

Vangenot et al. (2019)

Humans and chimpanzees display opposite patterns of diversity in *arylamine N-acetyltransferase* genes

Supplementary File S1, containing :

Supplementary information on Materials and Methods

1. Description of the great apes DNA samples sequenced in this study
2. DNA amplification, PCR product purification and sequencing of great ape samples
3. Retrieval of unphased *NAT* polymorphic positions from the Great Ape Genome Project (GAGP)
4. Inference of *Pan NAT* haplotypes
5. Retrieval of phased human *NAT* haplotypes from the 1000 Genomes Project
6. Constitution of Western chimpanzee (*P. troglodytes verus*) samples of unrelated individuals

Supplementary information on Results

7. *NAT* genotypes obtained by Sanger sequencing
8. *NAT* polymorphisms in *Pan*
9. Potential hybrid individual in the CARTA collection
10. In silico prediction tools of the functional impact of mutations in coding sequences
11. Evaluation of prediction adequacy of functional consequence for human haplotypes at the *NAT1* and *NAT2* loci
12. Prediction of enzymatic activity of *Pan NAT1* and *NAT2* basal haplotypes
13. Comparison of nucleotide diversity in gorillas and orangutans at the three *NAT* loci with *Pan* and humans

14. References

Supplementary information on Materials and Methods

1. Description of the great apes DNA samples sequenced in this study

Among the individuals of the Biomedical Primate Research Centre (BPRC) colony (Supplementary Figure S1A), 24 were the founders of the colony (in yellow, orange and blue), and even if their relatedness is unknown mitochondrial genetic data and family study suggest that they are unrelated (de Groot et al. 2005). The 20 other individuals (in green, grey and beige in Supplementary Figure S1A) are first level descendants related through five genealogies.

The Center for Academic Research and Training in Anthropogeny (CARTA) Western chimpanzees consist in 10 unrelated individuals and 16 individuals related through two separate genealogies (in purple in Supplementary Figure S1B). The two Basel zoo Western chimpanzees (Supplementary Figure S1C) are unrelated. Among these 68 Western chimpanzees, 21 are wild, 22 were born in captivity and the status of the remaining 25 is unknown. Other *Pan* samples from the CARTA collection include the bonobo male, and the Central (individual C327) and Eastern (individual Harriet) chimpanzee females, the latter being also part of the Great Ape Genome Project (see below). Seven of the orangutan and four of the gorilla samples are from the Basel zoo, and the other samples are from the CARTA collection.

2. DNA amplification, PCR product purification and sequencing of great ape samples

DNA amplification, PCR product purification and sequencing of samples from the CARTA research center were performed by Retrogen (San Diego, California), using their 3730 sequencers with ABI BD3.1 sequencing chemistry. Amplifications were done on 50 ng of genomic DNA, 10 pmol of each primer, 1 unit HoTaq DNA Polymerase (MCLAB), 10 X PCR buffer supplied with the polymerase, and dNTP mix, for a total final volume of 20 µl. For *NAT1* and *NATP*, 1 ul additional 50mM MgSO₄ was added. For *NAT2*, 2 ul of PCRx Enhancer Solution (Invitrogen) was added. For *NAT1* and *NAT2*, samples were denatured at 94°C for 10 min, followed by 40 cycles of

94°C for 30 s, 55°C for 1 min, and 72°C for 2 min. For *NATP*, samples were denatured at 95°C for 10 min, followed by 38 cycles of 94°C for 1 min, 57°C for 1 min, and 72°C for 2 min, followed by a final elongation phase at 72°C for 10 min.

DNA amplification, PCR product purification and sequencing of samples from the BPRC research center and from the Basel zoo were performed by Macrogen (Seoul, South Korea), using their Standard Seq platform with the same protocol as used by Retrogen, except for gorillas and orangutans for which specific primers were developed (Supplementary Table S1) from the reference genomes of gorilla (GorGor3.1, May 2011) and orangutan (ponAbe2, July 2007). Because of unsatisfactory results for some chimpanzee samples, PCR conditions were slightly modified in a second round: samples were denatured at 94°C for 5 min, followed by 40 cycles of 94°C for 30 s, 55°C for 1 min, and 72°C for 2 min, followed by a final elongation phase for 10 min. For two DNA samples (Oscar and Gerda) new primers were designed for *NAT2* (Supplementary Table S1) and PCR conditions were adapted as follows: samples were denatured at 95°C for 5 min, followed by 35 cycles of 95°C for 1 min, 45-68°C for 1 min, and 72°C for 1 min, followed by a final elongation phase for 10 min.

Note that for the sequencing of *NATP* in the orangutan samples, both primers pairs Locus3Forward-C/13405 and Locus3Forward-D/Locus3Reverse-D were used but provided less satisfactory results, due to little overlapping of the forward and reverse fragments, which could be due to the quality of the reference sequence (ponAbe2, July 2007) from which they were defined.

3. Retrieval of unphased *NAT* polymorphic positions from the Great Ape Genome Project (GAGP)

We retrieved *NAT* unphased genotypes of 79 great apes from the Great Ape Genome Project (GAGP, Prado-Martinez et al. 2013, data downloaded in March 2014, at <https://eichlerlab.gs.washington.edu/greatape/data/VCFs/SNPs/>). The sequencing coverage of genomes in GAGP varies between 7.40 and 49.81 (average = 27.24, sd = 11.17). We extracted, from the available unphased VCF files, the sections corresponding to the coding exon of each of the two functional *NAT* genes, and the

corresponding homologous section of the *NAT* pseudogene, namely (numbering of positions according to the human chromosome 8 reference sequence hg19/GRCh37):

- for *NAT1*, the homologous stretch spanning from positions 18'079'545 to 18'080'447 (coding exon spans 18'079'557 to 18'080'426),
- for *NAT2*, the homologous stretch spanning from positions 18'257'489 to 18'258'603 (coding exon spans 18'257'514 to 18'258'383),
- for *NATP*, the homologous stretch spanning from positions 18'228'116 to 18'229'117 (the stretch of homology with the coding exons of *NAT1* and *NAT2* spans 18'228'116 to 18'228'986).

The GAGP data retrieved represents 25 chimpanzees (4 *P. t. troglodytes*, 6 *P. t. schweinfurthii*, 10 *P. t. ellioti* and 4 *P. t. verus*, and one individual that is a *P. t. verus/troglodytes* hybrid), 13 bonobos, 31 gorillas (27 *Gorilla gorilla gorilla*, 3 *Gorilla beringei graueri*, 1 *Gorilla gorilla dielhi*) and 10 orangutans (5 *Pongo abelii* and 5 *Pongo pygmaeus*).

All detected polymorphic positions in the Sanger sequenced samples of this study and retrieved from the GAGP VCF files are detailed in Supplementary Tables S3, S4 and S5.

For the data retrieved from GAGP, we checked for missing positions in the bed files containing the regions that did not pass the quality filters applied to the SNP data (bed files downloaded in March 2014 at https://eichlerlab.gs.washington.edu/greatape/data/VCFs/SNPs/Callable_regions/):

- for *P. troglodytes* 0.55%, 2.2% and 6.5% positions were uncallable for the genes *NAT1*, *NAT2* and *NATP*, respectively,
- for *P. Paniscus*, 0.55%, 0.28% and 2.8% positions were uncallable for *NAT1*, *NAT2* and *NATP*, respectively,
- for gorillas, 1.33%, 2.94% and 5.5% were uncallable for *NAT1*, *NAT2* and *NATP*, respectively,
- and for orangutans, 0.33%, 0.46% and 6.8% were uncallable for *NAT1*, *NAT2* and *NATP*, respectively.

Therefore, any potential variant overlapping with these positions would not have been present in the VCF SNP files. However, according to our Sanger sequenced chimpanzees and bonobo samples, none of the uncallable positions in the *Pan* genomes from GAGP were located in known polymorphic positions for this genus.

We thus tentatively considered those positions as identical to the panTro4 and panPan1 reference sequences, respectively, unless specified otherwise in Supplementary Tables S3, S4 and S5. We were unable to do so for gorillas and orangutans, as variants were overlapping with some unknown positions. For this reason, as well as due to the small number of gorillas (5 individuals) and orangutans (eight individuals) that were Sanger sequenced in this study (see Materials and Methods in main text), haplotype inference at each of the 3 *NAT* genes was performed only for individuals from the *Pan* genus.

4. Inference of *Pan NAT* haplotypes

Diploid haplotypes were inferred for all *Pan* individuals using PHASE version 2.1.1 (Stephens et al. 2001; Stephens and Scheet 2005). For each of the 3 *NAT* genes, 3 PHASE runs were performed with individuals grouped differently: once considering all *Pan* individuals together; once separating *Pan troglodytes* and *Pan paniscus* individuals in two groups; and once additionally separating Western chimpanzees (*Pan troglodytes verus*) from the other *Pan troglodytes* sub-species, since the former sample outnumbers the latter, thus considering 3 groups: *Pan troglodytes verus*, other *Pan troglodytes* sub-species and *Pan paniscus*. For *NAT1* and *NAT2*, the same haplotypes were inferred in the 3 runs. For *NATP*, different sets of haplotypes were returned for each run. When considering all *Pan* together, 19 haplotypes were given, five of which were different from the 19 returned when separating *Pan troglodytes* and *Pan paniscus* (number of pairwise differences = 5.29 and 5.08, respectively). When the program was run with three groups, 23 haplotypes were given (number of pairwise differences = 4.55), six and four being different from those returned with the other two grouping (all *Pan* together and separating *Pan troglodytes* and *Pan paniscus*, respectively). The difference between grouping considering all *Pan* together and separating *Pan troglodytes* and *Pan paniscus* was explained by some undefined positions in *Pan troglodytes* from the GAGP at otherwise fixed positions in all other *Pan* (on different bases in *Pan troglodytes* and *Pan paniscus*). Consequently, we chose the set of haplotypes inferred when *Pan troglodytes* and *Pan Paniscus* individuals were considered separately.

5. Retrieval of phased human *NAT* haplotypes from the 1000 Genomes Project

The human *NAT* sequences dataset assembled in this study was completed with *NAT1*, *NAT2* and *NATP* phased genotypes retrieved from the 1000 Genomes Phase 1 dataset (Abecasis et al. 2012). We verified that no known related individuals were present in the dataset.

Similarly to the treatment of data from the Great Ape Genome Project, we extracted the relevant part of the available VCF files from the 1000 Genomes Project (positions 18'079'557-18'080'426, 18'257'514-18'258'383, and 18'228'116-18'228'986 in the chromosome 8 human reference sequence GRCh37/hg19 for, respectively, *NAT1*, *NAT2* and *NATP*). We checked for missing positions in the pilot mask (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/) and no uncallable positions were found.

6. Constitution of two Western chimpanzee (*P. troglodytes verus*) samples of unrelated individuals

We separated the Western (*P. t. verus*) chimpanzees according to their provenance to form two samples of unrelated individuals.

The first one, called BPRC, includes the 20 founders of the BPRC colony for which DNA was available (in yellow in Supplementary Figure S1A), as well as four additional founders whose genotypes could be unambiguously deduced from those of their children: Izaak's genotypes, deduced for all three *NAT* genes (in blue in Supplementary Figure S1A), and Gerrit's genotype, deduced for *NATP* only (in orange, in Supplementary Figure S1A). We also included two unrelated first level descendants in the BPRC sample of genotypes (Oscar, in beige in Supplementary Figure S1A, successfully genotyped for the three *NAT* genes, and Annaclara, in gray for *NAT1* and *NAT2* only).

The second sample, called San Diego, comprises individuals from the CARTA collection and the two chimpanzees from the Basel zoo. The CARTA *P. t. verus* sample includes a large genealogy of 14 inter-related individuals (Supplementary Figure S1B). Estimation of frequency distributions from the San Diego sample was done through a re-sampling procedure leading to 122 sub-samples of 18 unrelated

individuals, so as to account for the multiple possible ways of choosing unrelated individuals among those of the CARTA collection. Indeed, the maximum number of unrelated individuals that can be chosen from this large genealogy is 6, and 10 distinct sub-samples of 6 unrelated individuals can thus be formed. However, individuals C316 (Tamblo) and C318 (Reuuh) are not included in any of these 10 sub-samples. In turn, 61 distinct sub-samples of 5 unrelated individuals can be chosen from the large genealogy, and all 14 inter-related individuals from the large genealogy are included in at least one sub-sample. We have thus chosen to use these 61 possible sub-samples of 5 unrelated individuals to generate a collection of non-independent *P. t. verus* sub-samples. We further added to each sub-sample one individual chosen among the mother-son pair C320-C319 (Supplementary Figure S1B), thus doubling the number of possible sub-samples. Finally, we completed each sub-sample with the remaining 10 unrelated individuals from the CARTA collection and with the two unrelated individuals from the Basel zoo. The San Diego cohort is thus made up of 122 non-independent sub-samples of 18 individuals (36 chromosomes).

When pooling together the BPRC sample with each of the 122 San Diego sub-samples, Hardy-Weinberg equilibrium was rejected for the *NATP* pseudogene in 45 of those pooled samples (37%). Moreover, although the two *P. t. verus* samples were not found to differentiate significantly at *NAT1* and *NAT2* (see Results), they did so at *NATP* in 75% of the sub-samples (Supplementary Table S6). We thus decided to analyse the two samples separately.

Supplementary information on Results

7. NAT genotypes obtained by Sanger sequencing

We obtained 247 *NAT* genotypes out of the 84 DNA samples of great apes available for Sanger sequencing in this study. For *NAT1*, DNA samples from 83 individuals were successfully sequenced over a segment of 903 bp (i.e. all but one orangutan), for *NAT2*, 81 DNA samples over 1,115 bp (all but one Eastern chimpanzee, one gorilla and one orangutan), and for *NATP*, 83 DNA samples over 1,000 bp (all but one gorilla). Because of restricted DNA availability, we were unable to further adapt the conditions so as to repeat unsuccessful amplification or sequencing.

8. NAT polymorphisms in *Pan*

Both polymorphic positions and fixed segregating sites between hominids, including all the *Pan* species and sub-species, at the three *NAT* genes are reported in Supplementary Tables S3, S4 and S5. Segregating sites in the *Pan* genus are also shown in Table 2.

Ten polymorphic positions in total were observed for *NAT1* in *Pan*, all located in the exon open reading frame, six of them being non-synonymous (Table 2, Supplementary Table S3 and main text). In the *P. t. verus* subspecies, only five SNPs were found, three of which are non-synonymous: G76A (D26N), C147T, T369C, T597G (I199M) and A789G (I263M). The polymorphism at position 597 was only observed in *P. t. verus* whereas the two SNPs G76A and C147T were shared with all other *Pan troglodytes* sub-species; SNP A789G was also found in Central (*P. t. troglodytes*) and Eastern (*P. t. schweinfurthii*) chimpanzees, and SNP A369G was also shared with bonobos (*P. paniscus*). Besides the five SNPs observed in Western (*P. t. verus*) chimpanzees, four of which were also shared with other *Pan* species or sub-species, five additional SNPs were only observed among other *Pan* individuals: C303T only detected in bonobos, T341C only in Central (*P. t. troglodytes*), C458T and A518C only in Eastern (*P. t. schweinfurthii*), and G760C only in Nigeria-Cameroun (*P. t. ellioti*) chimpanzees. Three of these five additional polymorphisms involve a modification of the protein: T341C (I114T), A518C (E173A), and G760C (E254Q).

For *NAT2*, nine segregating sites were observed (seven in the coding exon and two in the 3'UTR), one of which G145A (E49K) was apparently fixed on the derived amino

acid (K) in *P. paniscus* (bonobos) (Table 2 and main text). Two out of the four polymorphisms observed among the 68 *P. t. verus* chimpanzees occur within the *NAT2* open reading frame and one involves a protein modification: C578T (T193M). Three out of the four polymorphisms detected among the 68 Western (*P. t. verus*) chimpanzees were not observed in other *Pan* individuals (i.e. polymorphisms at positions 578, 789 and 949), whereas SNP 934, outside the coding exon, was found among all *Pan troglodytes* but not in bonobos (*P. paniscus*). Actually, G934A is the only *NAT2* polymorphism found to be shared among all chimpanzee (*P. troglodytes*) sub-species. Four additional SNPs, all situated within the coding exon, were observed in the other *Pan* individuals: non-synonymous SNP A514G (N172D) was observed among Eastern (*P. t. schweinfurthii*), Nigeria-Cameroun (*P. t. ellioti*) and Central (*P. t. troglodytes*) chimpanzees, synonymous SNP T36C polymorphism was only detected in Eastern chimpanzees, and the 2 non-synonymous SNPs A72C (L24F) and G191A (R64Q) only in bonobos.

Among the 24 *NATP* segregating sites observed in *Pan*, 17 were observed to be polymorphic within at least one chimpanzee sub-species, but only a few were shared among species and sub-species (Table 2 and main text). However, three of these 24 segregating sites had undefined nucleotides in some genotypes retrieved from the GAGP (Supplementary Table S5). Thus, while four positions (18'228'242, 18'228'304, 18'228'660, and 18'229'057) were apparently fixed differences between *P. troglodytes* (chimpanzees) and *P. paniscus* (bonobos), divergence between the two species at three other positions (18'228'279, 18'228'575, and 18'228'576) needs to be confirmed. The 17 other segregating sites were observed to be polymorphic within at least one chimpanzee sub-species, but only a few were shared among species and sub-species. Two SNPs (T/C at position 18'228'501 and T/G at 18'228'614) were shared by all chimpanzee (*P. troglodytes*) sub-species but were not observed in bonobos (*P. paniscus*), and two SNPs (T/A at 18'228'285 and C/T at 18'228'771) were shared by all chimpanzee (*P. troglodytes*) sub-species but Western (*P. t. verus*) chimpanzees. Four other polymorphisms were found in Western chimpanzees (A/G at 18'228'238, C/T at 18'228'368, C/A at 18'228'560, and G/T at 18'228'582), three of which were not observed in the other *Pan*, whereas SNP at 18'228'368 was also observed in Central (*P. t. troglodytes*) chimpanzees. One position was found polymorphic both in Eastern (*P. t. schweinfurthii*) chimpanzees and bonobos (G/A at 18'228'659); it was the only polymorphic position observed in

bonobos. Other polymorphisms that were observed only in one chimpanzee sub-species include: SNPs C/T at 18'228'404, A/G at 18'228'543, and T/C at 18'228'959 in Nigeria-Cameroun (*P. t. ellioti*), SNPs C/T at 18'228'189, G/A at 18'228'661, and A/T at 18'229'103 in Central (*P. t. troglodytes*), and SNP G/T at position 18'228'146 in Eastern (*P. t. schweinfurthii*) chimpanzees. One of the polymorphic positions in chimpanzees was found apparently fixed on the derived allele in bonobos (SNP 18'228'285). Finally, the C/T SNPs at positions 18'228'404 (polymorphic in Nigeria-Cameroun chimpanzees) and 18'228'771 (polymorphic in Eastern, Central and Nigeria-Cameroun chimpanzees) were also observed in the single *verus/troglodytes* hybrid individual, whereas an additional SNP, C/T at position 18'228'748, was only detected in this individual.

9. Potential hybrid individual in the CARTA collection

Besides the individual identified as a Western/Central (*P. t. verus/troglodytes*) chimpanzee hybrid in the GAGP collection, we identified a potentially hybrid individual also in the CARTA collection. Indeed, female C327, classified as *P. t. troglodytes* in the CARTA collection, is the only carrier of *NAT2*1* and *NAT2*6* among its sub-species. Considering the relatively recent discovery of the Nigeria-Cameroun (*P. t. ellioti*) sub-species (Gonder et al. 1997; Gonder et al. 2006), and the hybridisation zone between this sub-species and Central (*P. t. troglodytes*) chimpanzees near the Sanaga river, female C327 could be a Nigeria-Cameroun chimpanzee, a Central/Nigeria-Cameroun hybrid, or a descendant of such a hybrid. At present, in the absence of other information, we kept this individual classified as a Central chimpanzee (*P. t. troglodytes*).

10. In silico prediction tools of the functional impact of mutations in coding sequences

Phenotypic predictions of the functional impact of specific mutations in *NAT1* and *NAT2* haplotypes were performed with three online software tools (analysis done May 2017: PolyPhen (Adzhubei et al. 2010), SIFT (Sim et al. 2012) and PANTHER cSNP Scoring tool (Tang and Thomas 2016). To evaluate the confidence in the results

returned by the 3 prediction software tools, we first applied them on human haplotypes of known enzymatic activity. For that, we selected human haplotypes with a single substitution compared to the reference haplotypes *NAT1**4 and *NAT2**4 (GenBank accessions X17059 and X14672) following the standards of the official nomenclature of human *NAT* alleles (McDonagh et al. 2014).

We ran PolyPhen with the default query options. PolyPhen returns three results, namely the score, which is the probability that a substitution is damaging, as well as values of sensibility and specificity, and an associated prediction that can be “benign”, “possibly damaging” and “probably damaging”. For SIFT, we used the default parameters and searched the Uniprot-SwissProt + TrEMBL 2010_09 database. SIFT returns a score, which is the probability that a substitution is tolerated, and which translates into a prediction (either “tolerated”, or “affects protein function” if SIFT score <0.05). It also returns the number of sequences used for the prediction at the substituted position (not counting sequences with gaps at that position) and the median sequence information used to measure the diversity of the sequences used for prediction (a median sequence information between 2.75 and 3.5 is recommended). When using the PANTHER cSNP Scoring tool, we selected *Homo sapiens* as reference organism. The PANTHER cSNP Scoring tool returns the position-specific evolutionary preservation (PSEP) index and an associated prediction. The PSEP measures the length of time a position in the current protein has been preserved (in millions of years) by tracing it back to its reconstructed direct ancestors. The associated prediction can be “probably damaging” (PSEP > 450my, corresponding to a false positive rate of ~0.2 as tested on HumVar), “possibly damaging” (450my > PSEP > 200my, corresponding to a false positive rate of ~0.4) and “probably benign” (PSEP < 200my).

11. Evaluation of prediction adequacy of functional consequence for human haplotypes at the *NAT1* and *NAT2* loci

The phenotypic predictions outputted by the three online software tools (PolyPhen, SIFT, and PANTHER cSNP Scoring) are provided in Supplementary Table S15.

For *NAT1*, the three tools appear to predict adequately the effect on enzymatic expression/activity of those substitutions whose effect is well established. Indeed,

haplotypes *NAT1*17* and *NAT1*22*, two well-defined and agreed decreased-activity haplotypes whose defining substitutions (C190T and A752T, respectively) reduce both protein level and N- and O-acetylation activity below detection (Hein et al. 2006), are predicted as damaging by the three tools (“probably damaging” for PolyPhen and PANTHER cSNP Scoring and “affect protein function” for SIFT). On the contrary, another substitution, G560A, characterising the decreased-activity haplotype, *NAT1*14B*, is predicted as “possibly damaging” by PolyPhen, “probably damaging” by PANTHER cSNP Scoring (but with a low PSEP) and as “tolerated” by SIFT. However, Zhu and Hein (2008) showed that G560A reduces protein levels more moderately compared to C190T and A752T (4-fold compared to 50-fold and 40-fold for *NAT1*17* and *NAT1*22* respectively), which is thus consistent with the prediction results. Three *NAT1* haplotypes, although differing from the reference haplotype *NAT1*4* by a non-synonymous substitution, are described as “equivalent to *NAT1*4*” in terms of phenotype assignment by the official nomenclature of human *NAT* alleles, *NAT1*21*, *NAT1*24* and *NAT1*25*. Among these three haplotypes, only substitution A613G, defining *NAT1*21* is well predicted by the three tools as “benign” or “tolerated”. Substitution A787G of *NAT1*25* is predicted as “benign” and “tolerated” by PolyPhen and SIFT, respectively, but as “possibly damaging” by PANTHER cSNP Scoring. However, for the latter, the PSEP value of 220 is at the limit (PSEP=200) between “probably damaging” and “benign”. Finally, substitution G781A of *NAT1*24* is only predicted as “benign” by PolyPhen. These results are nevertheless congruent with the contradictory results returned by functional studies, such as those of (Lin et al. 1998), cited in Zhu and Hein (2008).

The adequacy of predictions is also straightforward for human *NAT2* when considering those substitutions whose effect on enzymatic function is well established. Indeed, among the seven haplotypes considered as associated with a slow phenotype, only the substitutions of four of them, C190T (defining *NAT2*19*), G191A (*NAT2*14A*), A434C (*NAT2*17*) and G590A (*NAT2*6B*) are well predicted as damaging by the three tools, and an additional one, T341C (*NAT2*5D*), by two of them (SIFT and PANTHER cSNP Scoring). However, while all five substitutions reduce N- and O-acetyltransferase activity (between 80% and 90% lower than *NAT2*4* for C190T, G191A, A434C and T341C, and between 36 and 70% for G590A, reviewed in (Hein et al. 2006)), and the protein level of expression as well (between 0 to 35% of *NAT2*4* for C190T, G590A, A434C and T341C and for G191A between 36

to 70%, (Hein et al. 2006)), C190T, G191A and G590A have also an effect on the thermostability of the protein. This could thus explain why T341C is predicted as damaging by only 2 out of the 3 tools, and A434C with less confidence by PANTHER cSNP Scoring (PSEP=456 compared to PSEP=4200 for C190T and G191A). Two others substitutions, G857A (defining *NAT2*7A*) and G499A (*NAT2*10*), were not predicted by any tool as damaging. However, the phenotypes associated with these two substitutions are reported as being potentially substrate-dependent in the official nomenclature of human *NAT* alleles, and this could explain the prediction results. Indeed, it was shown that G857A also reduces N-acetyltransferase and O-acetyltransferase activity compared to *NAT2*4* but to a lesser extent and only for some substrates. Moreover, it does not reduce the protein expression level, but rather affects the stability of the protein as much as C190T, G191A and G590A do (reviewed in (Hein et al. 2006)). The tools thus appear to detect well a substitution only when it has an effect on all three characteristics, i.e. acetylation, protein expression level and thermostability.

We also tested the prediction tools on two *NAT2* non-synonymous substitutions associated with a rapid acetylator phenotype (i.e. equivalent to *NAT2*4*). The A803G substitution (defining haplotype *NAT2*12A*) is well predicted as benign by the three tools. On the contrary, A845C (*NAT2*18*) is predicted as damaging by the three tools ("possibly damaging" by PANTHER cSNP Scoring). However, contradictory results exist for this latter substitution; while haplotype *NAT2*18* (A845C) is reported as rapid in the official nomenclature, it has also been suggested to be slow in publications (Hein et al. 2006).

Although for *NAT2* only two or three acetylation phenotypes have been defined (rapid, slow and intermediate, corresponding to the bi- or tri-modal distribution of responses to medication (Meyer 2004)), acetylation capacity is a continuous quantitative variable, thus allowing a wider number of different phenotypic responses to be recognized ((Ruiz et al. 2012; Selinski et al. 2013; Selinski et al. 2015). A similar observation applies to *NAT1* for which only two phenotypes are defined (rapid and slow, as well as absence of activity associated with substitution-induced truncated proteins). Moreover, for *NAT1* the correspondence between the phenotype and the genotype is less clear, and probably more complex, than for *NAT2*. Therefore, the results of the prediction tools, as described above (and reported in Supplementary Table S15), reflect this heterogeneity of phenotypes for both genes.

Substitutions in *NAT1* and *NAT2* that strongly impact acetylation activity are well detected as damaging by the three tools, and observed discordances between the prediction and the described phenotypes all correspond to documented variations of known phenotypes. We thus concluded that the three tools used in conjunction adequately predict the effect of single substitutions on acetylation activity.

12. Prediction of enzymatic activity of *Pan NAT1* and *NAT2* basal haplotypes

A recent study (Tsirka et al. 2014) has demonstrated that the function of the *NAT2* enzyme in the human and rhesus macaque species diverges in substrate selectivity, due to a G691A (V231I) substitution. This substitution mediates a shift in substrate affinity from substrates more *NAT2*-specific to substrates more *NAT1*-specific, but the study also demonstrated that neither the stability nor the overall activity of the 231V or 231I proteins were significantly altered. To the best of our knowledge, however, functional studies of chimpanzee NAT enzymes compared to humans have not yet been published. We thus used the three prediction tools to investigate whether the *Pan* basal *NAT1* (*NAT1*1*) and *NAT2* (*NAT2*4*) haplotypes could be considered functionally equivalent to the human reference haplotypes, i.e. human *NAT1*4* and *NAT2*4*. We first assumed that the common ancestral haplotype to humans and chimpanzees had an enzymatic activity equivalent to that of the human references *NAT1*4* and *NAT2*4*, i.e. that it conferred the rapid acetylation phenotype. Under this assumption, we used the predictions returned by the three online tools to infer the probable acetylation profile of each of the two *Pan* basal *NAT* haplotypes.

As deduced from the *NAT1* haplotypes network (Supplementary Figure S2), a single non-synonymous substitution (C529A, H177N) occurred on the lineage leading to the *Pan NAT1* haplotypes. The ancestral haplotype is embedded in a reticulation that includes another non-synonymous substitution (A583C, Q195K). Both C529A and A583C substitutions have been predicted as not affecting the protein function (Supplementary Table S16), and thus the *Pan* basal *NAT1*1* haplotype should also lead to an enzymatic activity equivalent to that of the ancestral protein. Note that only the results of PolyPhen and SIFT could be considered here, as PANTHER cSNP Scoring failed to test these substitutions.

For *NAT2*, the network of haplotypes (Supplementary Figure S3) indicates that the ancestral human-chimpanzee haplotype is differentiated from the *Pan* basal *NAT2*4* haplotype by five non-synonymous substitutions, i.e. T293C (V98A), T664C (F222L), C345A (D115E), G443C (C148S), and A595G (I199V), the first two being shared with the lineage of *Gorilla NAT2* haplotypes. Since the sequence of mutational events leading from the ancestral haplotype to the *Pan* basal *NAT2*4* haplotype is unknown, we could not test the functional impact of these substitutions in a sequential fashion, and opted to test them independently. Although impaired by this limitation, we note that none of the five substitutions is predicted as affecting the protein function (Supplementary Table S16). We thus tentatively assume that also the *Pan* basal *NAT2*4* haplotype should confer an enzymatic activity equivalent to that of the ancestral protein.

Then, we investigated whether assuming that the activity of the theoretical common human-chimpanzee ancestral *NAT* enzymes is equivalent to the rapid acetylation profile imparted by human reference haplotypes *NAT1*4* and *NAT2*4* is a plausible hypothesis. To this end, we used the three online tools to predict the impact of substitutions between the latter and the ancestral proteins.

For *NAT1*, as reported in Supplementary Table S16, the prediction tools did not return clear-cut results on the functional impact of the two non-synonymous substitutions differentiating the ancestral hominid haplotype (*ancestral_2*) and human *NAT1*4*. Indeed, the results of PolyPhen contradict those of SIFT and PANTHER cSNP Scoring for substitution A138T (E46D), whereas the output of this latter tool contradicts those of the formers for substitution G826C (E276Q). This raises the possibility that both mutations could have a moderately damaging effect on the protein activity, thus suggesting that the “normal” (i.e. reference) *NAT1* activity measured in humans could actually be decreased compared to that of the common ancestral human-chimpanzee *NAT1* enzyme, and by extension to that conferred by the *Pan* basal *NAT1* haplotype. We are aware of the highly speculative nature of this reasoning, but note nevertheless that the assumption of an equivalent enzymatic activity associated with the human reference haplotype (*NAT1*4*) and the chimpanzee basal haplotype (*NAT1*1*) is, in the light of the prediction results, a conservative hypothesis.

The results for *NAT2* (Supplementary Table S16) show that two out of the four non-synonymous substitutions differentiating the theoretical ancestral hominid

haplotype and the human reference haplotype (*NAT2*4*) are predicted as damaging by the three tools, i.e. C451T (R151C), and G834T (K278N). Thus, as for *NAT1*, we reason that assuming equivalent enzymatic activity for the human reference (*NAT2*4*) and *Pan* basal (*NAT2*4*) haplotypes is again a conservative hypothesis. Note that 451C is a rare variant in humans (rs747336755, from the Exome Aggregation Consortium), and the position was also found polymorphic in *P. abelii*.

13. Comparison of nucleotide diversity in gorillas and orangutans at the three *NAT* genes with *Pan* and humans

Supplementary Table S13 and Figure S5 report nucleotide diversity ($\pi \times 10^{-3}$) estimated in the two gorilla (western and eastern gorillas, *Gorilla gorilla* and *Gorilla beringei*, respectively) and the two orangutan (Sumatran and Bornean orangutans, *Pongo abelii* and *Pongo pygmaeus*, respectively) species. At the *NAT1* gene, a similar level of diversity to *Pan* was found in both species of gorillas (0.72 on average), thus also higher than in humans, whereas the highest value among all great apes was observed in orangutans. Actually, *NAT1* diversity markedly differs between the two orangutan species: $\pi (x 10^{-3})$ in *P. abelii* (3.94) is almost three times higher than in *P. pygmaeus* (1.36), in keeping with results from genomic studies (Prado-Martinez et al. 2013; Nater et al. 2015; Kuhlwilm et al. 2016; Nater et al. 2017). In turn, at *NAT2*, both species of orangutans have a comparable level of nucleotide diversity (1.52 on average) that is more similar to humans than to chimpanzees, whereas it differs markedly between the two gorilla species (1.72 and 0.31, in *G. gorilla* and *G. beringei*, respectively). Thus, while eastern gorillas (*G. beringei*) display, like *Pan*, a lower diversity for *NAT2* compared to *NAT1*, the reverse pattern is observed for western gorillas (*G. gorilla*). Contrasting results were also found at the *NATP* locus: while no nucleotide diversity was detected in *G. beringei* (genotypes of all three eastern gorillas were identical), that of western gorillas (*G. gorilla*, 2.77) and Bornean orangutans (*P. pygmaeus*, 2.29) is similar to the values found for Central (*Pan troglodytes troglodytes*) and Nigeria-Cameroon (*Pan troglodytes ellioti*) chimpanzees, while Sumatran orangutans (*P. abelii*, 3.76) display two to three times more diversity than chimpanzees and humans. We consider these contrasting results, both between the two gorillas species and between the two orangutan species as preliminary, and needing confirmation through the analysis of larger sample sizes, and possibly also

through a better representation of species and sub-species levels. Future investigations will also benefit from including extensive inter-gene sequencing within the *NAT* region on chromosome 8. Such a perspective is becoming more topical every day with new whole genome hominid sequences being produced at an unprecedented high pace (Nater et al. 2015; Xue et al. 2015; de Manuel et al. 2016; Nater et al. 2017).

14. References

- Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, and McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 491(7422):56-65. doi: 10.1038/nature11632
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods*. 7(4):248-249. doi: 10.1038/nmeth0410-248
- de Groot NG, Garcia CA, Verschoor EJ, Doxiadis GG, Marsh SG, Otting N, and Bontrop RE. 2005. Reduced MIC gene repertoire variation in West African chimpanzees as compared to humans. *Molecular biology and evolution*. 22(6):1375-1385. doi: 10.1093/molbev/msi127
- de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, Hernandez-Rodriguez J, Dupanloup I, Lao O, Hallast P et al. . 2016. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science (New York, NY)*. 354(6311):477-481. doi: 10.1126/science.aag2602
- Gonder MK, Disotell T, and Oates J. 2006. New Genetic Evidence on the Evolution of Chimpanzee Populations and Implications for Taxonomy. *Int J Primatol*. 27(4):1103-1127. doi: 10.1007/s10764-006-9063-y
- Gonder MK, Oates JF, Disotell TR, Forstner MR, Morales JC, and Melnick DJ. 1997. A new west African chimpanzee subspecies? *Nature*. 388(6640):337. doi: 10.1038/41005
- Hein DW, Fretland AJ, and Doll MA. 2006. Effects of single nucleotide polymorphisms in human N-acetyltransferase 2 on metabolic activation (O-acetylation) of heterocyclic amine carcinogens. *Int J Cancer*. 119(5):1208-1211. doi: 10.1002/ijc.21957
- Kuhlwilm M, de Manuel M, Nater A, Greminger MP, Krützen M, and Marques-Bonet T. 2016. Evolution and demography of the great apes. *Current Opinion in Genetics & Development*. 41:124-129. doi: 10.1016/j.gde.2016.09.005
- Lin HJ, Probst-Hensch NM, Hughes NC, Sakamoto GT, Louie AD, Kau IH, Lin BK, Lee DB, Lin J, Frankl HD et al. . 1998. Variants of N-acetyltransferase NAT1 and a case-control study of colorectal adenomas. *Pharmacogenetics*. 8(3):269-281
- McDonagh EM, Boukouvala S, Aklillu E, Hein DW, Altman RB, and Klein TE. 2014. PharmGKB summary: very important pharmacogene information for N-acetyltransferase 2. *Pharmacogenetics and genomics*. 24(8):409-425. doi: 10.1097/fpc.0000000000000062
- Meyer UA. 2004. Pharmacogenetics - five decades of therapeutic lessons from genetic diversity. *Nat Rev Genet*. 5(9):669-676. doi: 10.1038/nrg1428
- Nater A, Greminger MP, Arora N, van Schaik CP, Goossens B, Singleton I, Verschoor EJ, Warren KS, and Krutzen M. 2015. Reconstructing the demographic history of orang-utans using Approximate Bayesian Computation. *Mol Ecol*. 24(2):310-327. doi: 10.1111/mec.13027
- Nater A, Mattle-Greminger MP, Nurcahyo A, Nowak MG, de Manuel M, Desai T, Groves C, Pybus M, Sonay TB, Roos C et al. . 2017. Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. *Current biology : CB*. 27(22):3487-3498.e3410. doi: 10.1016/j.cub.2017.09.047
- 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. doi: 10.1038/nature12228
- Ruiz JD, Martinez C, Anderson K, Gross M, Lang NP, Garcia-Martin E, and Agundez JA. 2012. The differential effect of NAT2 variant alleles permits refinement in phenotype inference and identifies a very slow acetylation genotype. *PloS one*. 7(9):e44629. doi: 10.1371/journal.pone.0044629
- Selinski S, Blaszkewicz M, Getzmann S, and Golka K. 2015. N-Acetyltransferase 2: ultra-slow acetylators enter the stage. *Archives of toxicology*. 89(12):2445-2447. doi: 10.1007/s00204-015-1650-2
- Selinski S, Blaszkewicz M, Ickstadt K, Hengstler JG, and Golka K. 2013. Refinement of the prediction of N-acetyltransferase 2 (NAT2) phenotypes with respect to enzyme activity and urinary bladder cancer risk. *Archives of toxicology*. 87(12):2129-2139. doi: 10.1007/s00204-013-1157-7
- Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, and Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 40(W1):W452-W457. doi: 10.1093/nar/gks539
- Stephens M, and Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American journal of human genetics*. 76(3):449-462. doi: 10.1086/428594
- Stephens M, Smith NJ, and Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *American journal of human genetics*. 68(4):978-989. doi: 10.1086/319501
- Tang H, and Thomas PD. 2016. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*. 32(14):2230-2232. doi: 10.1093/bioinformatics/btw222
- Tsirka T, Boukouvala S, Agianian B, and Fakis G. 2014. Polymorphism p.Val231Ile alters substrate selectivity of drug-metabolizing arylamine N-acetyltransferase 2 (NAT2) isoenzyme of rhesus macaque and human. *Gene*. 536(1):65-73. doi: 10.1016/j.gene.2013.11.085
- Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN et al. . 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science (New York, NY)*. 348(6231):242-245. doi: 10.1126/science.aaa3952
- Zhu Y, and Hein DW. 2008. Functional effects of single nucleotide polymorphisms in the coding region of human N-acetyltransferase 1. *Pharmacogenomics J*. 8(5):339-348. doi: 10.1038/sj.tpj.6500483