SUPPLEMENTARY INFORMATION

SUPPLEMENTARY TABLES

Table S1 Genotyping of the *GSTM1* copy number by ddPCR of 17 individuals in YRI population in the 1KGP. Genotypes reported in 1KGP and validation by ddPCR were indicated.

	1000Genomes_CN	ddPCR_CN
NA18486	1	1
NA18507	0	0
NA18508	0	0
NA18853	1	1
NA18858	2	2
NA18870	1	1
NA18871	2	2
NA18912	1	1
NA19093	1	1
NA19096	2	2
NA19113	1	1
NA19114	2	2
NA19129	1	1

NA19144	1	1
NA19171	0	0
NA19172	2	2
NA19239	2	2

Table S2 Single nucleotide variants on the *Tanuki* haplogroup and their R^2 with the *GSTM1* deletion in CHB

rs number	**Location**	***F_{ST}*** **between CHB and YRI**	**R^2 with the *GSTM1* deletion in CHB**
rs35817611	110246810	0.66	0.46
rs11101986	110247240	0.66	0.46
rs11579576	110249176	0.66	0.47
rs4970772	110250168	0.66	0.46
rs35677255	110252091	0.62	0.46
rs3754446	110253241	0.66	0.46
rs12137743	110253555	0.66	0.46
rs11101989	110253741	0.66	0.46
rs4970773	110255596	0.67	0.39

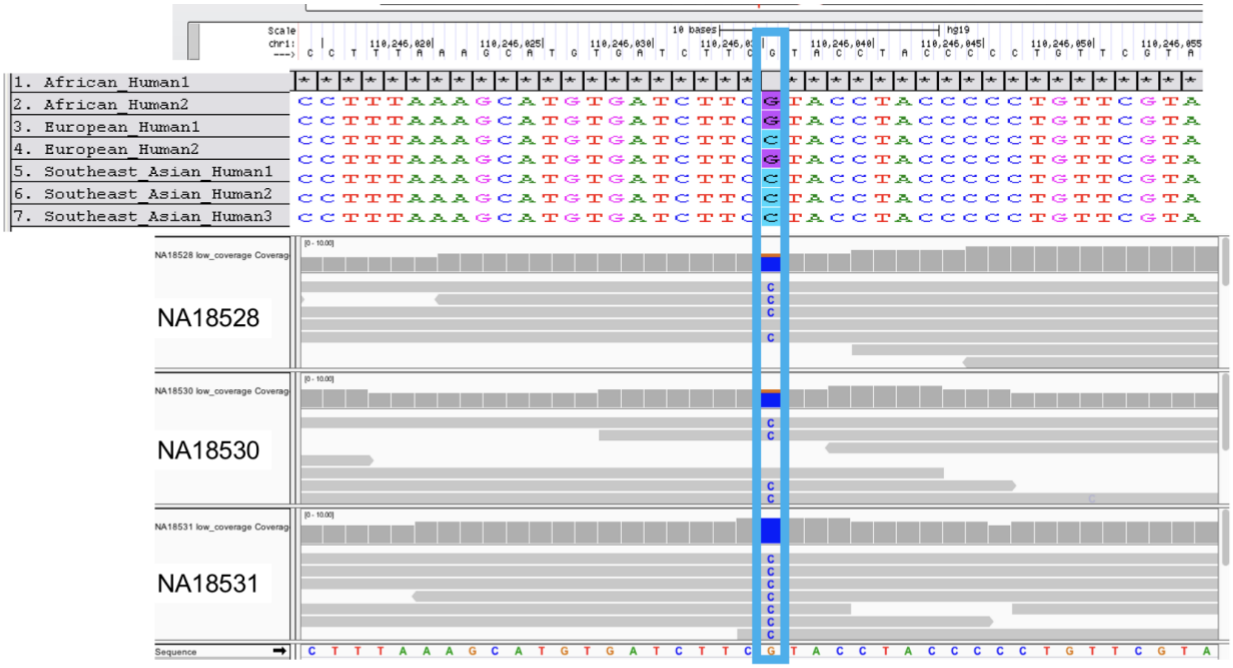
Table S3 Evolutionary conservation of the *GSTM1* gene. The number of common nonsynonymous polymorphisms and synonymous polymorphisms among humans are described in the human column. The number of nonsynonymous and synonymous substitutions of non-human hominins and apes compared with human reference genome are described in the other column.

	Nonsynonymous	Synonymous	Reference

Human	2	3	(Moyer et al. 2007)
Neanderthal	0	0	(Prüfer et al. 2014)
Denisovan	0	1	(Reich et al. 2010)
Chimpanzee	4	2	(The Chimpanzee Sequencing Consortium 2005)
Gorilla	4	2	(Scally et al. 2012)
Orangutan	5	12	(Locke et al. 2011)

SUPPLEMENTARY FIGURES

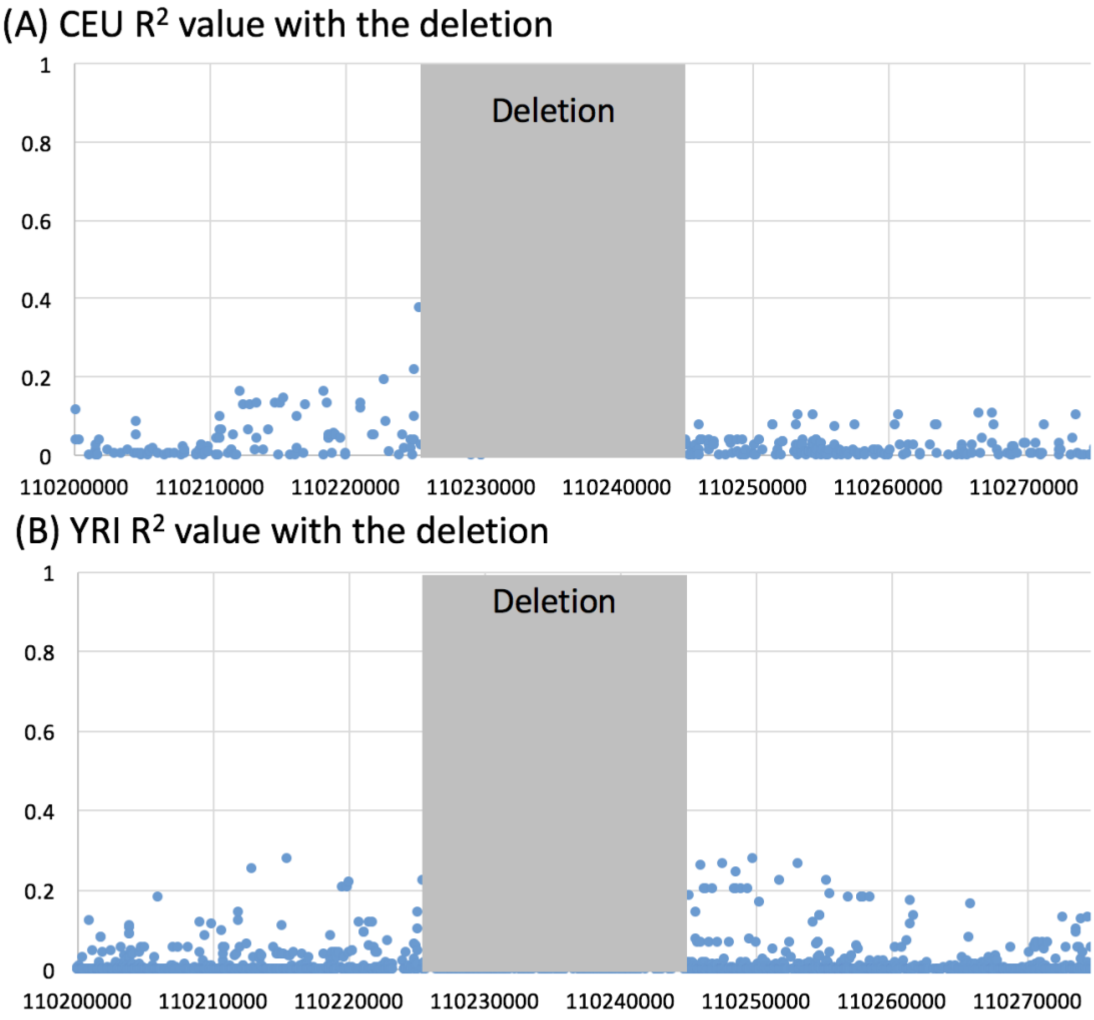
Figure S1 Screenshot of the validation of a flanking SNP of the *GSTM1* deletion by MEGA7.0 (Kumar *et al.* 2016) and Integrated Genome Viewer (Robinson *et al.* 2011; Thorvaldsdóttir *et al.* 2013). At the position chr:110245036, variants were found in both our sequences and data from 1000 genomes phase 3.



```
1. African_Human1            * * * * * * * * * * * * * * * * * * * *   * * * * * * * * * * * * * * * * * * *
2. African_Human2            C C T T T A A A G C A T G T G A T C T T C G T A C C T A C C C C C T G T T C G T A
3. European_Human1           C C T T T A A A G C A T G T G A T C T T C G T A C C T A C C C C C T G T T C G T A
4. European_Human2           C C T T T A A A G C A T G T G A T C T T C C T A C C T A C C C C C T G T T C G T A
5. Southeast_Asian_Human1    C C T T T A A A G C A T G T G A T C T T C G T A C C T A C C C C C T G T T C G T A
6. Southeast_Asian_Human2    C C T T T A A A G C A T G T G A T C T T C C T A C C T A C C C C C T G T T C G T A
7. Southeast_Asian_Human3    C C T T T A A A G C A T G T G A T C T T C C T A C C T A C C C C C T G T T C G T A
                             C C T T T A A A G C A T G T G A T C T T C C T A C C T A C C C C C T G T T C G T A
```

Scale
chr1:
--->
10 bases
hg19
110,246,020
110,246,025
110,246,030
110,246,035
110,246,040
110,246,045
110,246,050
110,246,055
C C T T T A A A G C A T G T G A T C T T C G T A C C T A C C C C C T G T T C G T A
NA18528 low_coverage Coverage
[0 - 10.00]
NA18528
C
C
C
C
NA18530 low_coverage Coverage
[0 - 10.00]
NA18530
C
C
C
C
NA18531 low_coverage Coverage
[0 - 10.00]
NA18531
C
C
C
C
C
C
C
Sequence
C T T T A A A G C A T G T G A T C T T C G T A C C T A C C C C C T G T T C G T A

Figure S2 R^2 value between the *GSTM1* deletion and flanking SNPs in (A) CEU and (B) YRI. Each dot represent each SNV. X-axis indicates chromosomal location and Y-axis indicates the R^2 value between the *GSTM1* deletion and flanking SNPs.


(A) CEU R^2 value with the deletion

(B) YRI R^2 value with the deletion

Figure S3 Neutrality tests on the target region (chr1:110246810-110255596, downstream the deletion, also represented in Figure. 2.), the 1000 randomly selected 9kb regions, and all the

segmental duplications in CEU, CHB and YRI populations in 1KGP. (A) Tajima's D value were calculated for 3kb intervals (B) iHS was calculated for 3kb intervals (C) F_{ST} value (D) pi was calculated for 3kb intervals (E) XP.CLR was calculated for 2kb intervals (F) XP.EHH was calculated for 2kb intervals

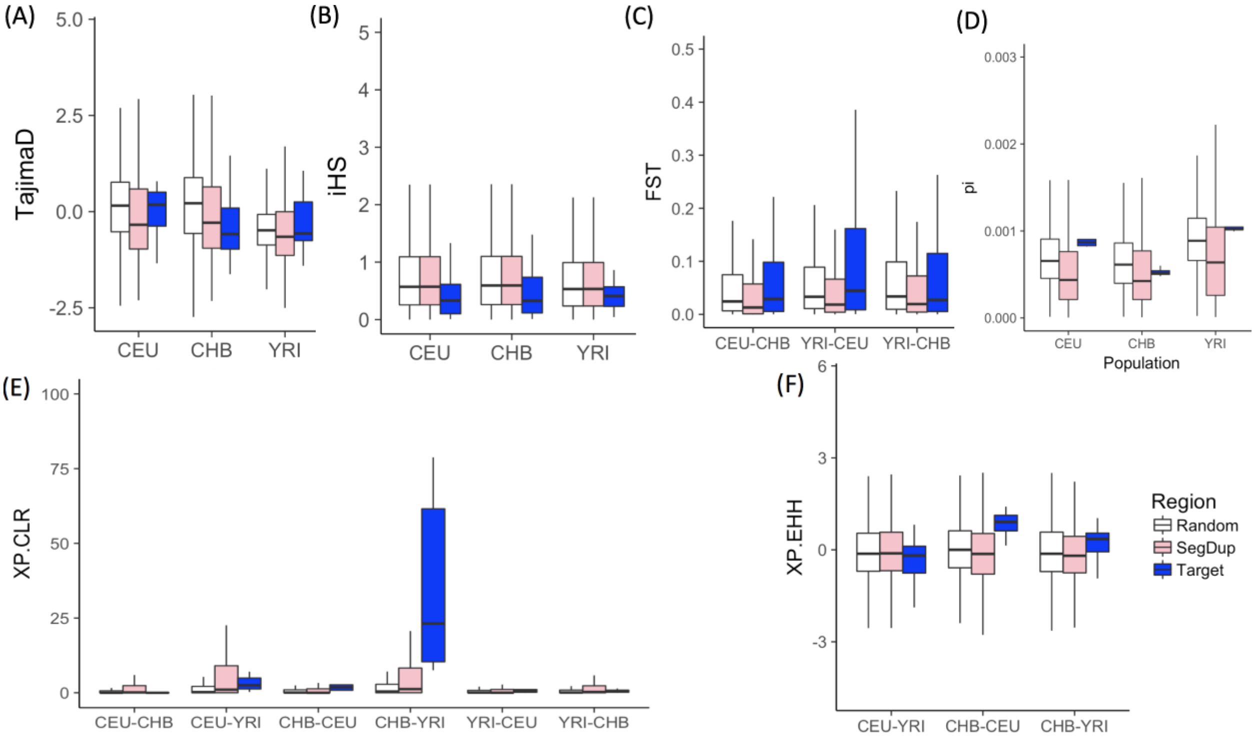


Figure S4. Calculating π values from the simulated dataset. We first created subgroups of haplotypes each defined by a simulated derived single nucleotide variant that is frequency-matched to *Tanuki* haplogroup. Then, these π was calculated based on the haplotypic variation within each of these subgroups.

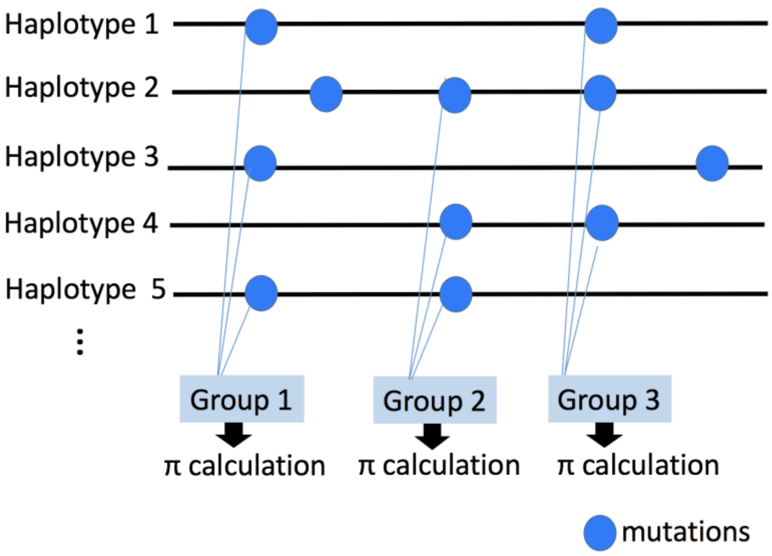

Haplotype 1
Haplotype 2
Haplotype 3
Haplotype 4
Haplotype 5
⋮
Group 1
π calculation
Group 2
π calculation
Group 3
π calculation
● mutations

Figure S5. Bootstrap resampling and calculation of ROH. Black line shows the distribution of the averaged length of ROH of the datasets with randomly resampled chromosomes and blue dashed line shows the observed length of ROH of the 238*Tanuki* homozygous individuals.

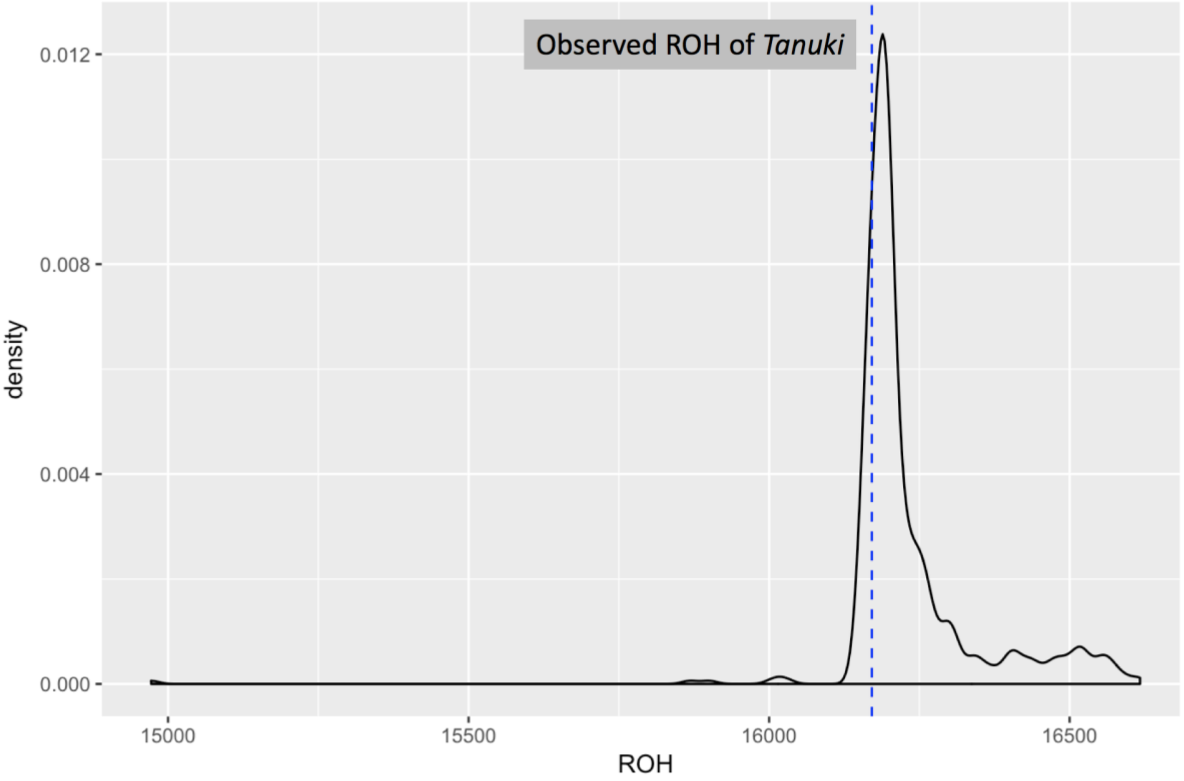


Figure S6. The phylogenetic tree of *Tanuki* haplotypes in CHB and CEU populations and haplotypes in Neanderthal and Chimpanzee by using ML method.

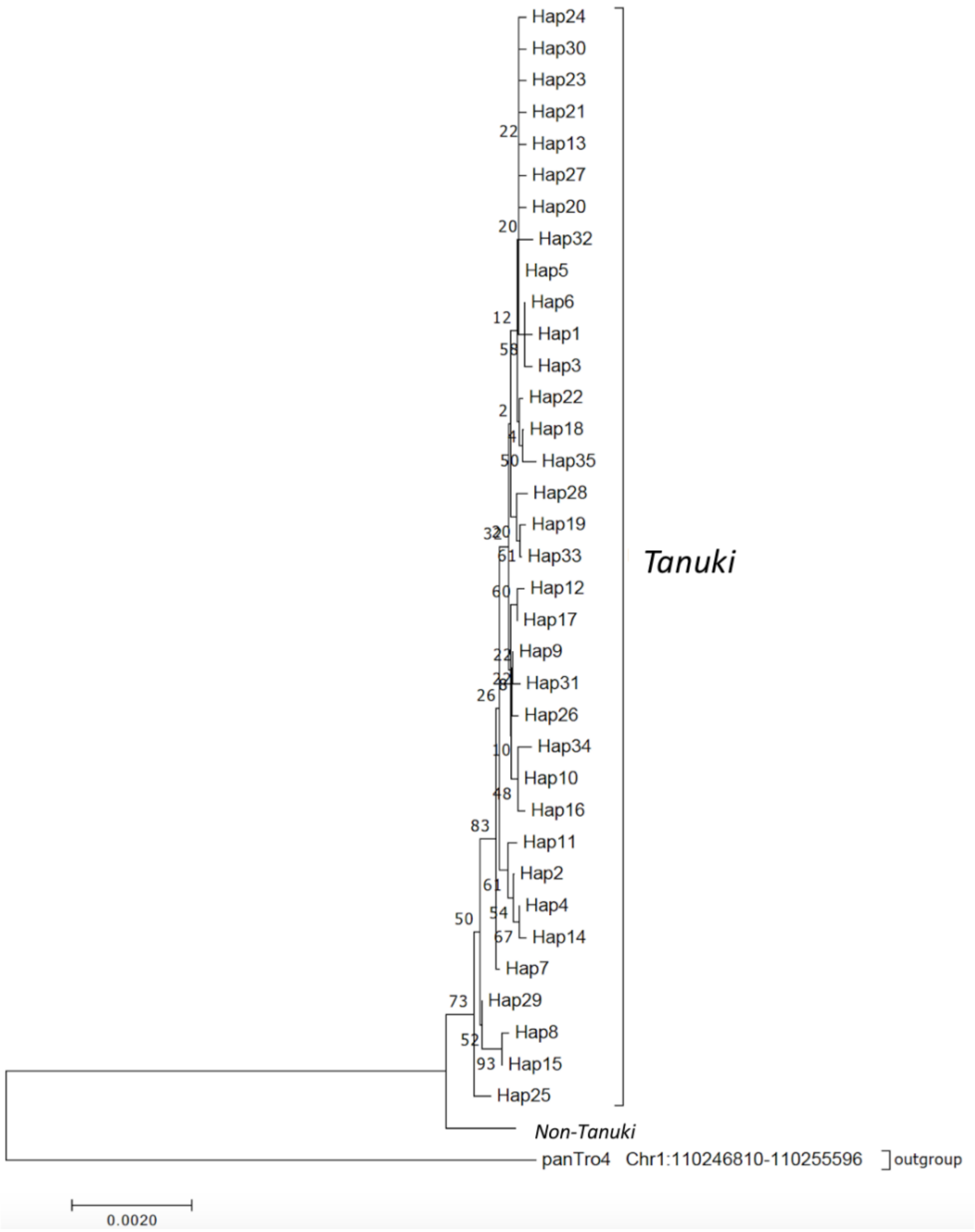


Figure S7. Haplotype network of *Tanuki* haplogroup. Orange represents CHB and light blue represents CEU, and brown represents *non-Tanuki* haplogroup.

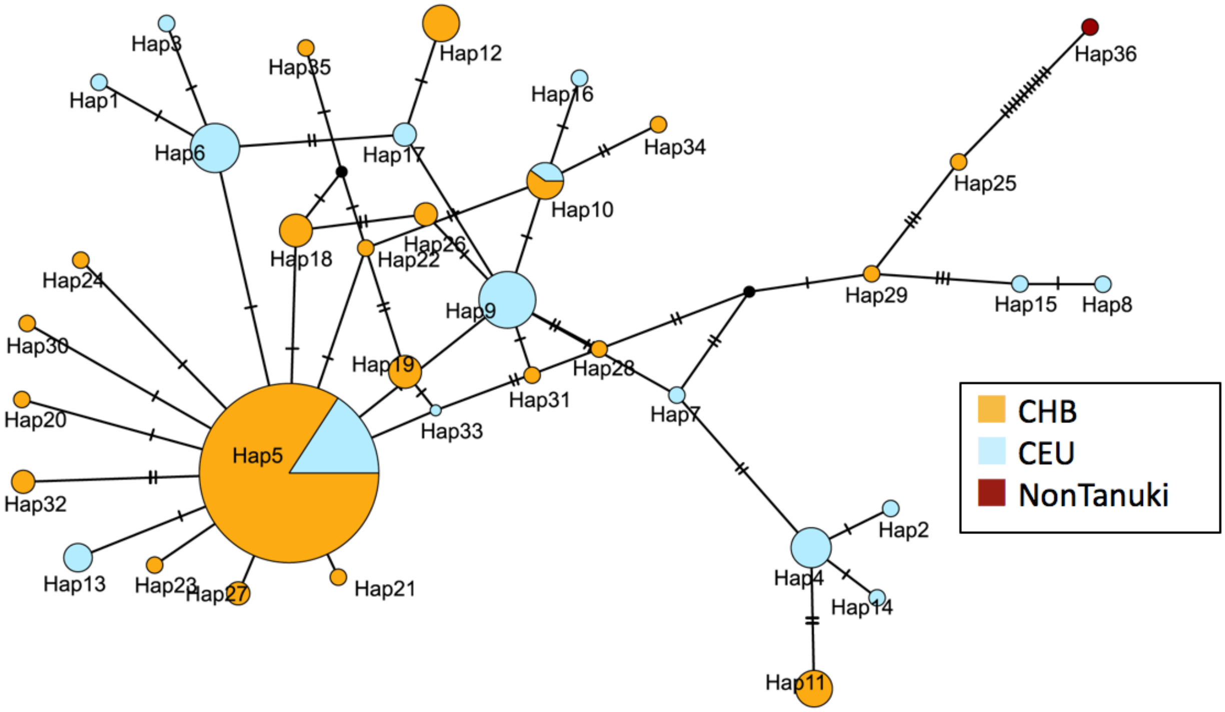
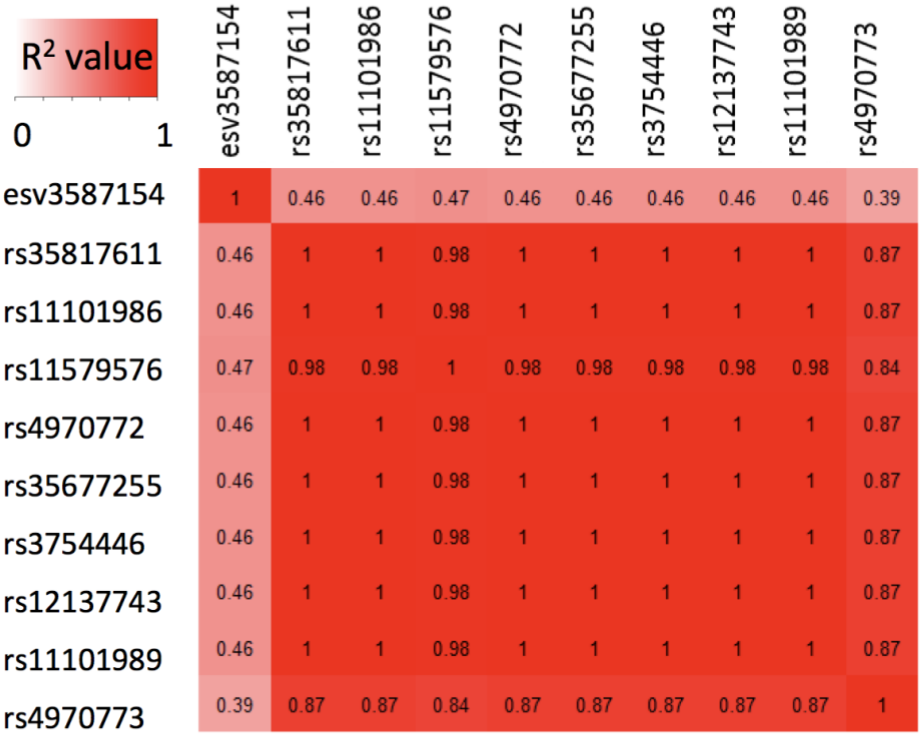
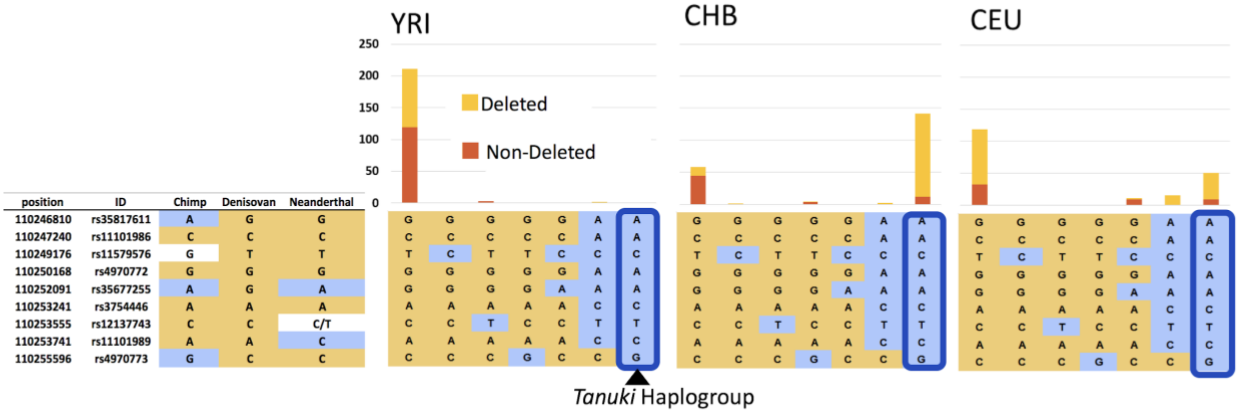


Figure S8. Linkage disequilibrium (R^2 value) between the high-F_{ST} SNPs on the *Tanuki* haplogroup and between the SNPs and the deletion in CHB. R^2 values between each SNPs are represented by color (from 0: white to 1: red). rs35817611 - rs4970773 locate on target2 region. esv3587154 is the *GSTM1* deletion.



	esv3587154	rs35817611	rs11101986	rs11579576	rs4970772	rs35677255	rs3754446	rs12137743	rs11101989	rs4970773
esv3587154	1	0.46	0.46	0.47	0.46	0.46	0.46	0.46	0.46	0.39
rs35817611	0.46	1	1	0.98	1	1	1	1	1	0.87
rs11101986	0.46	1	1	0.98	1	1	1	1	1	0.87
rs11579576	0.47	0.98	0.98	1	0.98	0.98	0.98	0.98	0.98	0.84
rs4970772	0.46	1	1	0.98	1	1	1	1	1	0.87
rs35677255	0.46	1	1	0.98	1	1	1	1	1	0.87
rs3754446	0.46	1	1	0.98	1	1	1	1	1	0.87
rs12137743	0.46	1	1	0.98	1	1	1	1	1	0.87
rs11101989	0.46	1	1	0.98	1	1	1	1	1	0.87
rs4970773	0.39	0.87	0.87	0.84	0.87	0.87	0.87	0.87	0.87	1

Figure S9. Evolutionary conservation of the nine variations in the *Tanuki* haplogroup. Blue: *Tanuki* haplogroup, Yellow: non-*Tanuki* haplogroup. Neanderthal and Chimpanzee also carried some alleles of the *Tanuki* haplogroup.Haplogroup counts of *Tanuki* SNPs in YRI, CHB and CEU populations. Bars above indicate counts of the *GSTM1* deletion for each haplotype.



position	ID	Chimp	Denisovan	Neanderthal
110246810	rs35817611	A	G	G
110247240	rs11101986	C	C	C
110249176	rs11579576	G	T	T
110250168	rs4970772	G	G	G
110252091	rs35677255	A	G	A
110253241	rs3754446	A	A	A
110253555	rs12137743	C	C	C/T
110253741	rs11101989	A	A	C
110255596	rs4970773	G	C	C

YRI

G	G	G	G	G	A	A
C	C	C	C	C	A	A
T	C	T	T	C	C	C
G	G	G	G	G	A	A
G	G	G	G	A	A	A
A	A	A	A	A	C	C
C	C	T	C	C	T	T
A	A	A	A	A	C	C
C	C	C	G	C	C	G

CHB

G	G	G	G	G	A	A
C	C	C	C	C	A	A
T	C	T	T	C	C	C
G	G	G	G	G	A	A
G	G	G	G	A	A	A
A	A	A	A	A	C	C
C	C	T	C	C	T	T
A	A	A	A	A	C	C
C	C	C	G	C	C	G

CEU

G	G	G	G	G	A	A
C	C	C	C	C	A	A
T	C	T	T	C	C	C
G	G	G	G	G	A	A
G	G	G	G	A	A	A
A	A	A	A	A	C	C
C	C	T	C	C	T	T
A	A	A	A	A	C	C
C	C	C	G	C	C	G

Figure S10. (A) π values for *Tanuki* and *nonTanuki* haplogroups and deleted and non-deleted haplotypes in East Asian populations (CHB, CHS, JPT and KHV) across a larger genomic region surrounding the *GSTM1* (x axis is chromosomal locations and y axis is the π values. Window size=9000 and step size=2000). Dashed line is the most frequent π values in the simulation result under neutrality. Gray bars under the genes are segmental duplications reported in the UCSC genome browser. (B) π values for *Tanuki* and *nonTanuki* haplogroups and deleted and non-deleted haplotypes in the target region of East Asian populations (CHB, CHS, JPT and KHV)

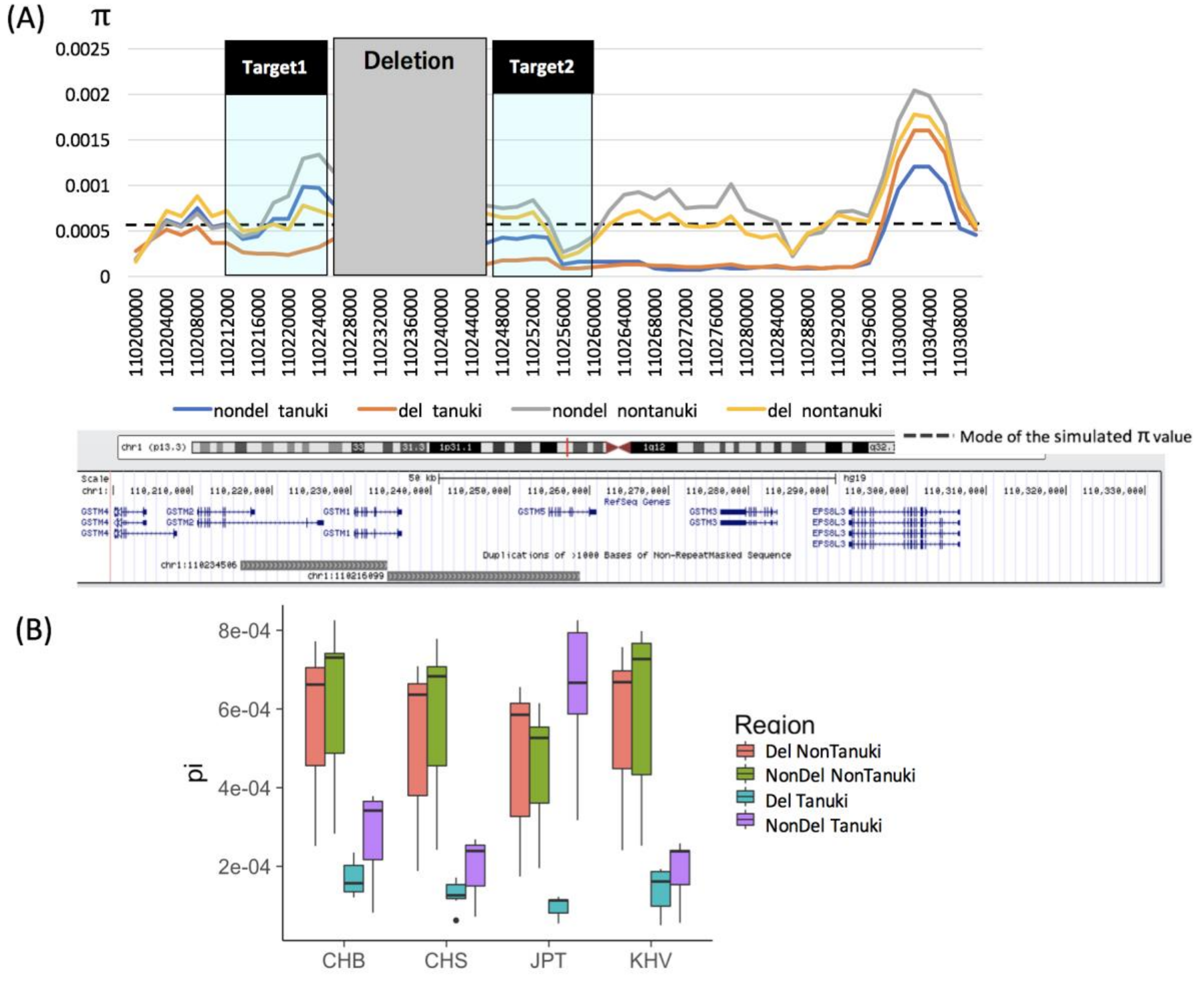


Figure S11. The linkage disequilibrium between the Tanuki haplotype and the flanking SNPs. GWAS positive SNP (MacArthur *et al.* 2017) was represented in yellow and NonSynonymous SNP was represented in red. SNPs with unusually high PHRED scores (Michailidou *et al.* 2017) were represented in blue.

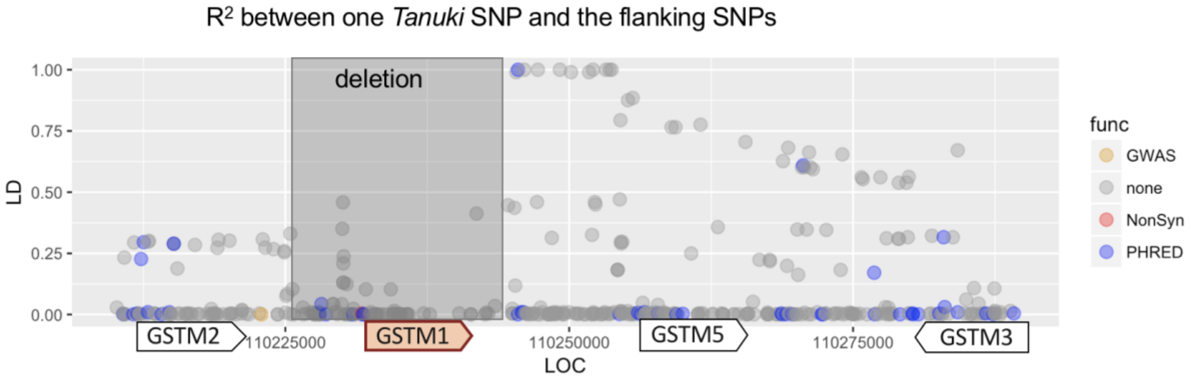
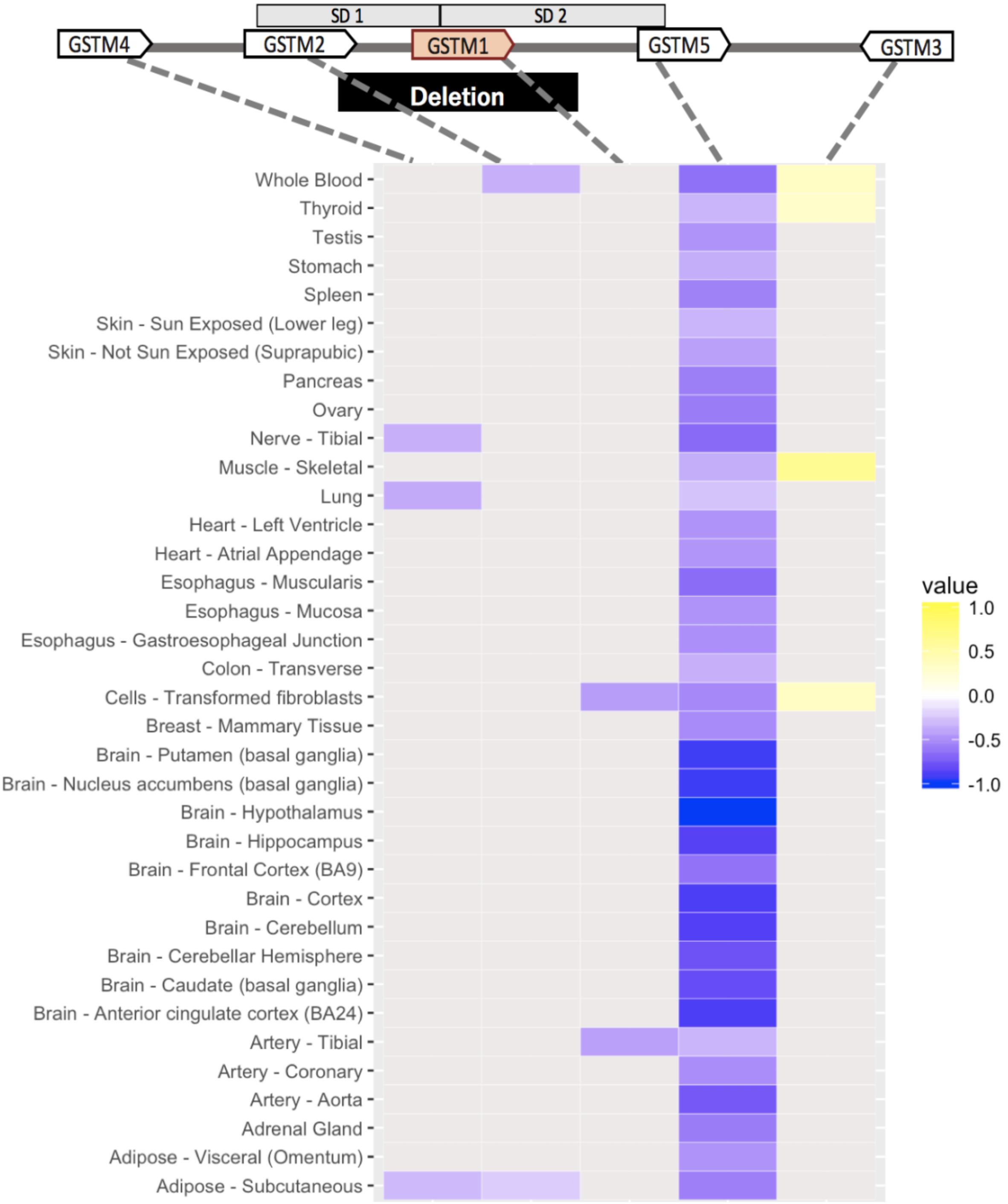


Figure S12. The effect size of the eQTLs of rs35817611 A (*Tanuki* allele) relative to the other allele on the *GSTM* gene expression from the Gtex portal. Effect size on each gene is represented by color (from -1: blue to 1: yellow). The effect size is defined as the slope of the linear regression.



References:

Kumar, S., G. Stecher, and K. Tamura, 2016 MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33: 1870-1874.

Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth et al., 2011 Comparative and demographic analysis of orang-utan genomes. Nature 469: 529–533.

MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall et al., 2017 The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 45: D896–D901.

Michailidou, K., S. Lindström, J. Dennis, J. Beesley, S. Hui et al., 2017 Association analysis identifies 65 new breast cancer risk loci. Nature 551: 92–94.

Moyer, A. M., O. E. Salavaggione, S. J. Hebbring, I. Moon, M. a. T. Hildebrandt et al., 2007 Glutathione S-transferase T1 and M1: gene sequence variation and functional genomics. Clin. Cancer Res. 13: 7207–7216.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman et al., 2014 The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43–49.

Reich, D., R. E. Green, M. Kircher, J. Krause, N. Patterson et al., 2010 Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053–1060.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander et al., 2011 Integrative genomics viewer. Nat. Biotechnol. 29: 24–26.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead et al., 2012 Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169–175.

The Chimpanzee Sequencing Consortium, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69–87.

Thorvaldsdóttir, H., J. T. Robinson, and J. P. Mesirov, 2013 Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14: 178–192.