Supplementary Table S5. Polymorphism and divergence in hominids at homologous positions in the 1'002 bp sequence including the *NAT* pseudogene (*NATP*).

The screened sequence spans from 18'228'116 to 18'229'117 on chromosome 8 in the human reference sequence GRCh37/hg19. The 77 substitutions corresponding to inter-species divergence (between two species at least) are highlighted by stars to the left part of the table. Polymorphisms shared between species or subspecies are highlighted in blue font. Deletions are indicated by del; (N) indicates that the position is undefined in some samples. Variants also recorded in sequenced ancient genomes of hominids (Denisova, Neanderthal or 45'000 years old Homo sapiens from Ust'-Ishim) are boxed (see Supplementary Table S8).

	Position			Common chimpanzee[2]					Gorilla[4]		Orangutan[5]	
	human reference sequence	SNP ID in Ensembl	Human[1]	Western	Nigeria-Cameroon	Eastern	Central	Bonobo[3]	*Gorilla gorilla*	*Gorilla beringei*	*Pongo abelii*	*Pongo pygmaeus*
1	18228116	rs10088180	A/G	A	A	A	A	A	A	A	A	A
2	18228117		T	T	T	T	T	T	T	T	T/A(N)	T/A
* 3	18228126		C	C	C	C	C	C	C(N)	N	T	T
4	18228146		G	G	G	G/T	G	G	G	G	G	G
5	18228151	rs190040562	T/G	T	T	T	T	T	T	T	T	T
6	18228159		C	C	C	C	C	C	C	C	C	C/A
7	18228162	rs34107584	G/del	G	G	G	G	G	G	G	G	G
8	18228172	rs143635794	A/G	G	G	G	G	G	G	G	G	G
9	18228189		C	C	C	C	C/T	C	C	C	C	C
10	18228213		A	A	A	A	A	A	A	A	A/T	N
* 11	18228214		G	G	G	G	G	G	C	C	G	G
* 12	18228224		C	C	C	C	C	C	C	C	T	T
13	18228238		A	A/G	A	A	A	A	A	A	A	A
* 14	18228242		A	A	A	A	A	G	A	A	A	A
15	18228246	rs73590295	T/C	T	T	T	T	T	T	T	T	T
* 16	18228256		A	A	A	A	A	A	A	A	G	G
17	18228267	rs181628748	C/T	C	C	C	C	C	C	C	C	C
18[6]	18228277	rs549576704	A/AT	A	N	A(N)	A(N)	A(N)	A(N)	A(N)	A(N)	N
19[7]	18228278		T	del(N)	N	del(N)	del(N)	del(N)	del(N)	N	del(N)	N
* 20	18228279		T	T(N)	N	T(N)	T(N)	C(N)	T(N)	N	T(N)	N
* 21	18228285		T/A	T	T/A(N)	T/A(N)	T/A	A	T	T	T	T
* 22	18228286	rs546046491	T	A	A	A	A	A	A/T	T	A/G(N)	A
23	18228291	rs2898473	C/A	A	A	A	A	A	A	A	A	A
24	18228294	rs527820494	T/C	T	T	T	T	T	T	T	T	T
* 25	18228297		A	A	A	A	A	A	A	A	G	G

	26	18228303		A	A	A	A	A	A	A	A	A/T	A/T
*	27	18228304		C	C	C	C	C	T	C	C	C	C
	28	18228306		C	C	C	C	C	C	C	C	C/T	C
*	29	18228315		G	G	G	G	G	G	G	G	T	T
*	30	18228320		T	G	G	G	G	G	T	T	T	T
*	31	18228327		G	G	G	G	G	G	G	G	C	C
	32	18228329	rs547876229	G/A	G	G	G	G	G	G	G	G	G
*	33	18228332		G	G	G	G	G	G	G	G	A	A
*	34	18228335		C	C	C	C	C	C	C	C	G	G
*	35	18228336		G	A	A	A	A	A	A	A	A	A
*	36	18228367		C	T	T	T	T	T	C	C	C	C
	37	18228368		T	T/C	T	T	T/C	T	T	T	T	T
	38	18228373		G	G	G	G	G	G	G	G	G/T(N)	G/T
*	39	18228382		A	G	G	G	G	G	G	G	G	G
*	40	18228385		G	G	G	G	G	G	G	G	A	A
	41	18228389	rs561082251	G/A	G	G	G	G	G	G	G	G	G
*	42	18228392		A	T	T	T	T	T	T	T	T	T
*	43	18228395		G	G	G	G	G	G	A	A	G	G
	44	18228404		C	C	C/T	C	C	C	C	C	C	C
	45	18228405	rs530061116	C/T	C	C	C	C	C	C	C	C	C
	46	18228409		G/A	G	G	G	G	G	G	G	G	G
*	47	18228418		A	A	A	A	A	A	A	A	G	G
*	48	18228419		T	C	C	C	C	C	T	T	T	T
*	49	18228420		A	A	A	A	A	A	A	A	G	G
*	50	18228424	rs184770517	C/T	T	T	T	T	T	C	C	T	T
	51	18228425	rs570049254	G/A	G	G	G	G	G	G	G	G	G
*	52	18228437		C	C	C	C	C	C	A	A	C	C
	53	18228442	rs538826711	T/A	T	T	T	T	T	T	T	T	T
*	54	18228458	rs372738250	G/A	A	A	A	A	A	G/A	G	G	G
*	55	18228463		C	C	C	C	C	C	T	T	T	T
*	56	18228470		T	T	T	T	T	T	A	A	T	T
*	57	18228480		C	C	C	C	C	C	C	C	T	T
	58	18228483	rs188523053	A/T	A	A	A	A	A	A	A	A	A
	59	18228501		C	C/T	C/T	C/T	C/T	C	C	C	C	C
	60	18228502	rs534373318	C/T	C	C	C	C	C	C	C	C	C
*	61	18228519		C	T	T	T	T	T	T	T	T	T
	62	18228543		A	A	A/G	A	A	A	A	A	A	A
	63	18228545	rs554671092	C/T	C	C	C	C	C	C	C	C	C
*	64	18228552		A	A	A	A	A	A	A	A	G	G
*	65	18228560		C	C/A	C	C	C	C	C	C	T	T
	66[8]	18228571	rs180998104	C/T	C(N)	C(N)	C(N)	C(N)	C	C	C	C(N)	C

*	67	18228574		G	G(N)	G(N)	G(N)	G(N)	G	G	A	G(N)	G
*	68	18228575		A	A(N)	A(N)	A(N)	A(N)	G	A	A	A(N)	A
*	69	18228576		G	G(N)	G(N)	G(N)	G(N)	A	G	G	G(N)	G
*	70	18228580		G	G(N)	G(N)	G(N)	G(N)	G	C	C	C/G(N)	G
	71	18228582		G	G/T	G	G	G	G	G	G	G	G
*	72	18228606		C	C	C	C	C	C	T	T	T	T
	73	18228608	rs537077588	T/G	T	T	T	T	T	T	T	T	T
	74	18228612		A	A	A	A	A	A	A	A	A/T	A
	75	18228614		G	G/T	G/T	G/T	G/T	G	G	G	G	G
	76	18228616	rs35548819	T/C	T	T	T	T	T	T	T	T	T
*	77	18228622		C	C	C	C	C	C	T	T	T(N)	T
*	78	18228628		T	T	T	T	T	T	G	G	T	T
	79	18228636	rs577255154	A/C	A	A	A	A	A	A	A	A	A
*	80	18228639		T	A	A	A	A	A	A	A	A	A
*	81	18228658		C	C	C	C	C	C	C	C	T	T
	82	18228659		G	G	G	G/A	G	G/A	G	G	G	G
*	83	18228660		G	C	C	C	C	T	G	G	G	G
	84	18228661	rs546009408	G/A	G	G	G	G/A	G	G	G	G	G
*	85	18228666		A	C	C	C	C	C	T	T	C	C
	86	18228669	rs73590297	G/C	G	G	G	G	G	G(N)	N	G	G
	87[9]	18228673	rs115350875	T/C	T(N)	N	T(N)	T(N)	T(N)	T/C(N)	T(N)	T(N)	N
	88	18228675		A/AA	A(N)	N	A(N)	A(N)	A(N)	A(N)	N	A(N)	N
	89	18228676	rs200058745	C/CA	C(N)	N	C(N)	C(N)	C(N)	C(N)	N	C(N)	N
*	90	18228678		A	A(N)	N	A(N)	A(N)	A(N)	ACAACAAA(N)	N	ACA/AAA(N)	N
*	91	18228679		A	del(N)	N	del(N)	del(N)	del(N)	A(N)	N	A/del(N)	N
*	92	18228685		A	C(N)	N	C(N)	C(N)	C	C(N)	N	A	A
*	93	18228686		T	C(N)	N	C(N)	C(N)	C	C(N)	N	T	T
*	94	18228701		C	T(N)	N	T(N)	T(N)	T	T	T	T(N)	T
	95	18228713		C	C	C	C	C	C	C	C	C/T	C/T
	96	18228716	rs112556119	T/C	T	T	T	T	T	T	T	T	T
	97	18228721	rs190996104	G/T	G	G	G	G	G	G	G	G	G
	98	18228727	rs530022558	G/A	G	G	G	G	G	G	G	G/A	G
*	99	18228732		T	T	T	T	T	T	T	T	C	C
	100	18228736		A	A	A	A	A	A	A	A	A/G	A
	101	18228738	rs543440451	T/G	T	T	T	T	T	T	T	T	T
	102	18228746		T	T	T	T	T	T	T	T	T/C	T
*	103	18228771		T	C	C/T	C/T	C/T	C	T	T	T	T
	104	18228783		A	A	A	A	A	A	A	A	A/G(N)	A
	105	18228788		G	G	G	G	G	G	G/C	G	G	G
*	106	18228810		T	T	T	T	T	T	T	T	G	G

	107	18228829		T/C	T	T	T	T	T	T	T	T	T
*	108[11]	18228838		A	A	A	A	A	A	A/G(N)	G(N)	A	A
*	109[11]	18228841		T	T	T	T	T	T	T/G(N)	G(N)	T	T
	110	18228851	rs73666630	T/C	T	T	T	T	T	T	T	T	T
*	111[11]	18228852		C	C	C	C	C	C	C/T(N)	T(N)	C	C
*	112	18228853		T	T	T	T	T	T	T	T	C	C
*	113	18228854		G	A	A	A	A	A	A	A	A	A
*	114	18228857		A	T	T	T	T	T	A	A	A	A
*	115	18228867		T	T	T	T	T	T	C	C(N)	C	C
	116	18228869		T/G	T	T	T	T	T	T	T	T	T
*	117[12]	18228876		C	C	C	C	C	C	G	G(N)	G(N)	N
	118	18228878		T	T	T	T	T	T	A/T(N)	T	T(N)	N
	119	18228880	rs532708242	G/C	G	G	G	G	G	G	G	G(N)	N
*	120	18228886		T	T	T	T	T	T	T	T	del(N)	N
*	121	18228889		A	A	A	A	A	A	A	A	G	G
*	122	18228917		G	G	G	G	G	G	T	T	T	T
*	123	18228927		T	T	T	T	T	T	T	T	C	C
*	124	18228939		A	T	T	T	T	T	T	T	T	T
*	125	18228943		T	T	T	T	T	T	T	T	C	C
*	126	18228944		T	C	C	C	C	C	C	C	T	T
*	127	18228949		G	G	G	G	G	G	A	A	A	A
	128	18228959		T	T	T/C	T	T	T	T/C(N)	T	T	T
*	129	18228960		G	A	A	A	A	A	A	A	G	G
	130	18228967		C	C	C	C	C	C	C	C	C	C/T
	131	18228971		G	G	G	G	G	G	G/A(N)	G	G	G
	132[13]	18228999	rs145709578	G/C	G	G	G	G	G	G	G	G(N)	N
*	133[11]	18229005		A	A	A	A	A	A	A/C(N)	C	A	A
*	134	18229011		A	A	A	A	A	A	G	G	A	A
*	135	18229016		T	T	T	T	T	T	C	C	T	T
	136	18229020	rs2172426	C/T	T	T	T	T	T	T	T	T	T
	137	18229034	rs528699467	C/A	C	C	C	C	C	C	C	C	C
	138	18229048		A	A	A	A	A	A	A/G(N)	A	A	A
*	139	18229050		C	T	T	T	T	T	T	T	T	T
	140	18229051	rs565728206	G/A	G	G	G	G	G	G	G	G	G
*	141	18229057		C	G	G	G	G	A	C	C	C	C
*	142	18229076		C	C	C	C	C	C	C	C	T	T
	143	18229077		C	C	C	C	C	C	C	C	C(N)	C/T
	144	18229085		T	T	T	T	T	T	T/A	T	T	T
*	145	18229087		T	C	C	C	C	C	T	T	T	T
	146	18229098		A/G	A	A	A	A	A	A	A	A	A

	147	18229103		A	A	A	A	A/T	A	A	A	A	A
	148	18229104	rs74444655	T/C	T	T	T	T	T	T	T	T(N)	T
	149	18229105		A/C	A	A	A	A	A	A	A	A(N)	A
*	150	18229106		T	T	T	T	T	T	T	T	C	C
	151	18229111		A	A	A	A	A	A	A	A	A/G	A

[1] Human polymorphisms are those recorded by the consensus gene nomenclature of human *NAT* alleles (http://nat.mbg.duth.gr/), complemented with haplotype data from 1000 Genomes Phase 1 data (The Genomes Project Consortium 2012), (Patin et al. 2006a) and (Mortensen et al. 2011), as well as 1000 Genomes SNPs and indels reported for GRCh37.p13 (hg19) in Ensembl (Yates et al. 2016) (http://grch37.ensembl.org/Homo_sapiens/Info/Index, accessed October 2016). Human positions not recorded as polymorphic but associated with a SNP identifier are reported with a highest population MAF < 0.01 in Ensembl (http://www.ensembl.org/Homo_sapiens/Info/Index). Note that the list is not exhaustive: a polymorphic position (A/G) at 18'228'182 is reported for Denisova in the ancient genome browser of the Max Planck Institute for Evolutionary Anthropology (see Supplementary Table S8); it is not recorded in the table since all humans and other hominoids are fixed on A, and an additional polymorphic position (T/C) at 18'228'748, which defines the *Pan NATP*13* haplotype, is not recorded in the table either as it was only observed in the hybrid *P. t. verus/troglodytes* individual (all humans and other hominids are fixed on T).

[2] Western, Niger-Cameroon, Eastern and Central chimpanzees are *Pan troglodytes verus*, *P. t. ellioti*, *P. t. schweinfurthii* and *P. t. troglodytes*, respectively. Polymorphism recording is based on the individuals of the present study (Supplementary Figure S1) and the chimpanzees of Prado-Martinez et al. (2013) cross-checked with the *P. t. verus* assembly reference sequence (panTro4, February 2011).

[3] Based on the individual of this study (Bonobo), the bonobos of Prado-Martinez et al. (2013) and the *Pan paniscus* draft assembly reference sequence (panPan1, May 2012).

[4] Based on the individuals of this study, the gorillas of Prado-Martinez et al. (2013) and the *Gorilla gorilla gorilla* draft assembly reference sequence (gorGor4, December 2014).

[5] Based on the individuals of this study, the orangutans of Prado-Martinez et al. (2013) and the *Pongo pygmaeus abelii* draft assembly reference sequence (ponAbe2, July 2007).

[6] All individuals from the *Pan* (*troglodytes* and *paniscus*) *Gorilla* and *Pongo* species and subspecies sequenced by Prado-Martinez et al. (2013) have undefined positions in the stretch between 18'228'271 and 18'228'282 (included). For this reason, positions in the stretch 18'228'271-18'228'282 were not considered for median-joining network construction.

[7] We speculate that a T insertion at position 18'228'278 probably occurred on the human lineage, as it was absent in all Sanger sequenced samples of chimpanzees, bonobos, gorillas and orangutans in this study (not represented in the median-joining network, see footnote 6).

[8] All individuals from the *Pan troglodytes* and *Pongo abelii* species and subspecies sequenced by Prado-Martinez et al. (2013) have undefined positions in the stretch between 18'228'567 and 18'228'581 (included). For this reason, positions in the stretch 18'228'567-18'228'581 were not considered for median-joining network construction.

[9] From positions 18'228'670 onwards, all individuals from the *Pan* (*troglodytes* and *paniscus*) and *Pongo abelii* species and subspecies sequenced by Prado-Martinez et al. (2013) have undefined positions: up to 18'228'701 (included) for *P. troglodytes*, up to 18'228'681 (included) for *P. paniscus* and up to 18'228'682 (included) for *Pongo abelii*. For *Gorilla*, positions 18'228'668 and 18'228'669 are undefined, as well as all positions between 18'228'672 and 18'228'689 (included). For *P. pygmaeus*, all positions between 18'228'672 and 18'228'682 (included) are undefined. Polymorphisms in this region were detected for the individuals sequenced in this study. However,

these are all indel polymorphisms of repetitive motives (CA, CAA, etc.) and we chose not to consider the stretch running from 18'228'670 to 18'228'701 for median-joining network construction.

[10] At position 18'228'678, an insertion of a CAACAAA motif was observed in all Sanger sequenced gorillas, and an insertion of a CA/AA polymorphic motif in all Sanger sequenced orangutans. In the latter, a T deletion was also found in all Sanger sequenced individuals at position 12'228'886 (not represented in the median-joining network, see footnote 9).

[11] Position 18'228'838 is polymorphic (A/G) in *Gorilla gorilla*, but not in *Gorilla beringei*, for which only nucleotide G (or N) is reported. At this position, the gorilla reference sequence (gorGor4) indicates nucleotide A, similarly to the human, chimpanzee, bonobo and orangutan reference sequences (hg19, panTro4, panPan1, ponAbe2). The same ambiguous pattern is found at positions 18'228'841, 18'228'852 and 18'229'005, although for the latter two, the gorilla reference sequence differs from that of the other species (at position 18'228'852, gorGor4 is T, whereas it is C in all other reference sequences, and at position 18'229'005, gorGor4 is C, whereas it is A in all other reference sequences). Thus none of these four positions, 18'228'838, 18'228'841, 18'228'852 and 18'229'005, were considered for median-joining network construction.

[12] All *Pongo* individuals sequenced by Prado-Martinez et al. (2013) have undefined positions between 18'228'876 and 18'228'887 (included). For this reason, positions in the stretch 18'228'876-18'228'887 were not considered for median-joining network construction.

[13] All *Pongo* individuals sequenced by Prado-Martinez et al. (2013) have undefined positions between 18'228'993 and 18'229'003 (included). For this reason, positions in the stretch 18'228'993-18'229'003 were not considered for median-joining network construction.

References

Mortensen HM, Froment A, Lema G, Bodo JM, Ibrahim M, Nyambo TB, Omar SA, and Tishkoff SA. 2011. Characterization of genetic variation and natural selection at the arylamine N-acetyltransferase genes in global human populations. Pharmacogenomics 12(11):1545-1558.

Patin E, Barreiro LB, Sabeti PC, Austerlitz F, Luca F, Sajantila A, Behar DM, Semino O, Sakuntabhai A, Guiso N et al. . 2006. Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. American journal of human genetics 78(3):423-436.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G et al. . 2013. Great ape genetic diversity and population history. Nature 499(7459):471-475.

The Genomes Project C. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L et al. . 2016. Ensembl 2016. Nucleic Acids Res 44(D1):D710-D716.