## Table of contents

- **Table S1**  Subfamilies recovered in this analysis and by Levy et al. [1]
- **Note S1**  Split of subfamilies between communities
- **Table S2**  Composition of chimpanzee communities D1 and D1b (θ 0.8)
- **Table S3**  Composition of human communities D1b, D1c and D1d (θ 0.9)
- **Figure S1**  Consensus sequence of the SINE-R of the sub-group of SVA_C elements split between communities 99 and 139.
- **Note S2**  Analysis of gorilla and orangutan short read archives to detect the SVA in the *NPLOC4* locus
- **Figure S2**  The SVA_D element in the human *NPLOC4* locus is surrounded by interspersed repeats.
- **Table S4**  Summary statistics of the “anchored” Blast search for the *NPLOC4* element in gorilla
- **Figure S3**  Multiple alignment of read pairs retrieved from gorilla “Kolo” short read archives
- **Additional References**
**Table S1**  Subfamilies recovered in this analysis and by Levy et al. [1]

<table>
<thead>
<tr>
<th>Species</th>
<th>subfamily</th>
<th>subfamily Levy et al. [1]</th>
<th>observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. gorilla</td>
<td>gg_SVA_DR</td>
<td>SVA_D1_Gg (+113G)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gg_SVA_D2a</td>
<td>SVA_D2_Gg</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gg_SVA_D5b</td>
<td>SVA_D3_Gg</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gg_SVA_D4c</td>
<td>SVA_D4_Gg</td>
<td></td>
</tr>
<tr>
<td>P. troglodytes</td>
<td>pt_SVA_DR</td>
<td>SVA_D1_Pt/ SVA_D1b_Pt*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pt_SVA_D2b</td>
<td>SVA_D2_Pt</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pt_SVA_D4a</td>
<td>SVA_D3_Pt</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pt_SVA_D6</td>
<td>SVA_D4_Pt</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pt_SVA_D5b</td>
<td>SVA_D5_Pt</td>
<td></td>
</tr>
<tr>
<td>H. sapiens</td>
<td>hs_SVA_DR</td>
<td>SVA_D1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hs_SVA_D2a</td>
<td>SVA_D2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hs_SVA_D3</td>
<td>SVA_D3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hs_SVA_D4</td>
<td>SVA_D4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hs_SVA_D5</td>
<td>SVA_D5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hs_SVA_DRb</td>
<td>SVA_D6</td>
<td></td>
</tr>
<tr>
<td>G. gorilla</td>
<td>gg_SVA_D1</td>
<td>D1a</td>
<td>all-hominid network</td>
</tr>
<tr>
<td></td>
<td>gg_SVA_DRb</td>
<td>(SVA_E)</td>
<td>as part of E in all-hominid network</td>
</tr>
<tr>
<td></td>
<td>gg_SVA_DRc</td>
<td>D6</td>
<td>as part of D6 in all-hominid network</td>
</tr>
<tr>
<td>P. troglodytes</td>
<td>pt_SVA_D1</td>
<td>D1a</td>
<td>all-hominid network</td>
</tr>
<tr>
<td></td>
<td>pt_SVA_DRd</td>
<td>part of D4</td>
<td>all-hominid network</td>
</tr>
<tr>
<td></td>
<td>pt_SVA_D3</td>
<td>D3</td>
<td>all-hominid network</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>hs_SVA_D1</td>
<td>D1a</td>
<td>Θ 0.9</td>
</tr>
<tr>
<td></td>
<td>hs_SVA_D4</td>
<td>D4</td>
<td>all-hominid network</td>
</tr>
<tr>
<td></td>
<td>hs_SVA_D6a-f</td>
<td>D1d</td>
<td>recovered at Θ 0.9 as part of D1d; in all-hominid network as separate SF in D1d/Pt_SVA_D5; no split into constituent SFs</td>
</tr>
<tr>
<td></td>
<td>hs_SVA_D7</td>
<td>D1d</td>
<td>combined in 69-member-D1d community 170 in all-hominid network</td>
</tr>
<tr>
<td></td>
<td>hs_SVA_D8</td>
<td>D1d</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hs_SVA_Fa</td>
<td>SVA_F_1</td>
<td>Θ 0.9; not recovered in all-hominid network</td>
</tr>
</tbody>
</table>

The upper part of the table lists subfamilies for which consensus sequences are provided by Levy et al. [1]; the lower part lists subfamilies for which identity was established by analysis of the data provided in the supplementary information to Levy et al. [1]. SF – subfamily; * subfamilies with identical consensus sequences.
Note S1  Split of subfamilies between communities

Communities identified by the network-based approach [1] were re-analyzed by retrieving the sequences of the elements based on the positions provided in the supplementary material, aligning them and sorting them into the subfamilies identified by diagnostic substitution-based sorting. Chimpanzee communities D1 and D1b (θ 0.8; Fig. 2 in [1]) both contain members of subfamilies pt_D1, pt_DRa-d and pt_D3 (Table S2).

Table S2  Composition of chimpanzee communities D1 and D1b (θ 0.8)

<table>
<thead>
<tr>
<th>SF (Levy)</th>
<th>SF</th>
<th>Total nr</th>
<th>pt_SVA_D1</th>
<th>pt_SVA_DRa</th>
<th>pt_SVA_DRb</th>
<th>pt_SVA_DRc</th>
<th>pt_SVA_DRd</th>
<th>pt_SVA_D3</th>
<th>pt_SVA_D5a</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1_Pt</td>
<td></td>
<td>328</td>
<td>66</td>
<td>10</td>
<td>27</td>
<td>5</td>
<td>4</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>D1b_Pt</td>
<td></td>
<td>243</td>
<td>45</td>
<td>9</td>
<td>40</td>
<td>5</td>
<td>5</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

Table S3 shows the composition of human communities D1b, D1c and D1d identified by Levy et al. at higher resolution (θ 0.9 [1]).

Table S3  Composition of human communities D1b, D1c and D1d (θ 0.9)

<table>
<thead>
<tr>
<th>SF (Levy)</th>
<th>SF</th>
<th>Total nr</th>
<th>hs_DRa</th>
<th>hs_D7</th>
<th>hs_DRa</th>
<th>hs_DRc</th>
<th>hs_D6a</th>
<th>hs_D6b</th>
<th>hs_D6c</th>
<th>hs_D6d</th>
<th>hs_D6e</th>
<th>hs_D6f</th>
<th>hs_SVA_D5b/D5a</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVA_D1b</td>
<td></td>
<td>65</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1D5</td>
</tr>
<tr>
<td>SVA_D1c</td>
<td></td>
<td>242</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1D5b</td>
</tr>
<tr>
<td>SVA_D1d</td>
<td></td>
<td>353</td>
<td>5</td>
<td>7</td>
<td>18</td>
<td>3</td>
<td>26</td>
<td>73</td>
<td>24</td>
<td>27</td>
<td>32</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

The last example of subfamilies split between communities is found in the all-hominid network where a clearly defined sub-group in C is split between communities 99 and 139. Figure S1 shows the consensus sequence for this subgroup in comparison to SVA_C. 31 members of the sub-group are found in community 99 and 32 in community 139.
Figure S1  Consensus sequence of the SINE-R of the sub-group of SVA_C elements split between communities 99 and 139. Variant residues are highlighted in yellow.

SVA_C      CAGCTCATTGAGAACGCGCATGATGAGCCGCTTTTGTGAAATAGAAAGGGGAAATGTGGGAAAAGATAGAGA
C_variant .............................................

SVA_C      AATCAGATTGTTGCTGTGTCTGTGTAGAAAGAAGTAGACATAGGAGACTCCATTTTTGCTCTTTACTAAGATAGATCTCTTC
C_variant .............................................

SVA_C      TGCCCTGGGATGCGTGTTGATCTGACCTTACACCAACCCCTCGCTCTGAAACTAGTGCTTGCCCTGACTCAGGGTTA
C_variant .............................................

SVA_C      AATGGATTAAGGGCGTGCAAGATGTGCTTTGTTAAACAGATGCTTTTGAGGACATGCTCGTATAGAAGATCATTACACTCC
C_variant .............................................

SVA_C      CCCTAATCTCAAGTACCAACCAAGGACACAAACTGCGGAAGGCCGCAGGGTCCTCTGGAAACACAGACCTTCTGT
C_variant .............................................

SVA_C      TCACCTTGTTATCTGCTGACCTCCTCCCTCCACTATTTGCTCTGCCCCTGCAAAACCCCTCTGCGAGAACACCAAGA
C_variant .............................................

SVA_C      ATGAT
C_variant .....
Analysis of gorilla and orangutan short read archives to detect the SVA in the NPLOC4 locus

Levy et al. [1] identified 13 SVA 3’ (SINE-R) read pairs with “one arm mapping uniquely to the flanking regions and the other to an SVA element” from gorilla short read archives by mapping the datasets to the human sequence flanking the SVA insertion. The analysis of the locus using short read archives is complicated by the fact that the target site is completely surrounded by repetitive sequence (mostly Alu): 42% of the sequence flanking the SVA at the 5’ end are repeat-masked; the 1kb downstream completely consist of interspersed repeats – except for the 3’ most section (Fig. S2). Human and gorilla sequences in the region 1kb downstream of the SVA insertion site are 98% identical (excluding gaps). If only the relevant 600bp (if one arm of the read should match the SVA and the other one the flanking region then, considering fragment lengths in the libraries, the 3’ read can map at maximum 600bp downstream) are taken into account, identity between human and gorilla is 97%. In Blast searches performed to detect the SVA in the gorilla NPLOC4 locus hits with less than 98% identity to the query were in general found to match to other Alus in the genome by BLAT. Thus, a read with a 100% match to the human sequence is likely to map to a different position in the gorilla genome.

To avoid false-positives resulting from retrieval of reads mapping to highly similar Alu sequences elsewhere in the genome I took advantage of the fact that approximately 400bp upstream of the integration site there is a stretch of non-repetitive sequence. I used this stretch to “anchor” a Blast search by using a query overlapping the non-repetitive sequence and the 5’ end of the AluJb. Provided that fragment length in the libraries is >350bp this approach should recover read pairs with the 5’ read “anchored” in non-repetitive sequence and the 3’ read localizing to the SVA – if it is present.

Paired-end short read archives are available for 26 gorilla individuals. For 18 of these individuals the average library fragment length is larger than 350bps. However, in none of the datasets a 3’ read covering SVA sequence could be identified, although the insertion site is covered: reads overlapping the empty
target site were recovered for all individuals; in some cases also 3’ reads mapping downstream of it could be identified – which, given library fragment size, would be impossible if the SVA is present. In parallel short read archives of 11 orangutans were analyzed using the same approach. In all of them all 3’ reads extending past the TSD contained SVA sequence matching that of the orangutan SVA_A (Table S4). No 3’ reads corresponding to the empty pre-integration site were recovered. For the remaining 8 gorilla individuals with library fragment length smaller 350bp Blast search was performed using a query spanning 100bp each up- and downstream of the target site. Mapping of the read pairs to the NPLOC4 locus was verified using BLAT. Blast hits below 98% identity to the query were found to match to other loci in the genome and discarded. For the 8 individuals between 5 and 17 reads confirming the absence of the SVA (i.e. matching the empty pre-integration site) were identified. No evidence was found for an SVA inserted into the target site. However, this approach is more likely not to detect the SVA-containing allele in a Blast search as longer hits with less similarity matching other Alu sequences in the genome might outnumber shorter hits matching only part of the query (the other part being SVA sequence) with higher similarity at a given maximum number of hits to return. Consequently, the shorter hits might not be reported.
### Table S4  Summary statistics of the “anchored” Blast search for the NPLOC4 element in gorilla

<table>
<thead>
<tr>
<th>individual</th>
<th>total nr of read pairs</th>
<th>3’ read 5’ of TSD</th>
<th>3’ read overlapping TSD</th>
<th>3’ read 3’ of TSD</th>
<th>average fragment length*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAROLYN</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>475</td>
</tr>
<tr>
<td>DIAN</td>
<td>13</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>524</td>
</tr>
<tr>
<td>MIMI</td>
<td>9</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>451</td>
</tr>
<tr>
<td>SANDRA</td>
<td>10</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>487</td>
</tr>
<tr>
<td>KATIE</td>
<td>11</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>426</td>
</tr>
<tr>
<td>DOLLY</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>464</td>
</tr>
<tr>
<td>OKO</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>376</td>
</tr>
<tr>
<td>HELEN</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>422</td>
</tr>
<tr>
<td>VILA</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>480</td>
</tr>
<tr>
<td>PORTA</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>388</td>
</tr>
<tr>
<td>ANTHAL</td>
<td>11</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>416</td>
</tr>
<tr>
<td>PAKI</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>436</td>
</tr>
<tr>
<td>CHOOMBA</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>478</td>
</tr>
<tr>
<td>AMANI</td>
<td>11</td>
<td>0</td>
<td>7</td>
<td>4</td>
<td>500</td>
</tr>
<tr>
<td>KOLO</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>485</td>
</tr>
<tr>
<td>COCO</td>
<td>15</td>
<td>0</td>
<td>11</td>
<td>4</td>
<td>473</td>
</tr>
<tr>
<td>DELPHI</td>
<td>17</td>
<td>3</td>
<td>4</td>
<td>10</td>
<td>483</td>
</tr>
<tr>
<td>BANJO</td>
<td>9</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>485</td>
</tr>
<tr>
<td>SUZIE</td>
<td>10</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>319</td>
</tr>
<tr>
<td>AZIZI</td>
<td>10</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>270</td>
</tr>
<tr>
<td>BULERA</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>289</td>
</tr>
<tr>
<td>AKIBA</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>411</td>
</tr>
<tr>
<td>TZAMBO</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>210</td>
</tr>
<tr>
<td>ABE</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>186</td>
</tr>
<tr>
<td>KOKOMO</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>264</td>
</tr>
<tr>
<td>KOWALI</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>284</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>individual</th>
<th>total nr of read pairs</th>
<th>3’ read 5’ of TSD</th>
<th>3’ read overlapping TSD</th>
<th>3’ read in SVA</th>
<th>average fragment length</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. abelii</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELSI</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>481</td>
</tr>
<tr>
<td>KIKI</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>518</td>
</tr>
<tr>
<td>DUNJA</td>
<td>10</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>454</td>
</tr>
<tr>
<td>BABU</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>450</td>
</tr>
<tr>
<td>BUSCHI</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>445</td>
</tr>
<tr>
<td>P. pygmaeus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAPOLEON</td>
<td>9</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>462</td>
</tr>
<tr>
<td>NONJA</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>476</td>
</tr>
<tr>
<td>SARI</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>493</td>
</tr>
<tr>
<td>TEMMY</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>405</td>
</tr>
<tr>
<td>TILDA</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>447</td>
</tr>
<tr>
<td>P. tapanuliensis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>28</td>
<td>2</td>
<td>1</td>
<td>326</td>
</tr>
</tbody>
</table>

* calculated based on the read pairs analyzed
Figure S3  Multiple alignment of read pairs retrieved from gorilla “Kolo” short read archives

The reference sequence on top of the alignment is composed of the SVA_D SINE-R consensus and the 3’ flanking sequence found in the gorilla genome. The extension of the read provided by Levy et al. [1] (Kolo_read) is shown in yellow; the extension of the corresponding 3’ reads in grey. Lower case indicates low sequence quality.

Additional References