	8274	0.38
3	TBG	HAM/TSP	PBMCs	64	2300	not available	not available	29429	2919	4892	0.45
3	TBK	HAM/TSP	PBMCs	75	1990	not available	not available	95290	1924	4505	0.60
3	TBW	HAM/TSP	PBMCs	60	940	not available	not available	37981	1610	3527	0.51
3	TW	HAM/TSP	PBMCs	34	430	not available	not available	41633	1727	3852	0.48
4	HAY	AC	PBMCs	67	360	810	42%	39593	1276	2511	0.53
4	HT/UV1	UV	PBMCs	81	150	751	44%	43623	591	1189	0.44
4	TAA	HAM/TSP	PBMCs	70	220	620	33%	21824	875	2041	0.46
4	TAL	HAM/TSP	PBMCs	71	870	670	50%	40315	1865	3269	0.45
4	TAW	HAM/TSP	PBMCs	81	140	350	27%	40045	566	1181	0.45
4	TAY	HAM/TSP	PBMCs	65	140	not available	not available	26511	581	996	0.36
4	TBC	HAM/TSP	PBMCs	67	2000	1220	not available	41063	4713	9521	0.47
4	TBG	HAM/TSP	PBMCs	73	1460	1220	54%	31079	2284	4098	0.50
4	TBK	HAM/TSP	PBMCs	82	2080	1490	39%	107509	1295	3638	0.87
4	TBW	HAM/TSP	PBMCs	65	730	260	18%	40735	1415	3245	0.52
4	TW	HAM/TSP	PBMCs	43	890	1350	59%	47244	2105	3877	0.49
5	HAY	AC	PBMCs	67	360	810	42%	23800	899	2038	0.47
5	HT/UV1	UV	PBMCs	81	150	751	44%	20111	523	1015	0.44
5	TAA	HAM/TSP	PBMCs	70	220	620	33%	18752	975	2350	0.45
5	TAL	HAM/TSP	PBMCs	71	870	670	50%	32749	1791	2942	0.44
5	TAW	HAM/TSP	PBMCs	81	140	350	27%	26460	547	1051	0.41
5	TAY	HAM/TSP	PBMCs	65	140	not available	not available	17236	591	1281	0.37
5	TBC	HAM/TSP	PBMCs	67	2000	1220	not available	22491	4002	7909	0.44
5	TBG	HAM/TSP	PBMCs	73	1460	1220	54%	18994	2007	3813	0.48
5	TBK	HAM/TSP	PBMCs	82	2080	1490	39%	63278	1173	3333	0.74
5	TBW	HAM/TSP	PBMCs	65	730	260	18%	41661	1942	3928	0.55
5	TW	HAM/TSP	PBMCs	43	890	1350	59%	29072	2157	4146	0.52

Flow cell no.	Patient code	Disease status	Sample type	Age (years)	PVL (copy number per 10,000 PBMCs)	CD4 count	percentage of CD4 in lymphocytes	Number of reads (human sequences only)	Number of UISs observed	Estimated total number of UISs in host	OCI
6	HAY	AC	PBMCs	67	360	810	42%	25194	1126	2331	0.51
6	HT/UV1	UV	PBMCs	81	150	751	44%	30736	541	1140	0.44
6	TAA	HAM/TSP	PBMCs	70	220	620	33%	30009	924	2163	0.46
6	TAL	HAM/TSP	PBMCs	71	870	670	50%	33302	1757	3145	0.43
6	TAW	HAM/TSP	PBMCs	81	140	350	27%	28024	551	1114	0.41
6	TAY	HAM/TSP	PBMCs	65	140	not available	not available	23194	640	1050	0.38
6	TBC	HAM/TSP	PBMCs	67	2000	1220	not available	21252	3364	7588	0.41
6	TBG	HAM/TSP	PBMCs	73	1460	1220	54%	26671	2132	3713	0.48
6	TBK	HAM/TSP	PBMCs	82	2080	1490	39%	92391	1243	3558	0.77
6	TBW	HAM/TSP	PBMCs	65	730	260	18%	42584	1664	3386	0.54
6	TW	HAM/TSP	PBMCs	43	890	1350	59%	36642	2206	3985	0.51
7	TAL	HAM/TSP	PBMCs	72	750	1046	51%	26058	1648	2955	0.43
7	TAN	HAM/TSP	PBMCs	52	400	760	54%	21024	1388	3427	0.27
7	TAS	HAM/TSP	PBMCs	61	300	1030	44%	13485	901	2147	0.32
7	TAT	HAM/TSP	PBMCs	74	960	1934	53%	10525	1282	2907	0.33
7	TBA	HAM/TSP	PBMCs	68	960	1270	59%	27784	1215	2887	0.47
7	TCO	HAM/TSP	PBMCs	54	180	1040	43%	17903	497	1346	0.31
7	TBJ	HAM/TSP	PBMCs	49	1580	1868	49%	16907	1442	4222	0.43
7	TCG	HAM/TSP	PBMCs	49	360	680	55%	28031	766	1471	0.43
7	TDA	HAM/TSP	PBMCs	42	620	1050	49%	7125	805	2370	0.27
7	TAZ	HAM/TSP	PBMCs	69	1250	1410	52%	13636	1965	4606	0.38
7	TBO	HAM/TSP	PBMCs	63	700	770	45%	13519	1066	2236	0.34
7	TBR	HAM/TSP	PBMCs	57	470	1540	63%	47148	1461	3028	0.39
7	TCR	HAM/TSP	PBMCs	48	590	1050	47%	38437	2048	4245	0.34
8	HDM-LFK	ATLL chronic	PBMCs	64	1570	4530	78%	27128	255	501	0.90
8	HBE	AC	PBMCs	75	360	650	34%	7860	888	2872	0.32
8	HBY	AC	PBMCs	52	160	1600	42%	10377	422	1637	0.19
8	HBT	AC	PBMCs	68	50	1050	41%	12183	126	191	0.42
8	KD5	ATLL chronic	PBMCs	56	1210	not available	not available	4659	17	47	0.82
8	TBX-LFI	ATLL lymphoma	PBMCs	35	270	850	36%	23472	353	770	0.61
8	HDA	AC	PBMCs	44	30	770	52%	10290	89	211	0.28
8	HFE	AC	PBMCs	68	650	871	30%	12767	702	1865	0.42
8	HDS	AC	PBMCs	59	250	1750	not available	18010	952	1993	0.36
8	HEZ	AC	PBMCs	51	340	490	40%	44252	1935	4020	0.38
8	TCT	HAM/TSP	PBMCs	56	550	1410	68%	26719	1943	4638	0.35
9	HDM-LFK	ATLL chronic	PBMCs	68	640	830	64%	25356	367	603	0.75
9	LFA	ATLL cutaneous	PBMCs	57	1060	1040	56%	30600	1025	1708	0.63
9	LFC	ATLL chronic	PBMCs	44	1230	4530	75%	29217	2426	5463	0.49
9	P7	ATLL chronic / polymyositis	PBMCs	72	1830	2430	80%	39283	3182	4885	0.57
9	KD5	ATLL chronic	PBMCs	58	1320	not available	not available	19801	27	73	0.92
9	TBX-LFI	ATLL lymphoma	PBMCs	36	210	390	36%	32102	238	717	0.83
9	HCG-LEY	ATLL chronic	PBMCs	37	1830	2060	87%	96662	269	1251	0.83
9	LEZ	ATLL acute	PBMCs	37	2670	1790	79%	86061	243	581	0.90
9	LFE	ATLL lymphoma	PBMCs	65	740	1430	56%	28028	1258	2018	0.49
9	TBS	HAM/TSP	PBMCs	69	1040	not available	not available	21127	2396	5379	0.45
9	LFP	AC	PBMCs	51	60	1280	53%	59035	242	658	0.25
10	A1 (~KD5)	ATLL acute	PBMCs	59	2120	not available	not available	39703	10	46	1.00
10	C1 (~KD5)	ATLL chronic	PBMCs	56	1650	not available	not available	22709	9	14	1.00
10	LN C2	ATLL chronic	Lymph node	unknown	1280	not applicable	not applicable	27356	105	310	0.93
10	C3	ATLL chronic	PBMCs	unknown	6400	not available	not available	15560	5	8	1.00

Flow cell no.	Patient code	Disease status	Sample type	Age (years)	PVL (copy number per 10,000 PBMCs)	CD4 count	percentage of CD4 in lymphocytes	Number of reads (human sequences only)	Number of UISs observed	Estimated total number of UISs in host	OCI
10	C4	ATLL chronic	PBMCs	unknown	2040	not available	not available	64408	30	114	1.00
10	S1	ATLL smouldering	PBMCs	59	1810	not available	not available	27415	830	2497	0.90
10	S2	ATLL smouldering	PBMCs	69	1080	not available	not available	7531	234	627	0.79
10	SKN S3	ATLL cutaneous	Skin	82	340	not applicable	not applicable	19760	276	749	0.76
10	S4	ATLL smouldering	PBMCs	unknown	1670	not available	not available	14621	629	2390	0.67
10	S5	ATLL smouldering	PBMCs	65	360	not available	not available	24410	53	176	0.99
11	LEU	ATLL lymphoma	PBMCs	64	4550	490	49%	19440	7	13	0.97
11	LEP	ATLL acute	PBMCs	73	2120	not available	not available	78233	276	864	0.97
11	AN	ATLL acute	PBMCs	47	2430	not available	not available	28545	45	119	0.97
11	JH	ATLL acute	PBMCs	44	3120	not available	not available	76432	107	409	0.99
11	LN TBX-LFI	ATLL lymphoma	Lymph node	36	7410	not applicable	not applicable	52959	5	6	0.85
12	S6 (~SKN S3)	ATLL cutaneous	PBMCs	82	1500	not available	not available	24823	282	649	0.94
13	HES	AC	PBMCs	68	1770	890	48%	31271	1981	4119	0.36
13	HDR	AC	PBMCs	48	270	1090	57%	44265	471	725	0.48
13	HCS	AC	PBMCs	38	540	1145	53%	35281	76	88	0.48

Table S1. Summary of data by individual sample

The patients HAY, HT, TAN, TAY, TAZ, TBA, TBG, TBK, TBO, TBR and TBW were also studied (using conventional linker-mediated PCR) in Meekings KN, Leipzig J, Bushman FD, Taylor GP, Bangham CR. HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP. PLoS.Pathog.2008;4:e1000027.

Mapping of UISs and quantification of UIS abundance

Ten micrograms of DNA from uncultured PBMCs in a volume of 100µl of Elution Buffer (EB, Qiagen) were sheared by sonication with a Covaris S2 instrument (Covaris). The operating conditions were the following: water bath at 6 to 8°C, 5 seconds at 20% duty cycle, intensity level 5 and 200 cycles per burst and 90 seconds at 5% duty cycle, intensity level 3 and 200 cycles per burst. DNA ends were then end-repaired using 15 units of T4 DNA polymerase (New England Biolabs), 5 units of DNA polymerase I Klenow fragment (NEB), 50 units of T4 polynucleotide kinase (NEB) and 0.8 mM of dNTP (Sigma) in T4 DNA ligase buffer (NEB) at 20°C during 30 min. DNA was then cleaned using a Qiaquick PCR purification kit (Qiagen) and eluted in 64µl of EB. Addition of an adenosine at the 3' ends of the DNA was performed by adding 0.2mM of dATP (Sigma) and 15 units of Klenow Fragment 3' to 5' exo- (NEB) in NEB2 buffer (NEB) at 37°C for 30 min. DNA was then cleaned using a Qiaquick PCR purification kit and eluted in 40µl of EB. One hundred pmol of a partially double stranded DNA linker was ligated to the DNA ends using a Quick ligation kit (NEB). Twenty six different linkers were constructed, each one with a specific 6bp tag (see primer list below) to allow multiplexing of DNA samples during the sequencing. DNA was cleaned using a Qiaquick PCR purification kit and eluted in 60µl of EB. The 60µl of ligated product was then split into 3 aliquots of 20µl and each aliquot was used in a separate PCR1 reaction. For each PCR reaction, 20µl of ligated product was mixed with 0.2mM of dNTP (Sigma), 50pmol of B3 primer (binds HTLV-1 LTR), 10pmol of B4 primer (which anneals to the strand of the linker generated by the amplification from Bio3), 1 unit of Phusion DNA polymerase (Finnzyme, NEB) in High Fidelity buffer (Finnzyme, NEB). The following thermal protocol was used: 96°C 30sec; (94°C 5sec, 72°C 1min) 7 cycles; (94°C 5 sec, 68°C 1 min) 23 cycles; 68°C 9 min; hold 4°C until user stops. The 3 PCR1 products, derived from the same sample were then pooled, the DNA cleaned using a Qiaquick PCR purification kit and eluted in 150µl of EB. To perform PCR2, 1ul of the cleaned PCR1 product was mixed with 0.2mM of dNTP (Sigma), 25pmol of P5B5 primer (binds HTLV-1 LTR), 25pmol of P7 primer (binds the linker), 1 unit of Phusion DNA polymerase in High Fidelity buffer. The following thermal protocol was used: 96°C 30sec; (94°C 5sec, 72°C 1min) 7 cycles; (94°C 5 sec, 68°C 1 min) 23 cycles; 68°C 9 min; hold 4°C until user stops. See primer list above for the sequences of B3, B4, P5B5 and P7 primers. DNA was then cleaned using a Qiaquick PCR purification kit and eluted in 50µl of EB. A library was constructed by pooling the different PCR2 products (each one possessing a specific tag). Quantification of the libraries was made by QPCR using primers P5 and P7 (see primer list) and a Light cycler-fast start DNA master ^{plus} SYBR green 1 kit following manufacturer's protocol (Roche). Three dilutions of the library (200pg, 66pg and 22pg) were amplified. Standard curves were generated using a library quantified on a titration flow cell previously run on a

Genome Analyzer II (Illumina). Stock libraries were diluted down to 0.5 pM and clustered on the flow cell. Paired-end reads (read1 and read2 each 50bp) plus a 6 bp tag read (read 3) were acquired on a GA II. The library construction pipeline was split into four steps: 1.DNA isolation and shearing; 2. Pre-PCR manipulations (ends repair and ligation); 3. PCR1 and PCR2 and library QPCR; 4. Library sequencing. Each step was carried out in a specific room and the sample flow was unidirectional, to minimize the risk of PCR contamination.

The following definitions were used. An "amplicon" is a molecule generated during PCR; "duplicates" are amplicons of a given insertion site having the same length (i.e. having the same shear site). "Sister cells" are cells where the HTLV-1 provirus is inserted at the same site in the cellular genome and a "clone" is a population of sister cells. Figure 1A illustrates the amplicon structure. Read 1 and read 2 were mapped against the HTLV-1 and the human genome (hg18 assembly) using CASAVA software 1.5. First, we excluded overlapping clusters, and then applied the following quality control filters: i) the single-read alignment scores of read 1 and read 2 must be higher than 10 (value attributed by CASAVA); ii) the strand orientation of the 2 reads must be opposite; iii) the length of the amplicons must be smaller than 1kb. Because the PCR is not specific to the 3'LTR, we discarded the amplicons generated from the 5'LTR that contained only HTLV-1 sequences. We identified read 1 sequences that started with ACACA (the five bases at the 3' terminus of the HTLV-1 LTR). Read 1 and read 2 sequences were used to map the insertion site and the shear site. Read 3 was used to allocate the insertion site to a particular sample. Together with the mapping of a large number of UISs, the goal of our approach is to quantify the abundance of each UIS as illustrated in Figure 1B. Because PCR preferentially amplifies short products, the number of amplicons cannot be used to quantify the abundance of a given UIS. Random DNA fragmentation by sonication is a key feature to allow the quantification of the UIS abundance because, unlike restriction enzymes, it is not biased to particular nucleotide sequences. For each UIS, we count the number of amplicons of different length. Additionally, misreading of the tag index during the sequencing could lead to the attribution of a particular insertion site to different samples. We solved this issue by taking into account not only the 6bp tag information but also the total number of distinct shear sites and the total number of reads for a given insertion site to attribute the insertion site to the correct sample. Finally, control DNA (from a human T-cell line (Jurkat) uninfected by HTLV-1) was run on every lane of each flow cell to assess the effectiveness of our quality control procedures. An example of the analysis pipeline from a representative experiment is given in Figure S1A. The number of distinct shear sites for a given insertion site was then corrected using a calibration curve (see Figure S1B) because the probability of DNA shearing at the same place in two distinct cells increases with the number of sister cells of that infected T-cell clone. To generate the calibration curve, 3 dilutions (10µg, 1µg and 0.1µg) of a unique infected genomic DNA were analyzed in duplicate. The calibration curve is a spline fit of the data and the coefficients in it were estimated using the `lm` function of the R language for statistical computing [1]:

s_{corr} =EXP(LN(MIN(50,s)) + 1.18*MAX(0,LN(s)-LN(50)) + 0.707*MAX(0, LN(s) - LN(50))^2)

where s is the number of distinct shear sites observed

s_{corr} is the number of distinct shear sites expected, also referred as the corrected number of distinct shear sites

The absolute abundance of a given UIS (number of a particular insertion site per 10,000 PBMCs) was calculated as follows:

$$absolute\ abundance\ of\ a\ given\ UIS\ i = \frac{(s_{corr})_i}{\sum_{i=1}^{S}(s_{corr})_i}\ PVL$$

where S is the total number of UISs identified in that sample and PVL is the proviral load (sum of the absolute abundances of all insertion sites in that sample).

The absolute abundance of a given UIS (in number of copies per 10,000 PBMCs) is equal to the absolute abundance of the T-cell clone carrying that provirus (in number of cells per 10,000 PBMCs) only when that clone carry one provirus per cell. It has been shown that the leukemic clones carry generally one provirus per cell [2-4] but large-scale systematic studies will be necessary to generalize this observation to the majority of infected clones.

Oligoclonality index and estimation of the total number of UISs in the entire body of the host

We wished to use a measure of the clonality of the UIS population, i.e. a measure of the non-uniformity of the frequency distribution of UIS abundance. We used two measures: the oligoclonality index and the Shannon evenness index.

Oligoclonality index (see Figure S2A for illustration)

The oligoclonality index (OCI) is based on the Gini coefficient [5] and is calculated as

$$OCI = 2 * \left\{ \sum_{k=1}^{S} \frac{X_k}{S} - 0.5 \right\}$$

Where

$(s_{corr})_i$ the corrected number of distinct shear sites for a given UIS i

S the total number of UISs identified in the sample

$N = \sum_{i=1}^{S} (s_{corr})_i$ the sum of the corrected number of distinct shear sites for all the UISs

$p_i = \frac{(s_{corr})_i}{N}$ the relative abundance of each UIS

$X_k = \sum_{i=1}^{i=k} p_i$ the cumulative relative abundance of UIS ranked in decreasing order of abundance

Shannon evenness index

We calculate the Shannon evenness index as

$$E_H = -\frac{\sum_{i=1}^{S} \left(p_i - \ln(p_i) - \left[\left(\frac{S-1}{2N} \right) \right] \right)}{\ln S}$$

We routinely use the OCI because oligoclonality is a critical biological feature of the persistence and replication of HTLV-1.

Estimation of total number of UISs in the entire body of the host (Chao1-bc)

The Chao1-bc estimator [6] is calculated as follows:

$$Chao1bc = S + \frac{f_1(f_1 - 1)}{[2(f_2 + 1)]}$$

Where

S the total number of UISs identified in the sample

f_1 the number of UISs that are represented exactly once in the sample (i.e. UISs with a corrected number of distinct shear sites = 1)

f_2 the number of UISs that are represented exactly twice in the sample (i.e. UISs with a corrected number of distinct shear sites = 2)

Genetic and epigenetic environment around the proviral insertion site

We used the Integration Site Pipeline and Database (INSIPID) from the Pr Bushman laboratory (Department of Microbiology, University of Pennsylvania School of Medicine, Pennsylvania, Philadelphia, United States of America) (http://microb215.med.upenn.edu/insipid/). The purpose of this web-based tool is to house sequences of newly inserted elements in vertebrate genomes and allow users to investigate their locations. In our experiment, we produced a set of Unique Insertion Sites of the HTLV provirus from several infected patients. INSIPID houses the UISs together with associated annotation, then allow us to call up the UIS and information about them.

The following table gives a concrete example of the INSIPID output and how we calculate the values shown in Fig. 4:

								Figure 4, panel A										Figure 4, panel B		Figure 4, panel D	
Patient code	Disease Status	Chromo some	Position	Strand	UIS abundance (in copies per 10,000 PBMCs)	inGene (Refseq)	inGene Orientation	nearest Gene Distance (bp)	nearest CpG Island Distance (bp)	H2BK120 ac Count 10k	H2BK20 ac Count 10k	H2BK5 ac Count 10k	H3K18 ac Count 10k	H3K27 ac Count 10k	H3K4 ac Count 10k	H3K9 me2 Count 10k	H3K9 me3 Count 10k	H4K20 me3 Count 10k	H4K5 ac Count 10k	H4K8 ac Count 10k	H4K91 ac Count 10k
C3	ATLL	chr5	94400662	+	6370	MCTP1	-	42639	244553	1	4	3	2	1	0	33	38	10	4	7	2
LEU	ATLL	chr1	20877564	+	4212	KIF17	-	-14471	-16739	7	4	6	16	8	6	12	15	12	18	24	5
JH	ATLL	chr2	138170969	+	3071	FALSE	NA	-19212	804534	3	5	6	2	1	2	55	24	17	4	4	2
LEP	ATLL	chr3	197773499	-	2025	WDR53	-	-6311	-6236	35	52	41	28	43	30	21	3	4	49	29	37
C4	ATLL	chr7	90098693	-	2005	FALSE	NA	-77954	34393	6	1	5	1	4	1	28	15	7	1	5	2
AN	ATLL	chr9	112538958	+	1950	MUSK	+	64141	-157108	1	3	3	6	4	2	58	32	24	4	2	1
LEZ	ATLL	chr12	129441003	+	1826	FALSE	NA	5630	33727	3	3	8	2	8	4	58	15	65	6	11	3
KD5	ATLL	chr9	15797049	+	1485	C9orf93	+	164848	-253715	2	1	2	6	2	2	44	28	8	1	3	0
HCG-LEY	ATLL	chrX	147227167	-	1472	FALSE	NA	-162663	-162386	68	56	56	77	65	61	27	11	29	29	40	60
S1	ATLL	chr13	35811059	-	1263	SPG20	-	-7587	-6678	14	18	6	9	7	11	42	9	8	13	22	7
S6	ATLL	chr8	142483050	-	1170	FALSE	NA	-18138	11374	194	181	194	219	184	182	12	4	13	141	98	267
S4	ATLL	chr1	182126012	+	924	RGL1	+	38277	-84973	60	60	28	84	55	52	36	7	11	67	51	37
S2	ATLL	chr10	51234264	+	359	FALSE	NA	849	813	66	61	82	83	91	52	25	9	12	54	71	79
S5	ATLL	chr16	64689163	+	354	FALSE	NA	268862	172623	2	3	6	5	3	13	50	14	10	5	7	2
TBX-LFI	ATLL	chr1	151510970	+	85	FALSE	NA	-9746	264078	4	7	3	6	2	5	30	20	10	7	5	5
LFC	ATLL	chr13	45858440	-	85	C13orf18	-	-1196	-245	13	26	16	36	19	19	28	12	11	31	36	28
LFA	ATLL	chr15	36147866	-	66	FALSE	NA	116951	-4421	20	18	7	21	6	19	55	35	15	26	24	12
P7	ATLL	chr11	31332995	-	38	DCDC1	-	-14902	-14668	8	9	3	8	3	4	13	15	4	9	8	1
LFE	ATLL	chr5	60952642	-	19	FALSE	NA	-16750	-4649	8	12	5	7	1	11	37	12	18	12	12	5

6 UISs out of 19 lie within +/- 10kb of a gene: Pr=32%
Pr random=19%

6 UISs out of 19 lie within +/- 10kb of a CpG island: Pr=32%
Pr random=15%

The mean count N=15
N random=16

The mean count N=24
N random=15

The first 6 columns are used as input for INSIPID and give for each UIS (each row), the patient code (from whom the blood sample is coming), the disease status, the position and orientation of the provirus (Chromosome, Position and Strand columns), and the UIS abundance.

The other columns are the output from INSIPID:

> "In Gene" column gives the name of the Gene when the provirus is inserted in it or FALSE if not. We meant here RefSeq gene.
>
> "inGene Orientation " column gives the orientation (+ or – strand) of the gene when applicable
>
> "nearest Gene Distance (bp)" column gives the distance in bp of the nearest gene (3' or 5' end)
>
> "nearest CpG Island Distance (bp)" column gives the distance of the nearest CpG island
>
> The columns "histoneX Count 10k" give the number of that particular histone mark in a 10kb window around the UIS. These data come from 2 resource papers:
>
> > Barski A, Cuddapah S, Cui K et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129:823-837.
> >
> > Wang Z, Zang C, Rosenfeld JA et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat.Genet. 2008;40:897-903.

Primer list

B3: 5' CCTTTCATTCACGACTGACTGCCG

B4: 5′ TCATGATCAATGGGACGATCA

P5B5: 5′ AATGATACGGCGACCACCGAGATCTACACTGGCTCGGAGCCAGCGACAGCCCAT

P5: 5′ AATGATACGGCGACCACCGAGAT

P7: 5′ CAAGCAGAAGACGGCATACGA

“upper arm” linker: 5'p-GATCGGAAGAGCGAAAAAAAAAAAAA

“lower arms” linker with different tag

LA1
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATCGTGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA2
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATGTCCTGCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA3
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATAGTGCACGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA4
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATTAGAGCCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA5
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATCTCTAGCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA6
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATACGGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA7
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATCGTACCCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA8
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATGGAATTCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA9
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATTATACGCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA10
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATGGGATTCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA11
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATCACACACGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA12
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATTGTGTGCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA13
5’TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATGTAGTACGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA14
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATAGGTCTCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA15
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATACAGGCCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA16
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATCTAGTCCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA17
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATCAGATCCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA18
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATCTGCTACGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA19
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATGTTCAGCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA20
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATGAGTTCCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA21
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATGCATAGCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA22
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATTCAGACCGGTCTCGGCATTCCTGCT
GAACCGCTCTTCCGATCT

LA23
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATTCTTGGCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA24
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATTGGTCCCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA25
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATATGATGCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

LA26
5'TCATGATCAATGGGACGATCACAAGCAGAAGACGGCATACGAGATTCGTGCCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

REFERENCES

1. R Development Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing; 2010.
2. Tamiya S, Matsuoka M, Etoh K et al. Two types of defective human T-lymphotropic virus type I provirus in adult T-cell leukemia. *Blood* 1996;88:3065–3073.
3. Ohshima K, Ohgami A, Matsuoka M et al. Random integration of HTLV-1 provirus: increasing chromosomal instability. *Cancer Lett*. 1998;132:203–212.
4. Kamihira S, Sugahara K, Tsuruda K et al. Proviral status of HTLV-1 integrated into the host genomic DNA of adult T-cell leukemia cells. *Clin.Lab Haematol*. 2005;27:235–241.
5. Gini C. Sulla misura della concentrazione e della variabilita dei caratteri. *Transactions of the Real Istituto Veneto di Scienze* 1914;LIII:1203.
6. Chao A. Species estimation and application. In: Balakrishman N, Read CB, Vidakovic B, eds. *Encyclopedia of Statistical Sciences*. Vol 12. New York: Wiley; 2005:7907–7916.