

Supplementary Material

The genetic diversity of Indonesian cattle has been shaped by multiple introductions and adaptive introgression

Xi Wang^{1*}, Casia Nursyifa^{1*}, Sabhrina Gita Aninta¹, Genís Garcia-Erill^{1,2}, Laura D. Bertola^{1,3}, Anubhab Khan^{1,4}, Josiah Kuja¹, Kristian Hanghøj¹, Jonas Meisner^{1,5,6}, Thomas Bøggild¹, Corey J. A. Bradshaw^{7,8}, Amal Al-Chaer¹, Alam Putra Persada⁹, Dwi Sendi Priyono¹⁰, Yuli A. Tribudi¹¹, Pita Sudrajad¹², Cynthia Dewi Gaina¹³, Yu Jiang¹⁴, Johannes A. Lenstra¹⁵, Reagan Cauble-Sims¹⁶, Benjamin D. Rosen¹⁷, Darren E. Hagen¹⁸, Michael P. Heaton¹⁹, Timothy P. L. Smith¹⁹, Laurent Frantz^{20,21}, Greger Larson²², Mikkel-Holger S. Sinding¹, Dedy Duryadi Solihin^{9,23}, Muhammad Agil²⁴, Bambang Purwantara²⁴, Rasmus Heller^{1#}

¹Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark

²Bioinformatics Research Centre, Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark

³National Centre for Biological Sciences (NCBS), Bangalore, India

⁴Centre for Ecological Sciences, Indian Institute of Science, Bangalore, Karnataka, India

⁵Biological and Precision Psychiatry, Mental Health Centre Copenhagen, Copenhagen University Hospital, 2100 København, Denmark

⁶Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 København, Denmark

⁷Global Ecology | *Partuyarta Ngadluku Wardli Kuu*, College of Science and Engineering, Flinders University, Adelaide, South Australia, Australia

⁸Australian Research Council Centre of Excellence for Australian Biodiversity and Heritage, Wollongong, New South Wales, Australia

⁹Genetic Conservation and Genome Laboratory, IPB University, Bogor, Indonesia

¹⁰Faculty of Biology, Universitas Gadjah Mada, Yogyakarta, Indonesia

¹¹Department of Animal Science, Faculty of Agriculture, Universitas Tanjungpura, Pontianak, Indonesia

¹²Research Center for Animal Husbandry, National Research and Innovation Agency, 16911, Indonesia

¹³Faculty of Medicine and Veterinary Medicine, Universitas Nusa Cendana, Kupang, Indonesia

¹⁴Department of Animal Genetics, Breeding and Reproduction, College of Animal Science and Technology, Northwest Agriculture and Forestry University, Yangling 712100, China

¹⁵Faculty of Veterinary Medicine, Utrecht University, Utrecht, 3508 TD, The Netherlands

¹⁶Ten Triple X Ranch, Glen Rose, Texas 76652, USA

¹⁷U.S. Department of Agriculture, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, Maryland, USA

¹⁸Department of Animal and Food Sciences, Oklahoma State University, Stillwater, Oklahoma, USA

¹⁹U.S. Department of Agriculture, Agricultural Research Service, U.S. Meat Animal Research Center, Clay Center, Nebraska, USA

²⁰Paleogenomics Group, Department of Veterinary Sciences, Ludwig Maximilian University, Munich, Germany

²¹School of Biological and Behavioural Sciences, Queen Mary University of London, London, United Kingdom

²²Palaeogenomics and Bio-Archaeology Research Network, School of Archaeology, University of Oxford, Oxford OX1 3QY, United Kingdom

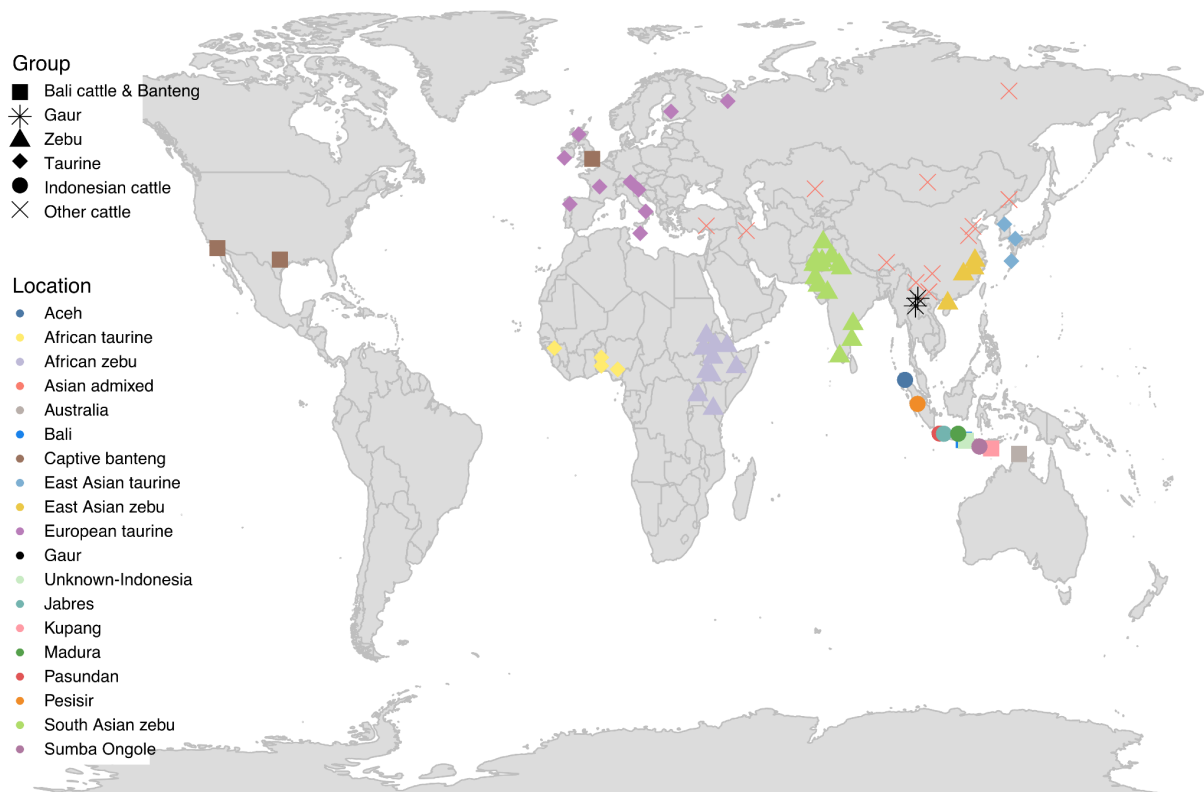
²³Department of Biology, Faculty of Science, IPB University, 16680, Bogor, Indonesia

²⁴Division of Reproduction and Obstetrics, School of Veterinary Medicine and Biomedical Sciences, IPB University, Bogor 16680, Indonesia

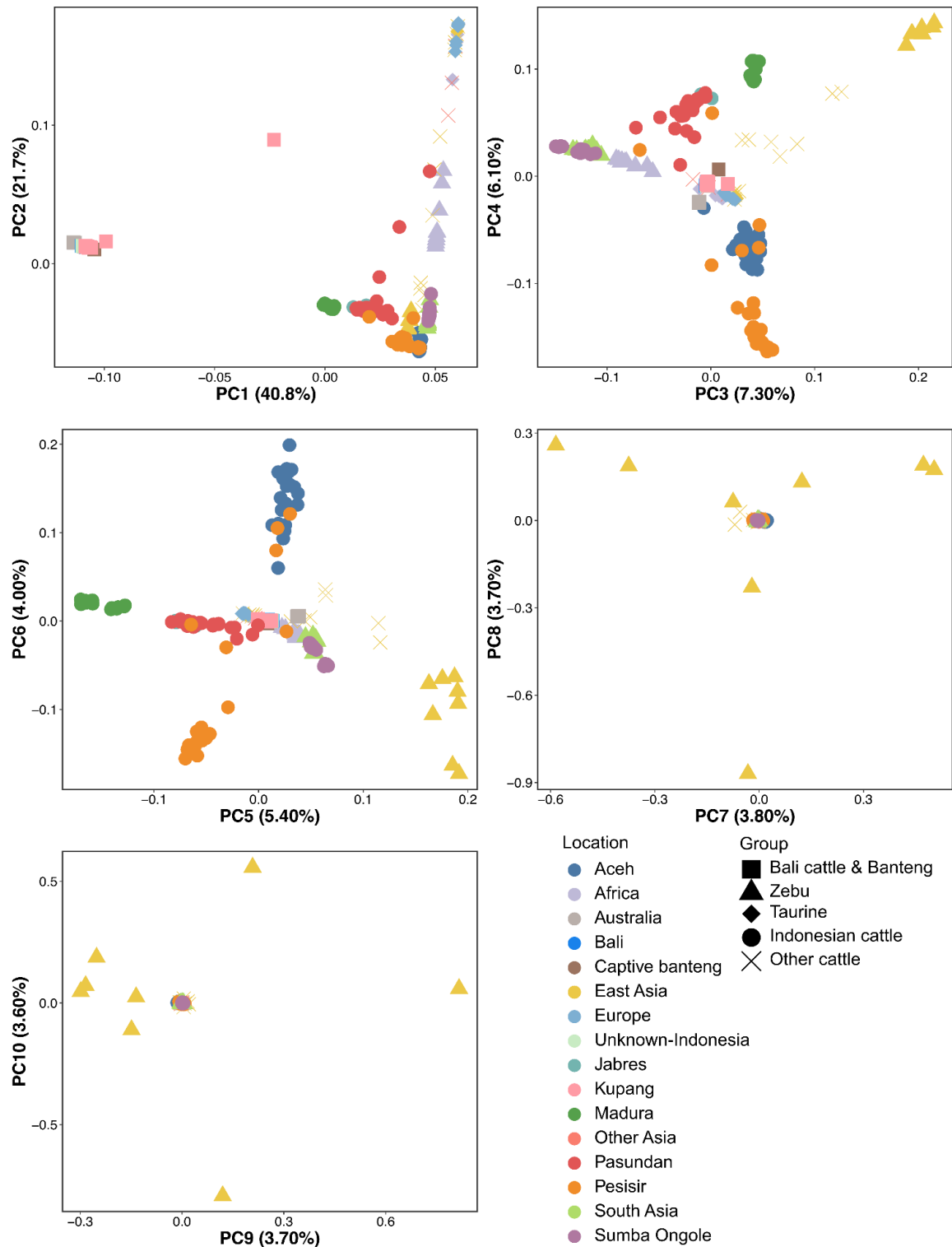
* Contributed equally

Corresponding author: rheller@bio.ku.dk

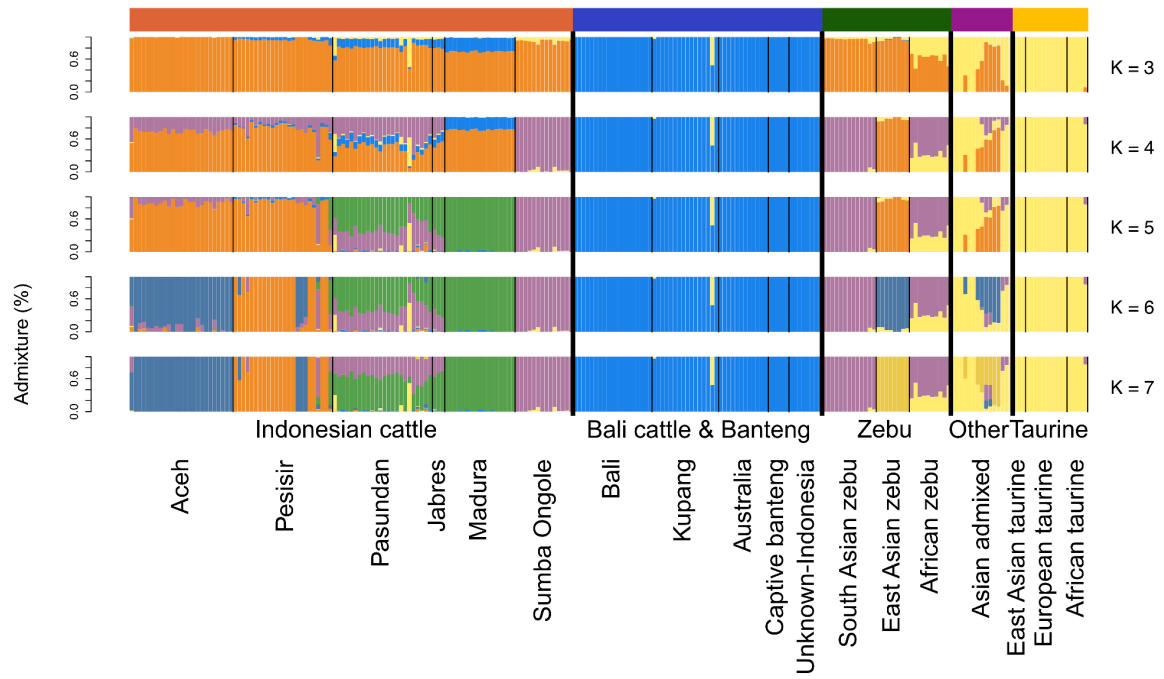
Supplementary Figures



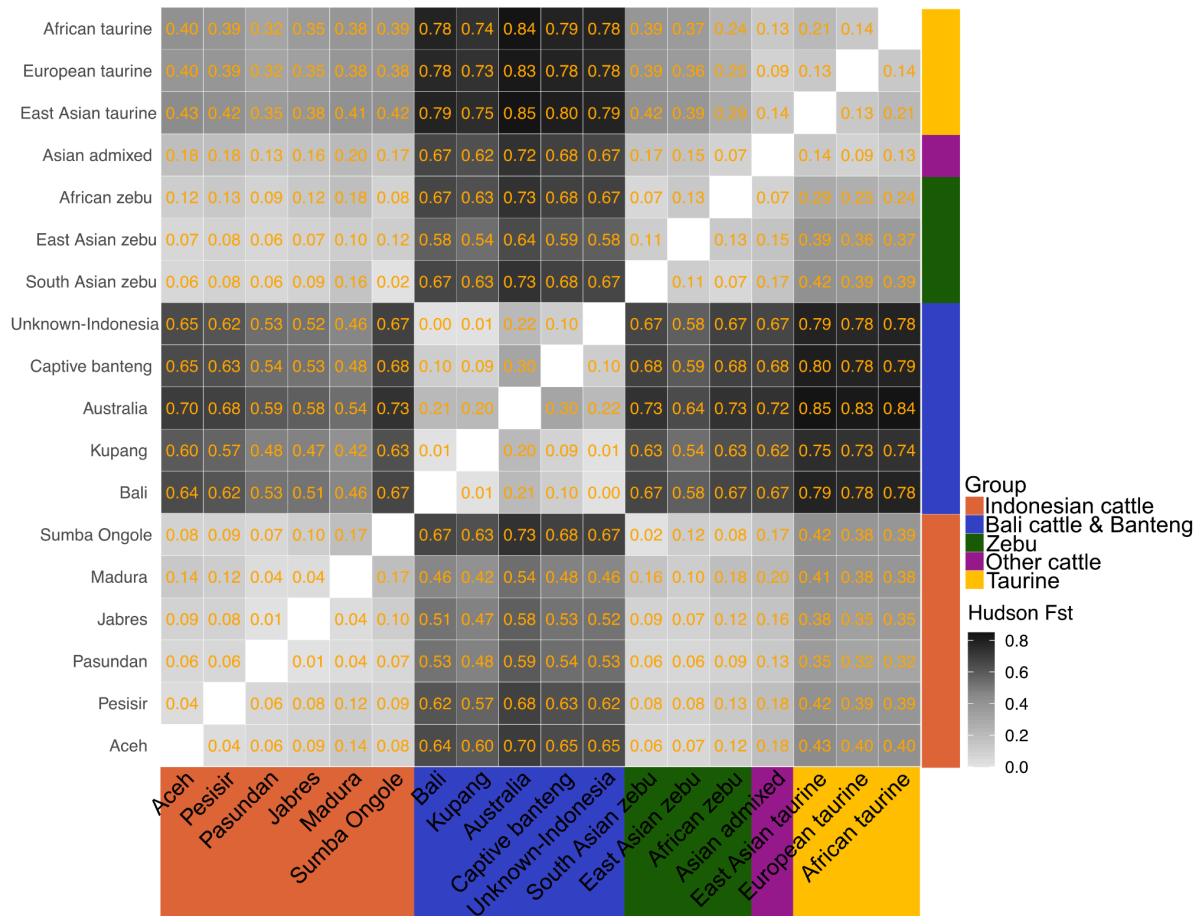
Supplementary Figure 1. Sampling Location for 314 individuals. A total of 233 new genomes, including 179 zebu cattle representing six Indonesia cattle breeds, 51 Bali cattle representing three breeds and 3 captive individuals of Javan banteng, were sequenced. 81 publically available genomes from 2 captive Javan banteng, 8 Bali cattle, 42 zebu, 27 taurine, and 2 gaur were downloaded, resulting in 314 samples in total. Detailed information was shown in Supplementary Data 1 and Supplementary Data 2. The world map is generated using an R package 'maps' (<https://CRAN.R-project.org/package=maps>).



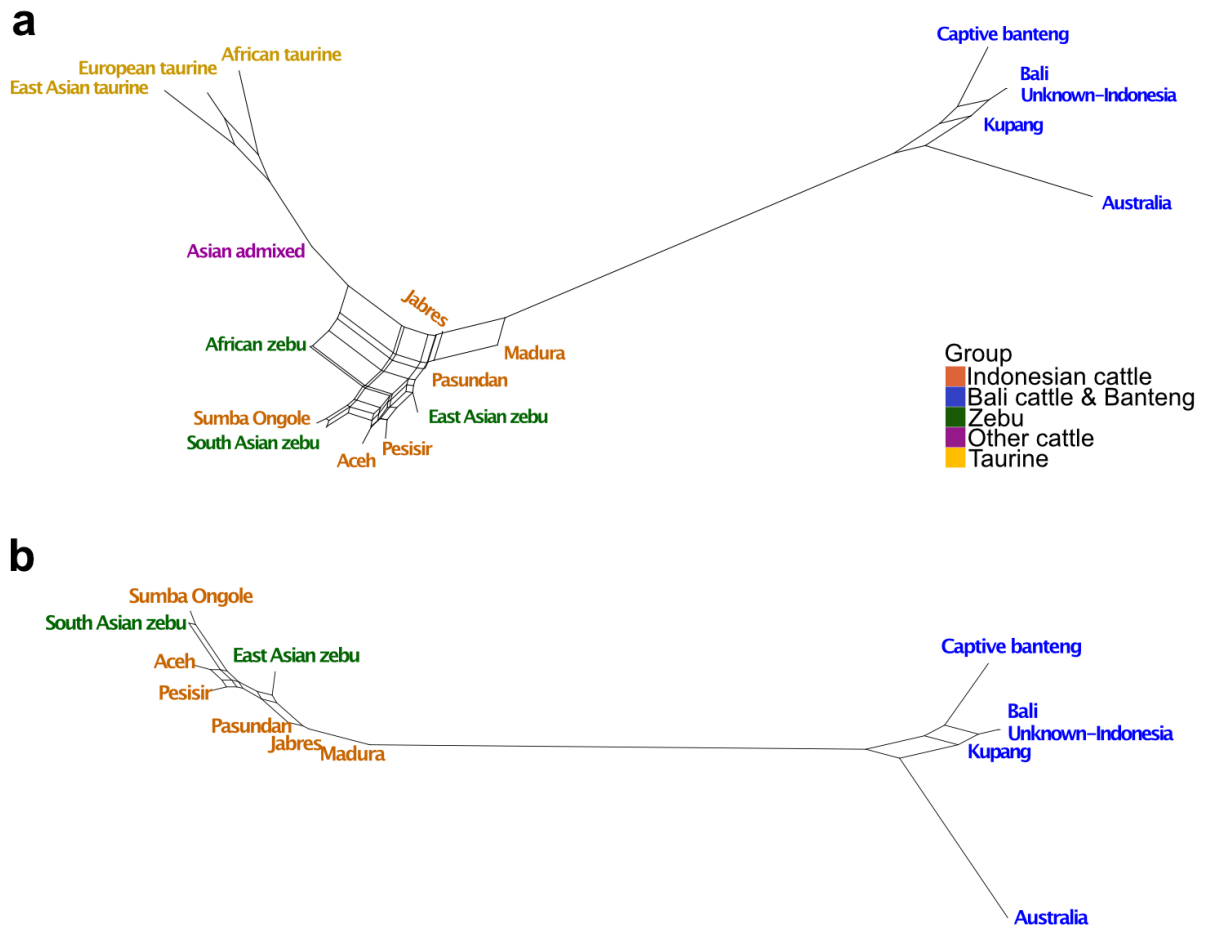
Supplementary Figure 2. PCA plot of 231 samples colored by sampling locality on ten principal components inferred by HaploNet.



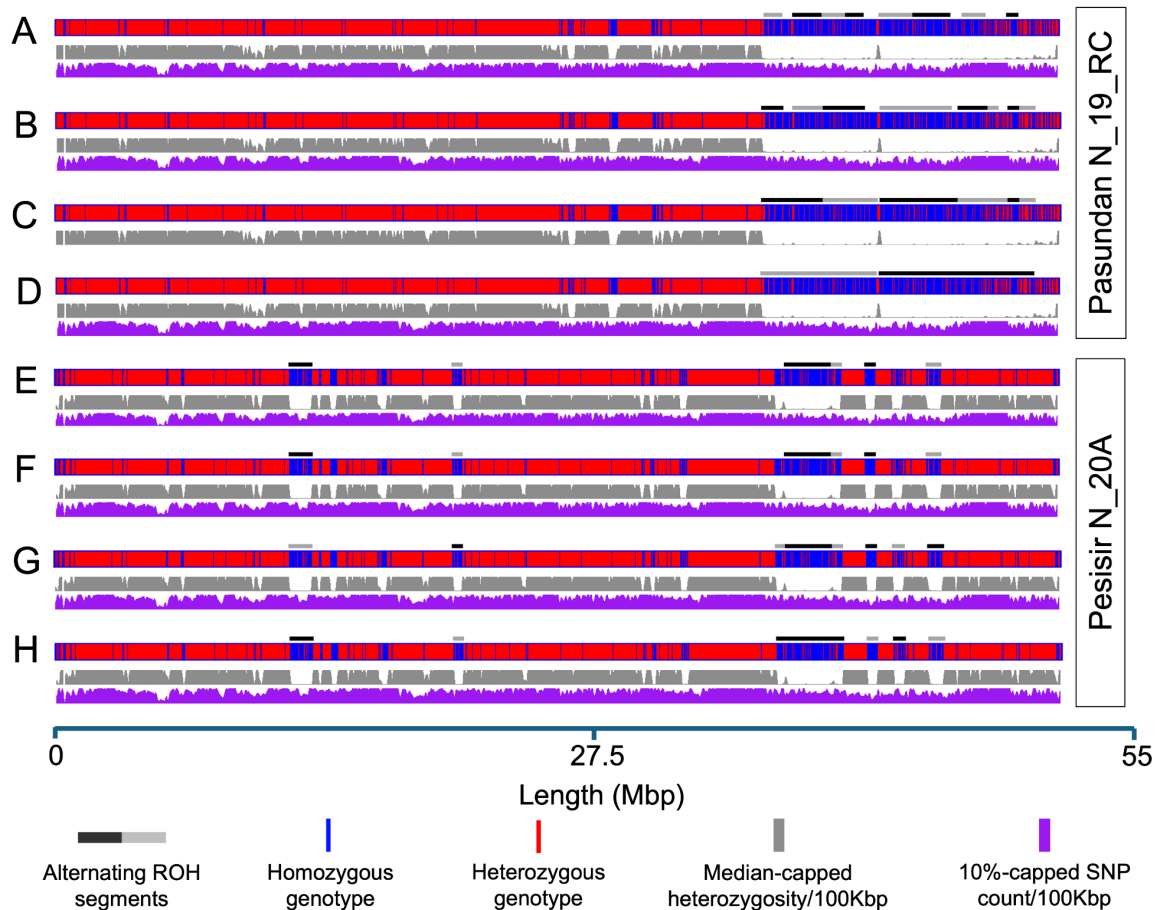
Supplementary Figure 3. Individual ancestry proportions estimated with HaploNet assuming from $K = 3$ to $K = 7$. Individuals are grouped by sampling locality.



Supplementary Figure 4. Global genetic differentiation measured as F_{ST} values using Hudson's estimator between all population pairs by PLINK2.

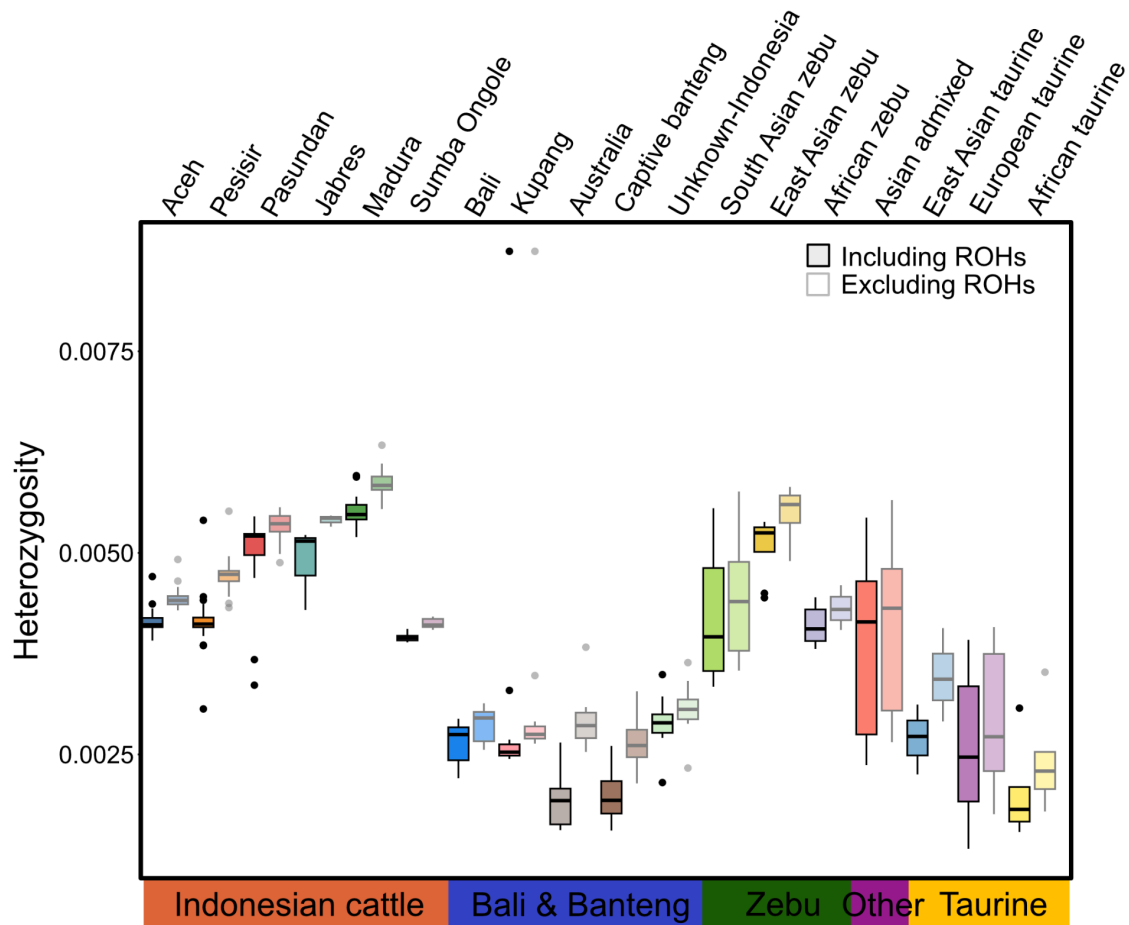


Supplementary Figure 5. Neighbornet visualization of **a**, all 18 populations and **b**, of the 13 banteng, zebu and Indonesian populations based on global F_{ST} values in Supplementary Fig. 4.

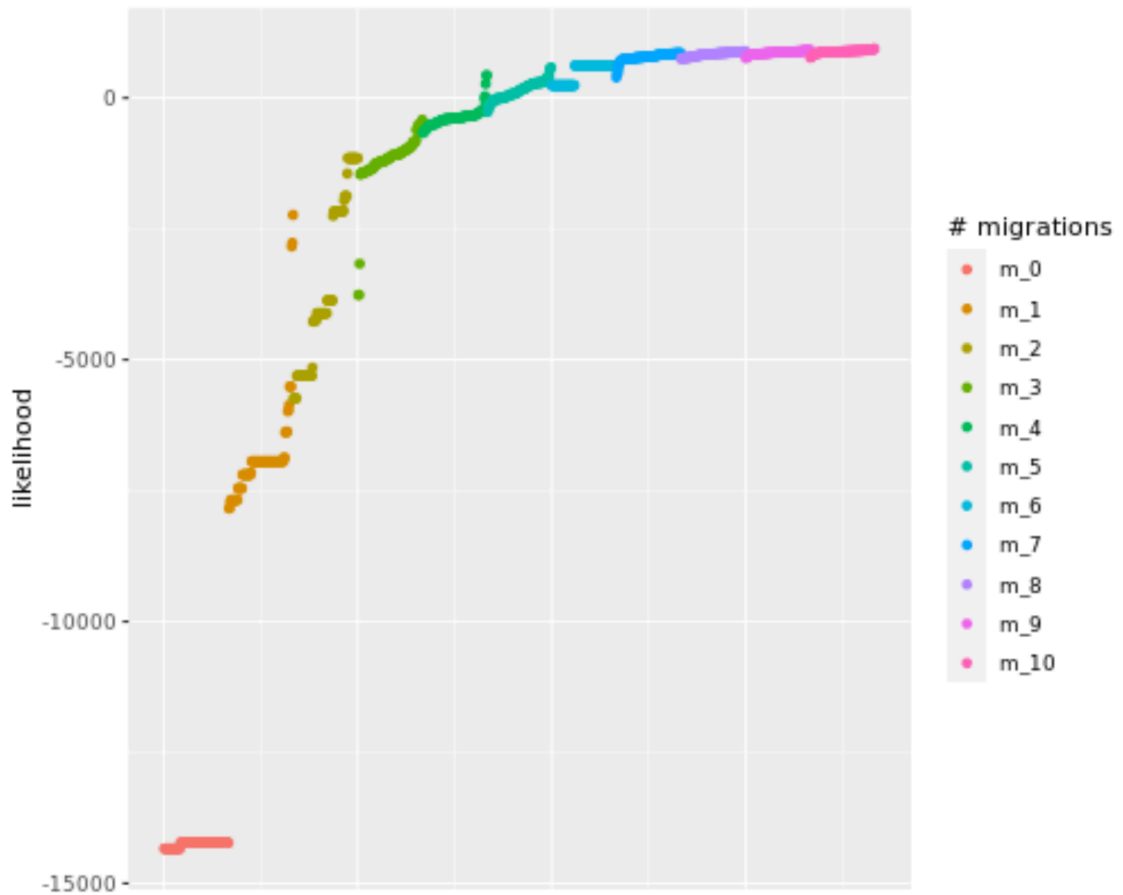
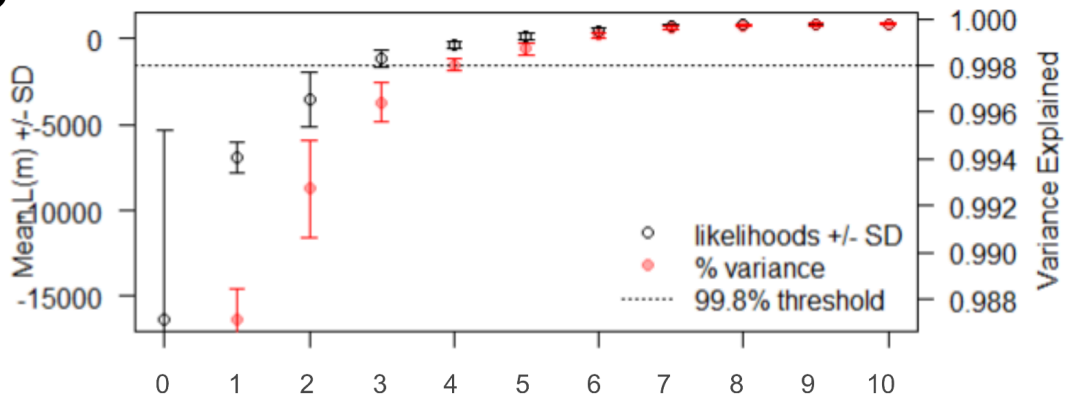


Supplementary Figure 6. ROH segments visualisation in chromosome 29 of one individual from Pasundan (N_19_RC, panel **A-D**) and one individual from Pesisir (N_20_A, panel **E-H**). These samples were chosen as an example because they were found to be exceptionally inbred within their respective population. Each panel presents the results from one ROH calling setting, with three tracks being illustrated for each setting. The top track plots first homozygous genotype calls as blue, overlaid with heterozygous genotype calls in red. This track provides a reasonably clear identification of the putative ROHs. The middle track shows in grey bars the heterozygosity per 100 kb window capped by the median value of this statistic across the genome for that individual, further illustrating the underlying patterns of strongly reduced heterozygosity. The lower track shows in purple bars the SNP count per 100 kb window, capped by 10% of genome-wide SNP count for that individual, to illustrate whether the ROH inference is affected by SNP density in certain regions. PLINK ROH segment calls are shown by alternating black and grey segments in the top of blue and red bars. ROH segments were called using PLINK with default parameters except for minimum ROH size (`--homozyg-kb 500`) and different numbers of heterozygous sites allowed per ROH segment (`--homozyg-window-het 3` [A, E], 4 [B, F], and 5 [C, G]) and only considering non-missing genotypes (`--geno 0`). Finally, we show the effect of post-hoc merging of PLINK inferred ROHs less than 100 kb apart (E and H), which appears to resolve several (but not necessarily all) cases of erroneous breaking of longer ROHs that could not be resolved by any explored PLINK settings or data filtering.

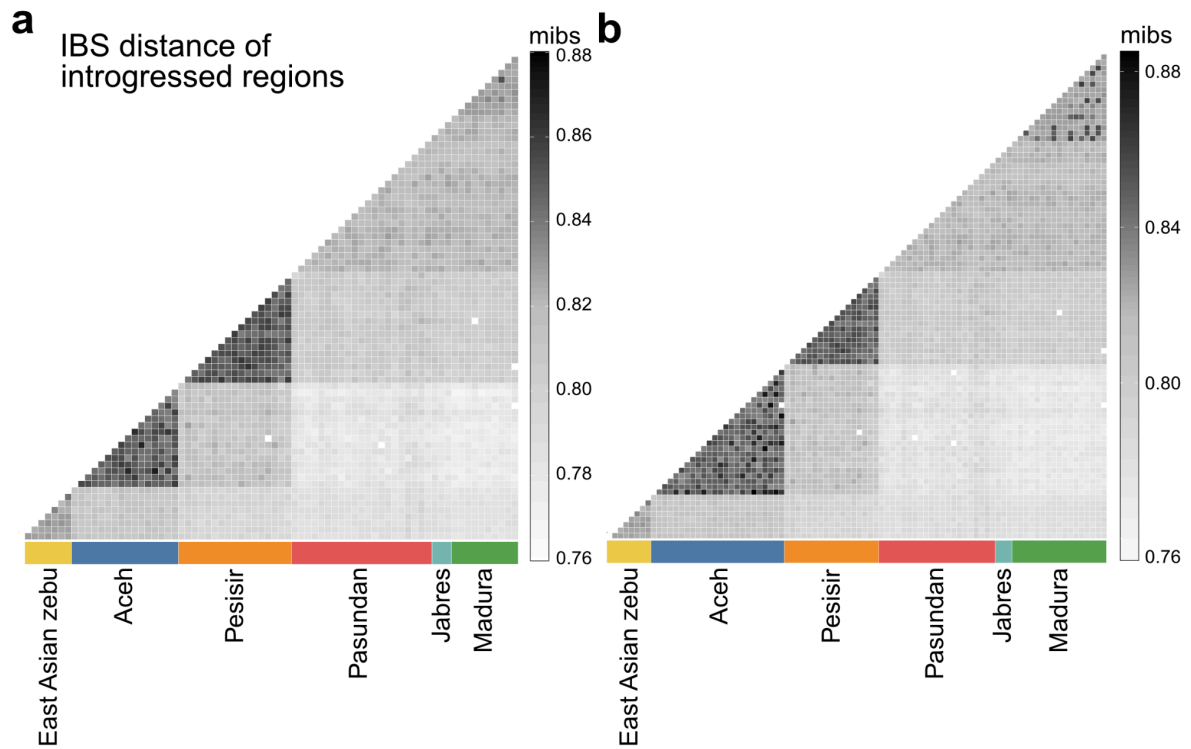
Supplementary Figure 7. Total ROH in genome fraction for each individual in the populations (different column). Colors indicate the category of ROH based on its length.



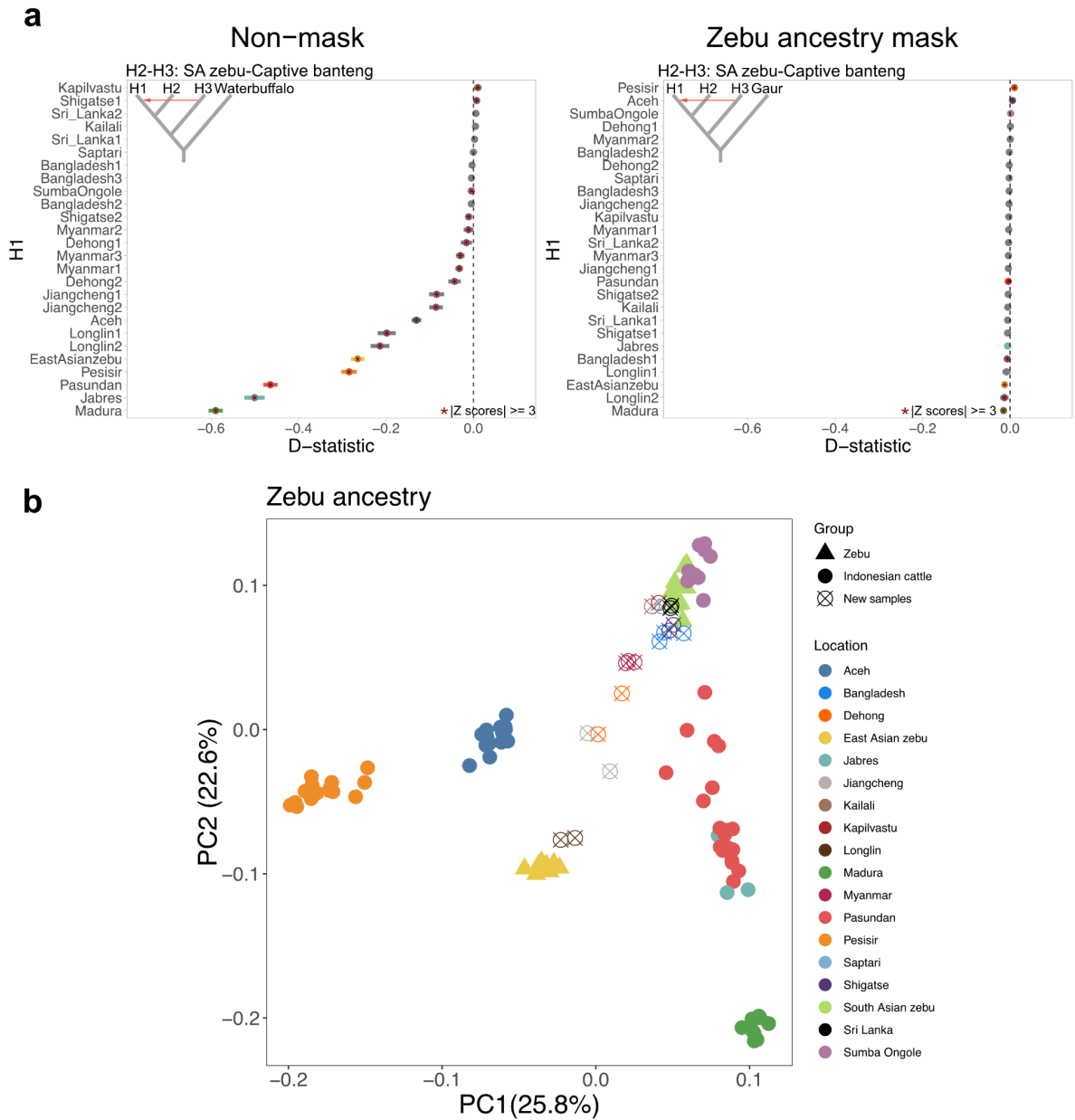
Supplementary Figure 8. Heterozygosity of all Indonesian cattle and other cattle based on genotype data with ROHs and excluding ROHs. Sample sizes of the populations are as follows: Aceh ($n = 25$), Pesisir ($n = 24$), Pasundan ($n = 24$), Jabres ($n = 3$), Madura ($n = 17$), Sumba Ongole ($n = 14$), Bali ($n = 19$), Kupang ($n = 16$), Australia ($n = 12$), Captive banteng ($n = 5$), Unknown-Indonesia ($n = 8$), South Asian zebu ($n = 13$), East Asian zebu ($n = 8$), African zebu ($n = 10$), Asian admixed ($n = 15$), East Asian taurine ($n = 3$), European taurine ($n = 10$), and African taurine ($n = 5$). Boxplots indicate median (centre line), the 25th and 75th percentiles (box), and the highest and lowest values within the upper and lower quartiles $\pm 1.5 \times$ interquartile range, respectively (whiskers).

a**b**

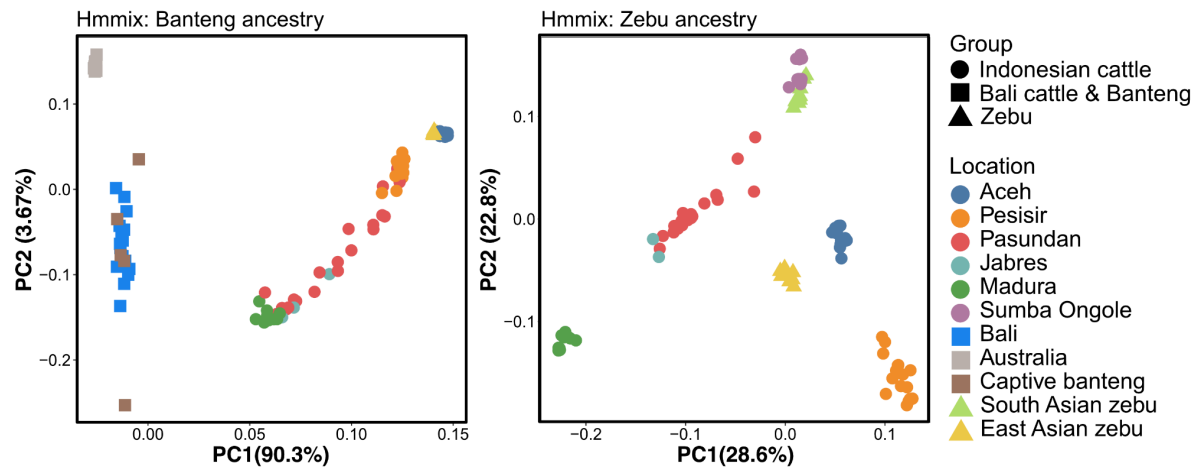
Supplementary Figure 9. Likelihood comparisons from different numbers of migration inferred by TreeMix. **a**, The likelihood of each model of phylogenetic network inferred by TreeMix, across all number of migration (m) scenarios (0-10). **b**, The mean and standard deviation of composite likelihood (left x-axis) and proportion of variance explained (right x-axis) based on 100 iterations, produced by OptM. The horizontal line represents the 99.8% threshold.



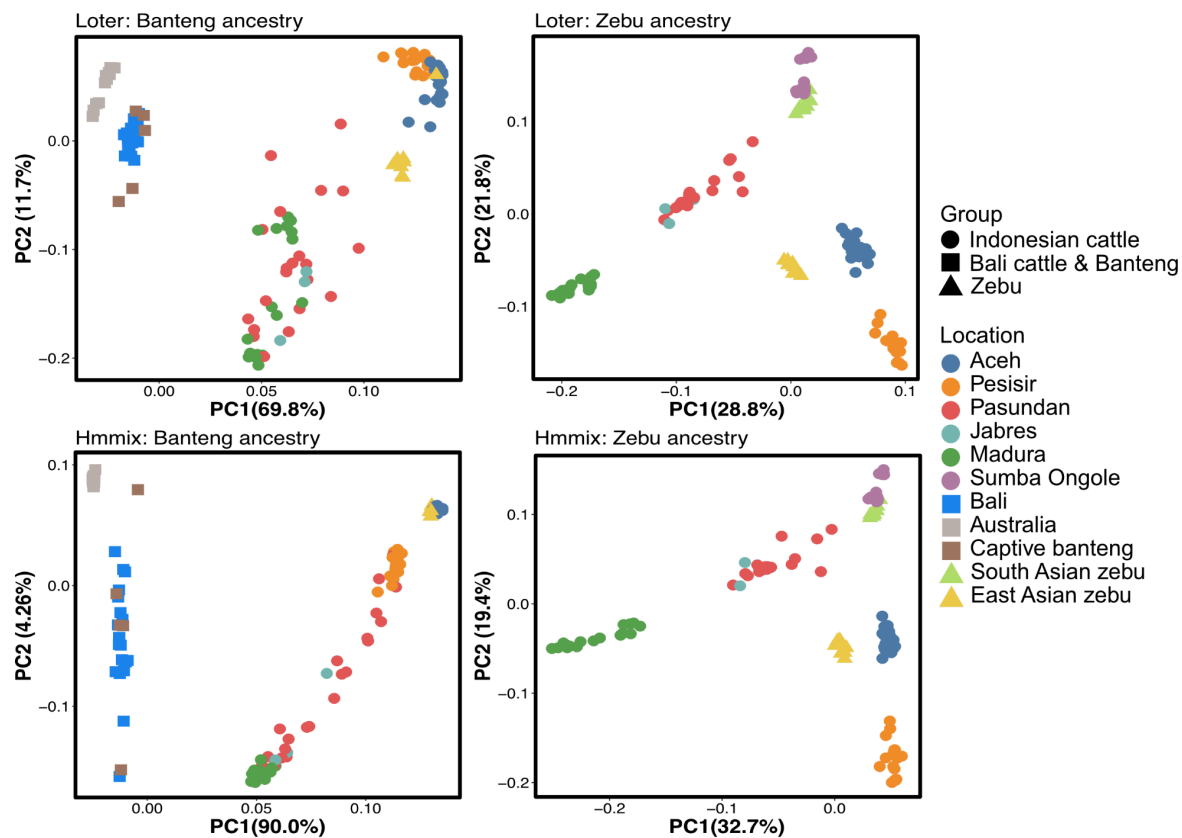
Supplementary Figure 10. Pairwise ibs matrix (mibs) calculated from overlapping archaic regions (probability > 0.9) between two individuals using 10 kb of window size, based on **a**, individuals removing one of each pair of individuals with $K1 > 0.2$ identified by ngsRelate in Supplementary Figure 31 and **b**, all individuals.



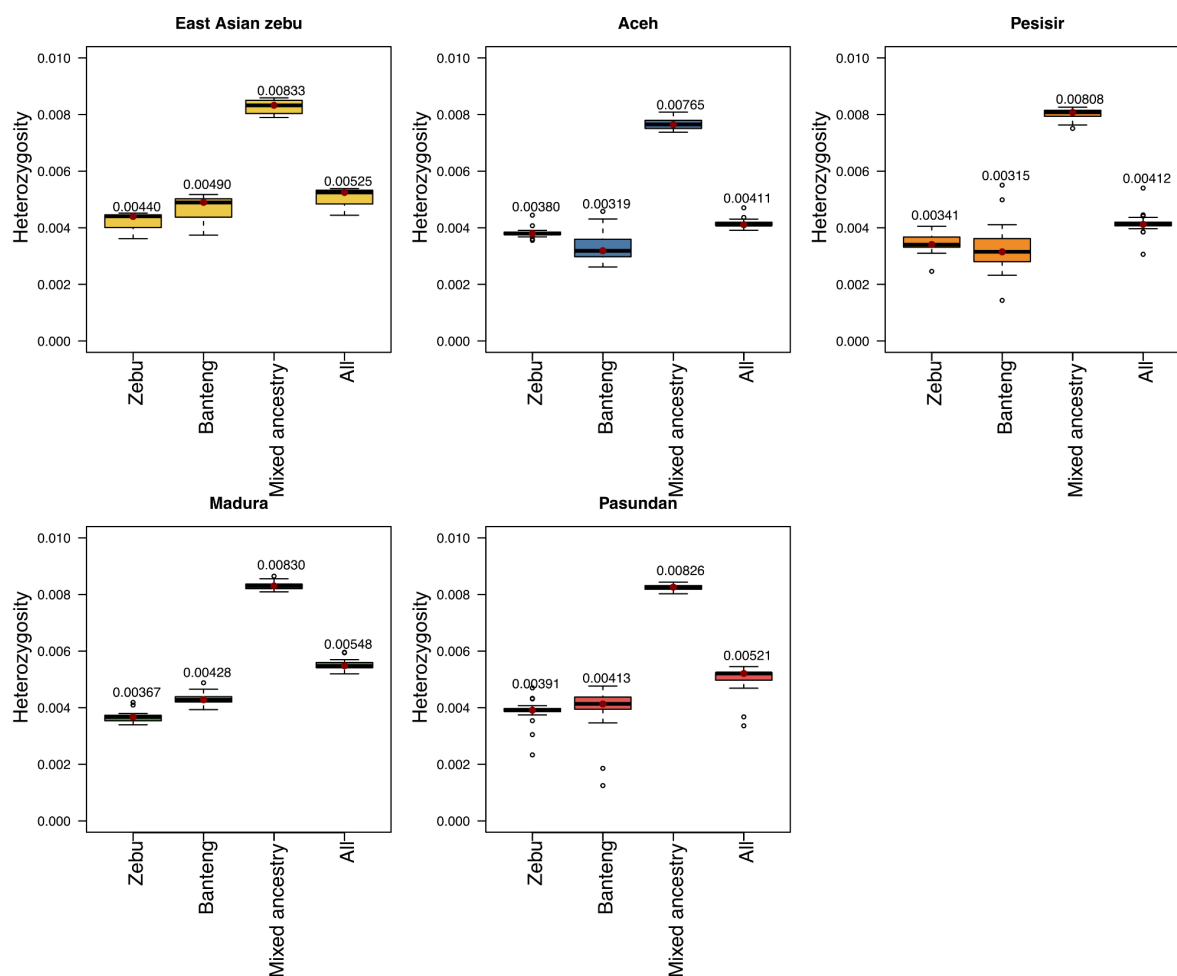
Supplementary Figure 11. D-statistics calculated by ADMIXTOOLS2 **a**, when using all sites (left panel), and masked genome segments of zebu ancestry origin inferred by LOTER (right panel). A significant negative non-zero value, as depicted by the red arrow in the graphic for each panel, provides evidence for gene flow between H3 and H1. Data are presented as the estimated D-statistic \pm 3 standard errors. Star represents significant allele sharing for each combination. **b**, PCA analysis on genome segments of zebu ancestry origin when adding newly downloaded additional samples from northernmost Southeast Asia and East Asia (e.g. Myanmar and southern China), as inferred by LOTER.



Supplementary Figure 12. Population structure of banteng ancestry origin and zebu ancestry origin inferred by Hmmix. One of each pair individuals with $K1 > 0.2$ identified by ngsRelate were removed.

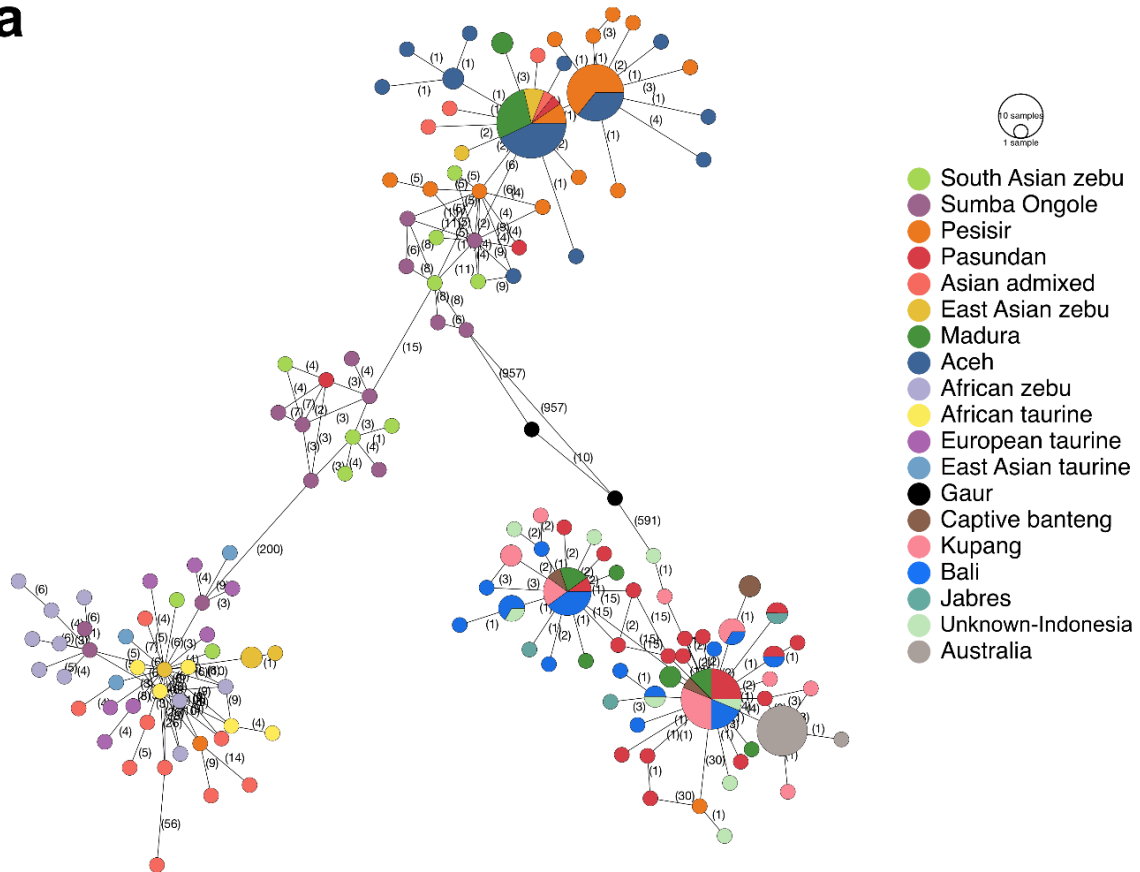


Supplementary Figure 13. Population structure of banteng ancestry origin and zebu ancestry origin inferred by Loter and Hmfix using all of individuals.

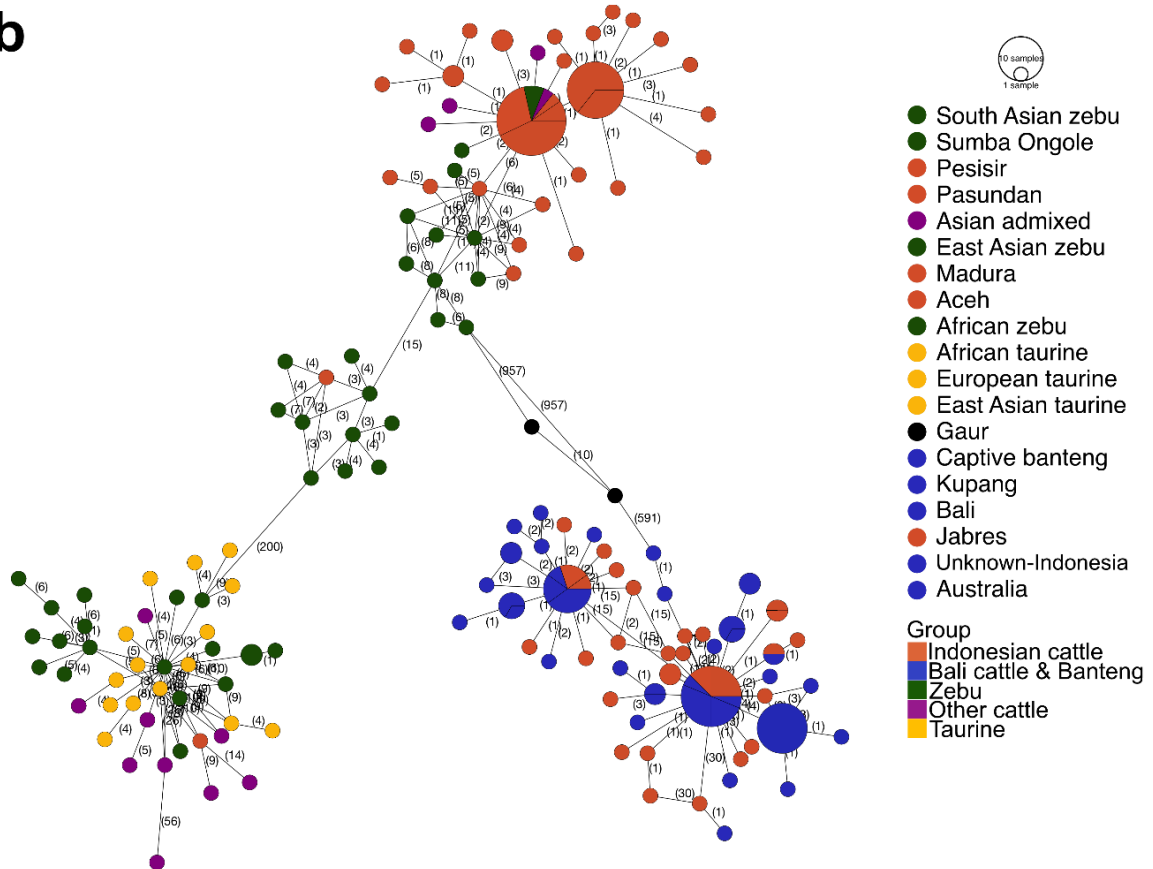


Supplementary Figure 14. Heterozygosity in admixed cattle populations according to the local ancestry tracts along each individual's genomes. Sample sizes for each population are as follows: Aceh (n = 24), Pesisir (n = 17), Pasundan (n = 21), Madura (n = 17), East Asian zebu (n = 8). Boxplots indicate median (centre line), the 25th and 75th percentiles (box), and the highest and lowest values within the upper and lower quartiles $\pm 1.5 \times$ interquartile range, respectively (whiskers).

a

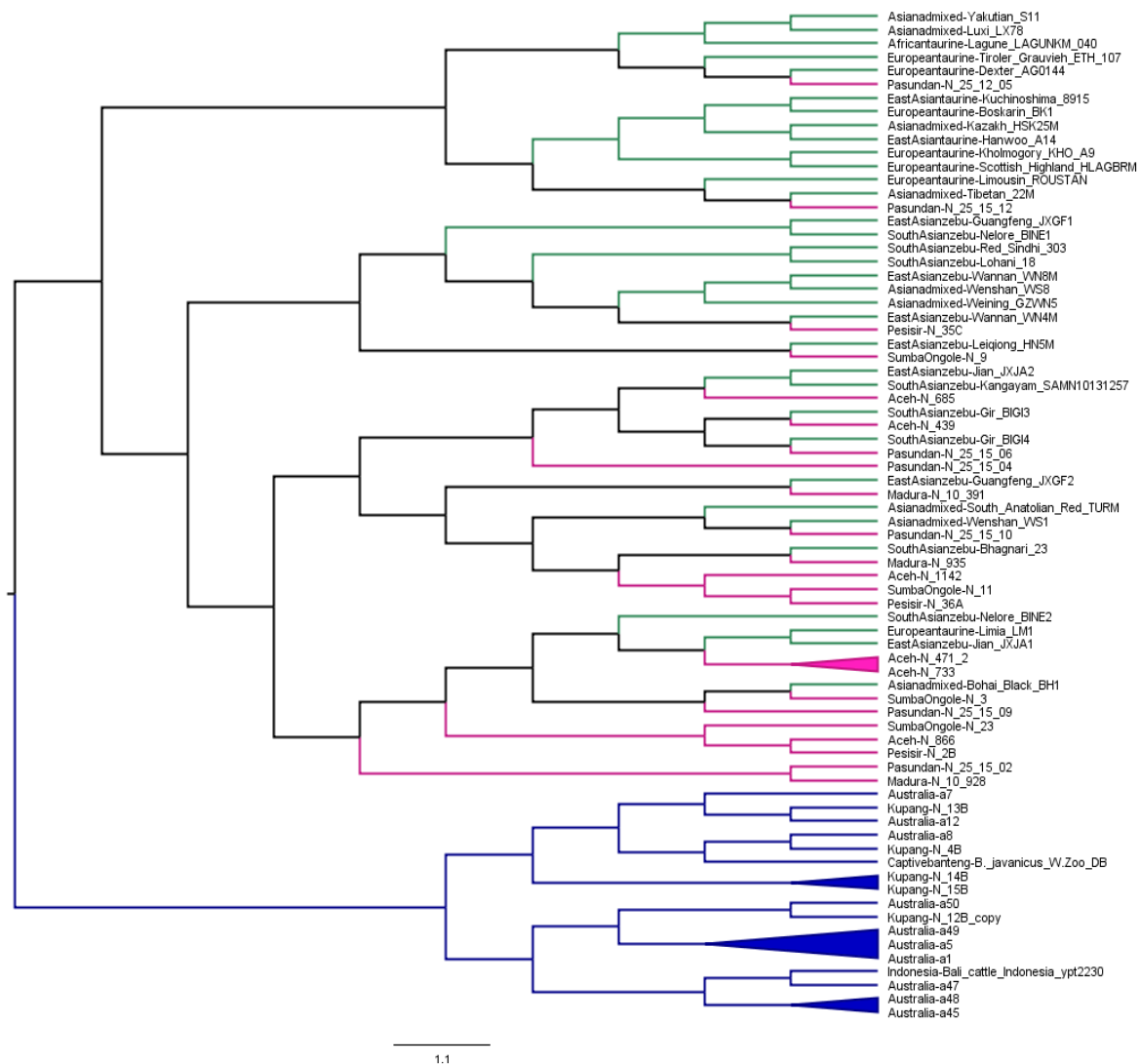


b

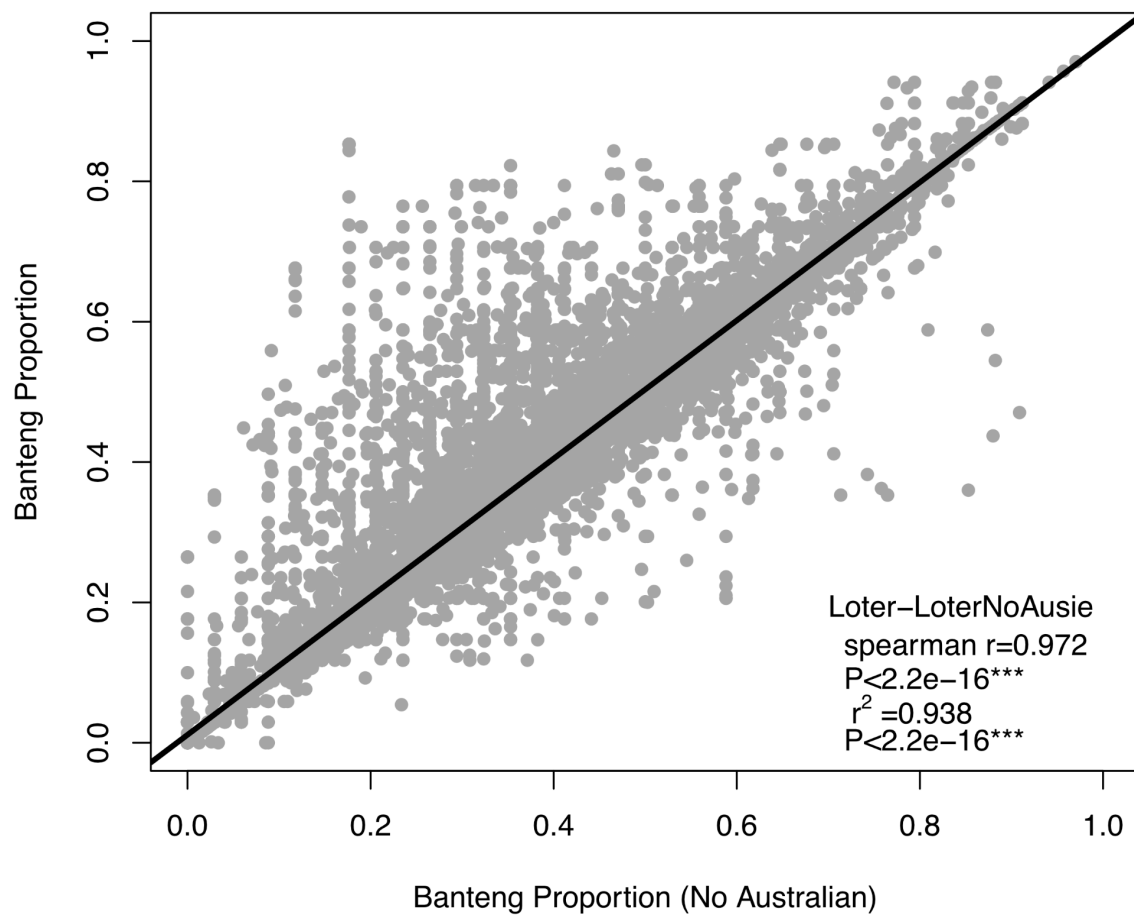


Supplementary Figure 15. Phylogenetic haplotype network for mitochondrial DNA (mtDNA) between

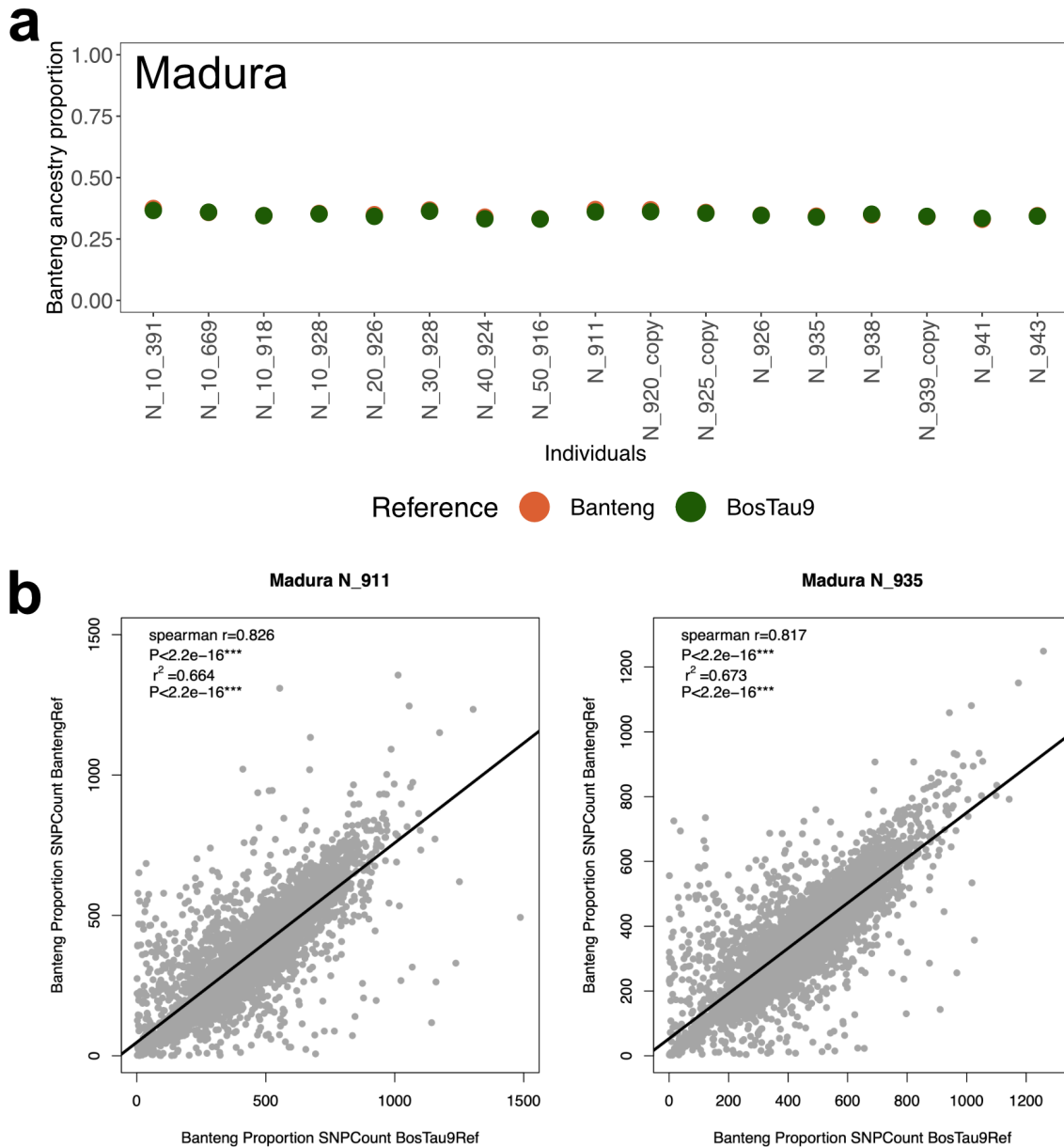
cattle illustrated by **a**, populations and **b**, groups using POPART. The number of mutations between each haplotype is shown in digits above each branch, which are not to scale.



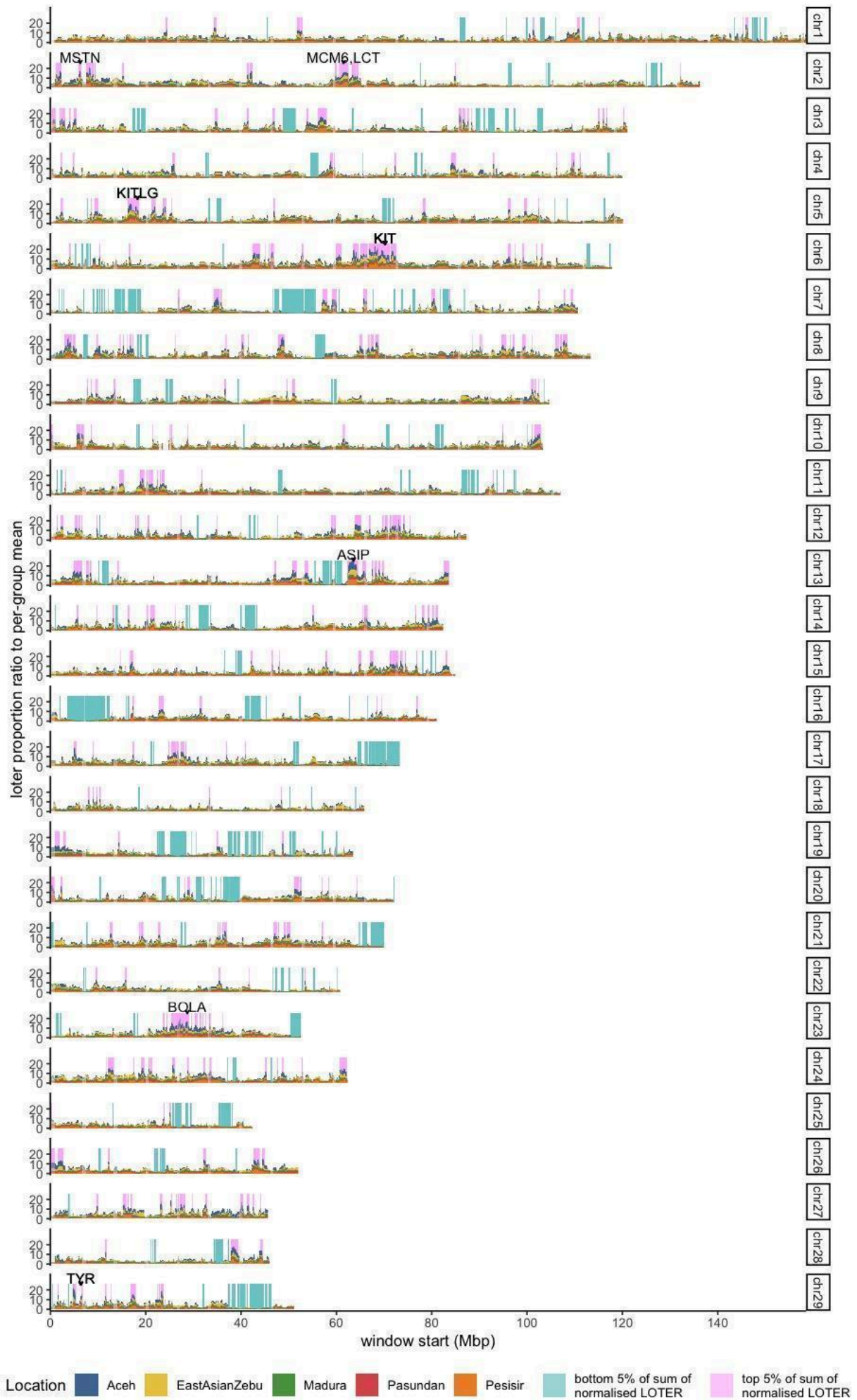
Supplementary Figure 16. Phylogenetic Maximum Clade Credibility (MCC) tree for Y chromosome between cattle using BEASTv1.10.4. We categorise all samples into 3 big clusters, banteng and Bali cattle (blue), Indonesian cattle: Aceh, Pesisir, Pasundan, Jabres, Madura and Sumba Ongole (pink), and other cattle (green).



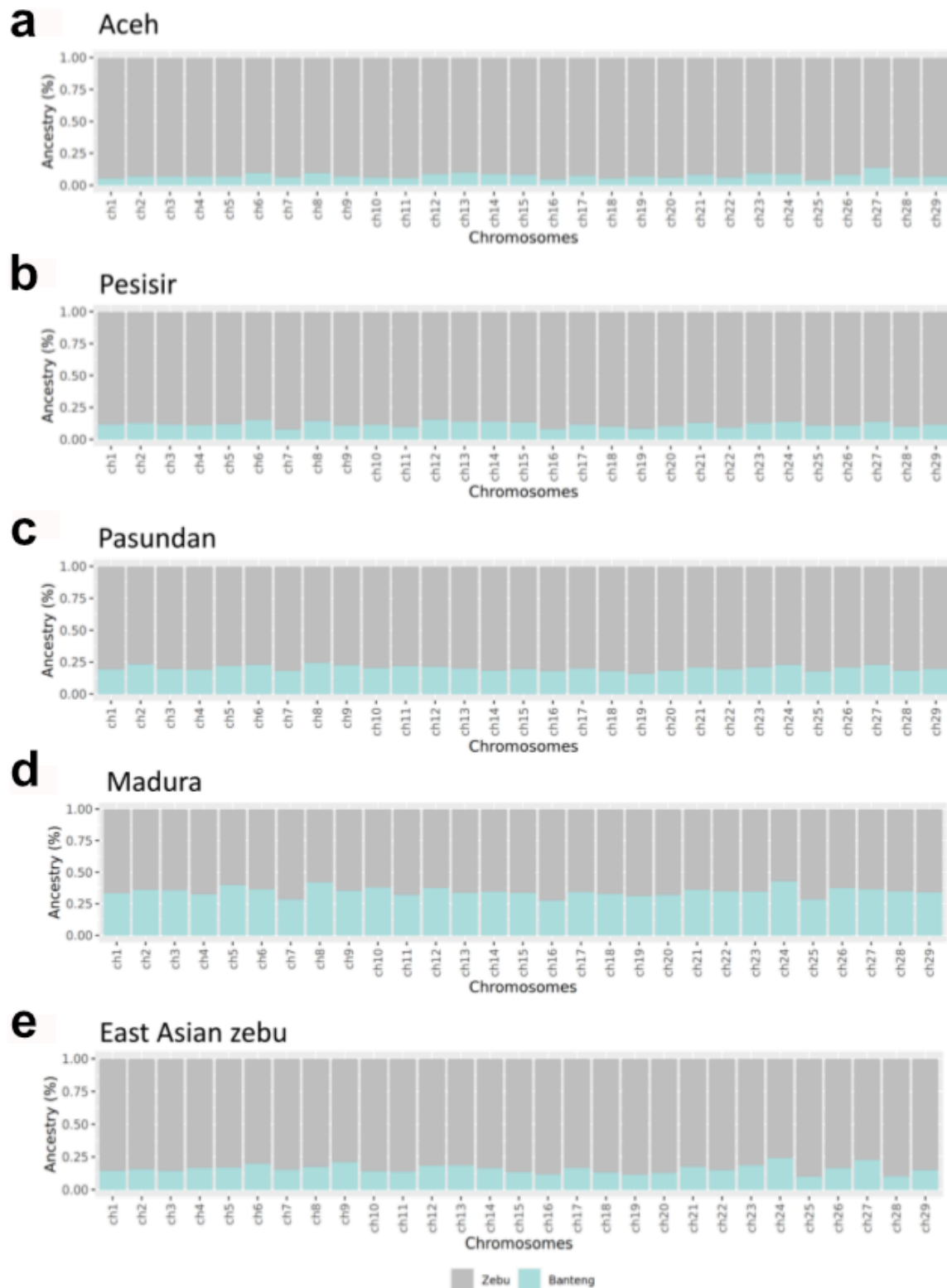
Supplementary Figure 17. Correlation between banteng ancestry proportion in Madura population when including and excluding Australia population from the ancestry source reference panel.



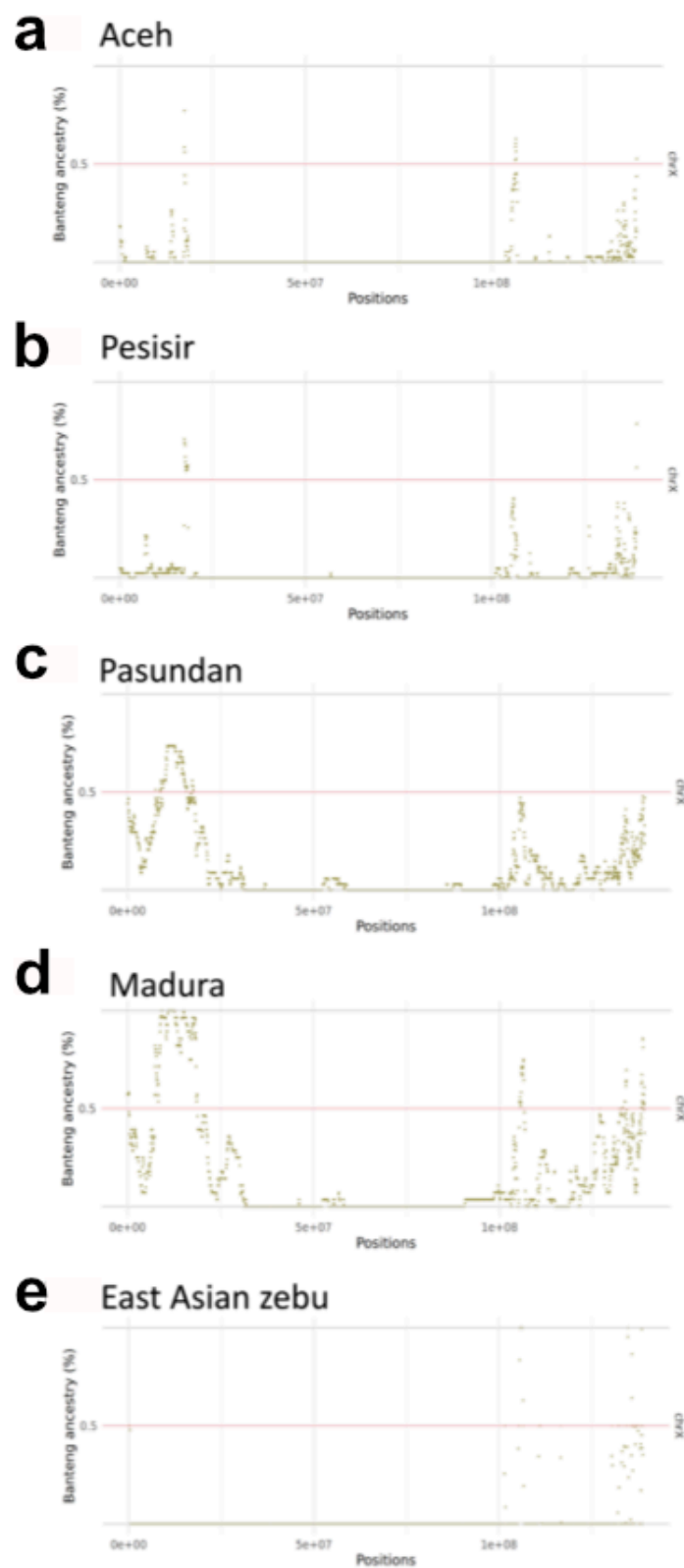
Supplementary Figure 18. Evaluation of how reference genome (*Banteng* and *BosTau9*) influence local ancestry inference by LOTER. **a**, Individual admixture proportion from LOTER when using *Banteng* and *BosTau9* as reference genome. **b**, Correlation between SNP counts inferred to be of banteng ancestry in each 50 kb genomic window for two individuals (N_911 and N_935) from Madura population.



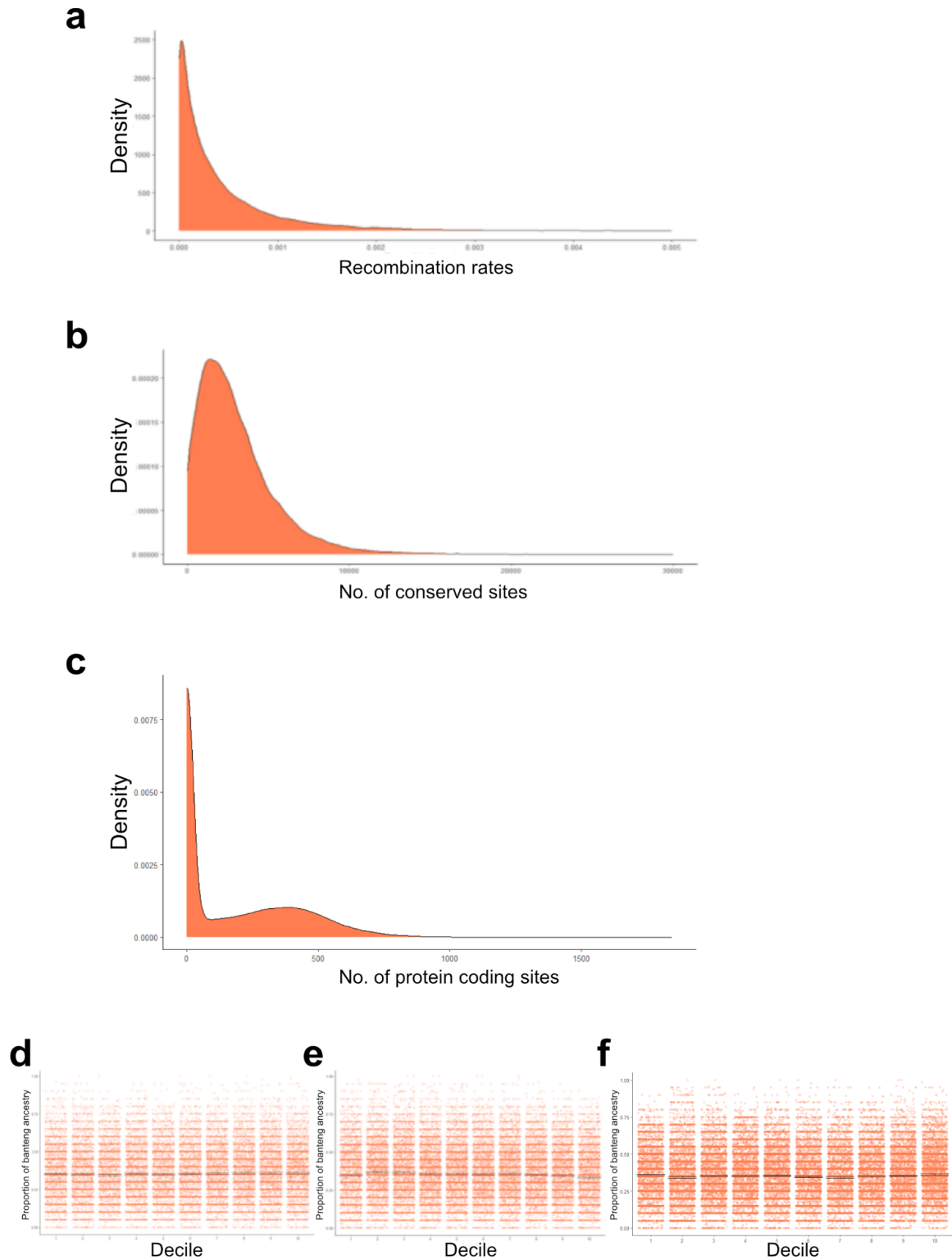
Supplementary Figure 19. Banteng ancestry across all of 29 chromosomes in five cattle groups (Aceh, East Asian zebu, Madura, Pasundan, and Pesisir). Window-based scan of regions with extreme banteng ancestry using the proportion of inferred banteng SNPs from LOTER for each cattle group divided by the mean proportion per group. Pink shade marks regions in the genome-wide top 5% of the normalized LOTER summed across all groups while the light blue shade marks the bottom 5% of the normalized LOTER summed across all groups indicating descent of introgression.



Supplementary Figure 20. Average window-based banteng proportions by chromosomes in **a**, Aceh, **b**, Pesisir, **c**, Pasundan, **d**, Madura, **e**, East Asian zebu.

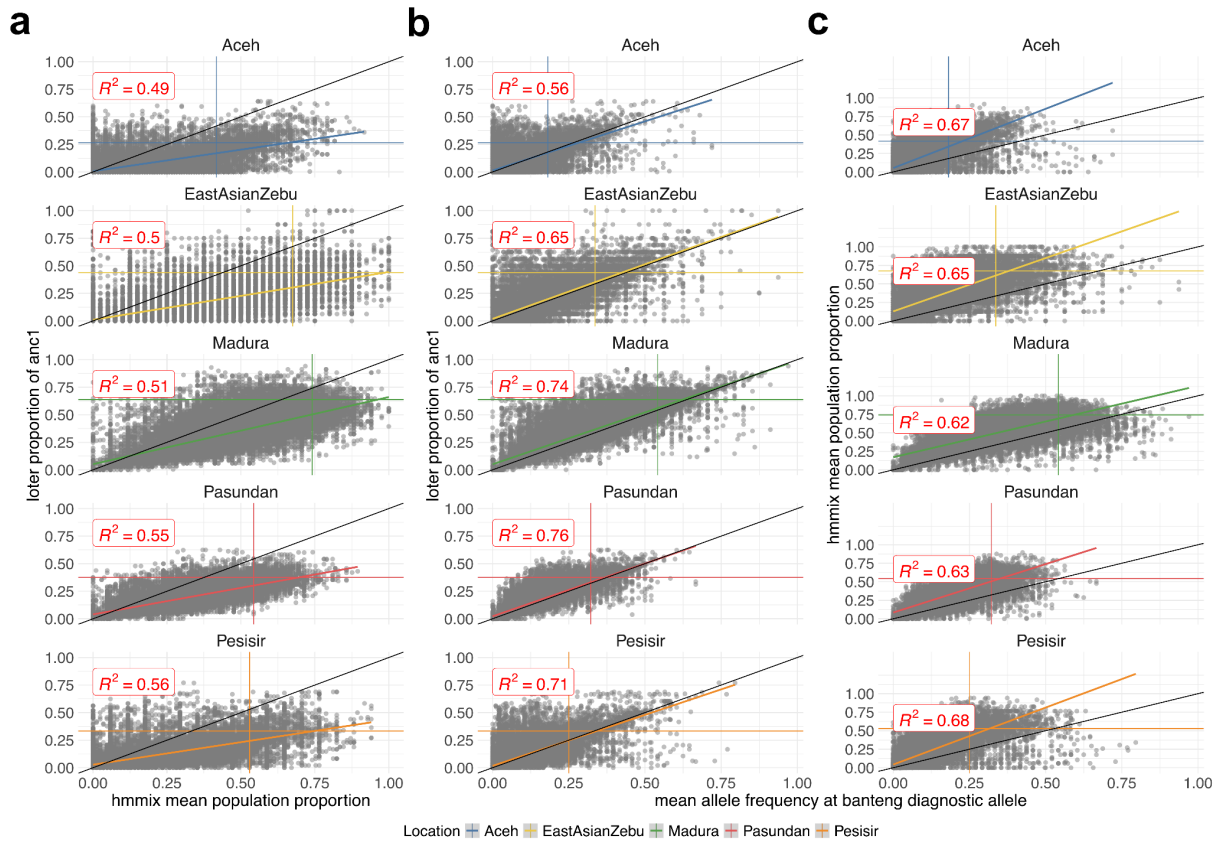


Supplementary Figure 21. Average window-based banteng proportions in chromosome X in **a**, Aceh, **b**, Pesisir, **c**, Pasundan, **d**, Madura, **e**, East Asian zebu.

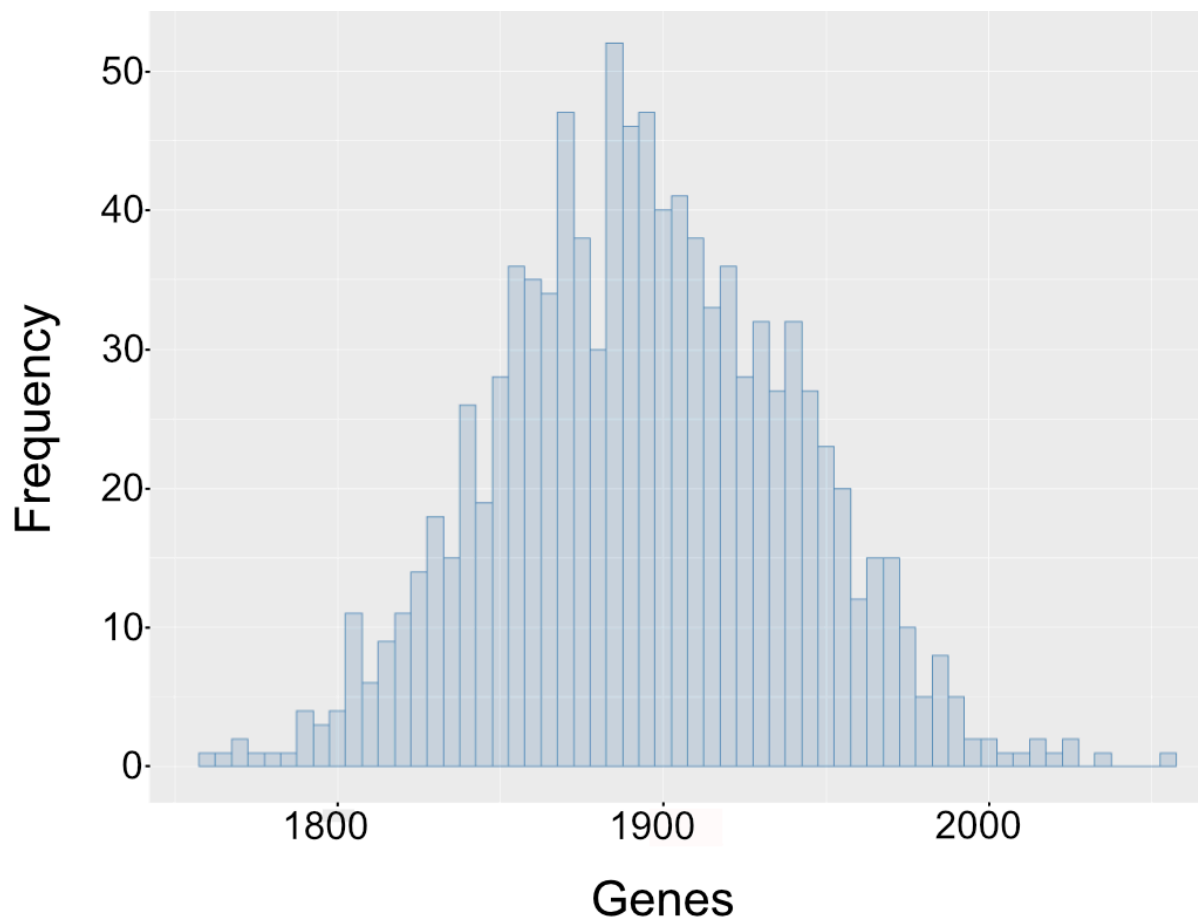


Supplementary Figure 22. Correlation between the mean banteng ancestry and genomic features recombination rate, conservation score, and coding region density in Madura cattle. **a**, Kernel density estimate of distribution of mean recombination rate in 50 kb windows. **b**, Kernel density estimate of distribution of mean conserved sites in 50 kb windows. **c**, Kernel density estimate of distribution of protein coding sites in 50 kb windows. **d**, Proportion of banteng ancestry in Madura by decile of recombination rate in 50 kb windows. Pearson's correlation coefficient between recombination rate and banteng ancestry: 0.01850509 (p-val: 0.000283). **e**, Proportion of banteng ancestry in Madura by

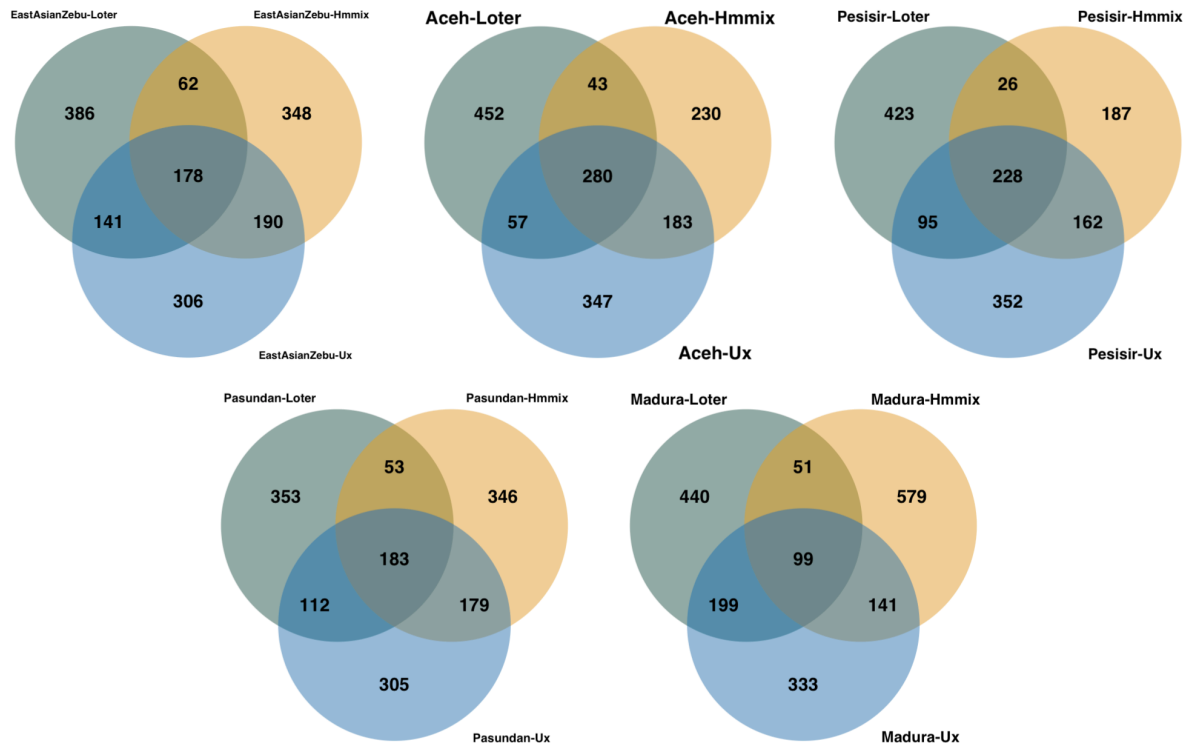
decile of conserved sites in 50 kb windows. Pearson's correlation coefficient between no. of conserved sites and banteng ancestry: -0.04284158 (p-val: 1.355628e-21). **f**, Proportion of banteng ancestry in Madura by decile of protein coding sites in 50 kb windows. Pearson's correlation coefficient between no. of protein coding sites and banteng ancestry: 0.01299004 (p-val: 0.003806845). The partial discreteness of the minor parent (banteng) ancestry proportion comes from many inferred local ancestry tracts spanning multiple whole windows, making the minor parent ancestry either 0 or 1 in many windows of each haplotype.



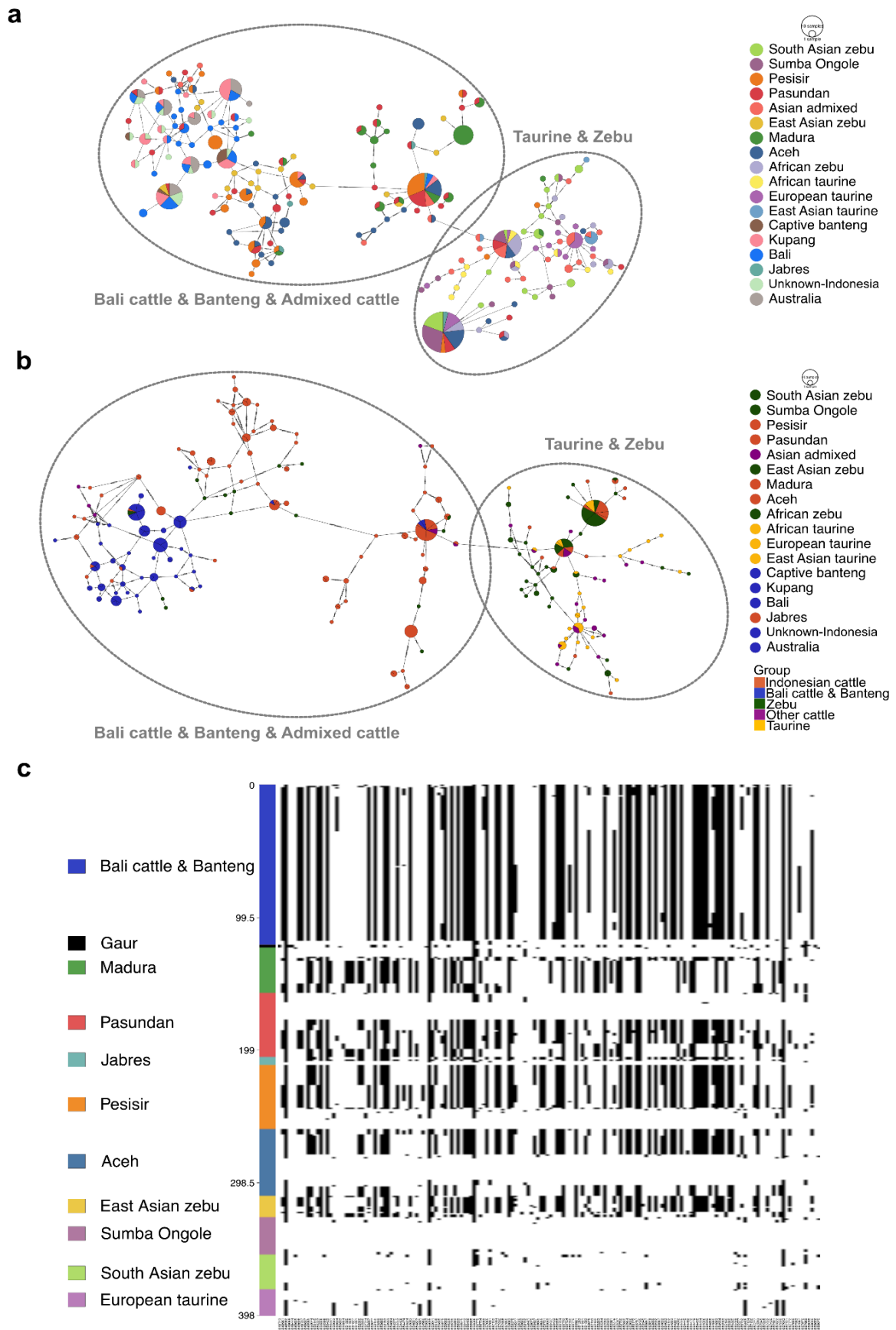
Supplementary Figure 23. Positive correlation between the three local ancestry inference methods estimated from association between **a**, LOTER proportion of banteng ancestry and Hmfix mean population proportion, **b**, LOTER and U_x , the mean allele frequency of banteng diagnostic allele, and **c**, Hmfix mean population proportion and U_x , all within 50 kb sliding windows across the genome. Regression line colors are corresponding to each cattle group (Aceh, East Asian zebu, Madura, Pasundan, and Pesisir) and solid black line showing area of slope equal to one. Positive correlations are all significant ($P < 0.001$) using the Spearman correlation test.



Supplementary Figure 24. Simulation number of genes from 5% samplings of windows randomly. As a comparison, the top 5% proportion of population-wise banteng ancestry as inferred from LOTER overlaps with 839 genes in Madura, 813 genes in Pasundan, 892 genes in Pesisir, 1041 genes in Aceh, and 914 genes in East Asian zebu, which are significantly lower gene density than expected by chance as shown here.

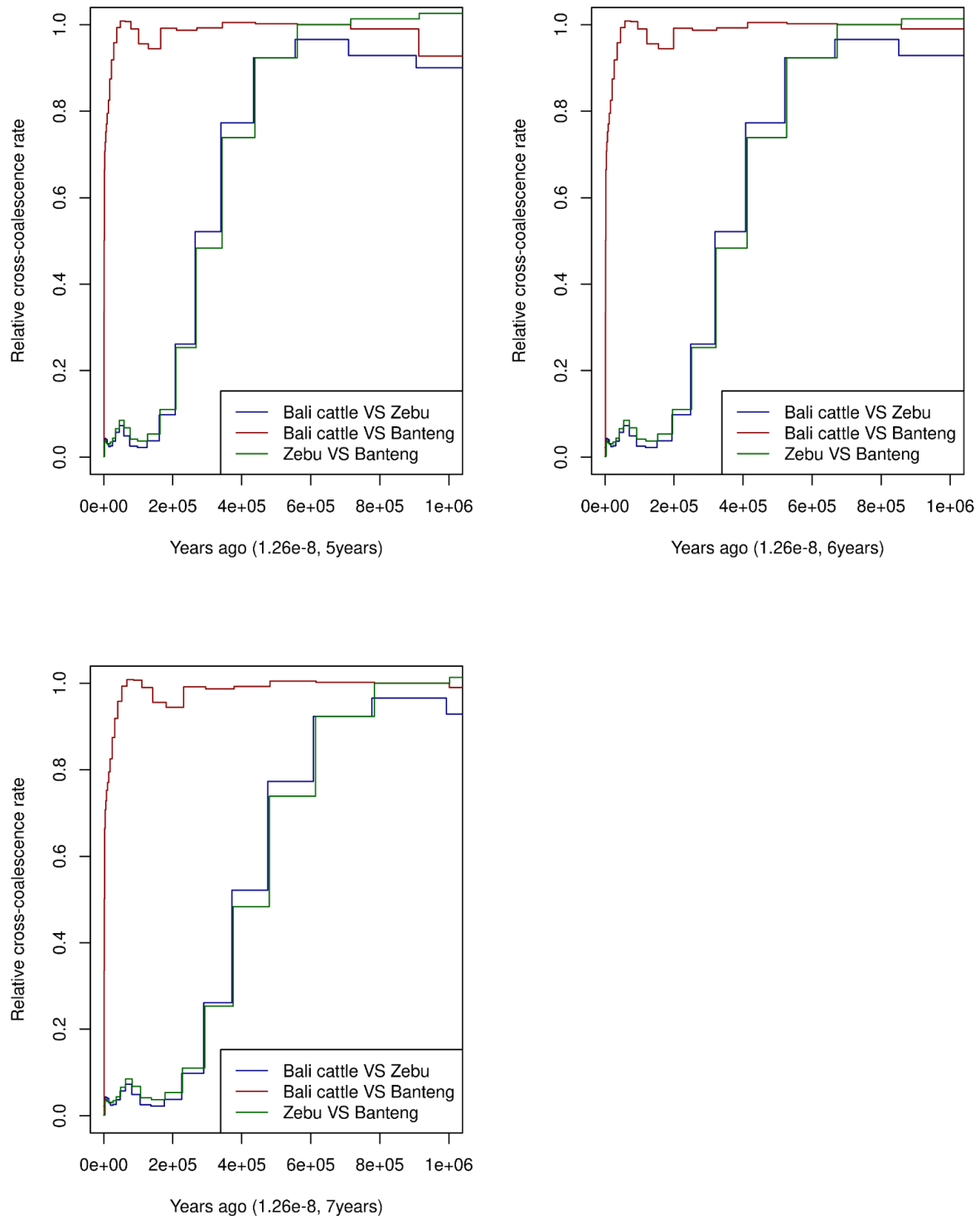


Supplementary Figure 25. Venn diagram plot showing the number of overlapping genes for each Indonesian breed between any two out of three methods (LOTER, Hmmix, and U_x). Gene ID lists are shown in Supplementary Data 9-11.



Supplementary Figure 27. Haplotype structure of all populations in the *ASIP* gene on chromosome 13 between 63.64 Mb and 63.67 Mb. **a**, Haplotype network illustrated by different populations. **b**,

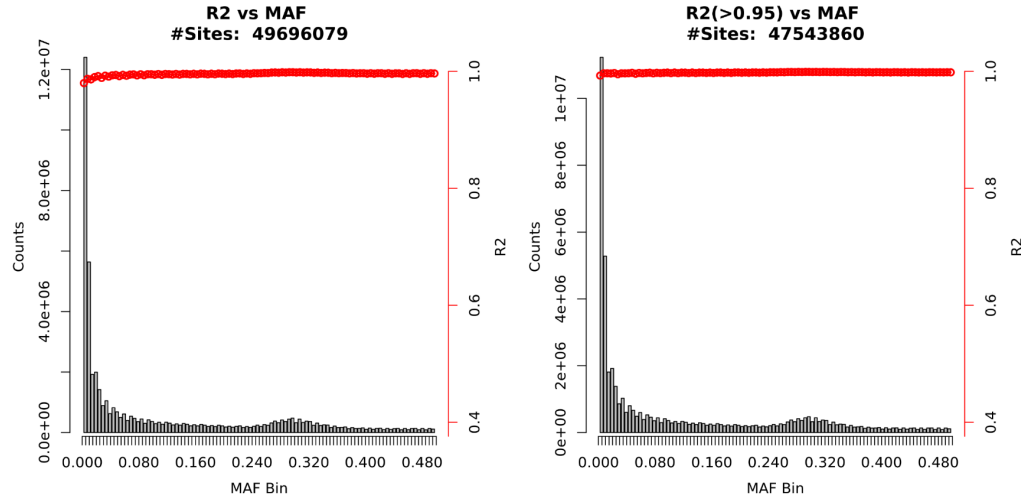
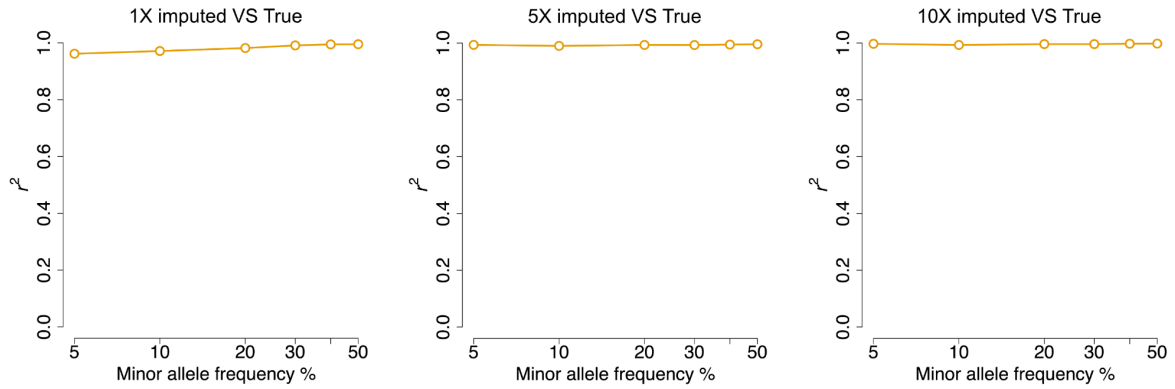
Haplotype network illustrated by different groups. **c**, Haplotype structure by haplostrips.



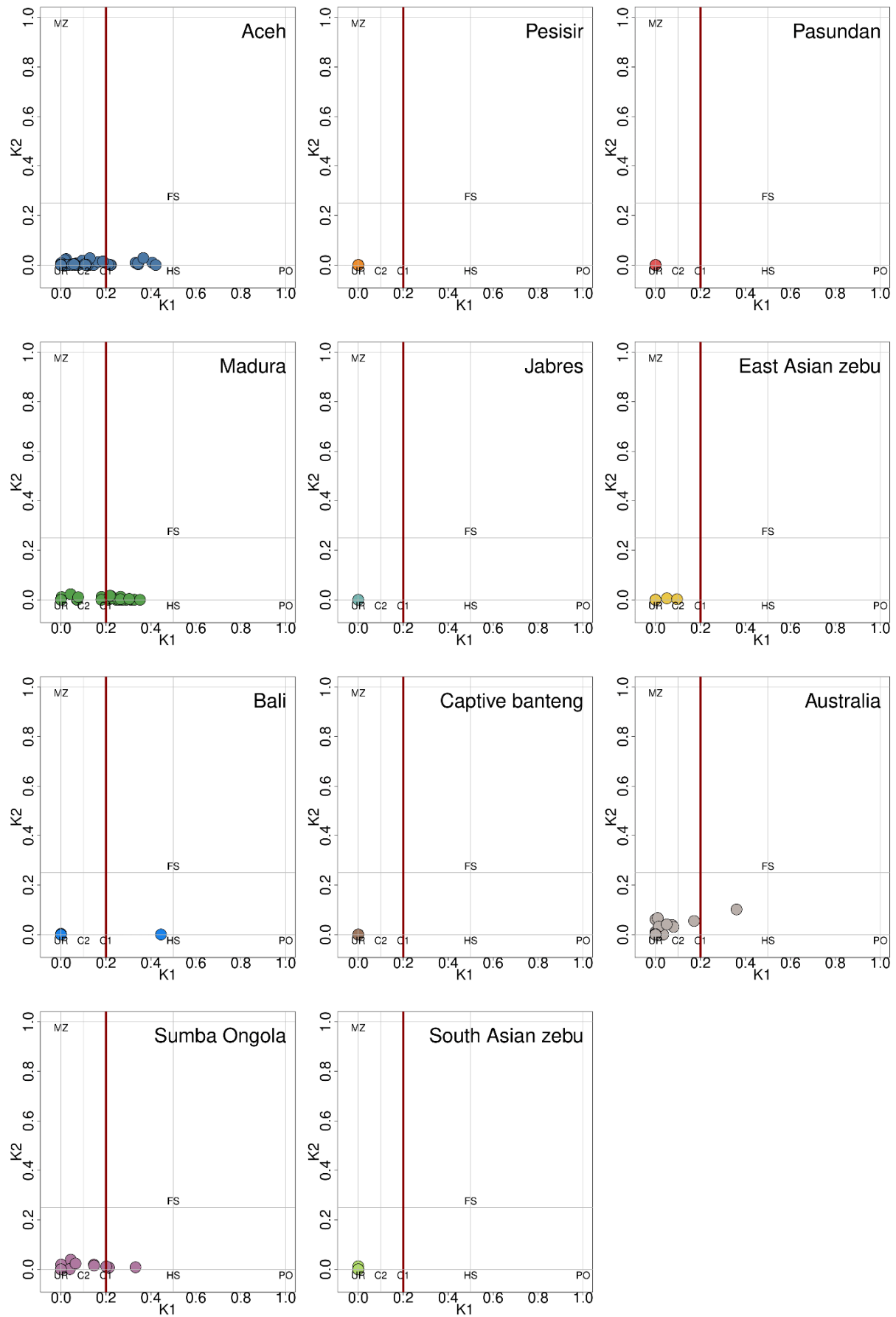
Supplementary Figure 28. The relative cross coalescence rate from two individuals per population between *Bos Indicus* (Zebu), *Bos Javanicus* (Banteng) and Bali cattle using MSMC2. Visualization were scaled from results by assuming generation time of 5-7 years and a mutation rate of 1.26×10^{-8} per generation.



Supplementary Figure 29. Spearman correlations of banteng proportion in top 5% windows of 50 kb. As the top 5% windows are different from one breed to another, the pairwise correlation should be read by column (not symmetrical triangle matrix).

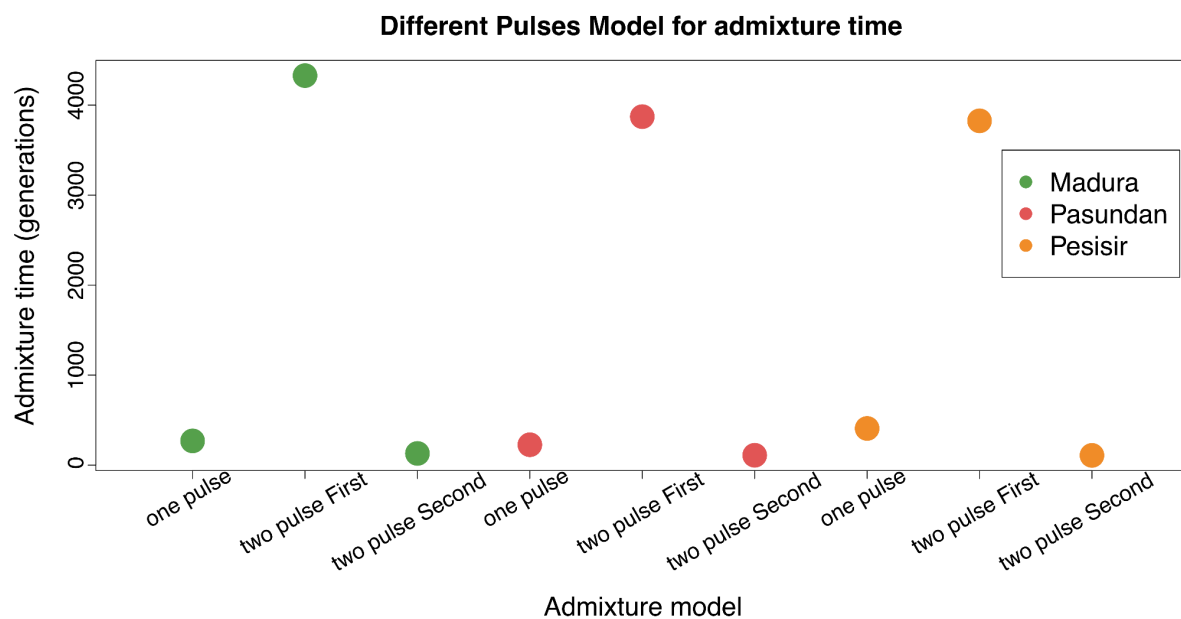
a**b**

Supplementary Figure 30. Evaluation of accuracy of imputation. **a**, Visualization of minor allele frequency (MAF) distribution of genotype discordance between the original vcf and imputed vcf genotype files. R^2 is the estimated squared correlation ($0 \leq R^2 \leq 1$) between the allele dosage with highest posterior probability in the genotype probabilities file and the true allele dosage for the marker. **b**, Visualisation of MAF distribution (chromosome 1) of individual LIB112407_Banteng_85B_Texas (depth of 34.86 X) between imputed genotypes (downsampled into depths of 1X, 5X and 10X) and the high-quality genotype calls for the full data from this sample. MAFs were calculated using all of the individuals in this study.



Supplementary Figure 31. Estimates of relatedness for cattle within each sampling location by

NGSrelate. One of each pair with $K1 > 0.2$ was removed for analysis of the matrix based on Hmfix and inference of ancestry PCA structure using EMU.



Supplementary Figure 32. Inference of admixture time using AdmixtureHMM with one pulse model and two pulse model as comparison. Extremely lower admixture proportion (~0.01%) and extremely higher admixture time suggest that the two pulse model is the spuriously model compared with one pulse model.

Supplementary Note. Gene ontology enrichment analysis in this study

Aceh. Gene ontology (GO) enrichment analysis identified a total of 110 GO terms that were significantly enriched for the candidate genes (Supplementary Data 8), including 43 terms of biological process related to positive regulation of biological process (GO:0048518; adj. P = 1.19×10^{-9}) and developmental processes (GO:0032502; adj. P = 6.69×10^{-4}), among others. There were also a large number of enriched GO categories and KEGG pathways related to immune functions (Supplementary Data 8).

Madura. Notable genes found among the outlier windows in Madura with well-known trait associations in cattle include *KITLG*, strongly linked to coat color¹, *LCT* encoding the lactase gene required for lactose digestion^{2,3}, *EGFR*, associated with embryo development⁴ and body weight⁵. In Madura, we observed enrichment (Supplementary Data 8) in e.g. developmental processes (GO:0032502; adj. p-val=0.0048), anatomical structure development (GO:0048856; adj. p-value=0.0112), glutamate metabolic process (GO:0006536; adj. p-val=0.0250) and chemokine activity (GO:0008009; adj. p-val=0.0236). These enriched GO categories are plausibly related to selection on morphological, neurological/behavioral and immune system traits in Madura cattle.

Pasundan. The coat color gene *ASIP* was also among the outlier genes in Pasundan. Notable enrichment results (Supplementary Data 8) include developmental processes (GO:0032502; adj. p-val=0.0021), anatomical structure development (GO:0048856; adj. p-value=0.0035) and multicellular organism development (GO:0007275; adj. p-val=0.0039), mirroring the enrichment of anatomy related GO categories found in Madura. In addition, we found enrichment for blood coagulation (GO:0007596; adj. p-val=0.0191) and hemostasis (GO:0007599; adj. p-val 0.0232), both of which were previously found enriched in an analysis of heat adaptation in cattle⁶.

Pesisir. Some of the best-known coat color genes were among the outliers in Pesisir, including *ASIP*, *KIT*, *KITLG* and *TYR*, suggesting strong and polygenic selection on this trait in Pesisir cattle. In Pesisir, we found enrichment of interleukin-1 receptor activity (GO:0004908; adj. p-val=0.01196) related to immune response.

East Asian zebu. East Asian zebu were also enriched for developmental processes (GO:0032502; adj. p-val=0.0000866), and in addition for several GO categories and KEGG pathways related to the immune system, such as regulation of natural killer cell mediated

cytotoxicity (GO:0042269; adj. p-val=0.003065) and natural killer cell mediated cytotoxicity (KEGG:04650 ;adj. p-val=0.004076).

Reference

1. Weich, K. *et al.* Pigment Intensity in Dogs is Associated with a Copy Number Variant Upstream of. *Genes* **11**, (2020).
2. Mattar, R., de Campos Mazo, D. F. & Carrilho, F. J. Lactose intolerance: diagnosis, genetic, and clinical factors. *Clin. Exp. Gastroenterol.* **5**, 113–121 (2012).
3. Ingram, C. J. E., Mulcare, C. A., Itan, Y., Thomas, M. G. & Swallow, D. M. Lactose digestion and the evolutionary genetics of lactase persistence. *Hum. Genet.* **124**, 579–591 (2009).
4. Zhang, X. *et al.* Bulk and mosaic deletions of Egfr reveal regionally defined gliogenesis in the developing mouse forebrain. *iScience* **26**, 106242 (2023).
5. Paredes-Sánchez, F. A. *et al.* Associations of SNPs located at candidate genes to bovine growth traits, prioritized with an interaction networks construction approach. *BMC Genet.* **16**, 91 (2015).
6. Freitas, P. H. F. *et al.* Genetic Diversity and Signatures of Selection for Thermal Stress in Cattle and Other Two Bos Species Adapted to Divergent Climatic Conditions. *Front. Genet.* **12**, 604823 (2021).