Supplementary Information

LINE-1 rewires *GPR52* regulation in the rapid evolution of canine tameness

Contents

Supplementary Methods

1. GAMA model and implementation

1.1 Model overview

We developed GAMA (Genomic Analysis of Multiple Admixture), an ancestry-aware framework designed to reconstruct high-resolution, pan-breed local-ancestry maps in domestic animals and plants with complex interbreeding histories and limited ancestral reference samples. GAMA advances beyond existing methods in three principal ways. First, it explicitly models an empirically estimated per-site mismatch parameter, ϵ, which captures residual mismatches between observed local haplotypes and reference-panel haplotype-frequency models, thereby improving robustness to sequencing, genotyping, phasing and mutation-related noise. Second, it implements Laplace (additive) smoothing when estimating haplotype frequencies from small ancestral panels, thereby reducing the influence of extreme frequency estimates. Third, it can infer and represent unsampled or extinct ancestral lineages by leveraging information from large descendant cohorts, a critical capability when true progenitor genomes are unavailable.

The genome of an admixed individual is a mosaic of ancestry tracts inherited from multiple source populations, with ancestry switches occurring at historical recombination breakpoints. The scale and structure of this mosaic depend on the time since admixture, local recombination rates and the proportional contributions of each source. GAMA models this process with three parameters: the number of generations since admixture λ, the ancestry-proportion vector $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)$ for N contributors, and an empirically estimated haplotype-error parameter ϵ that captures residual discrepancies arising from sequencing, genotyping, phasing and mutation-related noise. Although all three quantities enter the hidden Markov model likelihood and can in principle be estimated within a full EM framework, GAMA uses a hybrid estimation strategy in practice. The full EM formulation provides the likelihood-based

basis for parameter inference, whereas the implemented algorithm updates the ancestry-proportion vector $\boldsymbol{\pi}$ directly from the current decoded ancestry segments, estimates the admixture time λ from co-ancestry decay, and updates the mismatch parameter ϵ through the EM step.

1.2 Input data and hidden states

Input data. GAMA requires two inputs: phased haplotypes for ancestral reference panels and for admixed samples, and genetic distances between adjacent markers. Genetic distances should be obtained from an external linkage map when available. If no high-resolution map exists, distances may be approximated from a genome-wide recombination rate; however, local heterogeneity in recombination can reduce inference accuracy. Phased haplotypes $\mathbf{H} = \{h_1, h_2, ...h_L\}$ are partitioned into L consecutive, nonoverlapping windows of user-defined size (specified in base pairs, centimorgans, or number of SNPs). The distance between adjacent loci is denoted as $\mathbf{D} = \{d_1, d_2, ...d_{L-1}\}$, where d_l represents the genetic distance between the centers of loci l and $l+1$.

Hidden variables. The sequence of the hidden variables is denoted as $\mathbf{Z} = \{z_1, z_2, ...z_L\}$, $z_l \in \{1, ..., N\}$, where $z_l = n$ indicates that chromosome segment containing the local haplotype l originates from ancestral population n, and N represents the number of the ancestral populations.

1.3 Transition and emission probabilities

Transition matrix. The transition probability between hidden states is modeled as a function of genetic distance, generations since admixture, and ancestry proportions. The transition matrix at locus l is given by

$$p(z_{l+1} = j \mid z_l = i) = \begin{cases} (1-\tau_l) + \tau_l \pi_j, & for\ j = i; \\ \tau_l \pi_j, & for\ j \neq i. \end{cases} \quad (1)$$

where $\tau_l = 1 - e^{-\lambda d_l}$ denotes the probability that at least one recombination event has

occurred between loci l and $l+1$ since admixture, after which the ancestry state is redrawn according to the ancestry-proportion vector $\boldsymbol{\pi}$; λ denotes the number of generations since admixture; d_l denotes the genetic distance in Morgans between loci l and $l+1$; and π_j denotes the ancestry proportion of source population *j*. Because the probability of multiple recombination events over short, adjacent windows is negligible, we simplify the transition model by neglecting recurrent recombination within neighboring local-haplotype windows.

Emission probability. We modeled the emission probability as the local-haplotype frequency in ancestral population n, with an empirically estimated per-site mismatch parameter ϵ to account for residual mismatches between observed haplotypes and reference-panel haplotype-frequency models. Suppose a local haplotype segment in window l contains k_l biallelic SNPs. Although the theoretical haplotype space has size 2^{k_l}, the implementation does not enumerate all possible haplotypes. Instead, for each window, we enumerated only the distinct local haplotypes observed in the ancestral reference panels. Thus, the number of haplotypes evaluated in each window was bounded by the total number of sampled reference haplotypes rather than by 2^{k_l}.

Let $\mathcal{H}_l = \{h_{l1}, h_{l2}, \ldots, h_{lM_l}\}$ denote the set of M_l distinct local haplotypes observed in the ancestral reference panels at window l, and let $f_{n,l,j}$ denote the frequency of haplotype h_{lj} in ancestral population n, after Laplace smoothing as described below. For an observed local haplotype h_l in an admixed sample, the emission probability is

$$e(h_l \mid z_l = n) = \sum_{j=1}^{M_l} f_{n,l,j} (1-\epsilon)^{k_l - \Delta(h_l, h_{lj})} \epsilon^{\Delta(h_l, h_{lj})}, \quad (2)$$

where $\Delta(h_l, h_{lj})$ denotes the number of nucleotide differences between the observed haplotype h_l and reference-panel haplotype h_{lj}, and ϵ is the empirical haplotype

error parameter defined above. When the segment reduces to a single SNP $k_l = 1$ and $\epsilon = 0$, the emission probability simplifies to

$$e(h_l \mid z_l = n) = \begin{cases} f_{n,l}, & h_l = 1, \\ 1 - f_{n,l}, & h_l = 0. \end{cases} \tag{3}$$

where $f_{n,l}$ is the alternative allele frequency of ancestral population n at locus l.

1.4 Forward–backward algorithm and likelihood calculation

Forward-backward algorithm. Let $\boldsymbol{\theta} = (\boldsymbol{\pi}, \lambda, \epsilon)$. Denote $\alpha_l(i) = P(h_1, h_2, ..., h_l, z_l = i \mid \boldsymbol{\theta})$, $a_{ji}^l = p(z_{l+1} = i \mid z_l = j)$, $b_i^l(h_l) = e(h_l \mid z_l = i)$, and the forward function is given by:

$$\begin{aligned} &\alpha_1(i) = \pi_i b_i^1(h_1) \\ &\alpha_2(i) = \sum_j \alpha_1(j) a_{ji}^1 b_i^2(h_2) \\ &\vdots \\ &\alpha_{l+1}(i) = \sum_j \alpha_l(j) a_{ji}^l b_i^{l+1}(h_{l+1}) \end{aligned} \tag{4}$$

Denote $\beta_l(i) = P(h_{l+1}, h_{l+2}, ..., h_L, z_l = i \mid \boldsymbol{\theta})$, and the backward function is given by:

$$\begin{aligned} &\beta_L(i) = 1 \\ &\beta_{L-1}(i) = \sum_j a_{ij}^{L-1} b_j^L(h_L) \beta_L(j) \\ &\vdots \\ &\beta_l(i) = \sum_j a_{ij}^l b_j^{l+1}(h_{l+1}) \beta_{l+1}(j) \end{aligned} \tag{5}$$

The probability of the observations $\mathbf{H}$ given the parameters $\boldsymbol{\theta}$ can be written as,

$$\begin{aligned} P(\mathbf{H} \mid \boldsymbol{\theta}) &= \sum_i \alpha_L(i) \\ &= \sum_i \pi_i b_i^1(h_1) \beta_1(i) \\ &= \sum_i \sum_j \alpha_l(i) a_{ji}^l b_j^{l+1}(h_{l+1}) \beta_{l+1}(j). \end{aligned} \tag{6}$$

1.5 Expectation–maximization formulation for model parameters

The following EM formulation describes the likelihood-based basis of the GAMA hidden Markov model and is included for completeness. Let $\boldsymbol{\theta} = (\boldsymbol{\pi}, \lambda, \epsilon)$, where $\boldsymbol{\pi}$

denotes the ancestry-proportion vector, λ denotes the number of generations since admixture and ϵ denotes the mismatch parameter. In the implementation used for the analyses in this study, we used a hybrid estimator rather than a full EM update for all parameters: ϵ was updated through the EM step, whereas $\boldsymbol{\pi}$ was recalculated from the current decoded local-ancestry segments and λ was refined from co-ancestry decay, as described in section 1.7.

E-step:

$$\begin{aligned} Q(\boldsymbol{\theta}\,|\,\tilde{\boldsymbol{\theta}}) &= \sum_{Z} P(\mathbf{H},\mathbf{Z}\,|\,\tilde{\boldsymbol{\theta}}) \log P(\mathbf{H},\mathbf{Z}\,|\,\boldsymbol{\theta}) \\ &= \sum_{j} P(\mathbf{H}, z_1 = j\,|\,\tilde{\boldsymbol{\theta}}) \log \pi_j + \sum_{l=1}^{L-1}\sum_{i}\sum_{j} P(\mathbf{H}, z_l = i, z_{l+1} = j\,|\,\tilde{\boldsymbol{\theta}}) \log a_{ij}^{l} \\ &+ \sum_{l}\sum_{j} P(\mathbf{H}, z_l = j\,|\,\tilde{\boldsymbol{\theta}}) \log b_j^{l}(h_l). \end{aligned} \tag{7}$$

where $\tilde{\boldsymbol{\theta}}$ denotes the current estimation of the parameters.

M-step:

$$\begin{aligned} &\min\ -Q(\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}) \\ &s.t. \begin{cases} \sum_{j} \pi_j = 1; \\ \pi_j > 0,\ \forall j; \\ \lambda > 0, \\ 0 \le \epsilon \le 1. \end{cases} \end{aligned} \tag{8}$$

For convenience we denote

$$\begin{aligned} v_j^{l} &= P(z_l = j\,|\,H,\tilde{\boldsymbol{\theta}}) = \frac{\alpha_l(j)\beta_l(j)}{P(H\,|\,\tilde{\boldsymbol{\theta}})}, \\ \xi_{ij}^{l} &= P(z_l = i, z_{l+1} = j\,|\,H,\tilde{\boldsymbol{\theta}}) = \frac{\alpha_l(j) a_{ij}^{l} b_j^{l+1}(h_{l+1})\beta_{l+1}(j)}{P(H\,|\,\tilde{\boldsymbol{\theta}})}, \\ \xi_{jj}^{l} &= P(z_l = j, z_{l+1} = j\,|\,H,\tilde{\boldsymbol{\theta}}) = \frac{\alpha_l(j) a_{jj}^{l} b_j^{l+1}(h_{l+1})\beta_{l+1}(j)}{P(H\,|\,\tilde{\boldsymbol{\theta}})}. \end{aligned} \tag{9}$$

Then the above M-step can be written as minimize

$$\begin{aligned} F(\boldsymbol{\theta}) &= -\sum_{j} v_j^{1} \log \pi_j - \sum_{l=1}^{L-1}\sum_{i}\sum_{j\neq i} \xi_{ij}^{l} \log(\tau_l \pi_j) \\ &- \sum_{l=1}^{L-1}\sum_{j} \xi_{jj}^{l} \log(1 - \tau_l + \tau_l \pi_j) - \sum_{l=1}^{L}\sum_{j} v_j^{l} \log b_j^{l}(h_l), \end{aligned} \tag{10}$$

where $\tau_l = 1 - e^{-\lambda d_l}$.

1.6 Viterbi decoding and posterior ancestry assignment

Hidden states were inferred by applying the Viterbi algorithm to compute the single most probable sequence of latent states under the fitted model.

For $\forall i \in 1,2,...,N, l \in 1,2,...,L$, define,

$$\begin{aligned} \delta_l(i) &= \max_{z_1 \ldots z_{l-1}} P(z_l = i, z_{l-1}, ..., z_1, h_l, ..., h_1 \mid \boldsymbol{\theta}) \\ &= \max_j \delta_{l-1}(j) a_{ji}^{l-1} b_i^l(h_l), \\ \eta_l(i) &= \arg\max_j \delta_{l-1}(j) a_{ji}^{l-1} \end{aligned} \tag{11}$$

The inputs of Viterbi algorithm are $\mathbf{H} = (h_1, ..., h_L)$, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \lambda, \epsilon)$, and the output is $\hat{\mathbf{Z}} = (\hat{z}_1, ..., \hat{z}_L)$.

Step1: For $i \in 1,2,...,N$,

$$\begin{aligned} \delta_1(i) &= \pi_i b_i^1(h_1), \\ \eta_1(i) &= 0. \end{aligned} \tag{12}$$

Step2: For $l = 2,...,L$,

$$\begin{aligned} \delta_l(i) &= \max_j \delta_{l-1}(j) a_{ji}^{l-1} b_i^l(h_l), \\ \eta_l(i) &= \arg\max_j \delta_{l-1}(j) a_{ji}^{l-1}. \end{aligned} \tag{13}$$

Step3:

$$\begin{aligned} \Delta_L &= \max_i \delta_L(i), \\ \hat{z}_L &= \arg\max_i \delta_L(i). \end{aligned} \tag{14}$$

Step4: For $l = L-1, L-2, ..., 1$,

$$\hat{z}_l = \eta_{l+1}(\hat{z}_{l+1}). \tag{15}$$

And thus, we obtain the estimated hidden states $\hat{\mathbf{Z}} = \{\hat{z}_1, \hat{z}_2, ..., \hat{z}_L\}$.

Posterior probability of ancestral states. The marginal posterior probability that an allele at a given locus was inherited from a specific ancestral population is given by

$$\gamma_l(i) = P(z_l = i \mid \mathbf{H}, \boldsymbol{\theta}) = \frac{\alpha_l(i)\beta_l(i)}{\sum_{j=1}^{N} \alpha_l(j)\beta_l(j)}. \quad (16)$$

The local ancestry proportion of ancestor *i* at locus *l* is calculated as the mean of $\gamma_l(i)$ across all admixed samples.

1.7 Refinement of admixture parameters using co-ancestry decay

In the implemented GAMA algorithm, we used a hybrid parameter-estimation strategy rather than updating all parameters solely through EM. Although $\boldsymbol{\pi}$, λ and ϵ all enter the HMM likelihood and can in principle be estimated within a full EM framework, estimating only the mismatch parameter ϵ by EM substantially improves computational efficiency. The remaining two parameters can be updated more directly from the inferred ancestry mosaic. After each round of EM-based error-rate updating and Viterbi decoding, the ancestry-proportion vector $\boldsymbol{\pi}$ was calculated directly from the current inferred ancestry segments as the genome-wide fraction assigned to each ancestry component. The admixture time λ was then estimated from the decay of co-ancestry with genetic distance, which provided more accurate and stable time estimates than direct EM updates in our simulations. The mismatch parameter ϵ, which captures sequencing, genotyping, phasing and mutation-related discrepancies and is not directly observable from segment counts or co-ancestry decay, was therefore updated through the EM step. This hybrid procedure combines likelihood-based error-rate estimation with segment-based ancestry-proportion estimation and co-ancestry-based admixture-time calibration. The co-ancestry curve summarizes the probability that two haplotype chunks separated by a genetic distance d are assigned to the same or different ancestry components[1]. For clarity, we first describe the co-ancestry decay model for a two-ancestry pulse-admixture case. For a single admixture event with source ancestries *A* and *B*, we modeled the co-ancestry probabilities as follows:

$$
\begin{aligned}
p_{AA}(d) &= \pi_A(e^{-d\lambda} + [1-e^{-d\lambda}]\pi_A) = {\pi_A}^2 + \pi_A(1-\pi_A)e^{-d\lambda}, \\
p_{AB}(d) &= \pi_A(1-\pi_A) - \pi_A(1-\pi_A)e^{-d\lambda}, \\
p_{BB}(d) &= (1-\pi_A)^2 + \pi_A(1-\pi_A)e^{-d\lambda},
\end{aligned} \tag{17}
$$

where d denotes genetic distance (in Morgans), λ denotes generations since admixture, and π_A denotes the ancestry proportion contributed by source population A, p_{AB} denotes the ordered probability for A at the first locus and B at the second locus; the unordered different-ancestry probability is $2p_{AB}$.

The same formulation was extended to the multi-ancestry setting used in the GAMA analyses by treating each ancestry component against the remaining components, or equivalently by applying the same expected co-ancestry decay to each component pair. For ancestry components k and m with ancestry proportions π_k and π_m, the expected joint ancestry probabilities for two loci separated by distance d are

$$p_{kk}(d) = \pi_k^2 + \pi_k(1-\pi_k)e^{-d\lambda},$$

and, for $k \neq m$

$$p_{km}(d) = \pi_k\pi_m(1-e^{-d\lambda}).$$

Here $p_{km}(d)$ denotes the ordered probability that the first locus is assigned to component k and the second locus to component m. These expected decay curves were fitted to the decoded local-ancestry mosaic to refine the admixture-time parameter λ.

1.8 Laplace smoothing of local-haplotype frequencies

To reduce the influence of small ancestral reference-panel sizes on local-haplotype frequency estimates, GAMA applies Laplace smoothing separately within each genomic window and ancestry component. For window l, let M_l denote the number of distinct local haplotypes observed across all ancestral reference panels, $c_{n,l,j}$ the count of haplotype j in ancestral population n, and $n_{n,l}$ the total number of sampled reference haplotypes from population n. A total pseudo-count v is

distributed uniformly across the M_l observed haplotypes, giving

$$\hat{f}_{n,l,j} = \frac{c_{n,l,j} + \nu / M_l}{n_{n,l} + \nu}. \quad (18)$$

Thus, the smoothed frequencies sum to one within each population and window, while avoiding zero emission probabilities for rare local haplotypes in small reference panels. This smoothing is equivalent to a Dirichlet prior with parameter ν / M_l for each observed local haplotype, total concentration ν, and a uniform base measure over the M_l haplotypes. Unless otherwise specified, ν was set to 4.

1.9 Inference of unknown or unsampled ancestry components

Some ancestral sources may be unsampled, extinct, or otherwise unavailable. GAMA can infer such unknown ancestries from admixed descendants using a newly developed iterative procedure. Initially, GAMA is run with the available reference panels. For each local-haplotype segment across admixed samples we record the maximum posterior probability over sampled ancestors; segments are ranked by this maximum posterior, and the lower 50% (those with the smallest maxima) are pooled as an initial candidate set for the unknown ancestry. In each subsequent iteration, segments currently inferred as deriving from the unknown ancestry are aggregated to form an updated reference for that ancestry, and GAMA is re-run to update local-ancestry assignments. Iteration continues until the same convergence criteria used by the standard GAMA pipeline are met. The inferred unknown ancestry may represent one or more unsampled source populations.

1.10 Identification of ancestry-enriched regions

For each proxy ancestry component, ancestry-enriched regions were identified from genome-wide window-level local ancestry proportions inferred by GAMA. Let $x_{l,k}$ denote the mean local ancestry proportion of component k in genomic window l, calculated across all phased haplotypes in the analyzed cohort. To construct an empirical null distribution for each ancestry component, the genome-wide distribution of $x_{l,k}$ was modeled as

$$x_{l,k} \sim \mathrm{Beta}(\alpha_k, \beta_k).$$

The Beta distribution was fitted using all genomic windows for ancestry component k, and the fitted distribution was used as the empirical null model for ancestry-enrichment testing. For each window and ancestry component, ancestry-enrichment significance was calculated as the upper-tail probability

$$p_{l,k} = 1 - F_{\mathrm{Beta}}\left(x_{l,k}; \alpha_k, \beta_k\right). \quad (19)$$

Small $p_{l,k}$ values therefore indicate windows in which the local ancestry proportion of a given component is unusually high relative to its genome-wide empirical background. Genome-wide ancestry-enrichment scans were then used to identify candidate outlier regions for each proxy ancestry component.

1.11 Application of GAMA to Dog10K genomes

We applied GAMA to Dog10K genomes using 10-kb windows and a pedigree-based, canid-specific recombination map[2]. To reduce bias from inaccuracies in local recombination-rate estimates, chromosome-wide mean recombination rates were used. In the primary analysis reported in the main text, GAMA was applied using the unmasked basal reference panel, and all haplotypes from the same analyzed cohort were modeled jointly on a per-chromosome basis to estimate shared parameters, including admixture timing, ancestry proportions and the mismatch parameter.

As a sensitivity analysis, we additionally evaluated whether the leading ancestry-enrichment signal could be driven by recent introgression of modern-breed haplotypes into basal reference lineages. Six representative modern breeds were selected to cover major American Kennel Club breed groups, with one breed chosen from each group except the Non-Sporting Group: Labrador Retriever, Bloodhound, Boxer, Wire Fox Terrier, Cavalier King Charles Spaniel and German Shepherd Dog. Using GAMA, we screened 19 basal dog lineages and the wolf reference for genomic segments inferred to be of recent modern-breed origin. To reduce ambiguity in the direction of introgression and avoid masking short, potentially misassigned tracts,

only inferred modern-origin segments longer than 4 Mb were retained for masking. The retained modern-breed-derived fraction varied across basal dog lineages, ranging from 1.31% in Kintamani to 26.42% in Canaan Dog, with a mean of 11.43% across the 19 basal dog lineages; the corresponding fraction in the wolf reference was 0.61%. These inferred modern-origin segments were masked in the reference haplotypes, and GAMA was rerun using the masked reference panel. This masked-reference analysis was used only to assess robustness to potential recent modern-breed introgression in the reference panel, rather than as the primary ancestry scan. The masked-reference analysis recovered the same leading chromosome 7 ancestry-enrichment peak near *GPR52* (Supplementary Fig. 25), indicating that the focal signal was not driven by recent modern-breed-derived tracts in the basal reference panel.

2. Simulation benchmarks of GAMA performance

2.1 Simulation of ancestral and modern dog genomes

To generate realistic test data for method validation, we implemented a two-phase simulation that reflects key features of canine demographic history. In the first phase we simulated ancestral-breed evolution under plausible demographic parameters (population sizes, divergence times and bottlenecks), and in the second phase we constructed modern genomes as mosaics of ancestry tracts by recombining segments from those ancestral lineages. This two-phase design captures both ancient breed diversification and the admixture patterns that define contemporary dog genomes, providing a realistic substrate for benchmarking local-ancestry inference.

Ancestral population simulation. An ancestral dog population was simulated in forward time using SLiM v4.2.2 under a neutral Wright–Fisher model[3]. Simulations began with a burn-in of 100,000 generations to approach mutation–drift equilibrium. We then imposed a severe founder bottleneck, reducing effective population size (Ne) from 10,000 to 2,000, and split the population into multiple isolated lineages. Each lineage was assigned an effective size of Ne =300 and evolved independently under

genetic drift for 200 generations, yielding lineage-specific allele-frequency differences.

Modern genome simulation. Modern chromosomes were generated by concatenating haplotype segments sampled from the diverged ancestral lineages. Recombination breakpoints between adjacent markers were introduced with probability determined by the recombination rate and the simulated admixture time; at each breakpoint, ancestry was switched to another source population, producing contiguous haplotype blocks of uniform ancestry along each chromosome. The resulting synthetic genomes form mosaics of ancestry tracts that recapitulate realistic patterns of linkage disequilibrium and local ancestry, and are therefore suitable for benchmarking local-ancestry inference and selection scans. The admixture models used for simulation are illustrated in Supplementary Fig. 1.

2.2 Performance across admixture complexity, mismatch parameters, and timing

We simulated genomic data under 3-, 5-, 7-, 9- and 20-way admixture scenarios to benchmark GAMA in complex admixture settings. Simulations spanned admixture times of 5, 10, 20, 50 and 100 generations and per-site mismatch parameters of 0, 0.02 and 0.05; each dataset comprised 20 admixed haplotypes. GAMA recovered admixture timing and mismatch parameters with little systematic bias (Supplementary Fig. 2). As expected, ancestry-assignment accuracy declined with increasing numbers of ancestral sources, longer admixture times and higher mismatch parameters. Nonetheless, GAMA maintained high assignment accuracy (>0.85) even under the most challenging condition tested: 20-way admixture, mismatch parameter = 0.05 and 100 generations.

2.3 Impact of staggered admixture events

We simulated a series of staggered admixture events involving five source populations (model in Supplementary Fig. 1b): populations 1 and 2 admixed 100 generations ago, and populations 3, 4 and 5 admixed 80, 60 and 40 generations ago, respectively.

Across 100 replicate simulations, GAMA recovered local ancestry with a mean accuracy of 96%. When estimating admixture timing for the five sources, GAMA resolved the relative ordering of admixture events but produced time estimates with modest bias (see Supplementary Fig. 3). These results indicate that GAMA reliably distinguishes multiple historic admixture pulses while exhibiting small systematic errors in absolute time calibration under the simulated conditions.

2.4 Performance with direct and indirect ancestral sources

In the admixture scenario illustrated in Supplementary Fig. 1c, the focal composite population M1234567 receives ancestry through four more-recent direct contributors (M123, M345, M567 and population 7), which themselves trace back to seven ancient, indirect source populations (1–7). Using the indirect, unrelated ancestral sources as reference panels, GAMA achieves a mean local-ancestry inference accuracy of 0.9476 (Supplementary Fig. 4a). By contrast, using the closely related direct ancestral panels yields a lower mean accuracy of 0.7949 (Supplementary Fig. 4b). The reduction in accuracy reflects genuine ambiguity introduced by related reference panels: overlapping contributions from the same deep ancestor (for example, both M123 and M345 carry segments derived from ancestor 3) make it difficult to assign short tracts uniquely to one recent source or another. These results emphasize that using closely related or nested reference panels can reduce the accuracy of local ancestry inference, whereas employing indirect and unrelated ancestral sources improves performance in such scenarios.

2.5 Inference of unknown or unsampled ancestries

When genomic data for one or more source populations are unavailable (model illustrated in Supplementary Fig. 1d), GAMA employs an iterative procedure to reconstruct these unknown ancestries from admixed genomes. As shown in Supplementary Figs. 5 and 6, the iterative algorithm reliably distinguishes inferred unknown components across a range of admixture times (T = 10 and 50 generations) and ancestral-panel sizes (100 and 20 haplotypes), demonstrating robustness to both

recent and older admixture scenarios and to variation in reference-panel depth.

2.6 Impact of pseudo-ancestral panels

Under the pseudo-ancestral-panel simulation shown in Supplementary Fig. 1e, the focal admixed population M123 derives genetically from three true source populations (1–3), whereas populations 4–6 contribute no ancestry. In practice, however, non-contributing populations can be included as candidate reference panels because of phenotypic similarity, incomplete population records or uncertainty in source designation; we refer to these non-contributing reference panels as pseudo-ancestries. As illustrated in Supplementary Fig. 7, including such pseudo-ancestral panels had only a limited effect on the accuracy of local-ancestry inference with GAMA, indicating robustness to moderate reference-panel misspecification.

2.7 Performance under small reference-panel sizes

GAMA incorporates Laplace smoothing to mitigate sampling noise when ancestral reference panels are small. To evaluate its effectiveness, we tested reference-panel sizes of 2, 4, 10, 20, 40 and 60 haplotypes. As shown in Supplementary Fig. 8, Laplace smoothing substantially improves local-ancestry accuracy at the smallest panel sizes and has negligible impact for larger panels. However, inferred mismatch parameters are biased upward when reference panels contain fewer than 40 haplotypes (Supplementary Fig. 9), indicating that error-rate estimates should be interpreted cautiously for very small panels.

2.8 Benchmarking against FLARE

To evaluate the efficacy of GAMA, we benchmarked it against FLARE v0.5.2 (Fast Local Ancestry Estimation), a leading method for local ancestry inference in large-scale datasets[4]. FLARE is designed to scale to tens of thousands of sequenced individuals by extending the classical Li and Stephens haplotype copying model with composite reference haplotypes, facilitating high computational efficiency while retaining competitive accuracy[4]. We selected FLARE as the primary comparator given

its status as a widely recognized benchmark for rapid inference. This comparison provides a robust test of our model's capabilities, particularly in handling complex, multi-way admixture events where traditional methods may face limitations.

Our simulations demonstrate that GAMA consistently outperforms FLARE across all tested parameters, including varying admixture times, diverse reference panel sizes, and increasing numbers of ancestral sources (Supplementary Fig. 10). Notably, GAMA maintains superior accuracy even when reference panels are small, highlighting its robustness in scenarios typical of non-model organism genomics.

3. Designation of basal reference breeds

Basal reference breeds were designated on the basis of deep phylogenetic divergence, geographic provenance, historical records and distinctive phenotypic features. These breeds were used as proxy reference panels because they retain divergent haplotype variation that captures major older ancestry components represented in domestic dogs. They were not assumed to be direct progenitors of modern breeds, but served as geographically and phylogenetically structured references for downstream local-ancestry inference.

To infer relationships among populations, we constructed a maximum-likelihood tree using IQ-TREE 2.0[5]. For each breed we selected a single representative as follows. We performed principal-component analysis (PCA) on all samples and used the scores of the first four principal components as a four-dimensional feature vector for each sample. Breeds represented by a single sample retained that sample; for breeds with multiple samples we computed the centroid of their feature vectors and chose the sample with the smallest Euclidean distance to the centroid. We then inferred a maximum-likelihood tree with IQ-TREE 2 using model selection and branch-support options (-m MFP -o Wolf -T 10 -alrt 1000 -B 1000). Tree visualization and annotation were performed with the Interactive Tree Of Life (iTOL)[6].

We assessed population structure among 37 basal dog breeds by performing PCA with the smartpca program in EIGENSOFT v6.1.4[7]. Prior to PCA, SNPs were pruned for linkage disequilibrium using PLINK v1.9 (--indep-pairwise 200 50 0.001), and the resulting SNP list was retained with --extract. Monomorphic variants were removed using PLINK's --maf filter. The PCA was computed on the pruned SNP set, and the leading principal components were examined to characterize population structure.

The maximum-likelihood tree showed that wolves and 37 dog breeds occupied deep-branching positions relative to the more derived clade formed by other modern domestic breeds (Supplementary Fig. 11). We refer to these lineages as basal breeds, with "basal" denoting their phylogenetic position rather than the formal age of breed recognition. Most of these breeds also retain historical or geographic associations consistent with long-established regional origins. Eurasier and Elo were treated separately because they are recent breed constructs: the Eurasier was developed in the 1960s from crosses among Chow Chow, Keeshond and Samoyed, whereas the Elo was derived from crosses involving Eurasier, Old English Sheepdog and Chow Chow. Their basal placement is therefore best explained by inherited basal genetic components from their ancestral contributors rather than by recent breed antiquity.

After excluding the recent hybrid-derived Eurasier and Elo, the remaining 35 basal breeds were grouped into four broad evolutionary categories by integrating phylogenetic position, genetic clustering, geographic distribution and functional history (Supplementary Table 1, Supplementary Figs. 11 and 12). These comprised: **1) East–Southeast Asian landrace working dogs**, including Japanese indigenous breeds, Southeast Asian hunters and Chinese guardians; **2) Arctic working breeds**; **3) Buddhist–Imperial sacred dogs**, including Buddhist sacred dogs and Imperial sacred dogs; and **4) Central–West Asian guardians and sighthounds**, including Fertile Crescent guardians and Central–West Asian sighthounds.

The East-Southeast Asian Landrace Working Dogs comprise 12 breeds,

predominantly from Japan and China with two Southeast Asian outliers. Most are medium-sized, with the notable exceptions of the Japanese and American Akita, which are large. The Japanese indigenous subgroup shares a common origin and a Spitz-type phenotype characterized by triangular ears, a curled tail and a dense double coat. The American Akita derives from the native Japanese Akita Inu but was selectively bred in the United States for increased body size and a more imposing stature. We selected Shiba Inu (8–11 kg) and Japanese Akita (30–50 kg) as representative ancestral morphotypes to capture the group's body-size extremes. Five additional East-Southeast Asian Landrace Working Dogs, which show distinct biogeographic and morphological divergence, were also included as representatives of this trajectory.

The Buddhist–Imperial Sacred Dogs comprise six small-to-medium Asian companion breeds united by luxuriant double coats, compact builds and a shared history as monastic or courtly companions. Most exhibit marked brachycephaly (cranial index >65), with the Tibetan Terrier an exception (mesocephalic), and all display low-to-moderate exercise requirements consistent with an indoor companion role. Breeds in this trajectory trace to two cultural origins: Tibetan monastic lineages (Tibetan Spaniel, Lhasa Apso, Tibetan Terrier) and East Asian courtly lines (Shih Tzu, Pekingese, Japanese Chin). The Tibetan Terrier (9–14 kg) retains a transitional morphology between companion and working forms, whereas other members show extreme facial shortening historically linked to ceremonial status. We selected Tibetan Spaniel, Lhasa Apso and Tibetan Terrier as representative ancestral morphotypes to capture variation in cranial form, body size and coat architecture.

The Greenland Dog, Alaskan Malamute and Samoyed exemplify Arctic working breeds shaped by millennia of selection in extreme polar environments. These Spitz-type dogs share erect ears, curled tails and dense, insulating double coats that confer thermoregulatory resilience in subzero conditions. Historically bred for endurance tasks such as sled pulling, long-range hunting and draft work, they

combine exceptional strength, stamina and cold tolerance. Although each lineage is embedded in a distinct circumpolar cultural context (Greenland Dog: Inuit; Alaskan Malamute: Native Alaskan; Samoyed: Siberian nomads), their convergent morphologies and life-history traits reflect parallel evolutionary responses to Arctic niches and prolonged human partnership. Because of their geographic and genetic distinctiveness, we selected these three breeds as representative ancestral lineages for the Arctic working trajectory.

The Central–West Asian Guardians & Sighthounds comprise 14 breeds from the Middle East, Central Asia and the Caucasus and partition into two trajectories: Fertile Crescent guardians and Central–West Asian sighthounds. Fertile Crescent guardians are generally large to giant breeds, with the Turkish Zerdava the only medium-sized member in this cluster. By contrast, the sighthounds are lean, aerodynamic hunters ($\approx$ 20 – 30 kg) adapted for high-speed pursuit using acute vision. Principal-component analysis groups nine Fertile Crescent guardian breeds (Sarplaninac, Caucasian Ovcharka, Turkish Mastiff, Anatolian Shepherd Dog, Akbash, Turkish Zerdava, Kangal, Sarabi and Kars) into a tight genetic cluster. For this reason, we designate the Anatolian Shepherd Dog, which has the largest sample size, as the representative ancestral lineage. The Central Asian Shepherd Dog is genetically and geographically distinct from the Fertile Crescent cluster, and the Canaan Dog retains semi-wild pariah ancestry and desert-adapted physiology, functioning more as a medium-sized vigilance specialist than a combative guardian; accordingly, we also designate these two breeds as ancestral representatives of the Central–West Asian guardians. Within the sighthound trajectory, the Tazi and Saluki share near-identical ancestry while the Afghan Hound shows an earlier divergence; we therefore select the Saluki and Afghan Hound as representative ancestral sighthounds.

Supplementary Table 1. Classification of basal canid lineages used as proxy ancestry components.

Group	Sub-group	Breeds[1]
East-Southeast Asian Landrace	Japanese indigenous breeds	Shikoku

Working Dogs		Kishu
		Shiba Inu
		Kai Ken
		Hokkaido
		American Akita[2]
		Japanese Akita
	Southeast Asian Hunters	**Thai Ridgeback**
		Kintamani
	Chinese Guardians	**Chow Chow**
		Chinese Shar-Pei
	-	**Formosan Mountain Dog**
Arctic Working Breeds	-	**Greenland Dog**
		Alaskan Malamute
		Samoyed
Buddhist-Imperial Sacred Dogs	Buddhist Sacred Dogs	**Tibetan Spaniel**
		Lhasa Apso
		Tibetan Terrier
	Imperial Sacred Dogs	Shih Tzu
		Pekingese
		Japanese Chin
Central-West Asian Guardians & Sighthounds	Fertile Crescent Guardians[3]	Sarplaninac
		Caucasian Ovcharka
		Turkish Mastiff
		Anatolian Shepherd Dog
		Akbash
		Turkish Zerdava
		Kangal
		Sarabi
		Kars
	-	**Central Asian Shepherd Dog**
	-	**Canaan Dog**[4]
	Central-West Asian Sighthounds	**Afghan Hound**
		Tazi
		Saluki

[1] The 19 dog breeds highlighted in bold, together with the wolf, are designated as basal reference breeds.

[2] The American Akita, derived from the Japanese Akita Inu, was selectively bred in the United States for a larger, more imposing physique.

[3] The Fertile Crescent Guardians comprise nine guardian breeds that trace their origins to the Fertile Crescent and adjacent regions.

[4] The Canaan Dog originated in the Fertile Crescent but differs from other Fertile Crescent Guardians by retaining semi-wild pariah ancestry and exhibiting a moderate size.

4. Population-genetic analyses of L1 insertions and *GPR52*-linked L1-Cf

4.1 Allele-frequency spectra of full-length and truncated L1 insertions

We generated allele frequency spectrum (AFS) for full-length (>6 kb) and truncated (1–6 kb) L1 insertions using the Dog10K structural-variation dataset. Insertions were treated as derived alleles; allele frequency at each locus was computed as the number of chromosomes carrying the insertion divided by the total number of chromosomes successfully genotyped at that locus. The AFS of truncated L1s was used as a putative neutral control. To test for an excess of high-frequency insertions among full-length elements, we compared the proportion of loci with allele frequency >0.5 between full-length and truncated L1s using a chi-square test of homogeneity on a 2×2 contingency table (high-frequency vs. low-frequency × full-length vs. truncated).

4.2 Extended haplotype homozygosity analyses in Chinese village dogs

We merged SNPs, indels, and structural variants (SVs) from the Dog10K dataset and phased the combined genotype panel with Beagle v5.1[8] using default parameters. Only biallelic variants were retained for downstream haplotype-based analyses. The final phased dataset comprised 84 China village dogs (168 phased haplotypes), of which 81 haplotypes carried the *GPR52* L1-Cf insertion and 87 haplotypes lacked the insertion. We calculated Extended Haplotype Homozygosity (EHH)[9] and HaploSweep EHH[10], treating the *GPR52* L1-Cf insertion as the core variant. EHH and HaploSweep analyses were performed with HaploSweep v1.0 using default settings[10]. We report EHH decay curves for the region surrounding the core variant.

4.3 Model-based age and selection-intensity estimation using HMM-Sweep

We estimated the model-based age and selection intensity of the *GPR52* upstream L1-Cf-bearing background using HMM-Sweep[11], which infers allele age and selection intensity from local haplotype structure under a hard-sweep model. Analyses used the

same phased genotype panel employed for EHH calculations (see above). The recombination rate was fixed at 1.1032×10^{-8} per bp per generation, based on a pedigree-derived canid recombination map for chromosome 7 (23–27 Mb)[2]. HMM-Sweep was run with default settings unless otherwise indicated. Conversion of age estimates from generations to years assumed a generation time of 3 years for canids.

In Chinese village dogs, HMM-Sweep estimated a selection coefficient of $s = 0.0117$ (95% CI, 0.0089–0.0312) and a model-based age of 781 generations (95% CI, 294–1,023) for the selected L1-Cf-bearing background, corresponding to approximately 2,343 years before present. These results provide additional haplotype-based support for positive selection at this locus outside the context of closed modern breeds.

4.4 Nucleotide diversity across the *GPR52* region

Nucleotide diversity (π) for wolves, the Fertile Crescent Guardians, and modern dog breeds was calculated with VCFtools v0.1.16[12] using 50-kb nonoverlapping windows with a 5-kb step (--window-pi 50000 --window-pi-step 5000). Samples from all breeds assigned to the Fertile Crescent Guardians were combined into a single group for π estimation; samples from all non-basal modern dog breeds were combined similarly. Per-window π values are reported for each group.

4.5 Branch-specific T-statistic analysis

We performed branch-specific T-statistic analyses[13] to identify lineage-specific allele-frequency shifts, using wolves as the outgroup and 15 basal reference breeds that each had >4 samples: Shiba Inu, Japanese Akita, Thai Ridgeback, Chow Chow, Chinese Shar-Pei, Formosan Mountain Dog, Greenland Dog, Alaskan Malamute, Samoyed, Tibetan Spaniel, Lhasa Apso, Tibetan Terrier, Pekingese, Anatolian Shepherd Dog, and Central Asian Shepherd Dog. For each focal branch we computed the T-statistic following the procedure in[13] and report test statistics. The phylogenetic tree used for the branch-specific T-statistic analysis was identical to that shown in

Supplementary Fig. 11.

4.6 Haplotype-network construction around the L1-Cf insertion

We constructed and visualized a haplotype network for the 20-kb interval surrounding the *GPR52* L1-Cf insertion (chr7:24,780,000–24,800,000 bp) using HapNetworkView[14]. Genotypes comprising SNPs, indels, and structural variants within this interval were merged into a single variant set and phased as described above; phased haplotypes were used as input for network inference. The network shown in the main text was drawn from the phased haplotype alignment and annotated in HapNetworkView; node sizes reflect haplotype counts and edges indicate mutational steps.

4.7 Introgression-time estimation in L1-Cf-positive wolves

We estimated the time since introgression by maximum likelihood using the observed distribution of local-ancestry tract lengths under a single-pulse admixture model. Under the Markovian model of tract formation, the genetic length C of an introgressed tract, measured in centimorgans, follows an exponential distribution with rate parameter $\rho = g(1-\alpha)/100$, where g is the time since introgression in generations and α is the background ancestry proportion of the source component on chromosome 7, estimated from GAMA local-ancestry assignments. For n observed introgressed source-ancestry tracts with total genetic length S in centimorgans, the maximum-likelihood estimate is

$$\hat{g} = \frac{100n}{(1-\alpha)S}. \tag{20}$$

Confidence intervals were derived from the exact Gamma distribution of total tract length under the independent exponential-tract model. Given n source-ancestry tracts with total genetic length S measured in centimorgans, the total length follows

$$S \sim \Gamma\left(n, \frac{100}{g(1-\alpha)}\right),$$

where n is the shape parameter and $100/[g(1-\alpha)]$ is the scale parameter.

Equivalently, the pivotal quantity

$$W = \frac{Sg(1-\alpha)}{100}$$

follows $\Gamma(n,1)$. Using the maximum-likelihood estimate $\hat{g}$ defined in Eq. 20, the true introgression time can be written as

$$g = \hat{g}\frac{W}{n}.$$

Therefore, a $100(1-\gamma)\%$ confidence interval is

$$CI = \left[\frac{\hat{g}q_{\gamma/2}}{n}, \frac{\hat{g}q_{1-\gamma/2}}{n}\right], \tag{21}$$

where q_p denotes the p-th quantile of the $\Gamma(n,1)$ distribution.

We analyzed three wolf haplotypes carrying the *GPR52*-upstream L1-Cf insertion that were assigned to the Fertile Crescent Guardian–associated ancestry. Physical tract lengths of 1.12, 1.22 and 1.24 Mb were converted to genetic distances using the local recombination rate for chr7:23–27 Mb (1.1032×10^{-8} per bp per generation)[2], yielding genetic lengths of 1.24, 1.35 and 1.37 cM, respectively. The chromosome 7 background ancestry proportion of this source component in these wolf samples, estimated from GAMA local-ancestry assignments, was α = 0.0462.

Treating the three L1-Cf-bearing source-ancestry tracts jointly under a single-pulse introgression model, we estimated the time since introgression to be 79.65 generations (95% CI: 16.43–191.82), corresponding to 238.95 years (95% CI: 49.29–575.46) assuming a generation time of 3 years. As a sensitivity analysis, we also estimated the introgression time separately for each tract, yielding tract-specific estimates of 84.88 generations (95% CI: 2.15–313.10), 77.91 generations (95% CI: 1.97–287.39) and 76.65 generations (95% CI: 1.94–282.75), respectively. The wide confidence intervals for individual tracts reflect the limited information available from single-tract estimates.

5. Reporter assays of the L1-Cf 5′ UTR

5.1 Cell culture

HEK293T cells were obtained from the American Type Culture Collection (ATCC). Cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and maintained at 37°C in a humidified incubator with 5% CO_2.

5.2 Construction of L1-Cf 5′ UTR reporter plasmids

The canine L1-Cf ORF1p amino-acid sequence (UniProt: A0A7U3RY77) was used to delimit the ORF1 coding region. The sequence spanning the transcription start site to the ORF1 start codon (chr7:24,792,672 – 24,793,568) was designated the L1-Cf 5' untranslated region (5′ UTR). The 5′ UTR was synthesized (Sangon Biotech, Shanghai, China) and cloned into pEGFP-C3 and derivative reporter plasmids at multiple positions to assay regulatory activity — specifically immediately upstream of the CMV promoter and downstream of the EGFP coding sequence.

5.3 Cell transfection

Cell transfection was performed according to the manufacturer's protocol for Lipofectamine 3000 (Thermo Fisher Scientific, USA). Briefly, HEK293T cells were seeded in 96-well plates at a density of 2×10^4 cells per well one day prior to transfection. When cells reached 70-80% confluence, transfection was carried out using 100 ng pEGFP-C3 plasmids and its derivative constructs, with 0.3 μL Lipofectamine 3000 and 0.2 μL P3000 reagent added to each well according to the optimized protocol.

5.4 Time-lapse live-cell imaging and GFP-signal quantification

Following cell transfection, 96-well culture plates were placed in the Incucyte S3 live-cell imaging system (Sartorius, Germany). Green fluorescence intensity was recorded every hour across five fields of view using a 10× objective. Quantification was performed using Incucyte software with the following segmentation settings in

the green image channel: segmentation mode, Top-Hat; radius, 100 μm; threshold, 2 GCU. Fluorescence intensity was calculated as the total integrated intensity divided by the total phase area across the five views. Each group was measured in five independent replicates. GFP signals were normalized to the promoter-only control, and construct groups were compared by two-tailed Student's t-tests.

6. Ancient DNA processing and dropout-adjusted likelihood modeling

6.1 Ancient canid sample panel and filtering

We collected raw sequencing data for 180 ancient canid genomes from 12 published studies listed in Supplementary Table 4[15-26], including 101 ancient dogs and 79 ancient wolves. These samples span a broad temporal range from the Late Pleistocene to near-modern periods, with the oldest wolf sample dating to ~100 ka and the oldest dog sample to ~16 ka. Three ancient dog samples and three ancient wolf samples lacking age information were excluded from downstream analyses. Among samples with age information, we further excluded those with zero local coverage across the 1-kb genomic region upstream of the 5′ UTR of the *GPR52*-linked L1-Cf insertion. After filtering, 73 ancient dog genomes and 70 ancient wolf genomes were retained for subsequent analyses. Chronological age, geographic origin, local coverage, L1-Cf detection status and source information for each sample are provided in Supplementary Table 4.

6.2 Read processing and alignment

Raw sequencing data were processed using the nf-core/eager pipeline, version 2.5.3[27]. Briefly, AdapterRemoval was used to trim adapter sequences and merge paired-end reads where applicable[28]. Processed reads were aligned to the CanFam4 reference genome using BWA aln[29], and PCR duplicates were marked and removed with Picard MarkDuplicates (Picard Toolkit, Broad Institute). Unless otherwise specified, all nf-core/eager steps were performed using default parameters, and the resulting BAM files were used for L1-Cf inspection and downstream likelihood-based analyses.

Because our objective was to detect the presence of an L1-Cf insertion near *GPR52* rather than to call single-nucleotide variants, terminal deamination damage typical of ancient DNA, such as C-to-T or G-to-A substitutions, was expected to have limited impact on the identification of insertion-supporting junction reads. We therefore did not apply additional terminal damage trimming after adapter removal and paired-end read merging. This strategy was applied consistently across non-UDG, partial-UDG, and UDG-treated libraries, preserving full read lengths to maximize the probability of detecting reads spanning the L1-Cf–genome junction.

6.3 Detection of L1-Cf using 5′ junction reads

To identify ancient samples carrying the *GPR52*-linked L1-Cf insertion, we manually inspected the *GPR52* locus in all processed BAM files using IGV, version 2.19.1[30]. We focused on reads spanning the 5′ boundary between the upstream genomic flank and the inserted L1-Cf sequence[31]. A high-quality 5′ junction read was required to cross this boundary and satisfy all three criteria: at least 8 bp aligned to the genome-side flanking sequence, at least 8 bp aligned to the L1-Cf-side sequence, and mapping quality MAPQ ≥25. We did not use the 3′ end of the insertion for primary junction-read calling because the long poly(A) tail at this end reduces alignment uniqueness and can compromise mapping quality. Thus, L1-Cf detection in ancient samples was based on high-quality 5′ junction-read evidence, and local coverage used in downstream likelihood analyses was measured in the 1-kb genomic region upstream of the L1-Cf 5′ UTR. Applying these criteria, we identified 32 high-confidence L1-Cf-positive ancient canid genomes (27 dogs and 5 wolves) for downstream analyses. The corresponding junction-read evidence is shown in Supplementary Figs. 34–36.

Because ancient genomes vary substantially in local sequencing depth and read-length distribution, absence of high-quality 5′ junction-read evidence was treated as non-detection rather than definitive absence of the insertion. The binary 5′ junction-read detection status, together with local depth in the 1-kb upstream region

and sample-specific junction-read detection efficiency, was used as input for the dropout-adjusted likelihood models described below.

6.4 Dropout-adjusted likelihood framework

6.4.1 Model rationale

Detection of the L1-Cf insertion in ancient genomes relies on junction reads spanning the boundary between the genome-side sequence and the inserted L1-Cf sequence[31]. However, ancient DNA genomes are often sequenced at low and heterogeneous depth, such that the absence of junction reads cannot be interpreted as definitive absence of the insertion. We therefore developed a dropout-adjusted likelihood framework to infer L1-Cf insertion frequencies while accounting for incomplete detection.

The framework accounts for four sources of uncertainty: the underlying population frequency of the insertion, uncertainty in unobserved diploid genotypes, false-negative detection caused by limited local sequencing depth and junction-read constraints, and regional differences in the timing of frequency change. We first used a bin-level version of the model to estimate dropout-adjusted insertion frequencies within discrete temporal bins. We then extended the same observation model to a region-specific trajectory framework to infer regional onset times and a shared selection coefficient.

6.4.2 Data and notation

For ancient sample i, let a_i denote the sampling age in years before present (BP), $r_i \in \{1,\ldots,R\}$ denote its geographic region, d_i denote local sequencing depth around the insertion-start region, and β_i denote the per-depth junction-read detection efficiency used in the model. The observed variable is $D_i \in \{0,1\}$, where $D_i = 1$ indicates that the L1-Cf insertion was detected by one or more qualifying 5′ junction reads, whereas $D_i = 0$ indicates that no qualifying 5′ junction read was detected. Sample ages used in the likelihood model are provided in Supplementary Table 4. For

samples reported with age ranges, the midpoint of the reported range was used as the sampling age a_i in the likelihood model.

The true diploid insertion genotype is unobserved and is denoted by $G_i \in 0,1,2$, where $G_i = 0$, $G_i = 1$, and $G_i = 2$ correspond to non-carrier, heterozygous carrier, and homozygous carrier genotypes, respectively.

6.4.3 Junction-read detection model

The L1-Cf insertion was detected using reads spanning the 5′ boundary between the upstream genomic flank and the L1-Cf 5′ UTR. High-quality junction reads were required to contain both genome-side and L1-Cf-side aligned anchors and to pass the mapping-quality threshold. We defined m_g and m_{TE} as the aligned anchor lengths on the genome side and L1-Cf side, respectively, and required:

$$m_g \geq 8, \quad m_{TE} \geq 8$$

$$\mathrm{MAPQ} \geq 25.$$

For a read of length L, the fraction of possible read-start positions satisfying the two-anchor junction-read requirement is:

$$\frac{\max(0, L - m_g^{\min} - m_{TE}^{\min} + 1)}{L},$$

where $m_g^{\min} = m_{TE}^{\min} = 8$ denotes the minimum required anchor lengths on the genomic flank and L1-Cf side, respectively.

Let $p_i(L)$ denote the empirical read-length distribution for sample i. The anchor-based component of the per-depth junction-read detection efficiency can be expressed as

$$\beta_i = \sum_L p_i(L) \frac{\max(0, L - m_g^{\min} - m_{TE}^{\min} + 1)}{L} \tag{22}$$

In practice, β_i was treated as a sample-specific effective detection-efficiency term

that captures variation in read-length distribution and anchor requirements. Breakpoint mappability and mapping-quality filtering were incorporated through the same junction-read calling criteria used for L1-Cf detection. Sample-specific β_i values used in the likelihood model are provided in Supplementary Table 4.

6.4.4 Genotype-specific detection probabilities

In the genotype-specific detection model, absence of junction reads was treated as non-detection rather than definitive absence of the insertion. Conversely, false-positive junction-read detection was assumed negligible after manual inspection and stringent junction-read filtering. We therefore set the detection probability for non-carrier genotypes to zero and modeled dropout only for carrier genotypes. For sample i, let d_i denote local sequencing depth. The expected number of effective junction reads in a homozygous insertion carrier was defined as

$$\lambda_i = \beta_i d_i.$$

Assuming Poisson sampling of detectable junction reads, the genotype-specific detection probabilities are

$$\begin{aligned} P(D_i = 1 \mid G_i = 0) &= 0 \\ P(D_i = 1 \mid G_i = 1) &= 1 - \exp\left(-\frac{\lambda_i}{2}\right) \\ P(D_i = 1 \mid G_i = 2) &= 1 - \exp(-\lambda_i) \end{aligned} \tag{23}$$

The heterozygous genotype has half the expected junction-read rate of the homozygous insertion genotype because only one of the two homologous chromosomes carries the insertion.

6.4.5 Temporal-bin frequency estimation

To estimate L1-Cf insertion frequencies within discrete temporal bins, we used a simplified version of the dropout-adjusted likelihood model. This bin-level model does not assume any temporal frequency trajectory, regional onset time, or selection-intensity parameter. Instead, all samples within a given temporal bin are assumed to share a single underlying insertion frequency.

For a temporal bin b, let I_b denote the set of ancient samples assigned to that bin. For sample $i \in I_b$, D_i denotes the observed detection status, d_i denotes the local sequencing depth, and β_i denotes the per-depth junction-read detection efficiency. The unknown true insertion frequency in the bin is denoted by f_b.

The bin-specific frequency was parameterized on the logit scale:

$$f_b = \frac{1}{1+\exp(-\theta_b)}$$

where θ_b is the optimized parameter.

Given f_b, unobserved diploid genotypes were marginalized under Hardy–Weinberg proportions, used here as a genotype prior within each temporal bin rather than as a claim of panmictic equilibrium in the ancient metapopulation:

$$P(G_i = 0) = (1 - f_b)^2$$

$$P(G_i = 1) = 2 f_b (1 - f_b)$$

$$P(G_i = 2) = f_b^2$$

Using the genotype-specific detection probabilities defined above, the genotype-marginalized probability that sample i is detected as L1-Cf positive is:

$$p_i = 2 f_b (1 - f_b)\left[1 - \exp\left(-\frac{\lambda_i}{2}\right)\right] + f_b^2 \left[1 - \exp(-\lambda_i)\right] \tag{24}$$

The likelihood for temporal bin b is therefore:

$$L_b(\theta_b) = \prod_{i \in I_b} p_i^{D_i} (1 - p_i)^{1 - D_i} \tag{25}$$

and the corresponding negative log-likelihood is:

$$\mathrm{NLL}_b(\theta_b) = -\sum_{i \in I_b} \left[D_i \log(p_i) + (1 - D_i) \log(1 - p_i) \right] \tag{26}$$

The maximum-likelihood estimate of the bin-level insertion frequency was obtained by minimizing this negative log-likelihood with respect to θ_b. To avoid boundary

estimates at exactly zero or one, the optimized frequency was constrained to a finite interval:

$$f_b \in [10^{-6}, 1-10^{-6}]$$

The fitted frequency estimate for bin b is denoted by:

$$\hat{f}_b = \frac{1}{1+\exp(-\hat{\theta}_b)}$$

For each bin, we also recorded the number of samples, the number of junction-read-positive samples, the raw detection rate, the mean local depth, and the mean detection-efficiency coefficient.

Uncertainty in bin-level frequency estimates was assessed by nonparametric bootstrap resampling. For each temporal bin, samples were resampled with replacement while preserving the bin-specific sample size. The bin-level frequency model was refitted to each bootstrap replicate, generating a bootstrap distribution of $\hat{f}_b$. Percentile-based 95% confidence intervals were calculated from the bootstrap distribution.

This bin-level model provides a dropout-adjusted estimate of insertion frequency within each temporal interval and was used to complement the raw observed detection rates. It corrects for false-negative detection caused by variation in local sequencing depth and junction-read detection efficiency, while avoiding assumptions about the shape of the temporal trajectory.

6.4.6 Region-specific selection-onset trajectory model

To model spatiotemporal frequency change across geographic regions, we extended the same dropout-adjusted observation model to a region-specific trajectory framework. The primary model assumes that all regions share a common post-onset selection intensity s, while each region has its own baseline frequency and onset time.

The trajectory model uses a forward-generation time scale, in which older samples have smaller values and more recent samples have larger values. When the input data are recorded in BP, sampling ages are converted as:

$$t_i = \mathrm{round}\left(\frac{a_{\max} - a_i}{g}\right) + 1$$

where $a_{\max}$ is the oldest sampling age in the analyzed dataset and g is the generation time. The region-specific onset time $t_{0,r}$ is defined on the same forward-generation scale. For display, $t_{0,r}$ can be converted back to BP as:

$$a_{0,r} = a_{\max} - g(t_{0,r} - 1).$$

Regional frequency trajectory with shared selection intensity. The primary model assumes that all regions share a common post-onset selection intensity s, while each region has its own baseline frequency and onset time. Let $f_r(t_i)$ be the L1-Cf insertion frequency in region r at forward-generation time t_i. For each region r, we define $f_{0,r}$ as the baseline frequency before the onset of frequency change, and $t_{0,r}$ as the regional onset time. The parameter s is shared across regions.

Before onset, the insertion frequency remains at the regional baseline frequency. After onset, the trajectory follows a logistic frequency-change model:

$$f_r(t_i) = \begin{cases} f_{0,r}, & t_i < t_{0,r}, \\ \dfrac{f_{0,r}\exp\{s(t_i - t_{0,r})\}}{1 - f_{0,r} + f_{0,r}\exp\{s(t_i - t_{0,r})\}}, & t_i \geq t_{0,r}. \end{cases} \quad (27)$$

In this formulation, $s > 0$ corresponds to a post-onset increase in insertion frequency, $s = 0$ corresponds to no directional frequency change after onset, and $s < 0$ corresponds to a post-onset decrease. The region-specific onset times $t_{0,r}$ capture temporal shifts among regional trajectories, whereas the shared s imposes a common post-onset rate of frequency change across regions.

Single-region trajectory model. The single-region model is a special case of this formulation with $R=1$, giving one baseline frequency f_0, one onset time t_0, and one selection coefficient s.

Genotype-marginalized observation likelihood. For sample i, the trajectory-predicted insertion frequency was $f_i = f_{r_i}(t_i)$. Using the same genotype-marginalized observation model as above, the detection probability was:

$$p_i = 2f_i(1-f_i)\left[1-\exp\left(-\frac{\lambda_i}{2}\right)\right] + f_i^2\left[1-\exp(-\lambda_i)\right] \tag{28}$$

Therefore, the probability of no junction-read detection is $P(D_i=0 \mid f_i, d_i, \beta_i) = 1-p_i$. The likelihood and log-likelihood across all ancient samples are:

$$\begin{aligned} L(\theta) &= \prod_i p_i^{D_i}(1-p_i)^{1-D_i}, \\ \ell(\theta) &= \sum_i \left[D_i \log(p_i) + (1-D_i)\log(1-p_i)\right], \\ \theta &= \left\{f_{0,r}, t_{0,r}\right\}_{r=1}^{R} \cup \{s\}. \end{aligned} \tag{29}$$

6.4.7 Penalized likelihood and parameter estimation

Penalized likelihood. For numerical stability under sparse ancient DNA sampling, we used a weak symmetric L2 penalty on the shared selection intensity:

$$\text{Penalty}(s) = \frac{1}{2}\left(\frac{s}{s_{\text{prior}}}\right)^2$$

The penalized log-likelihood is:

$$\ell_{\text{pen}}(\theta) = \ell(\theta) - \frac{1}{2}\left(\frac{s}{s_{\text{prior}}}\right)^2 \tag{30}$$

This penalty stabilizes estimation when sample size is small or when onset time and baseline frequency are partially confounded. It does not constrain the sign of s, but weakly shrinks extreme values toward zero. In all trajectory-based analyses, s_{prior} was set to 0.05.

Parameter estimation. Region-specific onset times $t_{0,r}$ were searched over

predefined grids. For each combination of regional onset times, the remaining continuous parameters were optimized, and the combination yielding the highest penalized log-likelihood was retained.

When baseline frequencies were estimated, the optimized parameter vector was:

$$\theta_{\mathrm{opt}} = \left\{\mathrm{logit}(f_{0,1}), \ldots, \mathrm{logit}(f_{0,R}), s\right\}$$

Baseline frequencies were constrained to a specified interval:

$$f_{0,r} \in [f_{\min}, f_{\max}]$$

The selection coefficient was optimized within finite bounds:

$$s \in [s_{\min}, s_{\max}], \quad s_{\min} < 0 < s_{\max}$$

In the implementation used here, s was allowed to vary within a symmetric signed interval. This bounded optimization avoids numerical instability while allowing both positive and negative post-onset frequency changes.

The model also supports fixed baseline-frequency modes. If f_0 is provided as a scalar, the same baseline frequency is used for all regions. If a vector is provided, each region uses its specified baseline frequency. If no fixed value is provided, region-specific $f_{0,r}$ values are estimated within the predefined bounds.

Because onset time and baseline frequency can be partially coupled in sparse temporal data, absolute onset-time estimates were interpreted together with their bootstrap uncertainty. We also emphasized the relative ordering of regional trajectories, which is more robust than any single point estimate of onset time.

6.4.8 Bootstrap resampling and regional robustness analyses

Uncertainty was assessed by bootstrap resampling. For the multi-region trajectory model, samples were resampled within each region with replacement so that the number of samples per region was preserved in each bootstrap replicate. The model was then refitted to each bootstrap dataset. Bootstrap distributions were used to

summarize uncertainty in regional onset times, the shared selection coefficient, and fitted frequency trajectories.

Sensitivity to regional definitions. To assess whether regional boundary choices affected the inferred frequency trajectories, we repeated the analysis under alternative regional assignments: excluding Balkan samples from the Fertile Crescent–Anatolian-related Central Eurasian region, assigning Balkan samples to Western Eurasia, and excluding Arctic samples from the Eastern/Northern Eurasian and Arctic North American group. The same likelihood framework and bootstrap procedure were applied in each case.

6.4.9 Real-data implementation settings for bin-level and trajectory analyses

The following settings were used for the empirical ancient-dog frequency analyses reported in this study. For the single-bin frequency analysis, ancient-dog samples were grouped into four temporal bins: <3,000 BP, 3,000 – <5,000 BP, 5,000 – <9,000 BP and ⩾9,000 BP. Within each bin, the insertion frequency was estimated independently using the dropout-adjusted likelihood framework described above. Uncertainty was assessed from 1,000 nonparametric bootstrap replicates generated by resampling samples with replacement within the corresponding temporal bin.

For trajectory fitting, sampling ages in BP were converted to a forward-generation scale using a generation time of 3 years, with the oldest modeled ancient-dog sample used as the temporal origin. The oldest modeled sample was assigned to generation 1, and younger samples were placed forward in time according to their age difference from this sample divided by 3 years. Estimated onset times on the forward-generation scale were converted back to BP using the same generation time.

Baseline insertion frequencies were estimated rather than fixed. For each region, the baseline frequency was constrained to 0.0001–0.05. The weak L2 penalty scale for the post-onset selection coefficient was set to 0.05. No ordering constraint was imposed

on regional onset times during model fitting.

For single-region trajectory fits, the model was fitted separately within each focal region. The onset time was searched over 50, 100, 150, …, 5,000 generations on the forward time scale. Uncertainty was assessed from 1,000 nonparametric bootstrap replicates generated by resampling samples with replacement within the analyzed region.

For the shared-selection model, all three regions were fitted jointly. This model allowed region-specific baseline frequencies and onset times, while sharing a single post-onset selection coefficient across regions. For each region, the onset time was searched over 100, 200, 300, …, 5,000 generations on the forward time scale. Bootstrap resampling was stratified by region, preserving the original regional sample sizes in each replicate. The shared-selection model was summarized from 1,000 stratified bootstrap replicates.

6.5 Simulation benchmarks of ancient DNA likelihood modeling

We performed three simulation benchmarks to evaluate the performance of the dropout-adjusted likelihood framework under sparse, low-coverage and uneven sampling conditions typical of ancient DNA datasets. In all simulations, local sequencing depth and per-depth junction-read detection efficiency were sampled from the empirical ancient-DNA dataset, and insertion detection was generated under the same genotype-specific dropout model used in the main analyses. Each parameter combination was simulated for 1,000 replicates. In Supplementary Figs. 37–39, boxplots summarize the full distribution of replicate estimates, gray points show 200 randomly sampled replicate estimates for visualization, and red diamonds indicate the true simulated values.

First, we assessed the single-bin frequency estimator across true insertion frequencies from 0.05 to 0.80 and sample sizes from 20 to 1,000 (Supplementary Fig. 37). The

estimator recovered the underlying insertion frequency across the simulated range, with uncertainty decreasing as sample size increased. As expected, estimates were most variable at low frequency and small sample size, where informative junction-read observations are sparse. These results indicate that the bin-level likelihood model can correct for false-negative dropout and recover insertion frequency from heterogeneous low-coverage observations.

Second, we evaluated recovery of selection coefficient in the trajectory model by fixing the onset time at 8,000 BP and varying the true selection coefficient from 0 to 0.05 (Supplementary Fig. 38). Estimated selection coefficients tracked the simulated values, and precision improved with increasing sample size. Weak selection was less readily separated from neutrality at small sample sizes, whereas moderate and strong selection produced more stable estimates. Thus, the trajectory model captured the magnitude of post-onset frequency increase under realistic dropout conditions.

Third, we evaluated recovery of selection-onset time by fixing the selection coefficient at 0.01 and varying the onset time from 4,000 to 12,000 BP (Supplementary Fig. 39). The model recovered the expected temporal ordering of onset times, with progressively narrower estimate distributions as sample size increased. These simulations show that onset-time inference is feasible under sparse ancient-DNA sampling, but its precision depends strongly on the number of informative samples and their temporal distribution.

Together, these benchmarks support the robustness of the dropout-adjusted framework used in the ancient DNA analyses. The single-bin model provides calibrated frequency estimates after accounting for detection dropout, whereas the trajectory model recovers both the rate and timing of frequency increase with sample-size-dependent precision. These results support the use of the framework for inferring the spatiotemporal dynamics of the *GPR52*-linked L1-Cf insertion from heterogeneously covered ancient genomes.

7. Canid pangenome construction and L1 insertion discovery

We obtained whole-genome sequences of 10 domestic dogs, 1 dingo, 2 gray wolves, and 1 coyote from NCBI databases (Supplementary Table 5) and performed comprehensive repeat element annotation using RepeatMasker (version 4.1.7) with canid-specific repeat libraries from Repbase (version 20181026). The Minigraph-Cactus (version 2.5.2) pangenome pipeline[32] was then employed to construct a canid pangenome from these 14 individuals, followed by allele-specific variant detection using VCFwave (version 1.0.12)[33]. We identified L1 retrotransposon variants based on two stringent criteria: a minimum length of 6 kb and RepeatMasker annotation as L1 elements. Genes located within 50 kb of these L1 insertions were then annotated, and the resulting gene set was collapsed to unique named gene symbols after excluding LOC-prefixed predicted or unnamed gene models, yielding 1,213 L1-adjacent genes for functional enrichment analysis. Enrichment analysis was conducted using KOBAS-i[34]. Because canine functional annotations remain limited, human GO terms and KEGG pathways were used as reference datasets.

Supplementary Table 5. Canid genome assemblies used for pangenome construction and L1 insertion discovery.

Assembly	GenBank	Breed/Isolate	Species
UMICH_Zoey_3.1	GCA_005444595.1	Great Dane	Dog
ASM4364393v1	GCA_043643935.1	Whippet	Dog
TAMU_N220234	GCA_031771975.1	Irish Wolfhound	Dog
mCanLor1.2	GCA_905319855.2	-	Gray Wolf
Clu-1	GCA_034620435.1	Wolf_2809A	Gray Wolf
Cla-1	GCA_034620425.1	Coyote_06102	Coyote
ASM325472v2	GCA_003254725.2	Sandy	Dingo
UNSW_CanFamBas_1.2	GCA_013276365.2	Basenji	Dog
BD_1.0	GCA_031010765.1	Bernese Mountain Dog	Dog
CA611_1.0	GCA_031010295.1	Cairn Terrier	Dog
LK_Cfam_Beagle_1.1	GCA_044048985.1	Beagle	Dog
CanFam_VD_v1_BIUU	GCA_004027395.1	BS72	Dog
UU_Cfam_GSD_1.0	GCA_011100685.1	German Shepherd	Dog
ROS_Cfam_1.0	GCA_014441545.1	Labrador retriever	Dog

8. Expression analyses and L1-associated transcriptional effects

8.1 Differential expression of *GPR52* and *RABGAP1L* by L1-Cf status

We obtained RNA-seq profiles from 20 central nervous system (CNS) tissues of six domestic dogs and four wolves from DoGA[35], together with available whole-genome sequencing data for L1-Cf genotyping. Whole-genome reads were aligned to the reference with BWA v0.7.17, and structural variants were called with Manta v1.6.0[36] to determine presence or absence of the *GPR52*-upstream L1-Cf insertion in each individual with genome-sequencing data. Two wolf individuals lacked matched whole-genome sequencing data and were treated as presumed non-carriers. This treatment was unlikely to materially affect the analysis because L1-Cf is rare in modern wolves, with a mean population-level frequency of 1.97%. RNA-seq reads were aligned with STAR v2.7.11b[37], and transcript-level abundances were estimated with RSEM v1.3.3[38]. Transcript estimates were summarized to gene-level counts using tximport and analyzed with DESeq2 v1.40.2[39]. DESeq2 size-factor normalization was applied, and differential expression between L1-Cf-bearing and non-carrier or presumed non-carrier individuals was tested using the Wald test.

Tissue-level differential-expression tests were performed only for CNS tissues with at least two samples in each comparison group, leaving 17 of the 20 CNS tissues testable. For each testable tissue, DESeq2 was used to perform genome-wide differential-expression analysis between L1-Cf-bearing and non-carrier or presumed non-carrier individuals. Within each tissue, P values were adjusted across all tested genes using the Benjamini–Hochberg procedure implemented in DESeq2. Adjusted P values for *GPR52* and *RABGAP1L* were then extracted from the corresponding tissue-level DESeq2 results.

8.2 Assessment of L1-associated regulatory effects across tissues and organ systems

We modeled the effect of L1 insertions on expression of neighboring genes using a

log-normal model. Let R_i denote the untransformed normalized expression level for sample i, and let $X_i = \ln(R_i + 1)$ denote its natural-log-transformed expression. Genotypes were coded as 0/0, 0/1 and 1/1 and represented numerically by $g_i \in \{0,1,2\}$, corresponding to the number of insertion alleles. We used an additive allelic contribution model on the raw-expression scale, with each insertion-bearing allele contributing *k*-fold the baseline allelic expression. Under this model, the expected expression levels for genotypes 0/0, 0/1 and 1/1 are 2μ, $(1+k)\mu$ and $2k\mu$, respectively, where μ represents the baseline expression contribution per non-insertion allele. The logarithmic gene-expression level was modeled as

$$X_i \sim N\left(\ln[g_i k\mu + (2-g_i)\mu + 1], \sigma^2\right), \tag{31}$$

where σ^2 denotes the variance of logarithmic gene-expression levels estimated from the data. Model parameters were constrained as $\mu > 0$ and $k > 0$, ensuring positive expected expression levels while allowing the L1-associated allele to either increase ($k > 1$) or decrease ($0 < k < 1$) expression.

The log maximum-likelihood function for parameters *k* and μ is given by

$$L(k, \mu \mid \mathbf{X}) = \sum_i \log\left(\mathrm{N}(X_i \mid \ln\left(g_i k\mu + (2-g_i)\mu + 1\right), \sigma^2)\right), \tag{32}$$

where $\mathbf{X} = (X_i)$, X_i is the log-transformed normalized expression value for sample *i*, $\mathrm{N}\left(X_i \mid \ln\left(g_i k\mu + (2-g_i)\mu + 1\right), \sigma^2\right)$ denotes the normal density evaluated at X_i. Genes adjacent to an L1 insertion that were monomorphic (only a single observed genotype class) were excluded from downstream analysis.

Tissue-level assessment of L1 effects on gene expression. We assessed the effects of L1 insertions on gene expression across 89 tissues in the DoGA dataset, comprising 759 RNA-seq samples, 654 of which had matched DNA-seq data. When multiple RNA-seq libraries were available for the same individual–tissue combination, only

the library with the greatest sequencing depth was retained. Samples with mean sequencing depth below 2× were excluded, as were tissues represented by fewer than two samples with matched DNA-seq data. After filtering, 401 samples from 84 tissues remained for analysis. For each tissue, we tested the null hypothesis of no mean effect on expression, defined as a mean log-fold change of zero, against one-sided alternatives corresponding to activation, mean log-fold change > 0, or inhibition, mean log-fold change < 0, using one-sample t-tests. Genes with a maximum expression level below 20 across samples were excluded. For each L1-adjacent gene, estimation of the parameter k in Eq. 32 was performed only when at least two distinct genotypes were present at the corresponding L1 locus. We therefore first retained genes satisfying this genotype-diversity requirement and then excluded tissues with fewer than 50 retained L1-adjacent genes from hypothesis testing. After this filtering step, 32 tissues were excluded, leaving 52 tissues for the tissue-level analyses reported in Fig. 4c and Supplementary Table 10. These retained tissues represented 11 organ systems. One organ system present in the original DoGA tissue panel was excluded because no constituent tissue satisfied the tissue-level retained-gene threshold.

Organ-system-level assessment of L1 effects on gene expression. We further evaluated the effects of L1 insertions on gene expression across 12 organ systems in the DoGA dataset, using the same analytical framework as applied to individual tissues. Each organ system encompassed multiple constituent tissues, and significance testing was performed after aggregating expression data across these tissues. Because organ-system-level tests were performed after aggregation, the organ system without a retained individual tissue in the tissue-level analysis still satisfied the organ-system-level retained-gene threshold and was included in Supplementary Table 11. For tissue-level and organ-system-level summaries, P values were adjusted separately across tested tissues or organ systems using the Benjamini–Hochberg procedure.

9. Genome-wide Hi-C processing and *GPR52* locus analysis

9.1 Hi-C library preparation

Hi-C libraries were generated from caudate nucleus tissue collected from two village dogs previously identified as homozygous carriers of the L1-Cf insertion near *GPR52*. Libraries were prepared following a previously published protocol, with minor modifications[40]. Briefly, approximately 50 mg of caudate nucleus tissue from each dog was finely minced on ice and cross-linked with 1% formaldehyde in 1× PBS. Nuclei were isolated, and chromatin was digested with 100 U of DpnII (New England Biolabs, cat. no. R0543) for 2 h at 37 °C. The resulting DNA overhangs were filled in and biotin-labelled using Klenow fragment (New England Biolabs, cat. no. M0210) in the presence of Biotin-14-dATP (Jena Bioscience, cat. no. NU-835-BIO14). Proximity ligation was performed with 4,000 U of T4 DNA ligase (New England Biolabs, cat. no. M0202) for 4 h at room temperature. Cross-links were reversed, and DNA was purified by ethanol precipitation. The ligation products were sheared by sonication to a fragment-size range of 250 – 500 bp, and biotinylated fragments were enriched using MyOne Streptavidin T1 Dynabeads (Thermo Fisher Scientific, cat. no. 65602). The captured DNA was end-repaired, A-tailed and ligated to MGI-compatible adapters (Vazyme, cat. no. NM35101). Libraries were amplified using FastPfu Fly DNA Polymerase (TransGen Biotech, cat. no. AP231) and sequenced on the MGI T7 platform (BGI) to generate 2 × 150-bp paired-end reads.

9.2 Hi-C data processing and loop detection

For each library, raw reads were processed using fastp v1.0.1[41] to remove adapter sequences and low-quality bases. Read pairs were aligned to the canFam4 reference genome using BWA-MEM v0.7.17 with the options -aSPM[42]. The -a option was used to retain all reported alignments, including secondary alignments, to improve the recovery of read pairs originating from repetitive genomic contexts. Sample-specific contact matrices were generated using Juicer v1.6[43]. Chromatin loops were called chromosome-wise using HiCCUPS implemented in Juicer Tools[44] at 5-, 10- and 25-kb

resolutions with KR normalization (hiccups --cpu -r 5000,10000,25000 -k KR). Both libraries yielded sufficient valid contacts to support locus-level analyses at resolutions from 5 to 25 kb.

To evaluate sample-level reproducibility at the *GPR52* locus, the same processing and loop-calling procedure was first applied to each homozygous L1-Cf-bearing caudate sample separately. The local interaction between the *GPR52*-overlapping bin and the L1-Cf-flanking region was detected independently in both samples under the same HiCCUPS settings, indicating that the locus-level signal was not driven by a single individual.

For the primary locus-level analysis, aligned reads from the two homozygous individuals were merged and name-sorted with samtools (v1.0.1; samtools sort -n)[29] to increase effective contact depth for read-pair processing and loop detection. Contact matrices were generated from the merged read set with Juicer, and chromatin loops were called using HiCCUPS at 5-kb, 10-kb and 25-kb resolutions with KR normalization. Loop calls from the three resolutions were merged into a unified BEDPE loop set for downstream locus-level analysis. Hi-C tracks and loop annotations were visualized using pyGenomeTracks (v3.9)[45].

For the *GPR52* locus, we focused on loops overlapping the 290-kb ancestry-enriched interval containing *GPR52* and the upstream L1-Cf insertion. *GPR52* is located at chr7:24,747,842–24,752,390 on the reverse strand, and the full-length L1-Cf insertion is located at chr7:24,791,364–24,797,664 on the forward strand, corresponding to an insertion approximately 39 kb upstream of *GPR52* relative to its transcriptional orientation. In the merged loop set, the relevant local interaction connected chr7:24,740,000–24,750,000, a reference bin overlapping *GPR52*, with chr7:24,800,000–24,810,000, a nearby reference-mappable bin flanking the L1-Cf insertion site. This loop had an observed contact count of 94 and was significant across all four HiCCUPS local background models (FDRBL was reported as 0 by

HiCCUPS, FDRDonut = 3.45×10^{-22}, FDRH = 7.55×10^{-16} and FDRV = 2.07×10^{-21}).

Because L1-Cf is a repetitive transposable-element insertion, uniquely mappable Hi-C contacts are expected to be assigned primarily to flanking reference-mappable sequence rather than to the inserted repeat sequence itself. The observed contact count therefore refers to the interacting reference-bin pair, not to reads mapped within the L1-Cf repeat body. We therefore interpreted the Hi-C signal as placing the L1-Cf insertion-site neighborhood within a local chromatin interaction context associated with *GPR52*.

10. Mouse behavioral assays

10.1 Animals, handling, and treatment

Behavioral tests were performed in 10-week-old male C57BL/6J mice. Animals were handled and habituated to the testing environment for 7 days before experiments. Two independent cohorts of 40 male mice each were used. In the first cohort, mice were randomly assigned to HTL0041178-treated[46] and vehicle-treated control groups (n = 20 per group) and were tested in the three-chamber social assay[47] followed by the looming-stimulus assay[48]. In the second cohort, mice were assigned using the same procedure and were tested sequentially in the elevated plus maze (EPM) assay[49] and the prepulse inhibition (PPI) assay[50]. In each cohort, the less aversive assay was performed before the assay involving strong threat or acoustic startle stimuli to minimize potential carryover effects on subsequent behavioral measurements.

HTL0041178, a *GPR52* agonist[46], was synthesized by Shanghai Yuanxi Biotechnology Co., Ltd. LC–MS analysis (SunFire C18 column, DAD detection at 214/254 nm and ESI+ MS) confirmed a purity of 99.65%. For administration, HTL0041178 (50 mg) was first dissolved in 200 µL DMSO and then diluted into 200 mL drinking water, yielding a final DMSO concentration of 0.1% (v/v). Treated mice

received this HTL0041178-containing drinking water as their sole drinking fluid for three days before testing and throughout the testing period. Vehicle-treated control mice received drinking water containing 0.1% DMSO without HTL0041178 on the same schedule. Based on an assumed average body weight of 25 g and daily water intake of 5 mL per mouse, the estimated HTL0041178 intake was approximately 1.25 mg per mouse per day, corresponding to ~50 $mg \cdot kg^{-1} \cdot day^{-1}$. Behavioral assays were conducted at the Behavior Analysis Center (BAC) of the Beijing Institute of Brain Science and Brain Creation (CIBR).

The compound HTL0041178 is a highly potent and selective *GPR52* agonist, characterized by nanomolar potency across human, canine, and murine orthologs (typical EC <10 nM in cAMP accumulation assays)[46,51]. Its specificity was rigorously validated through extensive profiling against a panel of >160 non-target GPCRs, where it demonstrated >1,000-fold selectivity over other receptors, including critical dopaminergic (D_1 –D_5), serotonergic, and adrenergic subclasses[46]. Furthermore, standard pharmacological safety screens (e.g., Eurofins Cerep and hERG assays) confirmed no significant off-target interactions at concentrations up to 10 μM. This selectivity profile supports the interpretation that the behavioral shifts observed in the looming-stimulus assay are primarily attributable to *GPR52* activation rather than broad collateral modulation of other aminergic or peptidergic neurotransmitter systems.

10.2 Looming-stimulus assay

The looming stimulus assay was performed by presenting an expanding black disc overhead to mimic an approaching predator.

Apparatus. Experiments were performed in a custom rectangular test chamber equipped with an overhead monitor used to present looming visual stimuli. Behavioral video and stimulus timing were recorded synchronously using OBS (Open Broadcaster Software).

Procedure. Testing comprised an adaptation day followed by the looming-stimulation session. On each test day mice were transferred to the behavioral room 30 min prior to testing.

Adaptation (day 1). With the overhead display set to a uniform gray background, the mouse was placed in the chamber and allowed to habituate for 5 min. The chamber was cleaned with 70% ethanol between animals.

Looming stimulation (day 2). With the display on a gray background, the mouse was placed in the chamber for a 2-min baseline period. While the animal was engaged in exploratory locomotion, a sequence of 15 looming trials was delivered. Each looming trial consisted of an expanding black stimulus whose visual angle increased from ~2° to ~80° over ~0.27 s, followed by a brief plateau (~0.25 s) before the screen returned to gray; trials were presented consecutively with a 1-s cycle (each trial lasting 1 s). Behavior was recorded continuously during and after the stimulus sequence. The chamber was cleaned with 70% ethanol after each session.

Scoring and analysis. Freezing (immobility) time was the primary measure of defensive responding to looming visual stimuli. Videos were analyzed with the custom behavior-analysis software of BAC, CIBR. Freezing episodes were identified as periods of immobility (excluding respiratory movements). Trial-by-trial freezing durations and aggregated epoch measures (e.g., trials 1, trials 2–5, trials 6–10, trials 11–15) were extracted for statistical comparison between groups. Freezing measures were quantified for each mouse and compared between HTL0041178-treated and vehicle-treated mice as described in section 10.6.

10.3 Elevated plus maze assay

Apparatus. A white elevated plus maze (EPM; two open arms and two closed arms) was used and placed inside a sound-attenuating room with no salient distal visual cues. Behavior was recorded and tracked using EthoVision XT16 (Noldus). Lighting and other environmental parameters were held constant across animals.

Procedure. Mice were transported to the behavioral room 30 min before testing for acclimation. Each mouse was placed in the center of the maze facing an open arm (preferably the arm opposite the experimenter) and allowed to freely explore for 5 min. At the end of the session the animal was returned to its home cage and the maze surfaces were wiped with 70% ethanol.

Analysis. Automated tracking (EthoVision XT16, Noldus) was used to extract total distance traveled, the number of entries into and time spent in the open and closed arms, and the frequency and duration of nose-in/head-dip events. These measures were used to assess locomotor/exploratory and anxiety-like behavior, with statistical comparisons described in section 10.6.

10.4 Three-chamber social assay

Apparatus. A custom three-chamber apparatus with opaque acrylic partition walls and small inter-chamber doors was used. Two identical wire-cup enclosures with removable lids were placed in diagonally opposite side chambers.

Procedure. Testing was carried out over three consecutive days. Mice were transferred to the behavioral testing room 30 min before the first trial. On Day 1 (habituation), the subject mouse was placed in the central chamber and allowed to explore the apparatus for 5 min with the side chambers closed; the wire cups were present but empty. On Day 2 (social-affiliation session), an unfamiliar sex- and age-matched conspecific (Stranger 1) was introduced into one wire cup, whereas the opposite cup remained empty. The subject mouse was placed in the central chamber and allowed to freely explore all three chambers for 10 min. On Day 3 (social-novelty session), Stranger 1 remained in the original cup and a second unfamiliar conspecific (Stranger 2) was placed in the opposite cup; the subject mouse was again allowed to freely explore for 10 min. Stranger mice were naïve to the subject and matched for background, age and sex for each test. Between trials, the apparatus and cups were cleaned with 70% ethanol to minimize olfactory cues.

Analysis. Behavioral scoring was performed using DeepLabCut-based tracking. Nose-tip, head and neck landmarks were annotated for each subject mouse. An interaction was scored as active social contact only when the subject's nose was directed toward the cup and entered a predefined social zone approximately 3–5 cm from the cup. Episodes in which the subject climbed onto the cup were excluded from the social-contact measure. Outcome measures included time spent in each chamber and time spent in active contact with each cup. Social affiliation was quantified as the difference in exploration time between the stranger-containing and empty cups and as a discrimination index, calculated as $(\text{social} - \text{object}) / (\text{social} + \text{object})$. Social novelty preference was quantified as the difference in exploration time between Stranger 2 and Stranger 1 and as a discrimination index, calculated as $(\text{novel} - \text{familiar}) / (\text{novel} + \text{familiar})$. These measures were used to assess social behavior and social novelty preference, with statistical comparisons described in section 10.6.

10.5 Prepulse inhibition assay

Apparatus. Acoustic startle response and prepulse inhibition (PPI) were measured using a four-channel MED Associates startle-response system (MED-ASR-PRO1). Before testing, speakers were calibrated to ensure accurate sound-pressure levels, and the sensitivity of the startle platforms was calibrated according to the manufacturer's instructions.

Procedure. Mice were transferred to the behavioral testing room before the experiment and allowed to acclimate to the testing environment. The PPI protocol consisted of an input/output (I/O) function phase, a short-term habituation phase and a prepulse-inhibition phase. During acclimation, mice were exposed to constant background white noise of 65–68 dB. In the I/O phase, 20-ms white-noise startle pulses of increasing intensity were presented to determine the pulse intensity that produced a near-plateau startle response in this cohort. Based on this calibration, a

110-dB white-noise pulse was selected as the main startle stimulus for the subsequent habituation and PPI phases.

During the short-term habituation phase, repeated 110-dB startle pulses were presented against the same constant background white noise, and startle responses were recorded across trials to assess habituation. During the PPI phase, pulse-alone trials and prepulse-plus-pulse trials were presented in a pseudorandom order. Four prepulse-plus-pulse conditions were used, consisting of two prepulse intensities (75 and 85 dB) crossed with two prepulse–pulse intervals (30 and 100 ms). Each main trial type was presented 10 times. Prepulse-alone trials were included as an additional control to verify that the prepulse stimuli themselves produced minimal startle responses. Startle responses were quantified using both peak amplitude and average response amplitude.

Analysis. For the I/O function, startle amplitude was plotted across increasing stimulus intensities to determine the main pulse intensity used for PPI testing. For short-term habituation, startle responses across repeated 110-dB pulse trials were normalized and compared between treatment and vehicle-treated control groups. For PPI analysis, responses in each prepulse-plus-pulse condition were averaged within each animal and expressed as a percentage of the pulse-alone baseline response. Percent PPI was calculated as

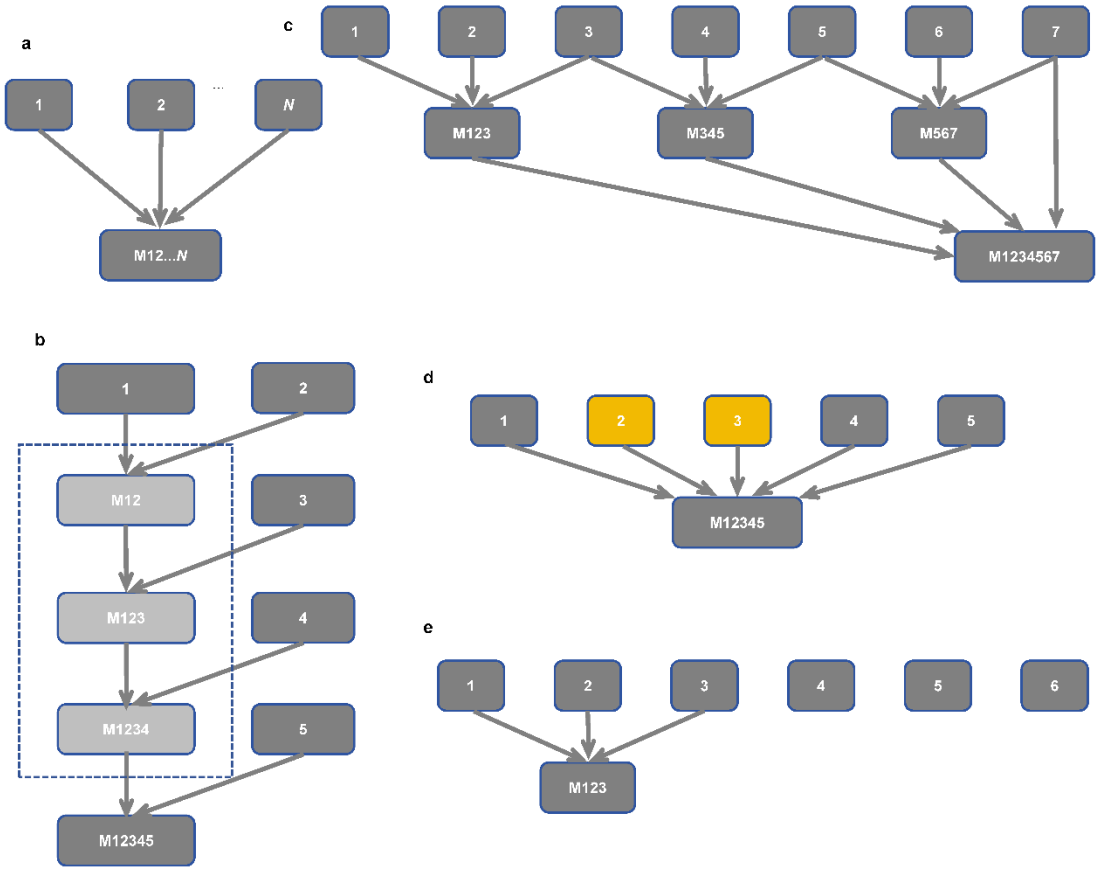
$$\%\text{PPI} = 100 \times \left[1 - \frac{\text{prepulse-plus-pulse response}}{\text{pulse-alone response}} \right].$$

Pulse-alone startle amplitudes and prepulse-alone responses were also examined to assess baseline startle reactivity and the effect of prepulse stimuli alone. Baseline startle amplitude, short-term habituation, prepulse-alone responses and PPI across the four prepulse conditions were used to assess acoustic startle reactivity and sensorimotor gating, with statistical comparisons described in section 10.6.

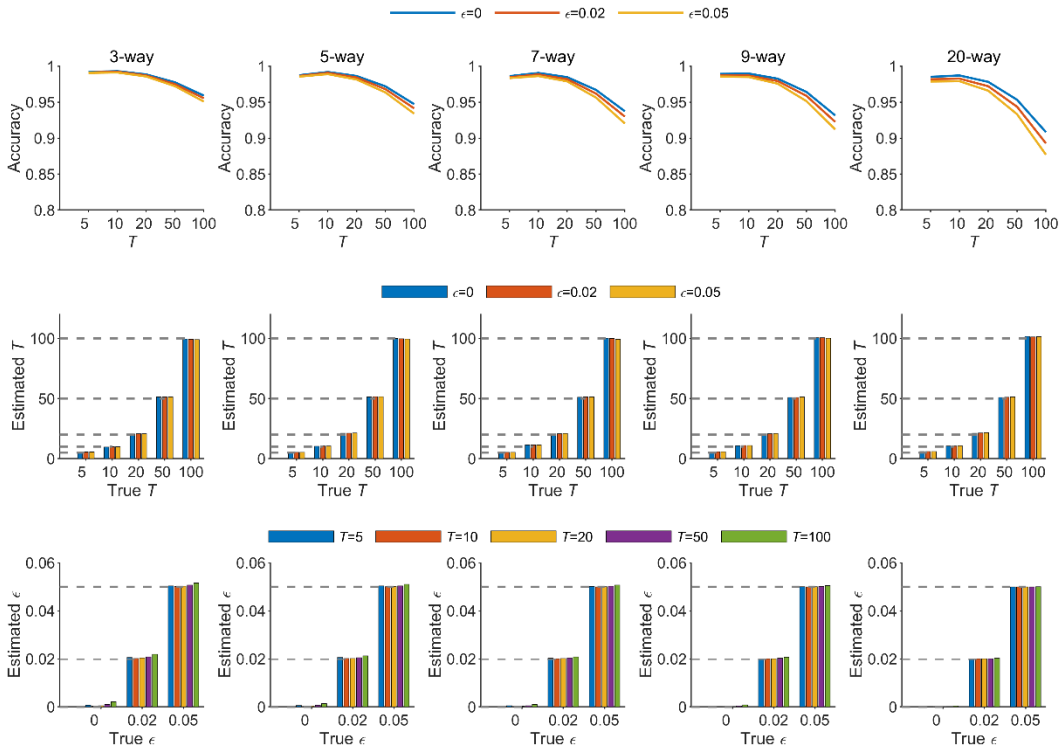
10.6 Statistical analysis of behavioral assays

For the looming-stimulus assay, the directional hypothesis was that *GPR52* activation would attenuate acute threat-evoked defensive freezing. One-sided independent-samples t tests were therefore used for predefined comparisons of looming-evoked freezing between HTL0041178-treated and vehicle-treated mice. The elevated-plus-maze, three-chamber social and prepulse inhibition assays were analyzed as complementary behavioral assays assessing locomotor/exploratory and anxiety-like behavior, social behavior and acoustic sensorimotor gating, respectively. For these assays, HTL0041178-treated and vehicle-treated mice were compared using two-sided independent-samples t tests for each measured variable. Non-significant results in these complementary assays were interpreted descriptively as no detectable differences under the tested dosing condition, not as formal evidence of equivalence or complete absence of an effect.

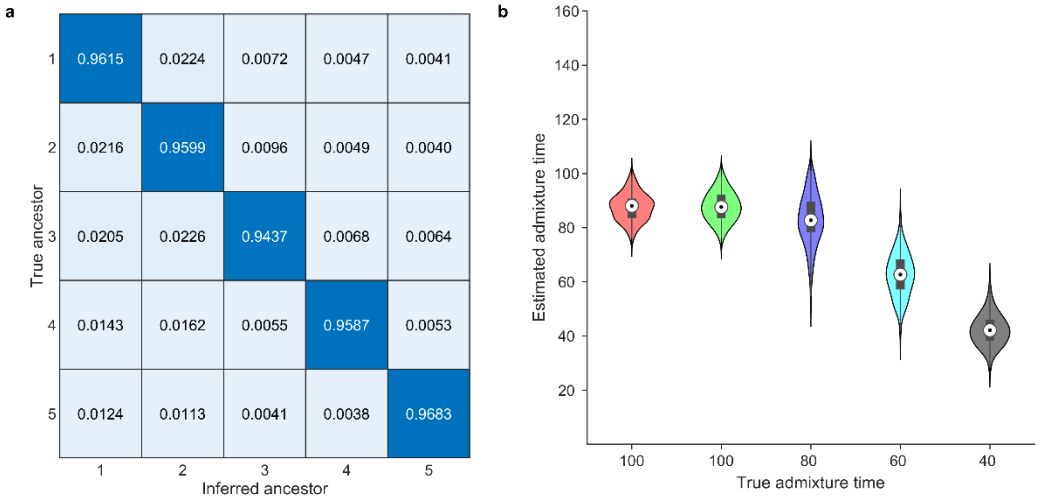
Supplementary Figures



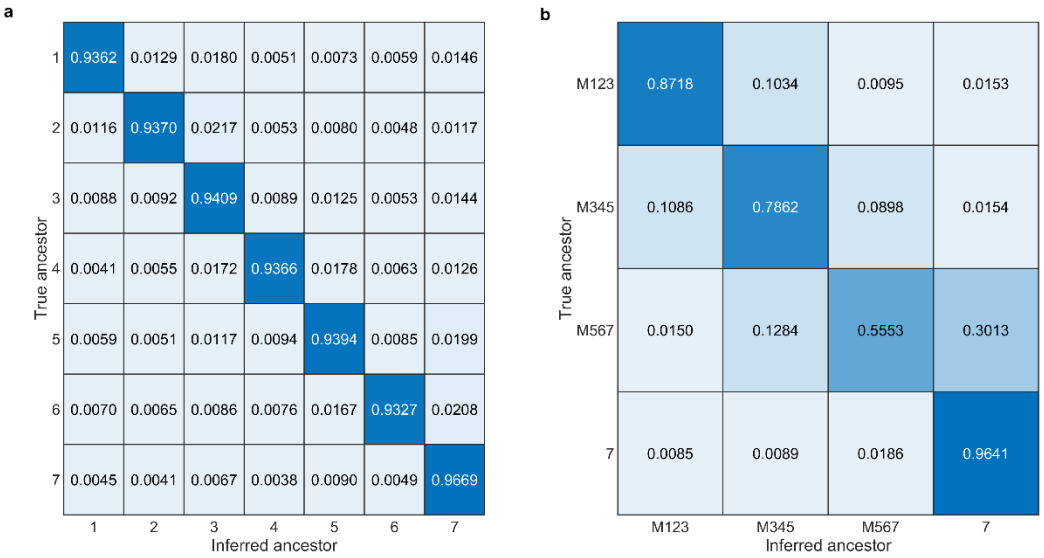
Supplementary Fig. 1 | Admixture models used for simulation. a, Simultaneous admixture of *N* source populations into a focal population. **b,** Continuous admixture from five source populations. Transient intermediate populations (M12, M123, M1234) are shown for illustration and have no extant descendants. **c,** Continuous admixture of seven source populations producing a composite population M1234567. Intermediate populations M123, M345 and M567 are extant and serve as direct contributors to M1234567; populations 1–6 act as indirect contributors, while population 7 contributes both directly and indirectly. **d,** Special case of model (**a**) with five ancestral sources but missing data for sources 2 and 3. **e,** Special case of model (**a**) comprising three true ancestral sources (1–3) plus three pseudo-ancestors that do not contribute genetically to M123.



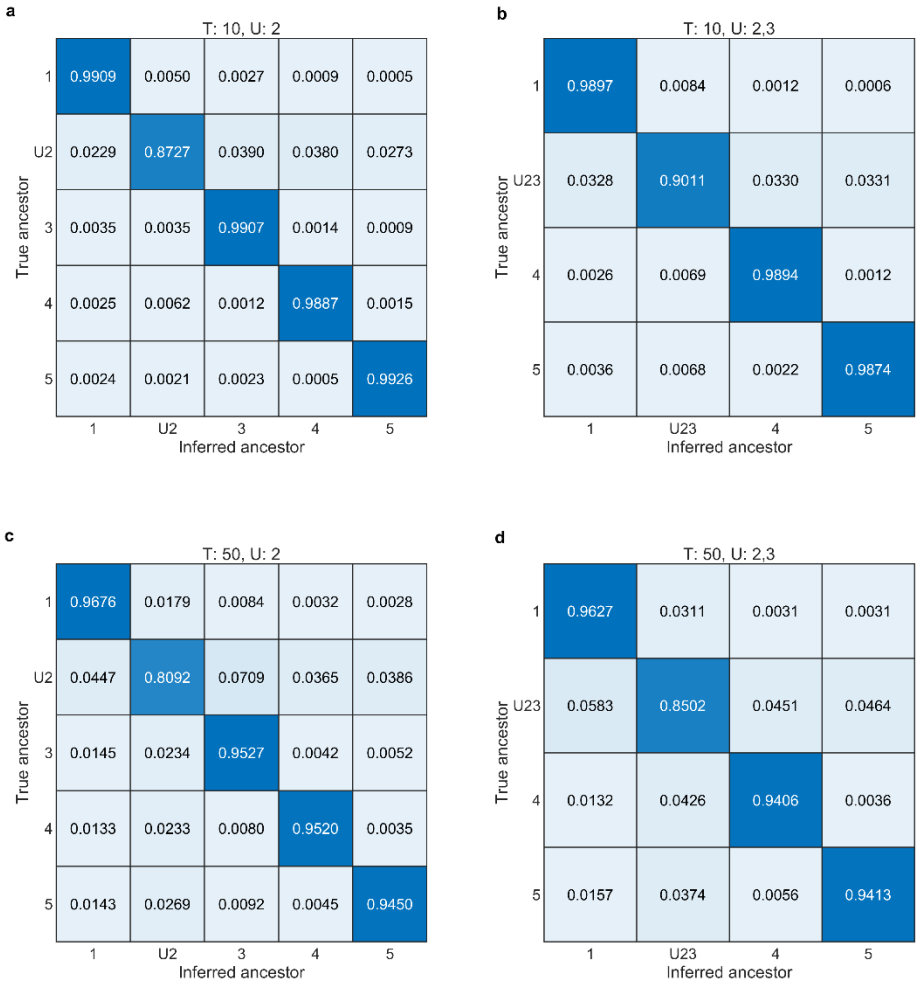
Supplementary Fig. 2 | Performance of GAMA for local-ancestry inference and parameter estimation under simulation model (a). We varied the number of ancestral sources (3, 5, 7, 9 and 20), the mismatch parameter (ϵ = 0, 0.02, 0.05) and the time since admixture (5, 10, 20, 50 and 100 generations). Each ancestral panel comprised 100 haplotypes and the admixed population comprised 20 haplotypes.



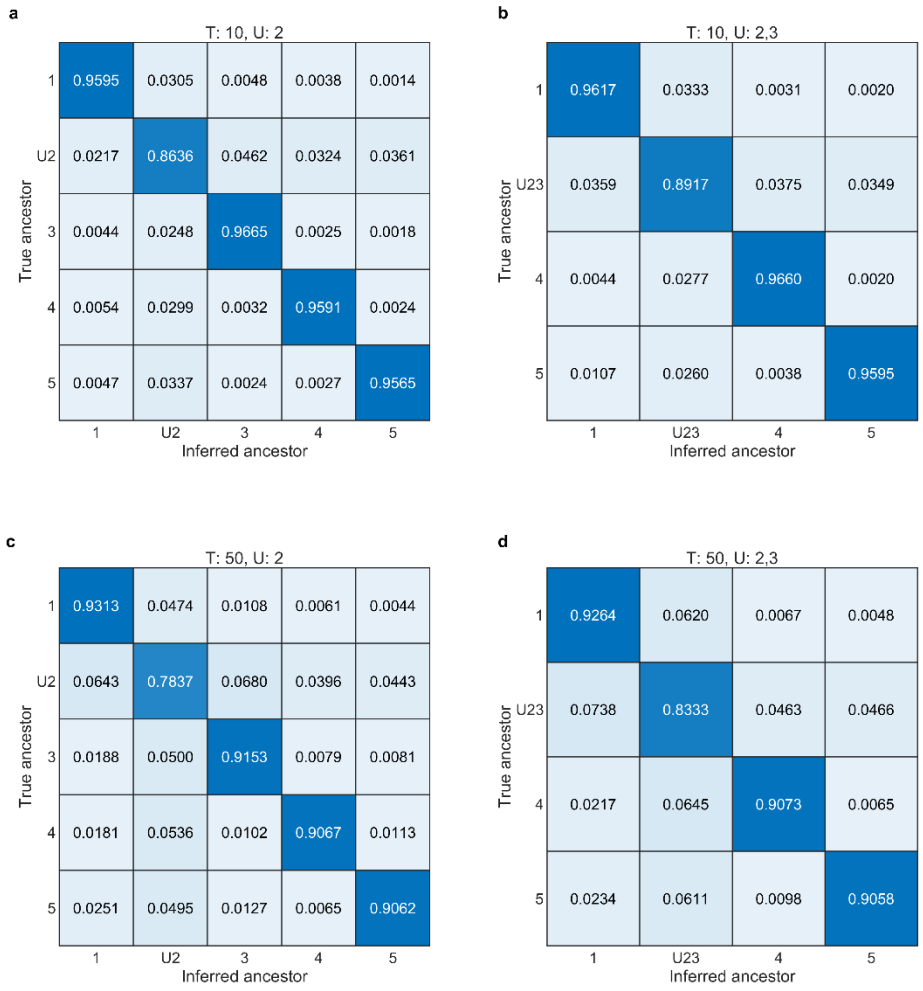
Supplementary Fig. 3 | Performance of GAMA in resolving staggered admixture events. The red and green violin plots (labeled '100') represent independent admixture time estimates for source populations 1 and 2, respectively, reflecting their simultaneous contribution to the initial admixture pulse 100 generations ago.



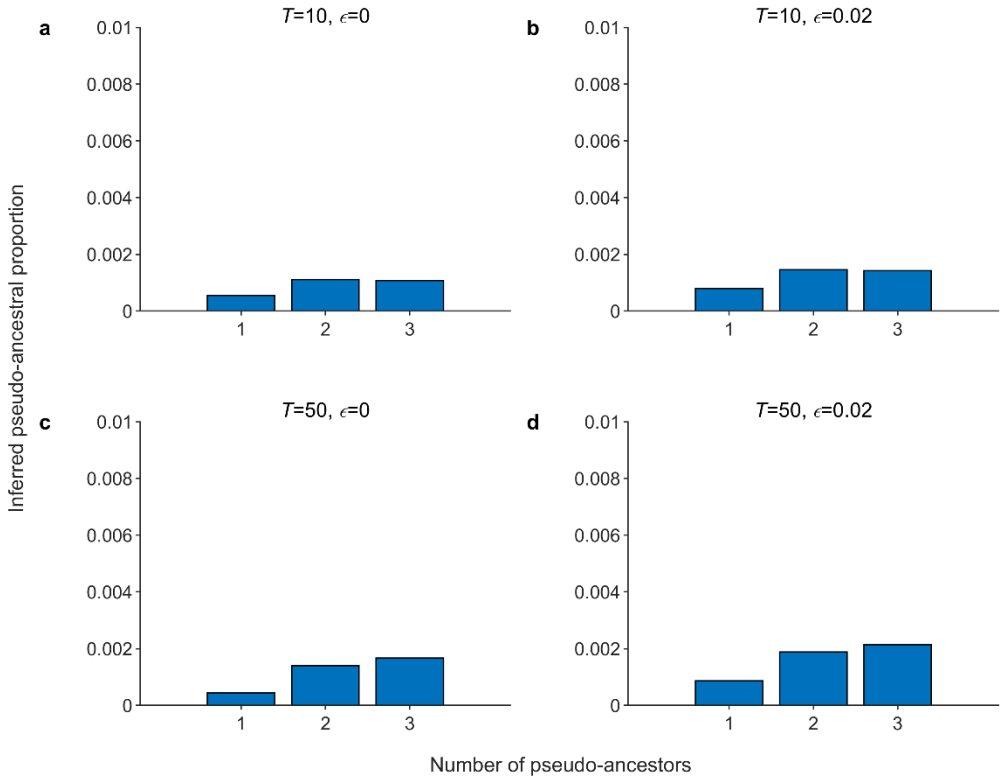
Supplementary Fig. 4 | Performance of GAMA under admixture model B using (a) indirect, unrelated ancestral sources and (b) direct, closely related ancestral sources.



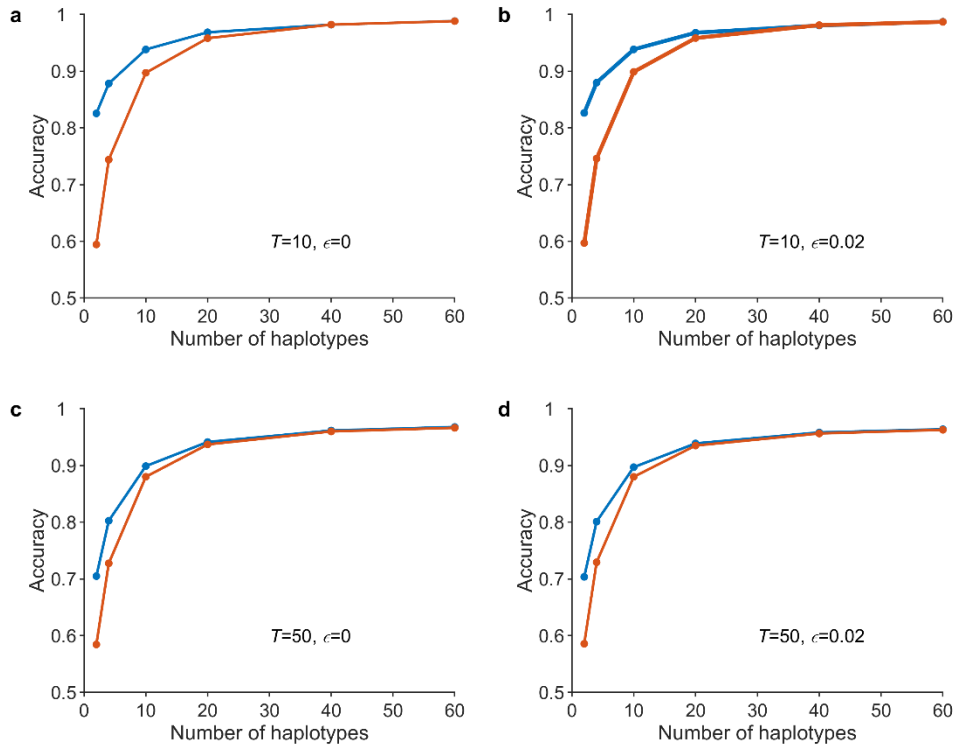
Supplementary Fig. 5 | Performance of GAMA under scenarios with unsampled ancestral sources. Ancestral reference panels comprised 100 haplotypes each; the admixed population comprised 20 haplotypes; mismatch parameter $\epsilon = 0$. "U2" indicates that ancestor 2 is unsampled; "U23" indicates that ancestors 2 and 3 are unsampled. Panels: **a,** admixture 10 generations ago; ancestor 2 unsampled. **b,** admixture 10 generations ago; ancestors 2 and 3 unsampled. **c,** admixture 50 generations ago; ancestor 2 unsampled. **d,** admixture 50 generations ago; ancestors 2 and 3 unsampled.



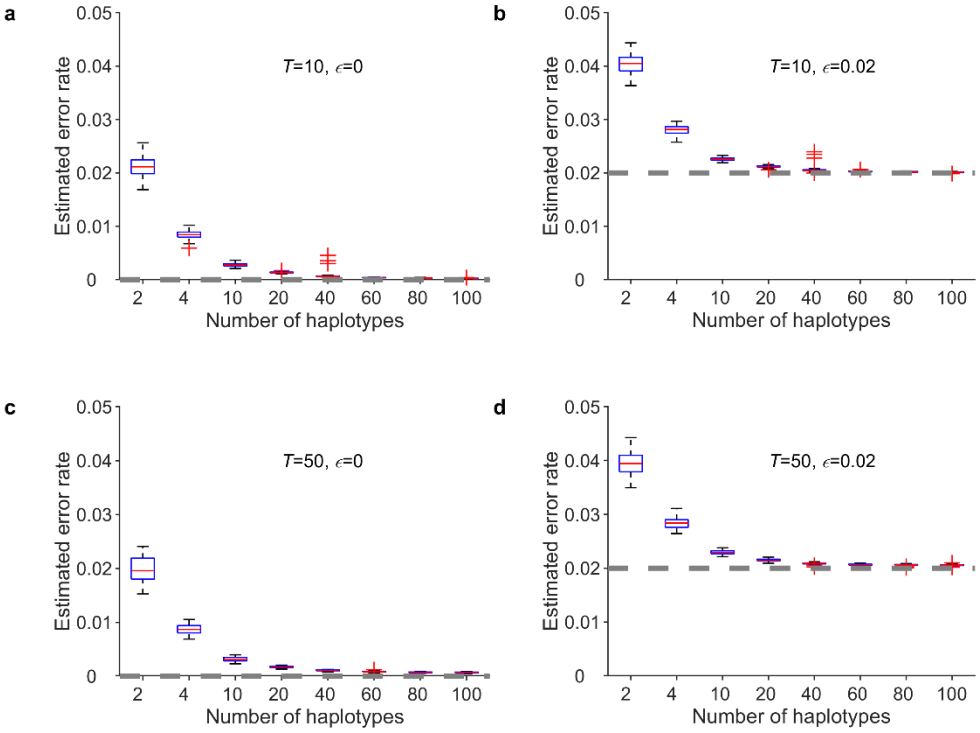
Supplementary Fig. 6 | Performance of GAMA under scenarios with unsampled ancestral sources. Each ancestral reference panel contained 20 haplotypes, and the admixed population comprised 20 haplotypes; mismatch parameter $\epsilon = 0$. "U2" indicates that ancestor 2 is unsampled; "U23" indicates that ancestors 2 and 3 are unsampled. Panels: **a,** admixture 10 generations ago; ancestor 2 unsampled. **b,** admixture 10 generations ago; ancestors 2 and 3 unsampled. **c,** admixture 50 generations ago; ancestor 2 unsampled. **d,** admixture 50 generations ago; ancestors 2 and 3 unsampled.



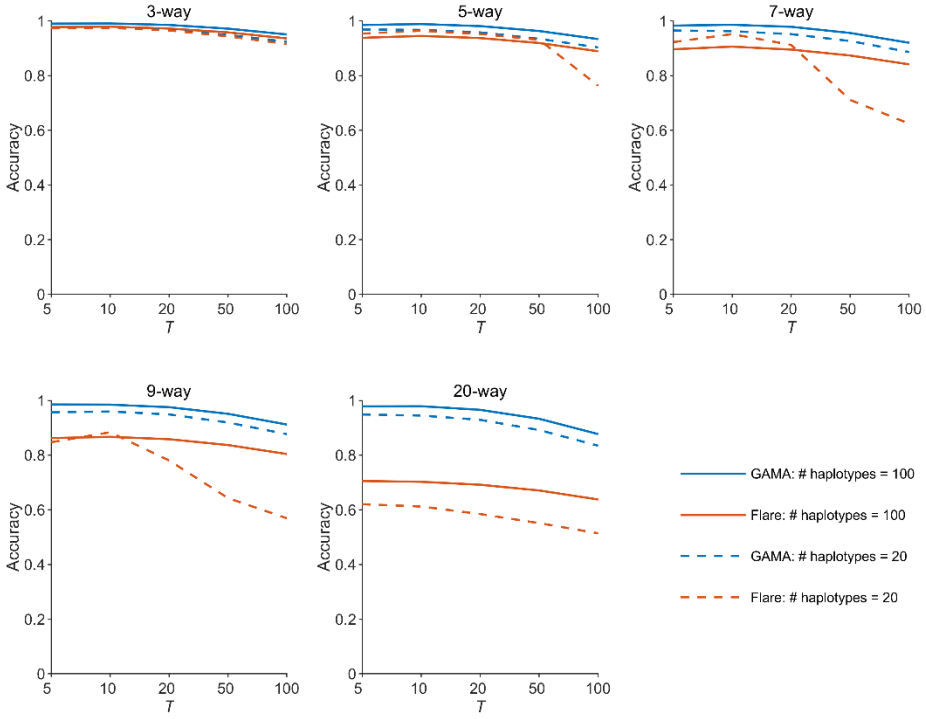
Supplementary Fig. 7 | Performance of GAMA in the presence of pseudo-ancestral panels. Ancestral reference panels comprised 100 haplotypes each and the admixed population comprised 20 haplotypes. Mismatch parameter (ϵ) and admixture time are indicated for each panel: **a,** 10 generations, α = 0; **b,** 10 generations, α = 0.02; **c,** 50 generations, α = 0; **d,** 50 generations, α = 0.02.



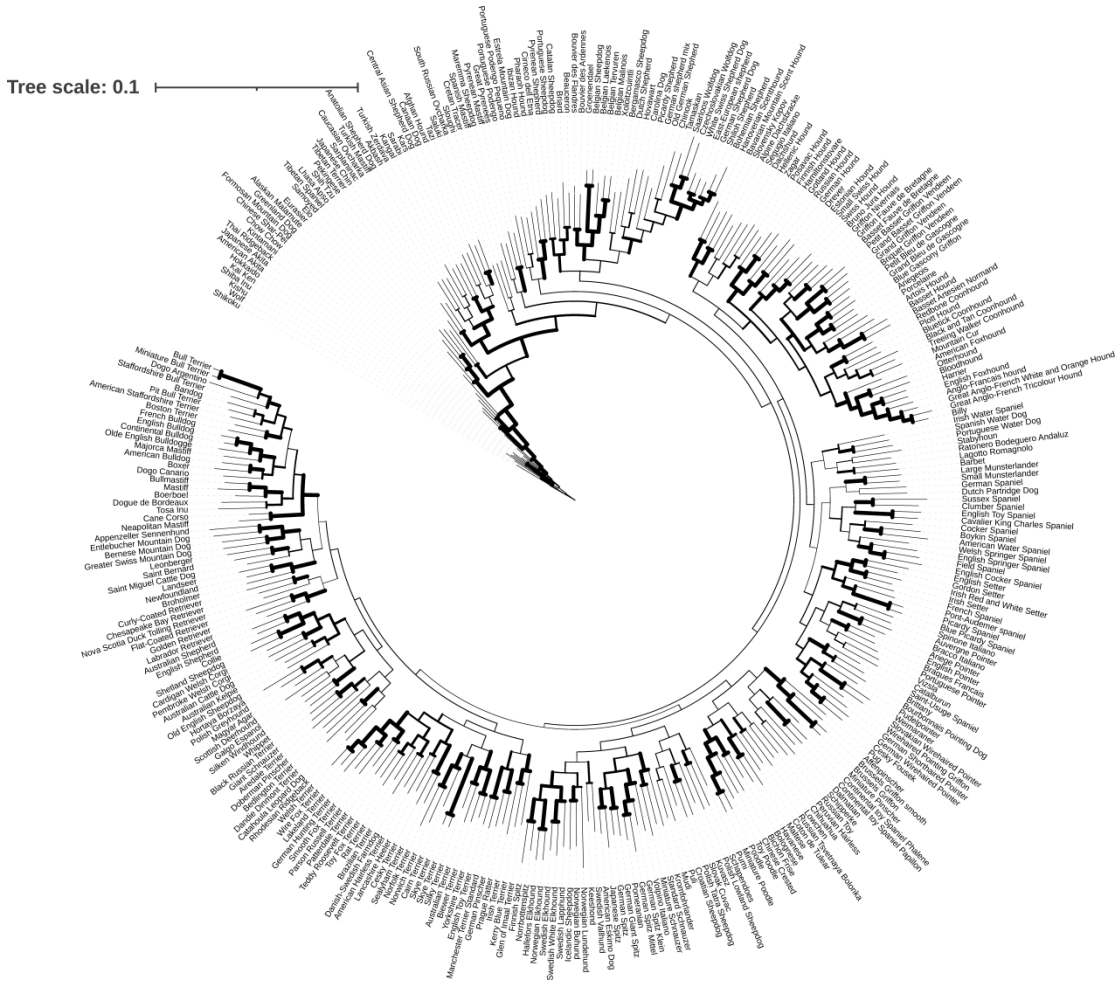
Supplementary Fig. 8 | Effect of Laplace smoothing on GAMA local-ancestry accuracy at small reference-panel sizes. Blue lines show results with Laplace smoothing; red lines show results without smoothing. Reference-panel sizes tested: 2, 4, 10, 20, 40 and 60 haplotypes. Panels: **a,** admixture 10 generations ago, mismatch parameter $\epsilon = 0$; **b,** admixture 10 generations ago, $\epsilon = 0.02$; **c,** admixture 50 generations ago, $\epsilon = 0$; **d,** admixture 50 generations ago, $\epsilon = 0.02$.



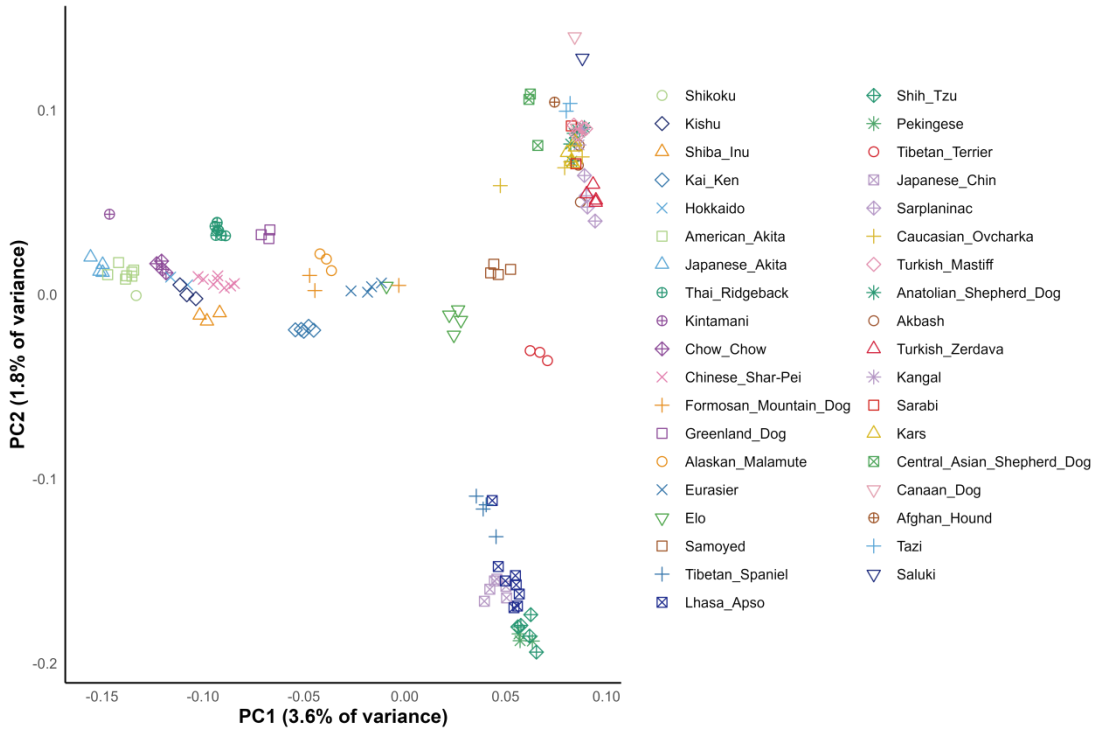
Supplementary Fig. 9 | Accuracy of inferred mismatch parameters across reference-panel sizes. Dashed horizontal lines denote the true mismatch parameters ($\epsilon = 0$ and $\epsilon = 0.02$). Reference-panel sizes tested: 2, 4, 10, 20, 40 and 60 haplotypes. Panels: **a,** admixture 10 generations ago, true $\epsilon = 0$; **b,** admixture 10 generations ago, true $\epsilon = 0.02$; **c,** admixture 50 generations ago, true $\epsilon = 0$; **d,** admixture 50 generations ago, true $\epsilon = 0.02$. Inferred mismatch parameters tend to be upwardly biased for reference panels smaller than 40 haplotypes.



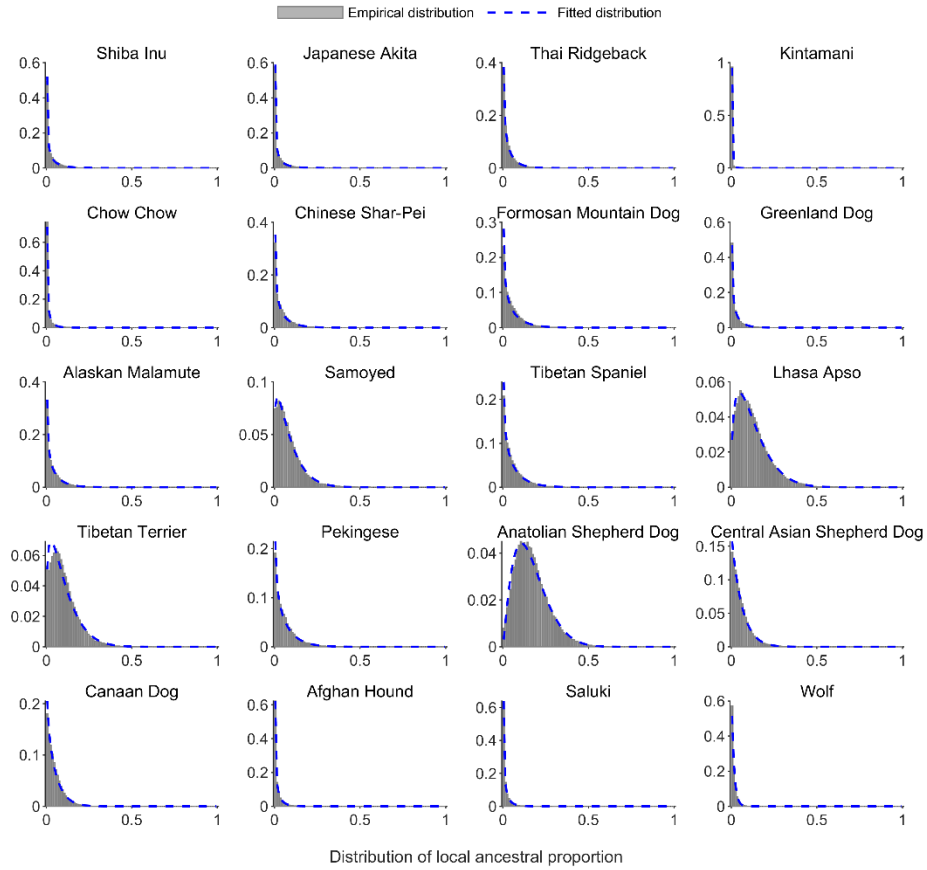
Supplementary Fig. 10 | Comparative accuracy of local ancestry inference between GAMA and FLARE under Simulation Model (a). Performance was evaluated by varying three key parameters: (i) the number of ancestral sources (3, 5, 7, 9, and 20); (ii) the time since admixture (5, 10, 20, 50, and 100 generations); and (iii) the reference panel size (20 and 100 haplotypes per ancestral population). The mismatch parameter was fixed at $\epsilon = 0.05$. The admixed population comprised 20 haplotypes. Data points represent the mean accuracy across 100 independent replicates for each parameter combination.



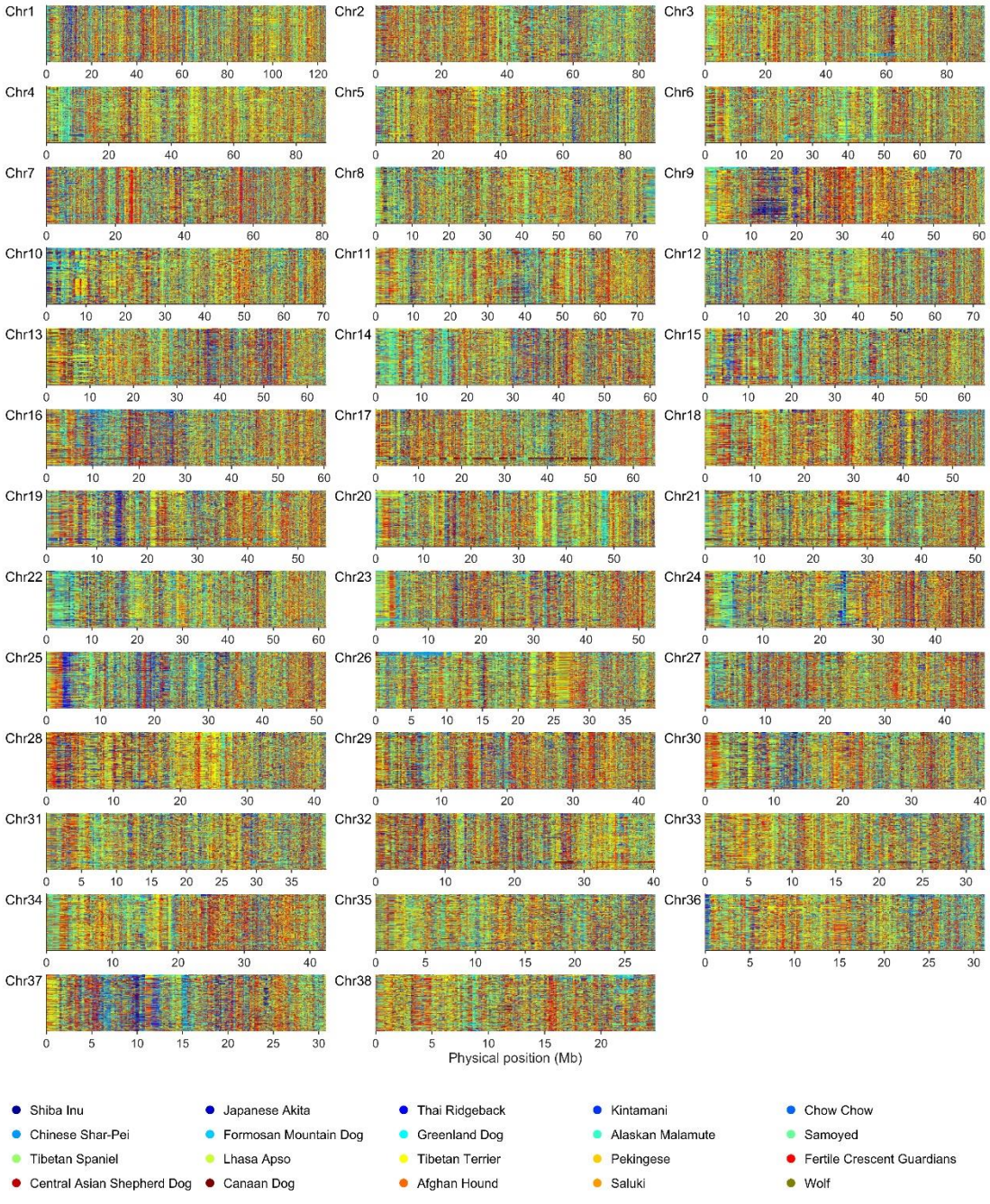
Supplementary Fig. 11 | Maximum-likelihood tree of domestic dogs and gray wolves based on Dog10K genomic data. Branch lengths are proportional to genetic distance.



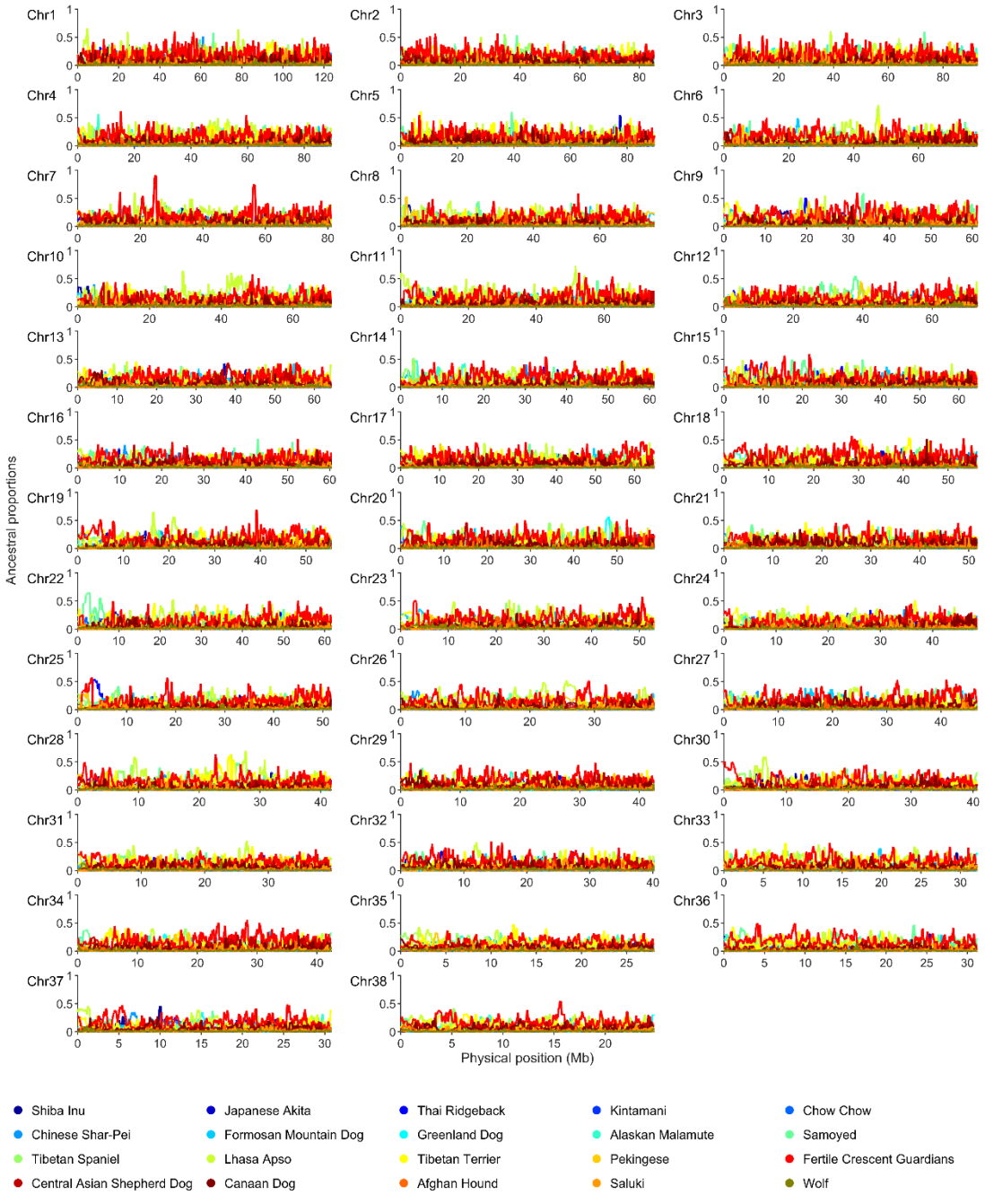
Supplementary Fig. 12 | Principal-component analysis of the 37 dog breeds identified as deep-branching in the maximum-likelihood tree. Each point represents an individual. Percent variance explained by PC1 and PC2 is shown on the axes.



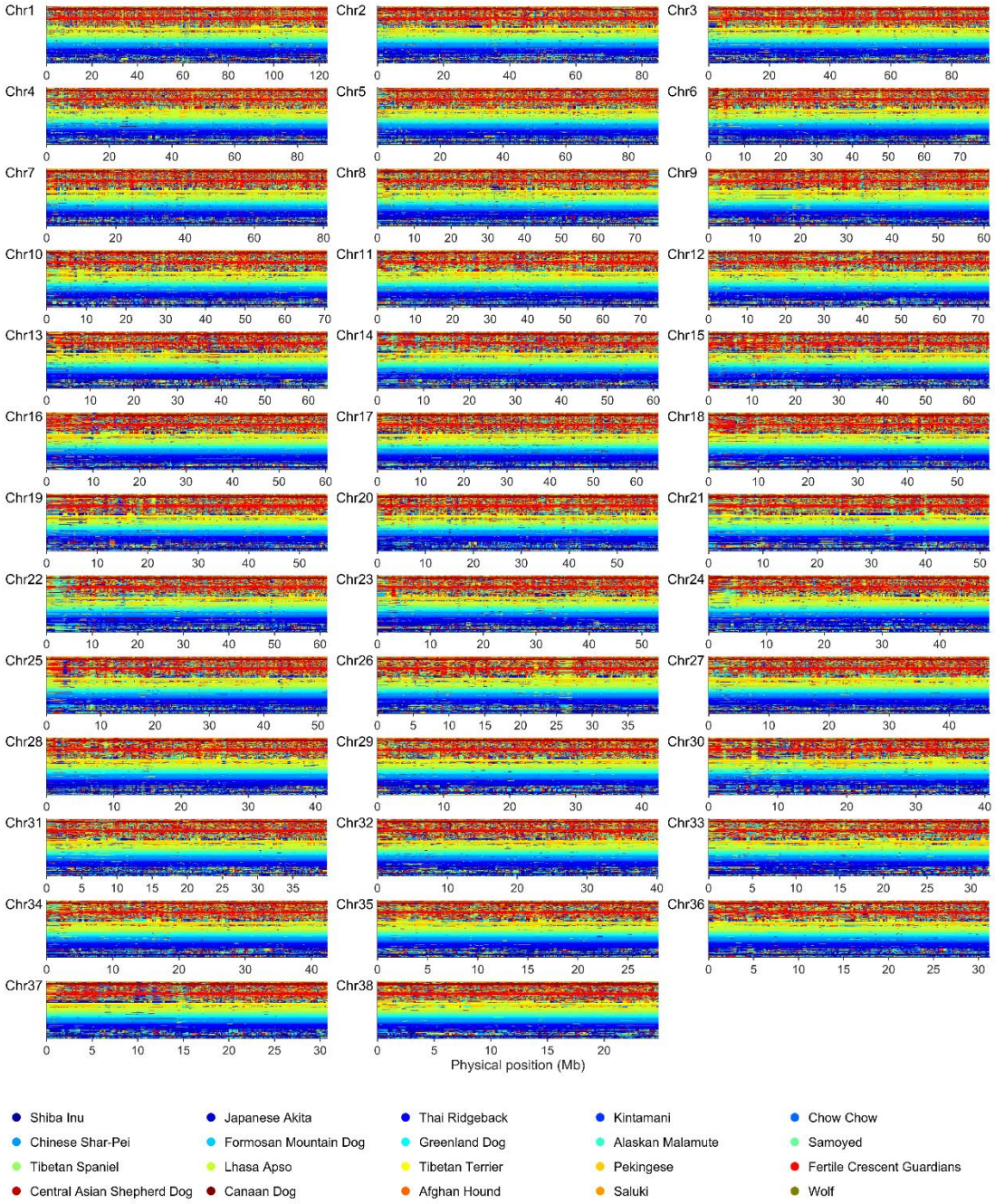
Supplementary Fig. 13 | Distributions of mean local-ancestry proportions across modern breeds. Distributions are shown separately for each of the 20 proxy ancestry components, together with fitted beta distributions. Basal breeds and mixed/other individuals were excluded.



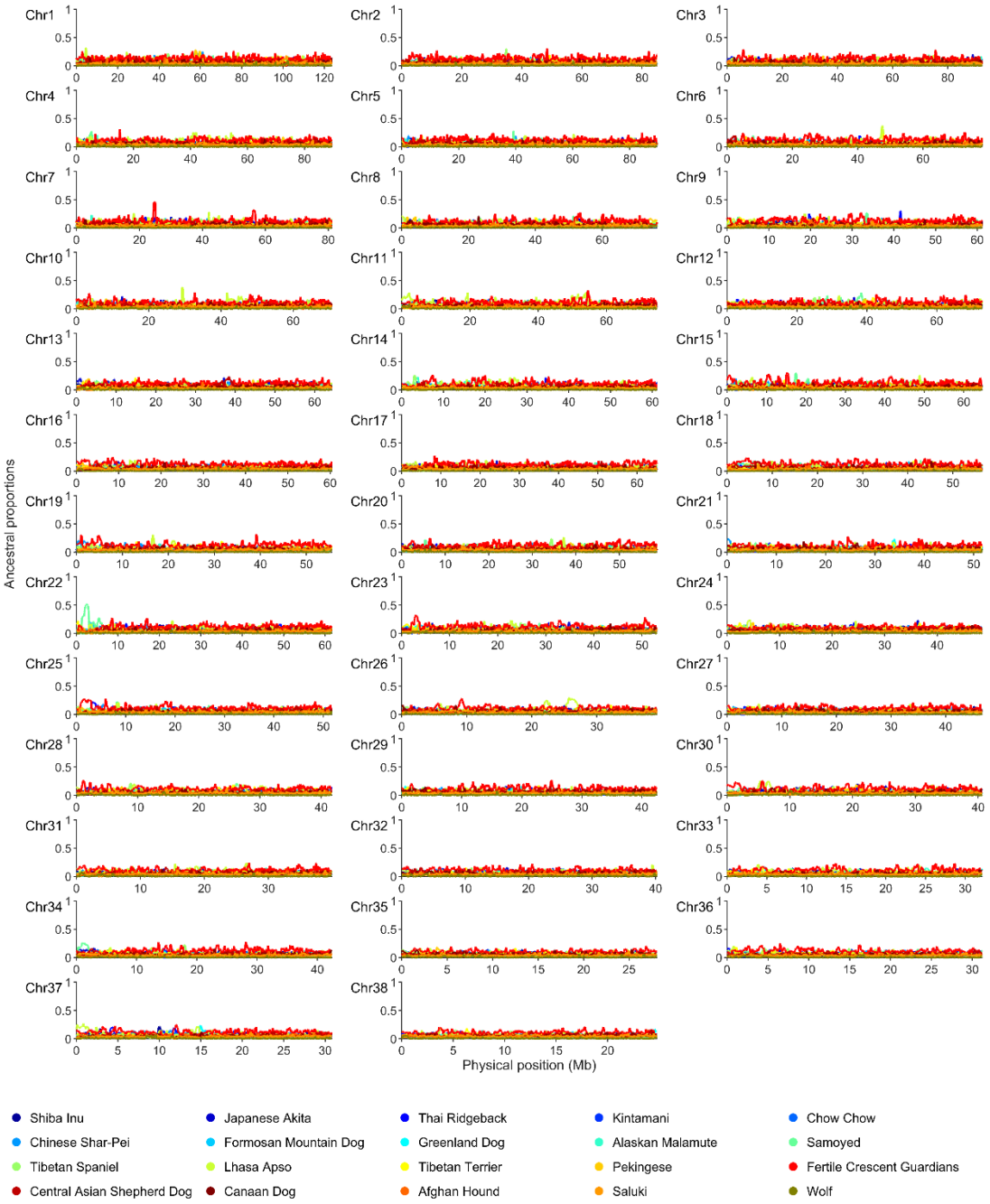
Supplementary Fig. 14 | High-resolution local-ancestry map for 1,437 non-basal modern-breed individuals from 284 breeds. Basal breeds and mixed/other individuals were excluded. Rows correspond to phased haplotypes and columns to SNP bins across the 38 canine autosomes; colors indicate inferred proxy ancestry components.



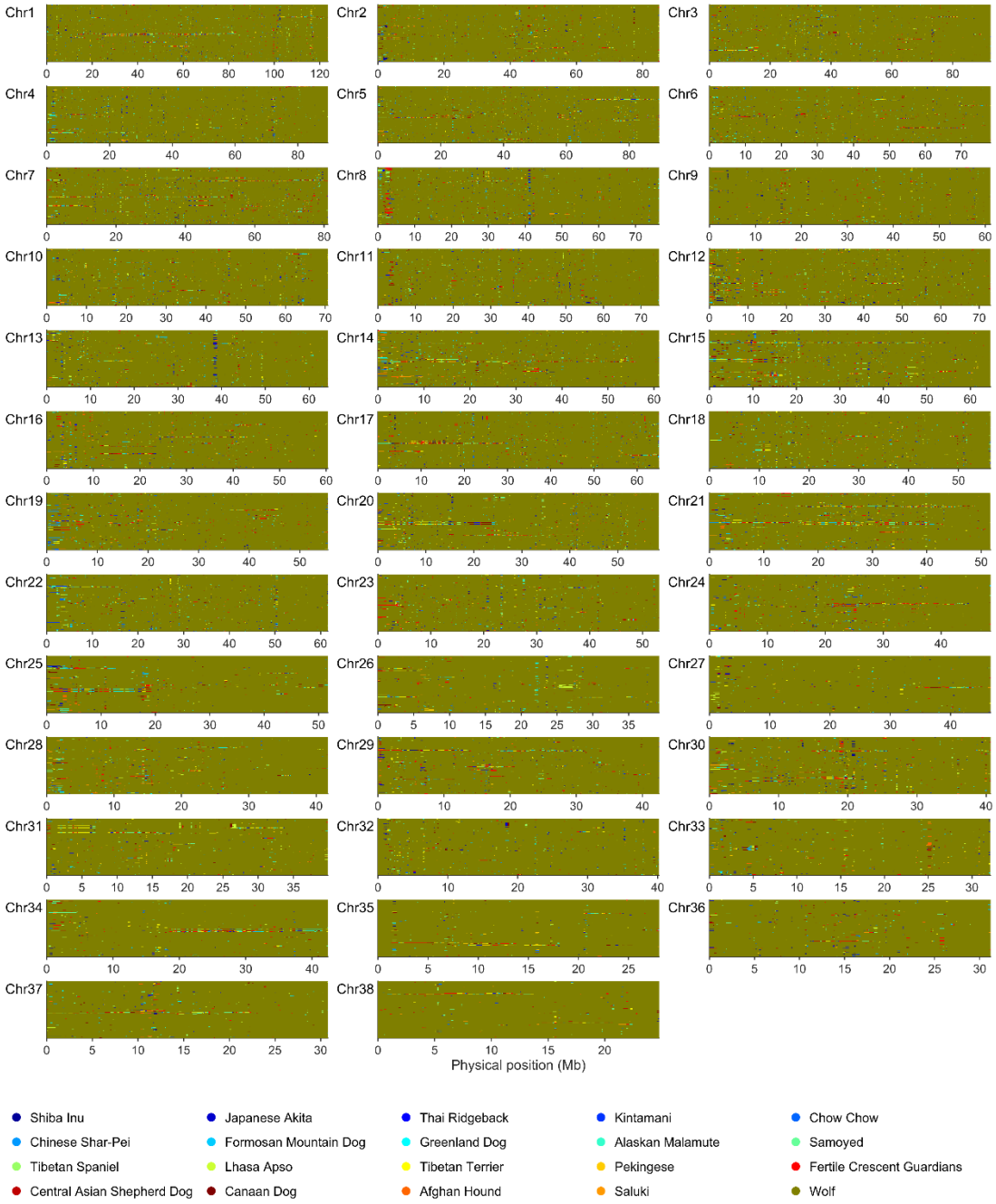
Supplementary Fig. 15 | Mean local-ancestry proportions assigned to each of the 20 proxy ancestry components. Analyses are shown for 1,437 non-basal modern-breed individuals from 284 breeds, excluding basal breeds and mixed/other individuals.



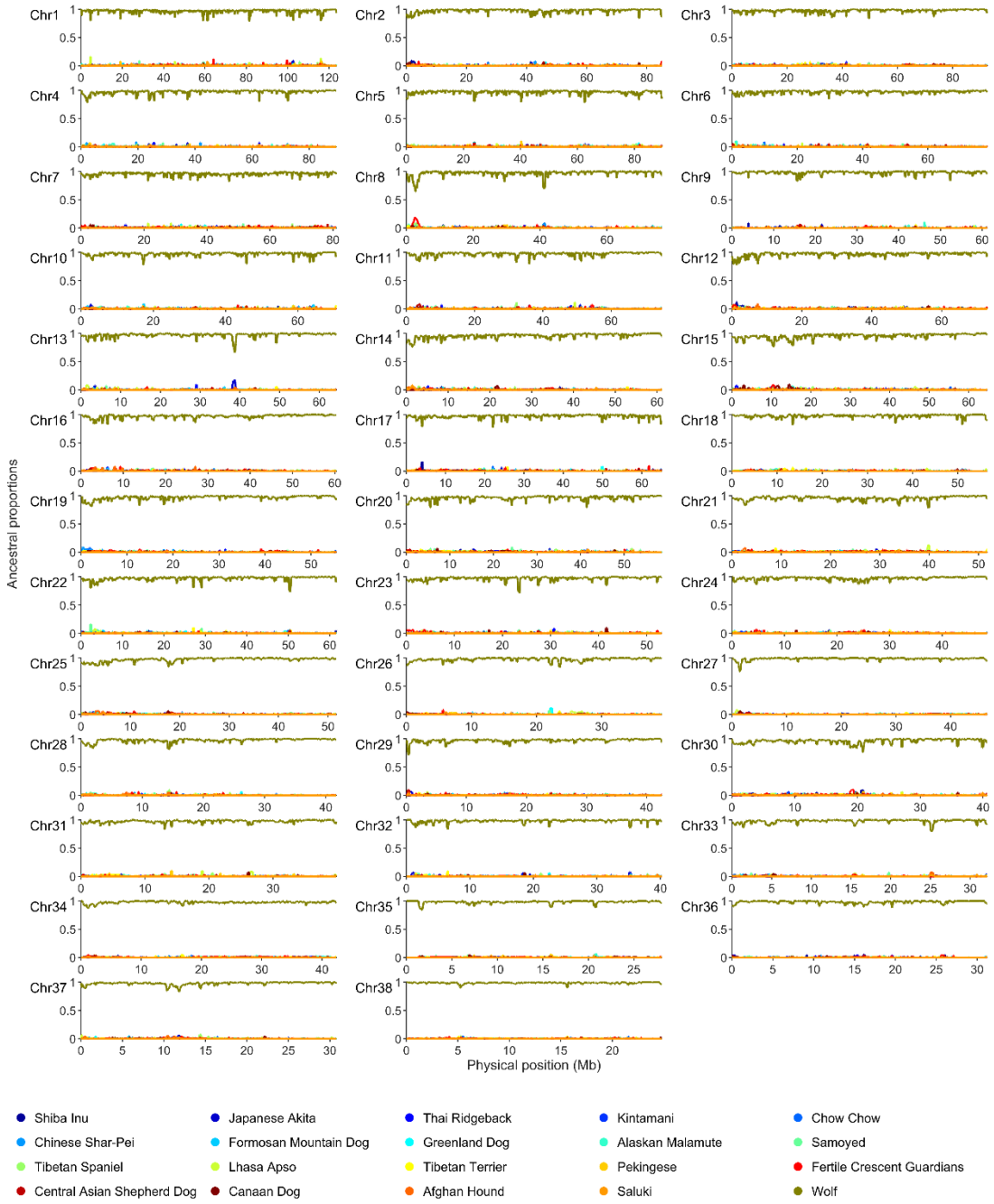
Supplementary Fig. 16 | High-resolution local-ancestry map for basal dog breeds. Rows correspond to phased haplotypes and columns to SNP bins across the 38 canine autosomes; colors indicate inferred proxy ancestry components.



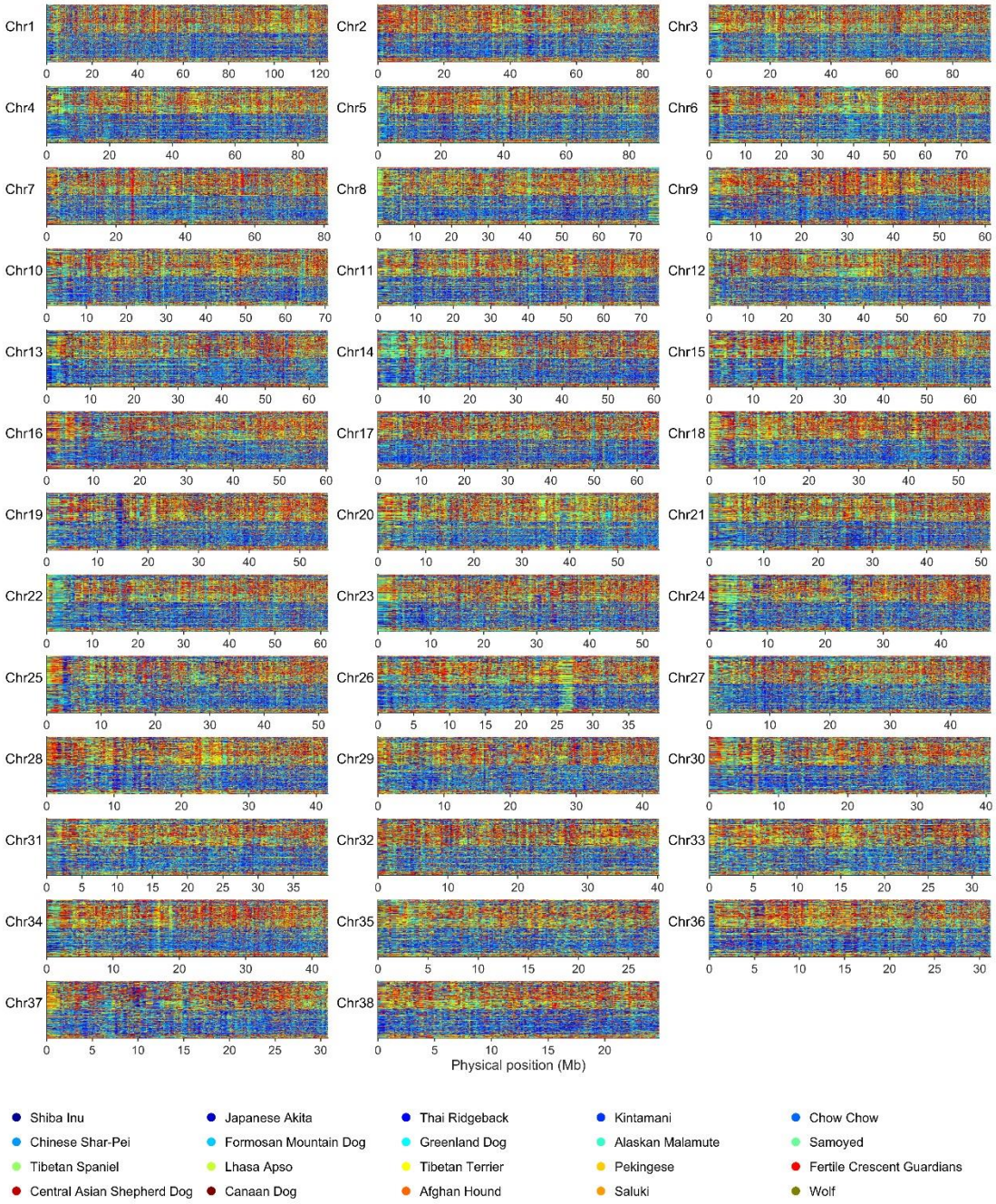
Supplementary Fig. 17 | Mean local-ancestry proportions across basal dog breeds assigned to each of the 20 proxy ancestry components.



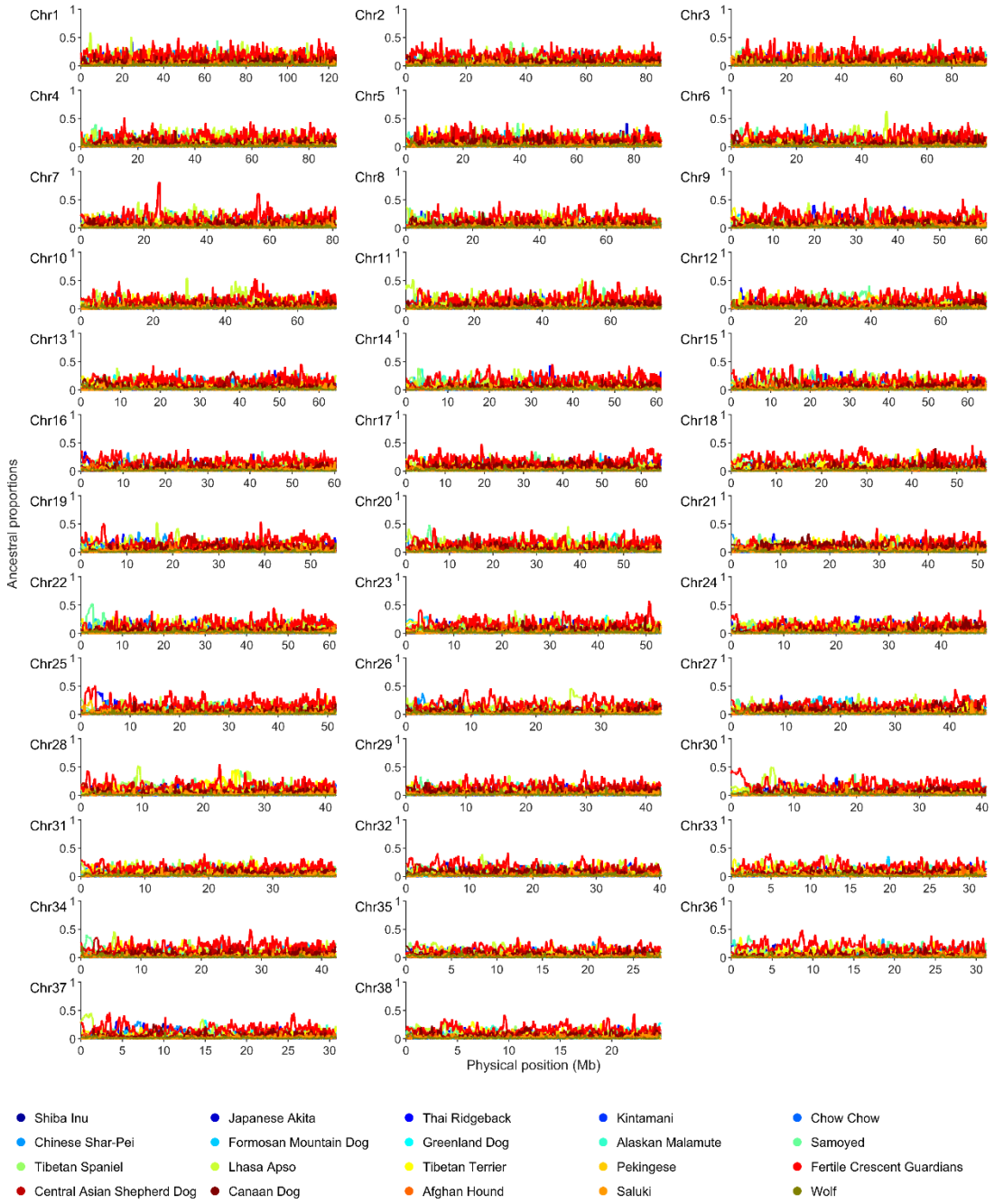
Supplementary Fig. 18 | High-resolution local-ancestry map for gray wolves. Rows correspond to phased haplotypes and columns to SNP bins across the 38 canine autosomes; colors indicate inferred proxy ancestry components.



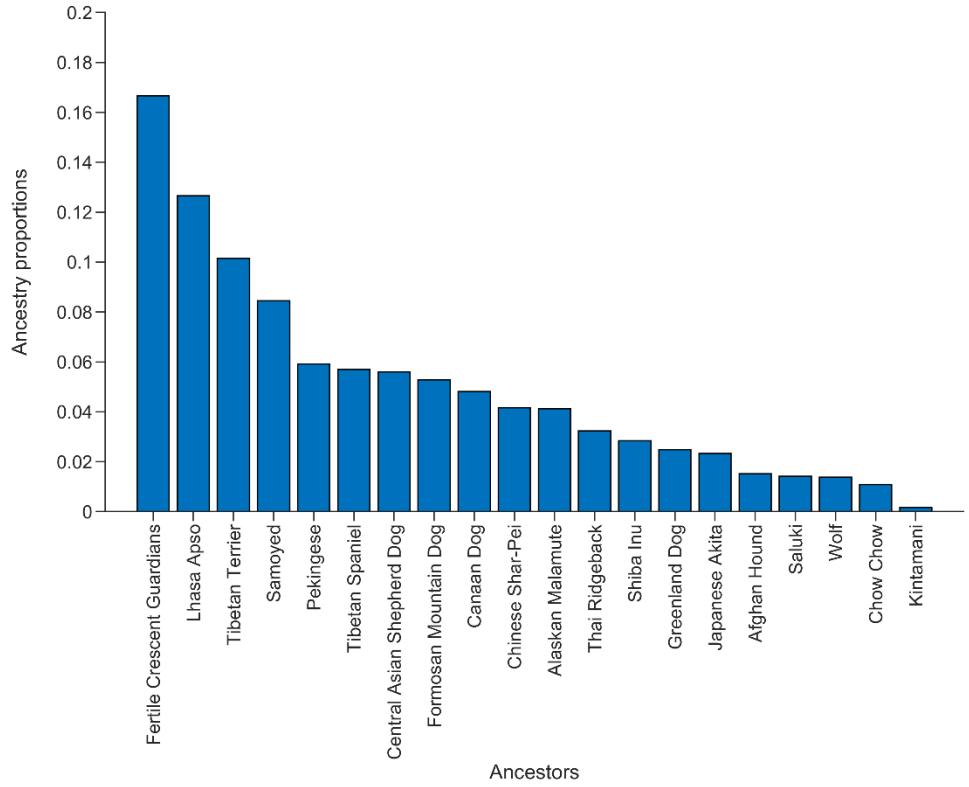
Supplementary Fig. 19 | Mean local-ancestry proportions in gray wolves assigned to each of the 20 proxy ancestry components.



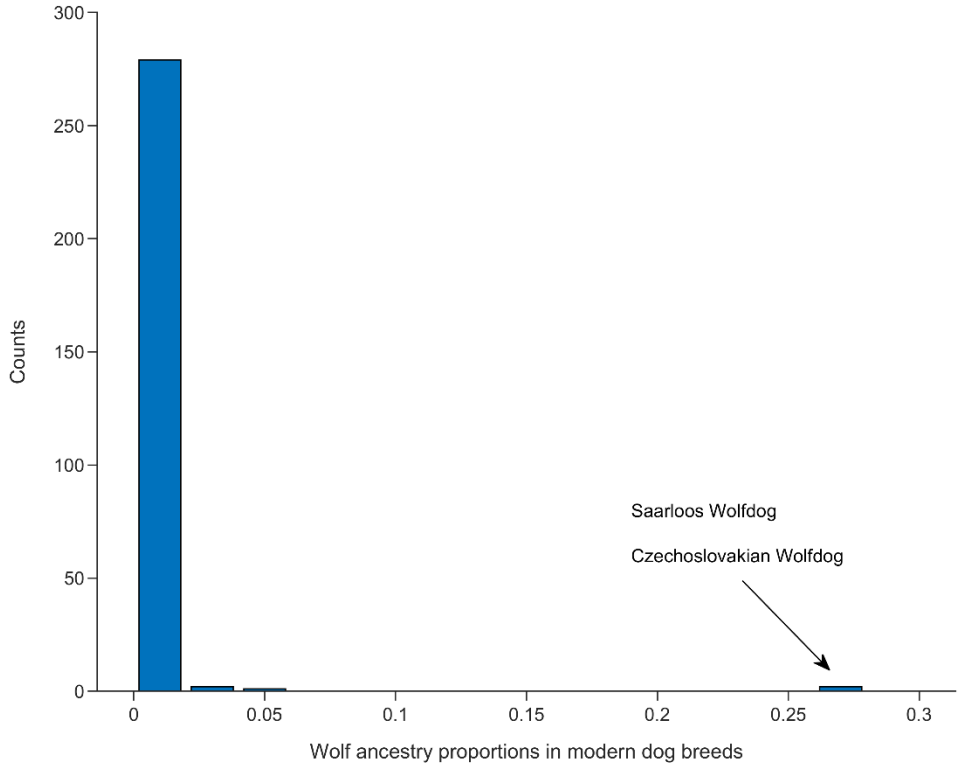
Supplementary Fig. 20 | High-resolution local-ancestry map for village dogs. Rows correspond to phased haplotypes and columns to SNP bins across the 38 canine autosomes; colors indicate inferred proxy ancestry components.



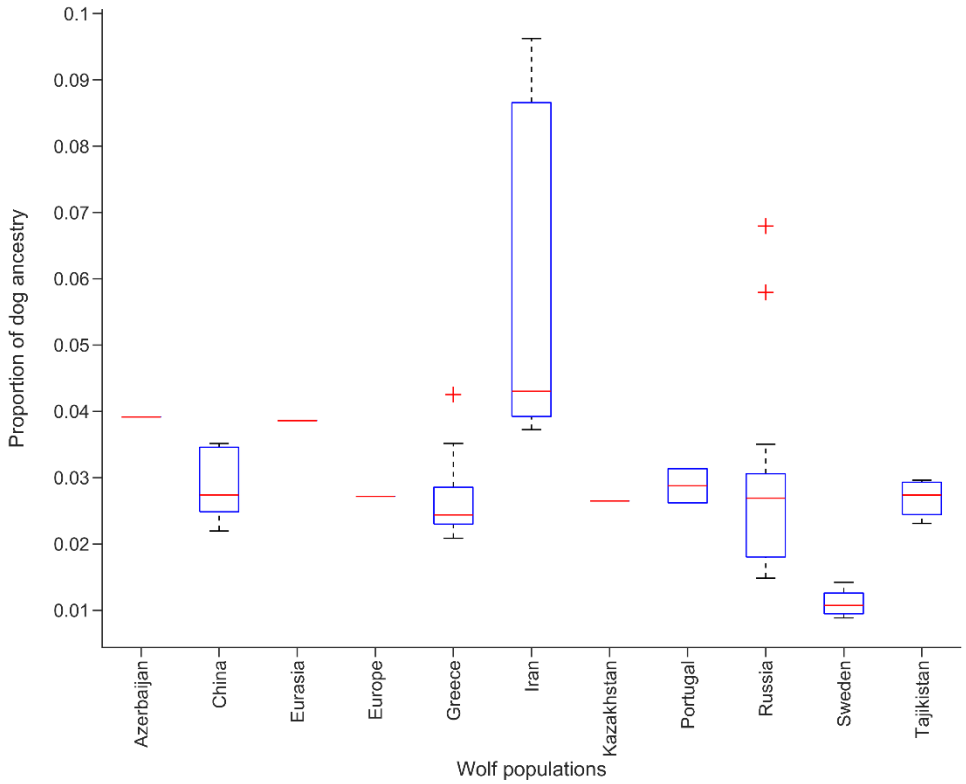
Supplementary Fig. 21 | Mean local-ancestry proportions across village dogs assigned to each of the 20 proxy ancestry components.



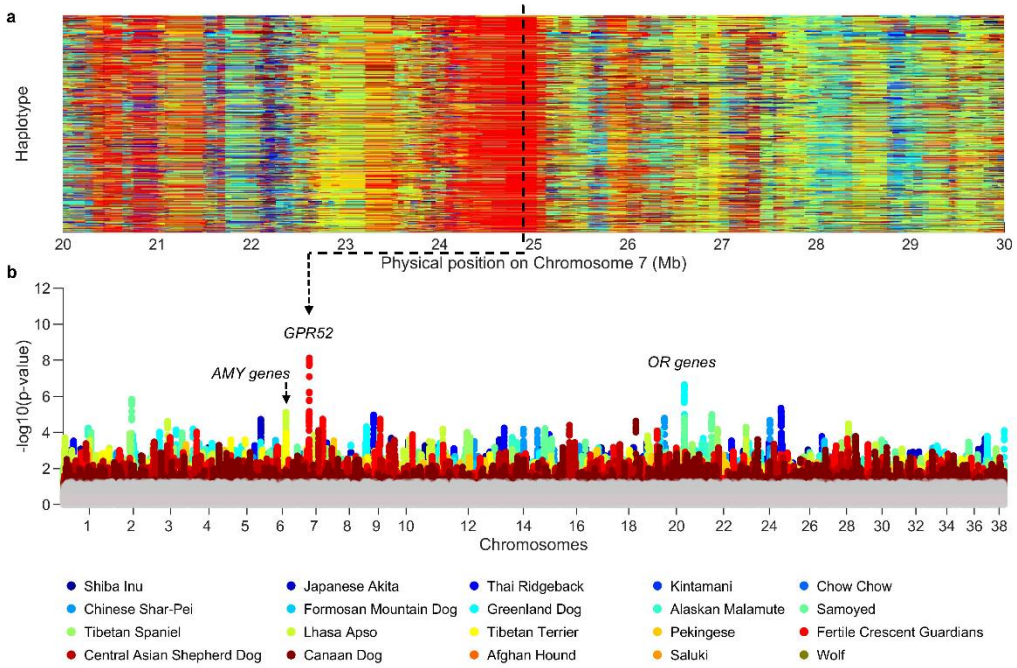
Supplementary Fig. 22 | Genome-wide local-ancestry proportions in non-basal modern breeds. Mean genome-wide contributions of the 20 proxy ancestry components were 0.1666, 0.1266, 0.1015, 0.0845, 0.0591, 0.0571, 0.0560, 0.0528, 0.0481, 0.0417, 0.0412, 0.0324, 0.0284, 0.0249, 0.0233, 0.0153, 0.0142, 0.0137, 0.0109 and 0.0017.



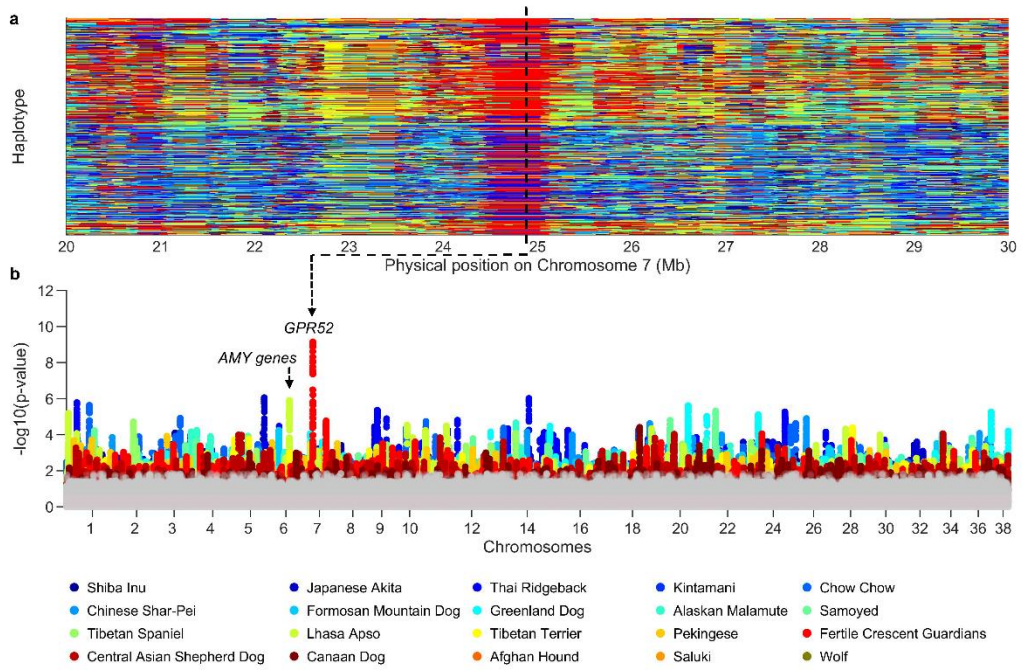
Supplementary Fig. 23 | Distribution of wolf-ancestry proportions across 284 modern breeds. Saarloos Wolfdog and Czechoslovakian Wolfdog retain 27.22% and 26.98% wolf ancestry, respectively. Both breeds derive from 20th-century cross-breeding programs that involved German Shepherds and Eurasian wolves, accounting for their elevated levels of wolf ancestry.



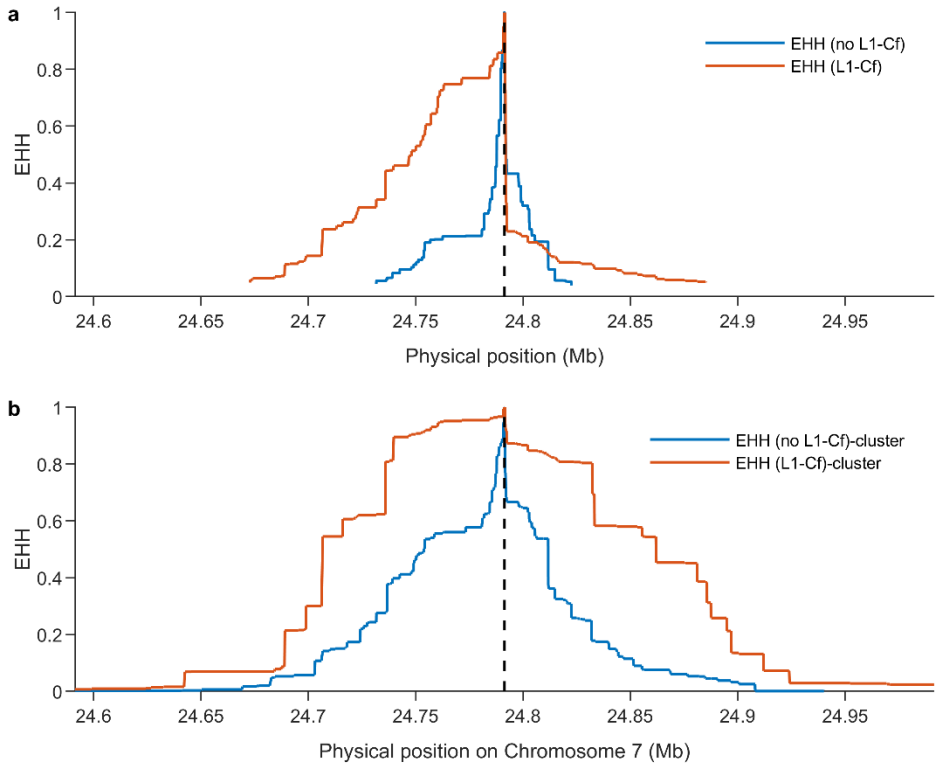
Supplementary Fig. 24 | Dog-ancestry proportions in 11 wolf populations from the Dog10K project. Mean dog-ancestry proportions for the 11 populations are 0.0392, 0.0286, 0.0386, 0.0272, 0.0267, 0.0576, 0.0265, 0.0288, 0.0298, 0.0111 and 0.0269, respectively.



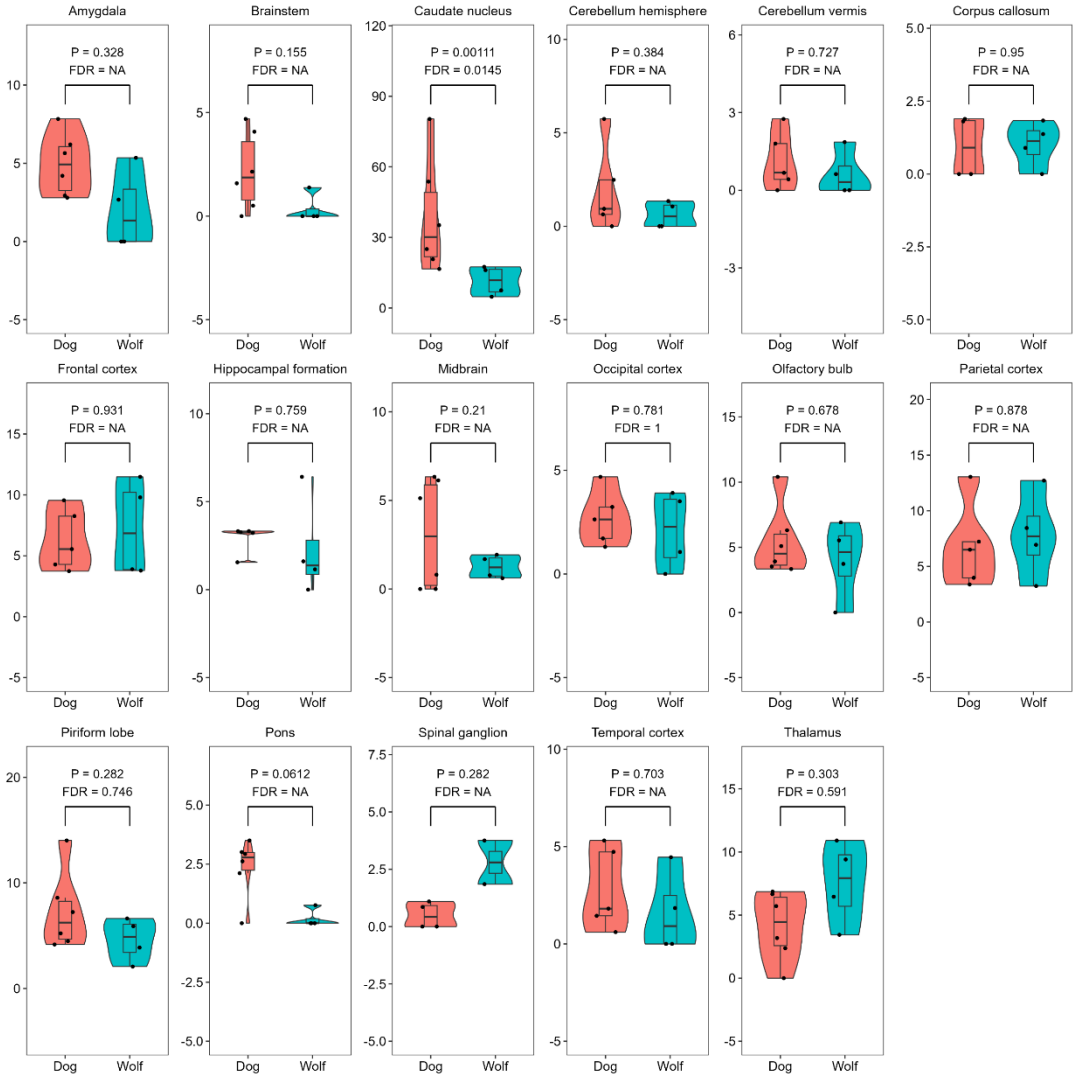
Supplementary Fig. 25 | GAMA ancestry-enrichment scan using a masked basal-lineage reference panel. a, Local ancestry reconstruction across chromosome 7:20–30 Mb in modern dog breed haplotypes using a basal-lineage reference panel after masking genomic segments inferred to derive from recent modern-breed introgression. Rows represent phased haplotypes and columns represent consecutive genomic windows; colors denote inferred proxy ancestry components. The *GPR52* region contains a prominent block of shared local ancestry. **b,** Genome-wide scan for ancestry-enriched regions using the same masked reference panel. Points are colored by inferred proxy ancestry component, and the y axis shows $-\log_{10}$(P value) for ancestry enrichment. Components with genome-wide mean ancestry proportion >0.02 are shown individually; windows with local ancestry proportion <0.1 are shown in gray. The strongest genome-wide signal maps to the *GPR52* locus on chromosome 7. Recovery of the same leading signal after masking supports the robustness of the *GPR52* ancestry-enrichment peak observed in the unmasked main analysis and argues against recent modern-breed introgression within the basal reference panel as the source of the signal.



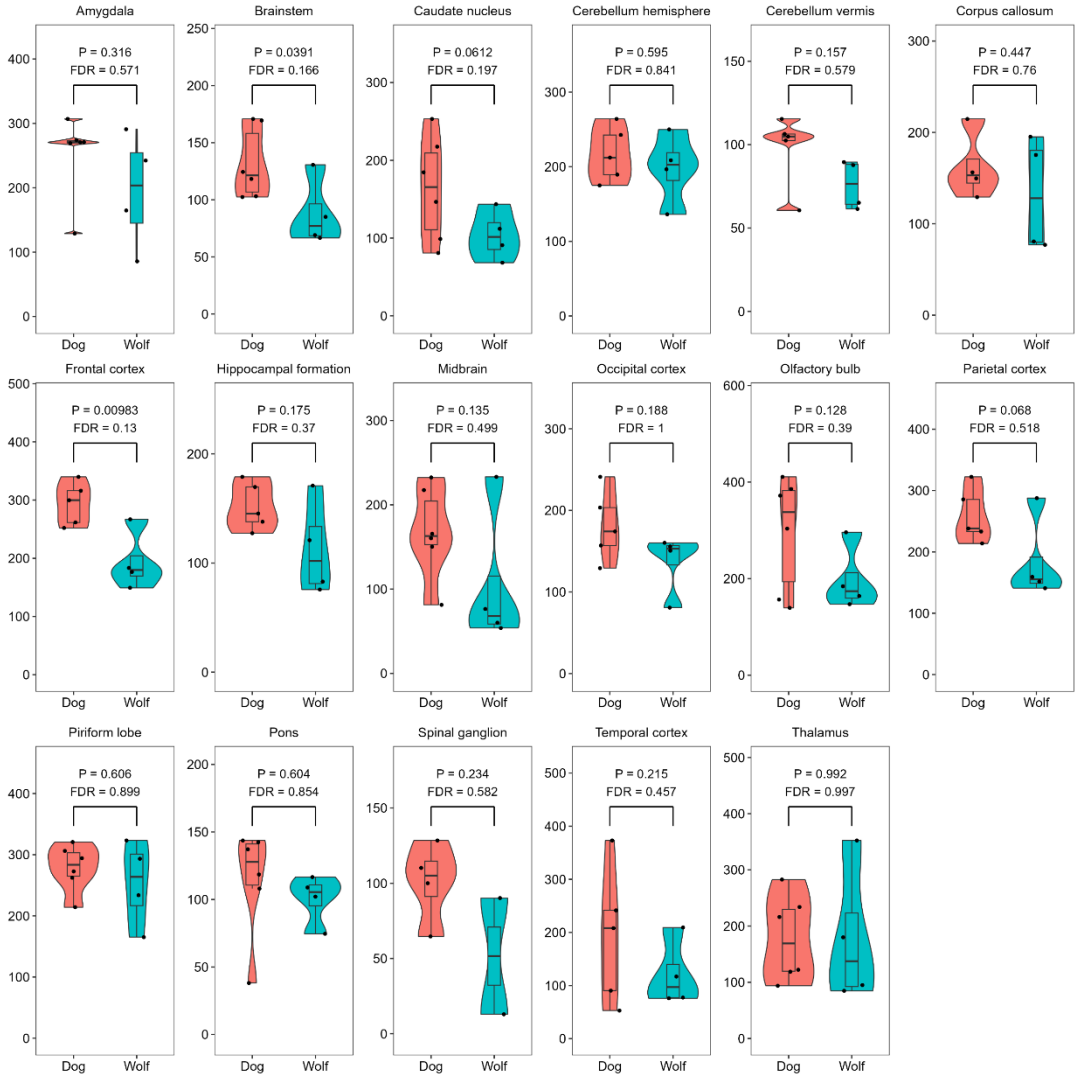
Supplementary Fig. 26 | Local ancestry enrichment at the *GPR52* locus in free-breeding village dogs. a, Local ancestry map across chromosome 7:20–30 Mb in 281 free-breeding village dogs from 26 populations. Rows represent phased haplotypes and columns represent genomic windows; colors denote inferred proxy ancestry components. The dashed line marks the ancestry-enriched interval near *GPR52*. **b,** Genome-wide scan for ancestry-enriched segments in village dog genomes. Points represent genomic windows and are colored by inferred proxy ancestry component. Components with genome-wide mean ancestry proportion >0.02 are shown individually; windows with local ancestry proportion <0.1 are shown in gray. A concordant ancestry-enrichment signal occurs near *GPR52* on chromosome 7 and is enriched for the Fertile Crescent Guardian–associated component, supporting the presence of the *GPR52*-linked ancestry signal outside closed modern breeds.



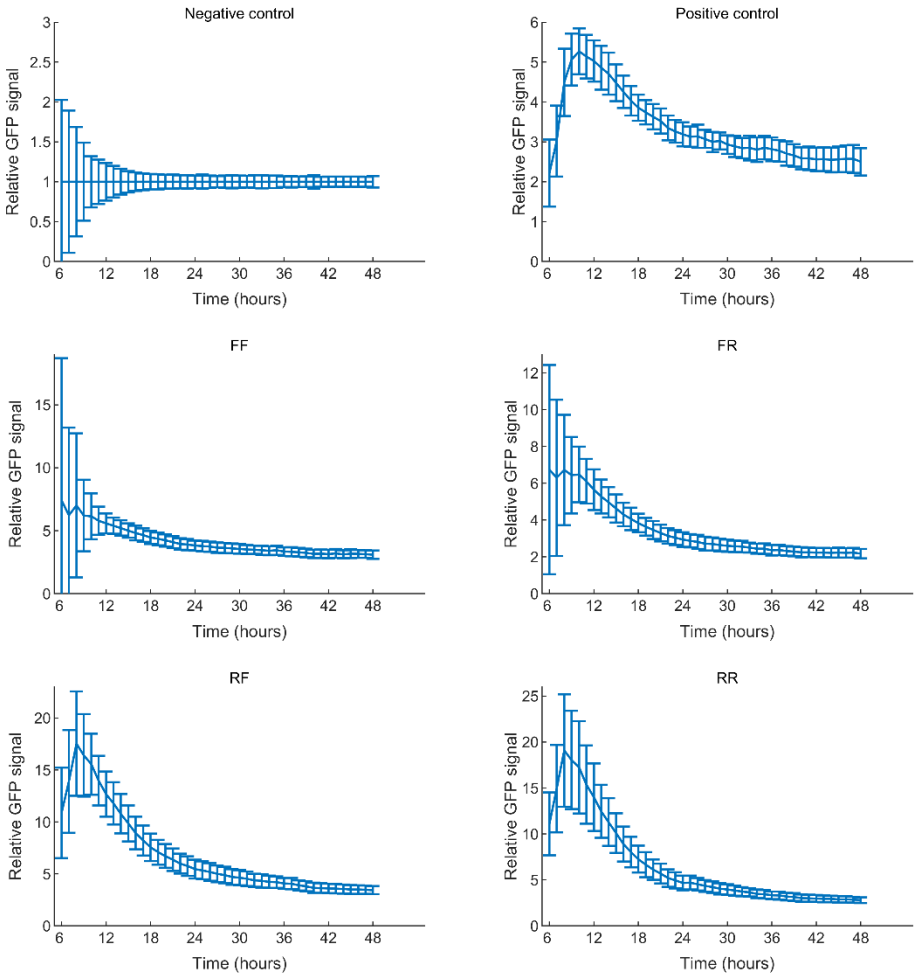
Supplementary Fig. 27 | Extended haplotype homozygosity (EHH) decay at the *GPR52*-upstream L1-Cf insertion. a, Traditional EHH curves. **b,** HaploSweep EHH curves. Red curves plot EHH for 81 haplotypes carrying the L1-Cf insertion; blue curves plot EHH for 87 haplotypes lacking the insertion. Horizontal axis: physical position along Chromosome 7 (Mb). Vertical axis: EHH. The vertical dashed line marks the focal insertion site.



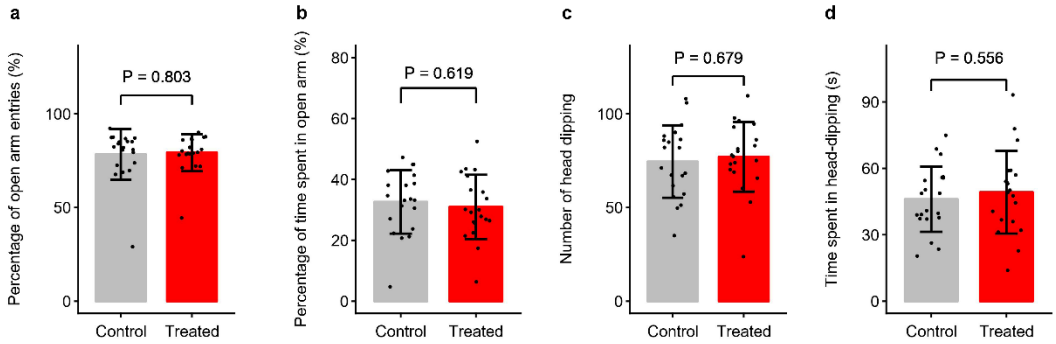
Supplementary Fig. 28 | *GPR52* expression across canine CNS tissues. Red, individuals harboring the L1-Cf insertion; blue, non-carrier or presumed non-carrier individuals. Expression values were normalized with DESeq2 and significance was assessed by the Wald test. False discovery rate was controlled using the Benjamini–Hochberg procedure. Meninges, neurohypophysis and spinal ganglion were excluded because of insufficient sample availability.



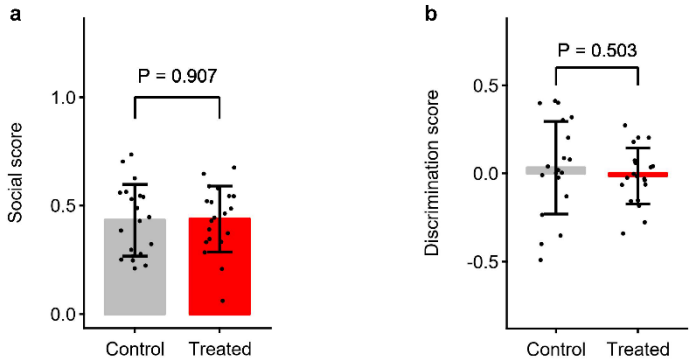
Supplementary Fig. 29 | *RABGAP1L* expression across canine CNS tissues. Red, individuals harboring the L1-Cf insertion; blue, non-carrier or presumed non-carrier individuals. Expression values were normalized with DESeq2 and significance was assessed by the Wald test. False discovery rate was controlled using the Benjamini–Hochberg procedure. Meninges, neurohypophysis and spinal ganglion were excluded because of insufficient sample availability.



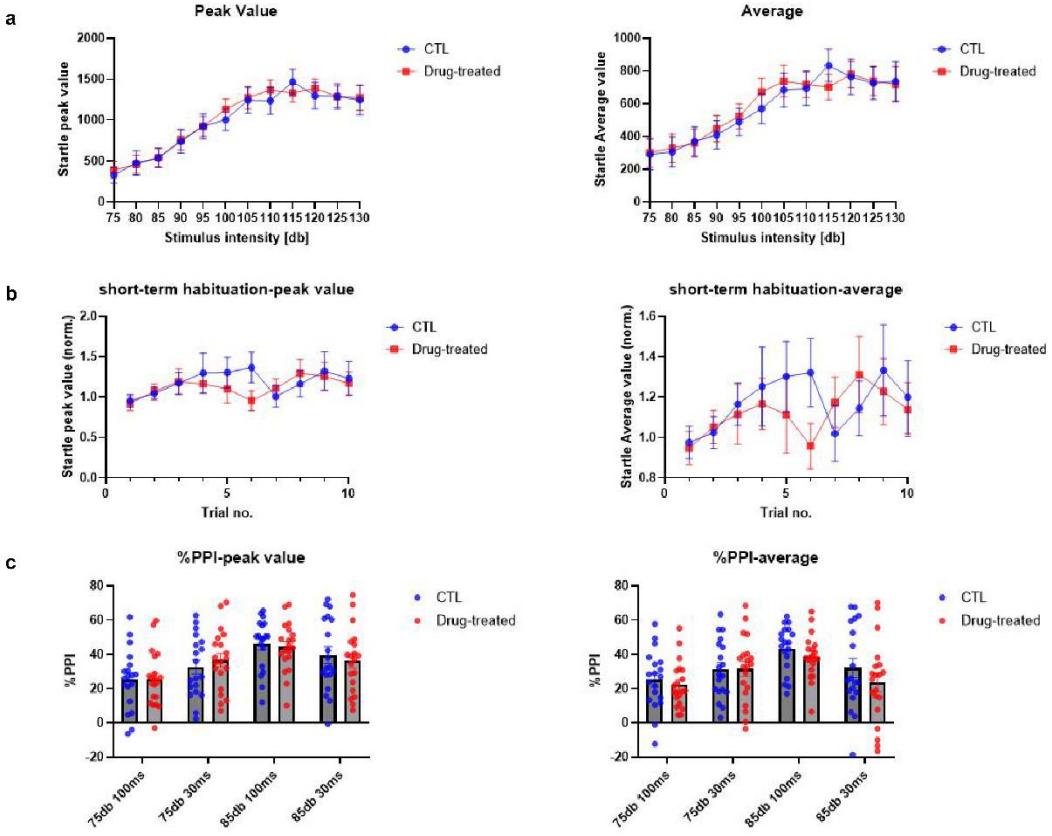
Supplementary Fig. 30 | Time course of relative GFP signal (6–48 h) from reporter constructs testing L1-Cf regulatory activity. FF, forward L1-Cf inserted upstream of the CMV promoter; FR, forward L1-Cf inserted downstream of the GFP coding sequence; RF, reverse L1-Cf inserted upstream of the CMV promoter; RR, reverse L1-Cf inserted downstream of the GFP coding sequence. PC, positive control with the CMV enhancer placed upstream of the CMV promoter; NC, negative control lacking the CMV enhancer. Each point represents one-hour interval measurements of GFP signal intensity.



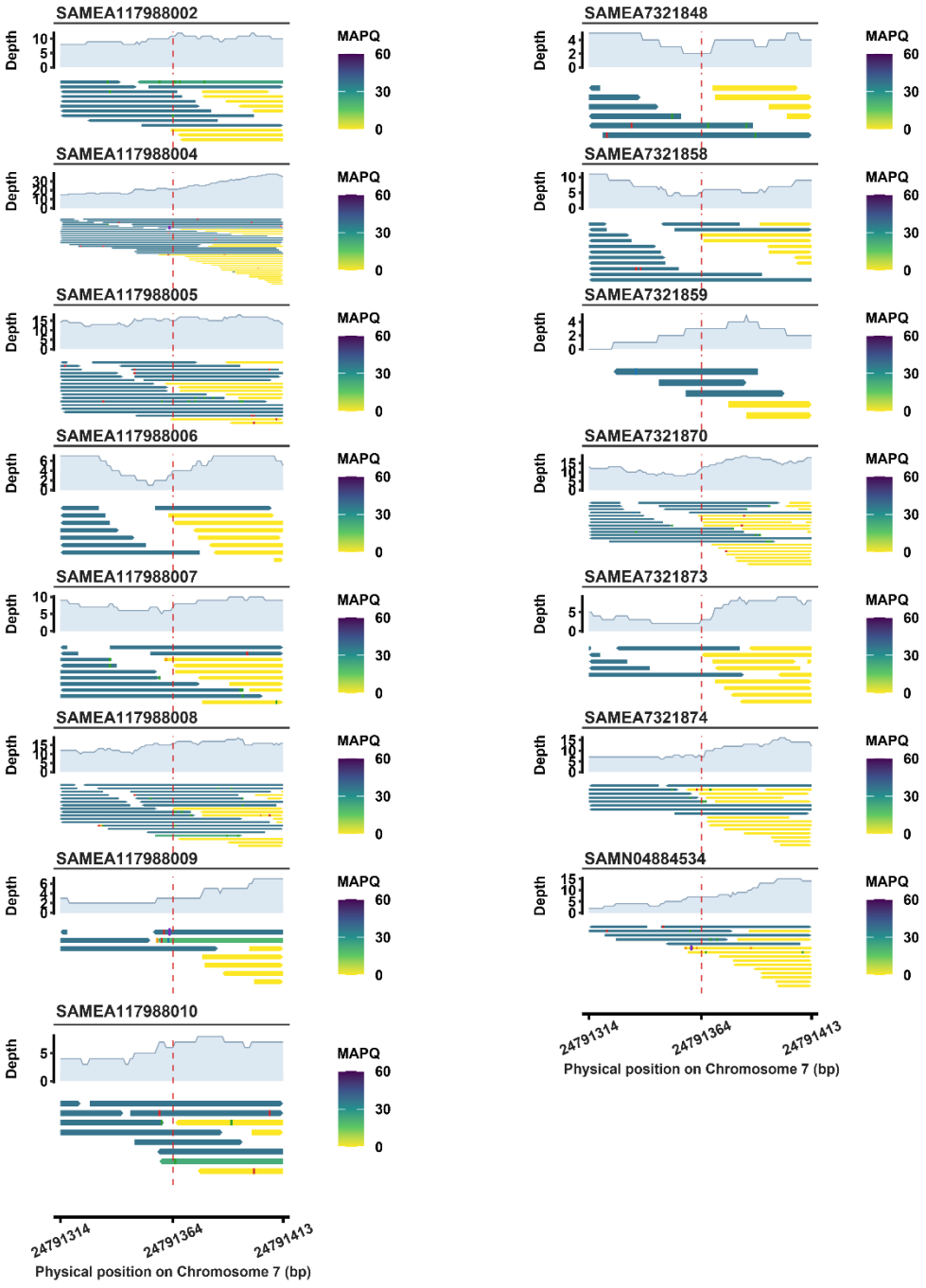
Supplementary Fig. 31 | Elevated-plus-maze performance after *GPR52* agonist treatment. C57BL/6J male mice (10 weeks old) received HTL0041178 in drinking water for 3 days or standard drinking water (controls), then were tested in the elevated-plus-maze (EPM) (n = 20 per group). Panels show: **a,** percentage of open-arm entries (open entries / total entries ×100); **b,** percentage of time spent in open arms (time in open arms / total test time ×100); **c,** number of head-dips; **d,** cumulative duration of head-dipping behavior. Data are presented as mean ± SD. No significant differences were observed between treated and control groups for any EPM metric



Supplementary Fig. 32 | Three-chamber social behavior following *GPR52* agonist treatment. C57BL/6J male mice (10-week-old) received HTL0041178 in drinking water for 3 days (treatment) or standard drinking water (control) before assessment in the three-chamber social assay (n = 20 per group). During testing, mice freely explored three interconnected chambers containing an unfamiliar conspecific confined under a wire cup (social target) and an empty wire cup (object). **a,** Social score, defined as time in the social chamber minus time in the object chamber. **b,** Discrimination index, calculated as (social − object) / (social + object). Data are presented as mean ± SD. No significant differences were detected between treated and control groups for either metric, indicating that *GPR52* agonism did not alter baseline sociability under the conditions tested.

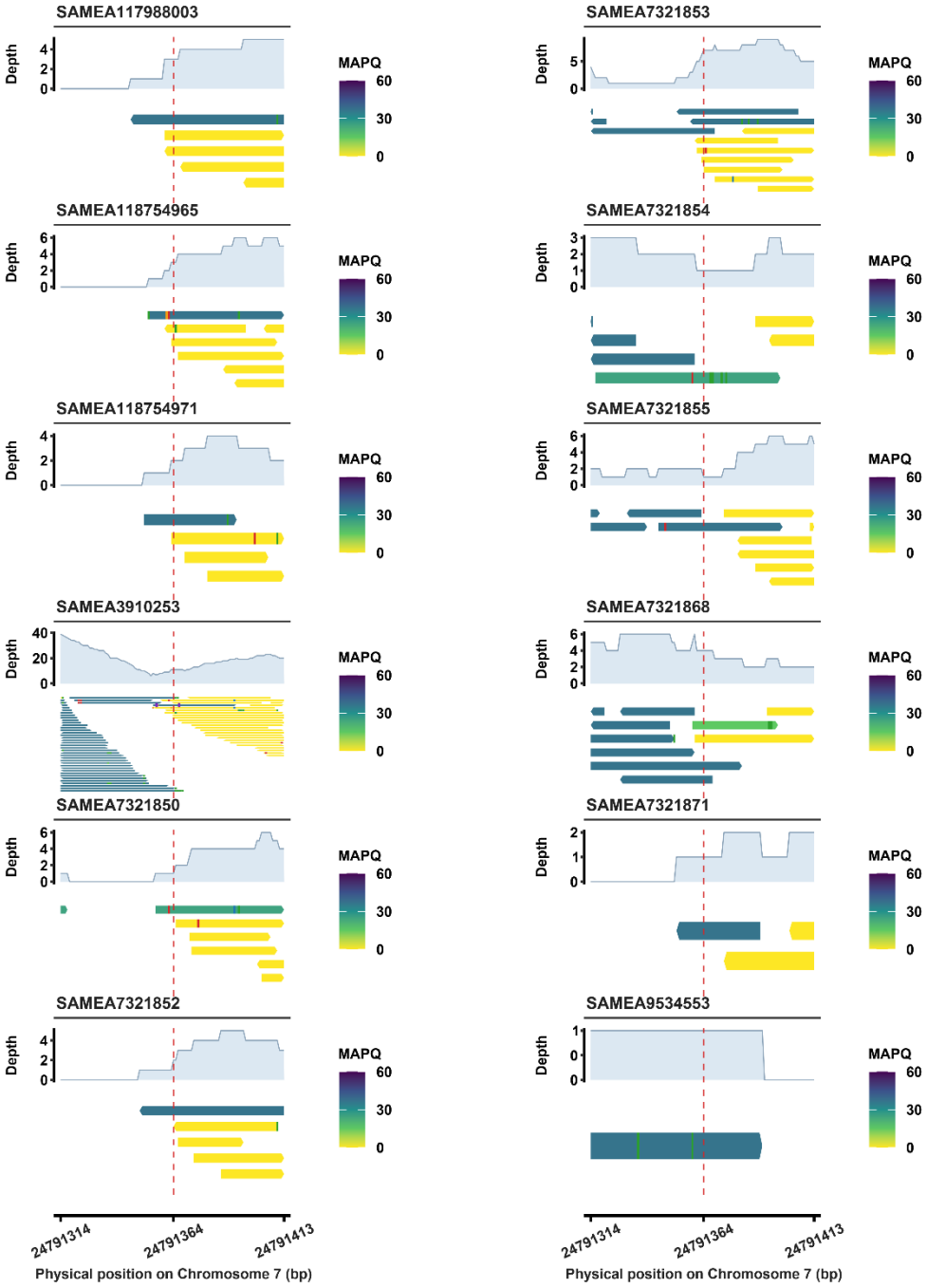


Supplementary Fig. 33 | HTL0041178 treatment does not detectably alter acoustic startle habituation or prepulse inhibition. a, Acoustic startle input/output function measured using peak and average response amplitudes across increasing stimulus intensities. The 110-dB pulse was selected for subsequent testing. **b,** Short-term habituation to repeated 110-dB startle pulses, shown as normalized peak and average responses across trials. **c,** Prepulse inhibition across four prepulse conditions, calculated as percent inhibition relative to the pulse-alone startle response. HTL0041178-treated mice did not differ detectably from controls in acoustic startle habituation or PPI under the tested dosing condition, arguing against broad disruption of acoustic sensorimotor gating. Points represent individual mice; bars show means ± SEM.



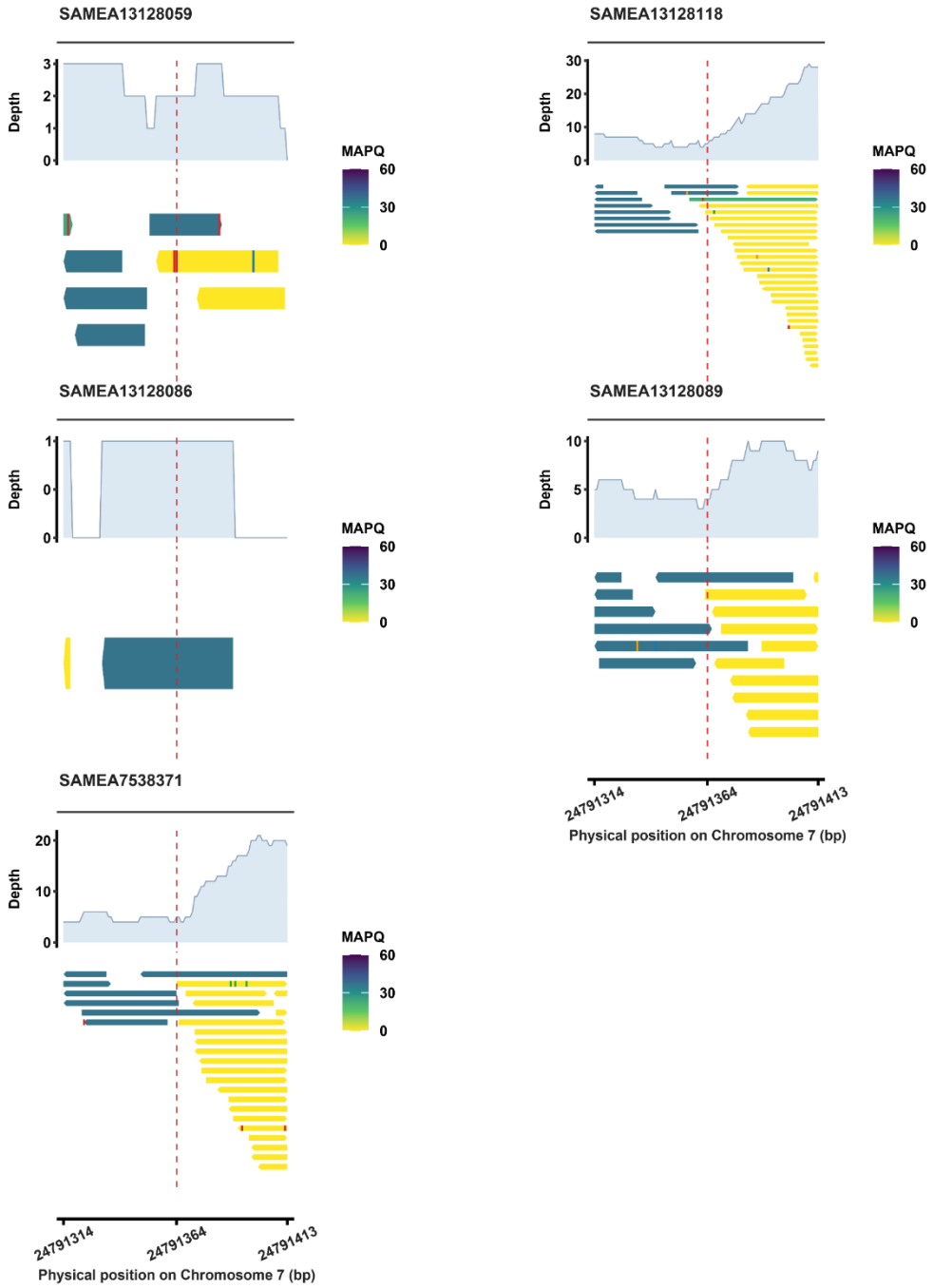
Supplementary Fig. 34 | L1-Cf insertion detection in ancient dog samples supported by multiple high-quality junction reads. Read alignments are shown for 15 of the 73 ancient dog genomes examined, each supported by two or more independent high-quality junction reads. The upper track in each panel shows local sequencing depth, and the lower track shows reads aligned around the 5′ boundary of the L1-Cf insertion on chromosome 7. The red dashed line marks the insertion start site. High-quality junction reads were defined as reads spanning this boundary with anchors of ≥8 bp on both the genome and L1-Cf sides and a mapping quality (MAPQ) ≥25. Read colors indicate MAPQ. Together with the 12 samples supported by a single high-quality junction

read shown in Supplementary Fig. 35, these comprise all 27 ancient dog genomes in which L1-Cf was detected.



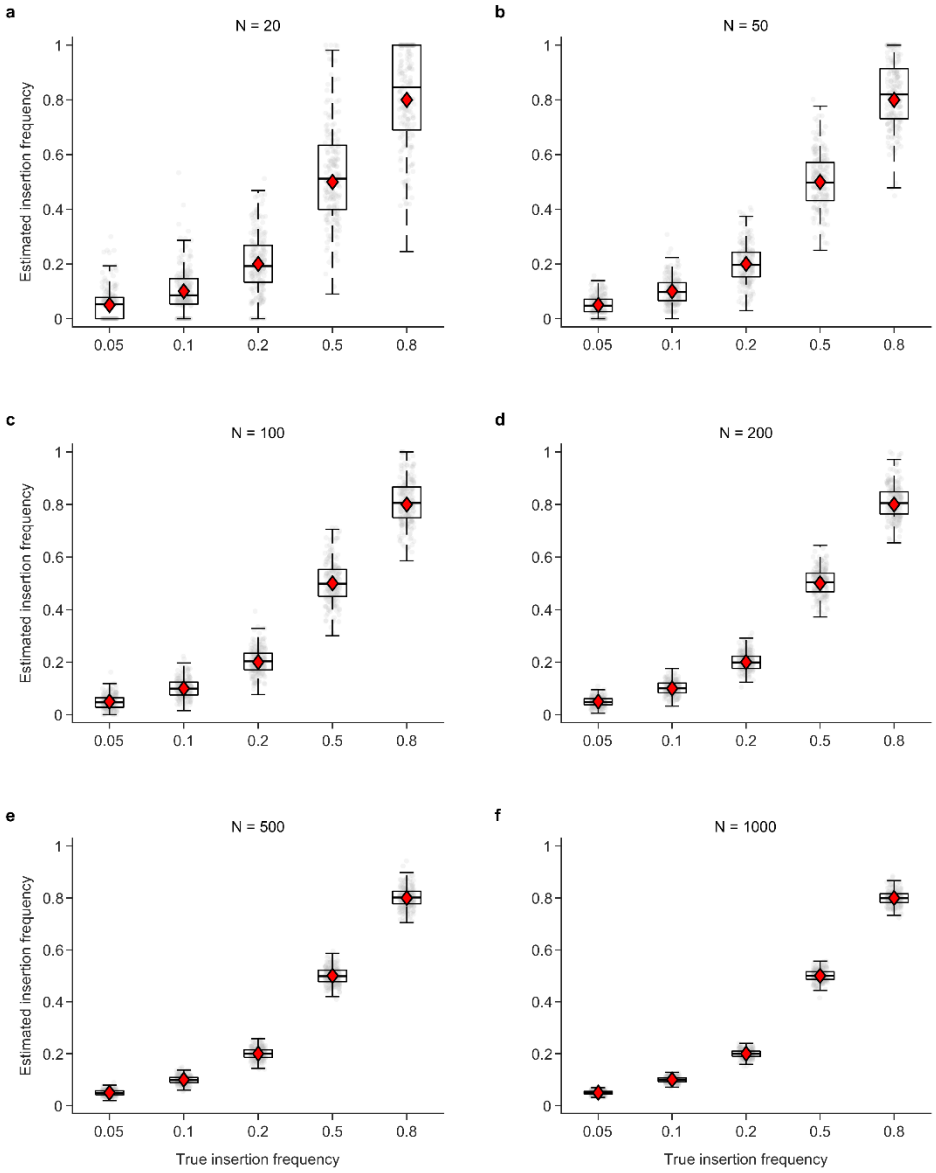
Supplementary Fig. 35 | L1-Cf insertion detection in ancient dog samples supported by a single high-quality junction read. Read alignments are shown for 12 of the 73 ancient dog genomes examined, each supported by exactly one high-quality junction read. The upper track in each panel shows local sequencing depth, and the lower track shows reads aligned around the 5′ boundary of the L1-Cf insertion on chromosome 7. The red dashed line marks the insertion start site. High-quality junction reads were defined as reads spanning this boundary with anchors of ≥8

bp on both the genome and L1-Cf sides and a mapping quality (MAPQ) ≥25. Read colors indicate MAPQ. Together with the 15 samples shown in Supplementary Fig. 34, these comprise all 27 ancient dog genomes in which L1-Cf was detected.



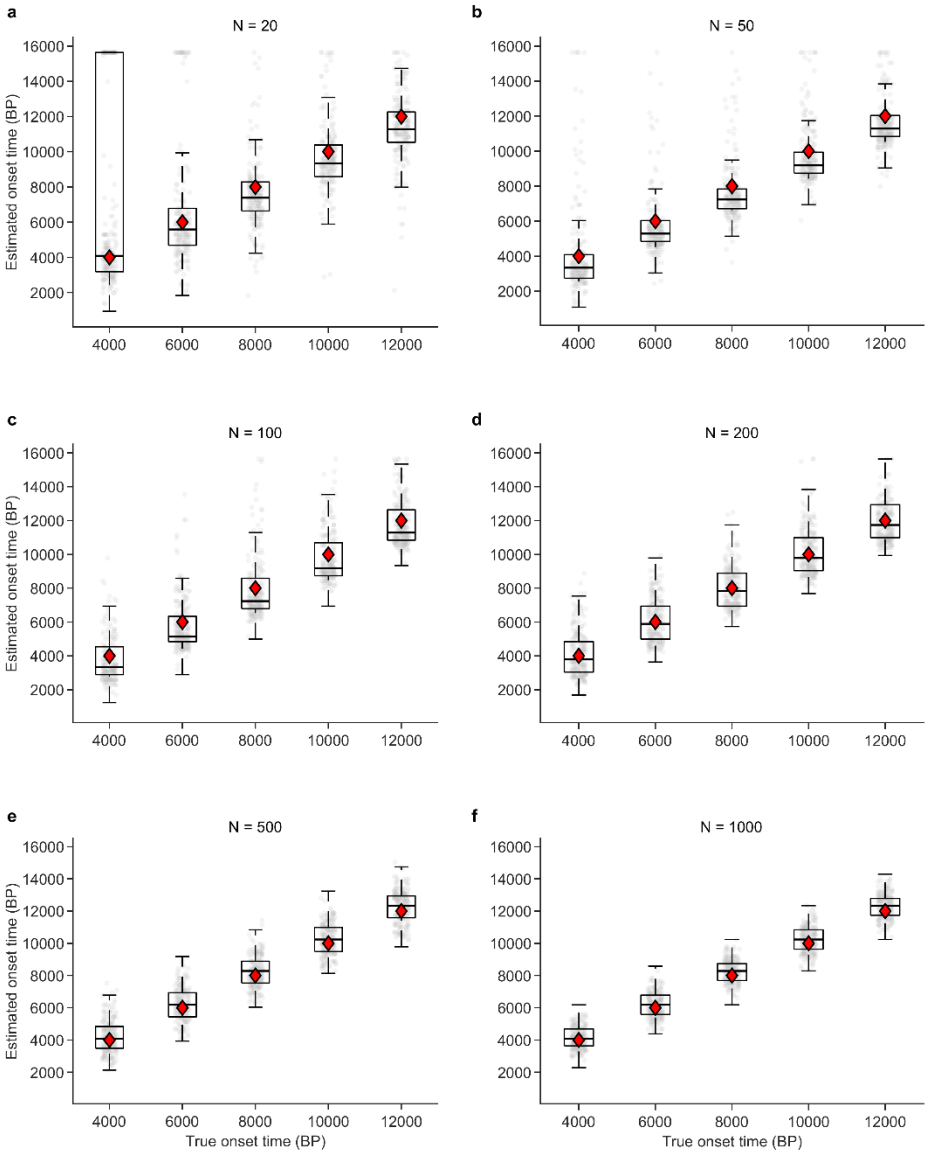
Supplementary Fig. 36 | Junction-read evidence for L1-Cf insertion detection in ancient wolf samples. Read alignments are shown for the 5 ancient wolf samples in which evidence for the L1-Cf insertion was detected among the 70 ancient wolf genomes examined. The upper track in each panel shows local sequencing depth, and the lower track shows reads aligned around the 5′ boundary of the L1-Cf 5′ UTR insertion on chromosome 7. The red dashed line marks the

insertion start site. High-quality junction reads were defined as reads spanning this boundary with both genome-side and L1-Cf-side anchors ≥8 bp and mapping quality MAPQ ≥25. Read colors indicate MAPQ. Of the 5 positive samples, 3 were supported by multiple high-quality junction reads and 2 were supported by only limited junction-read evidence. These results are consistent with the deep but low-frequency presence of L1-Cf in ancient wolves.



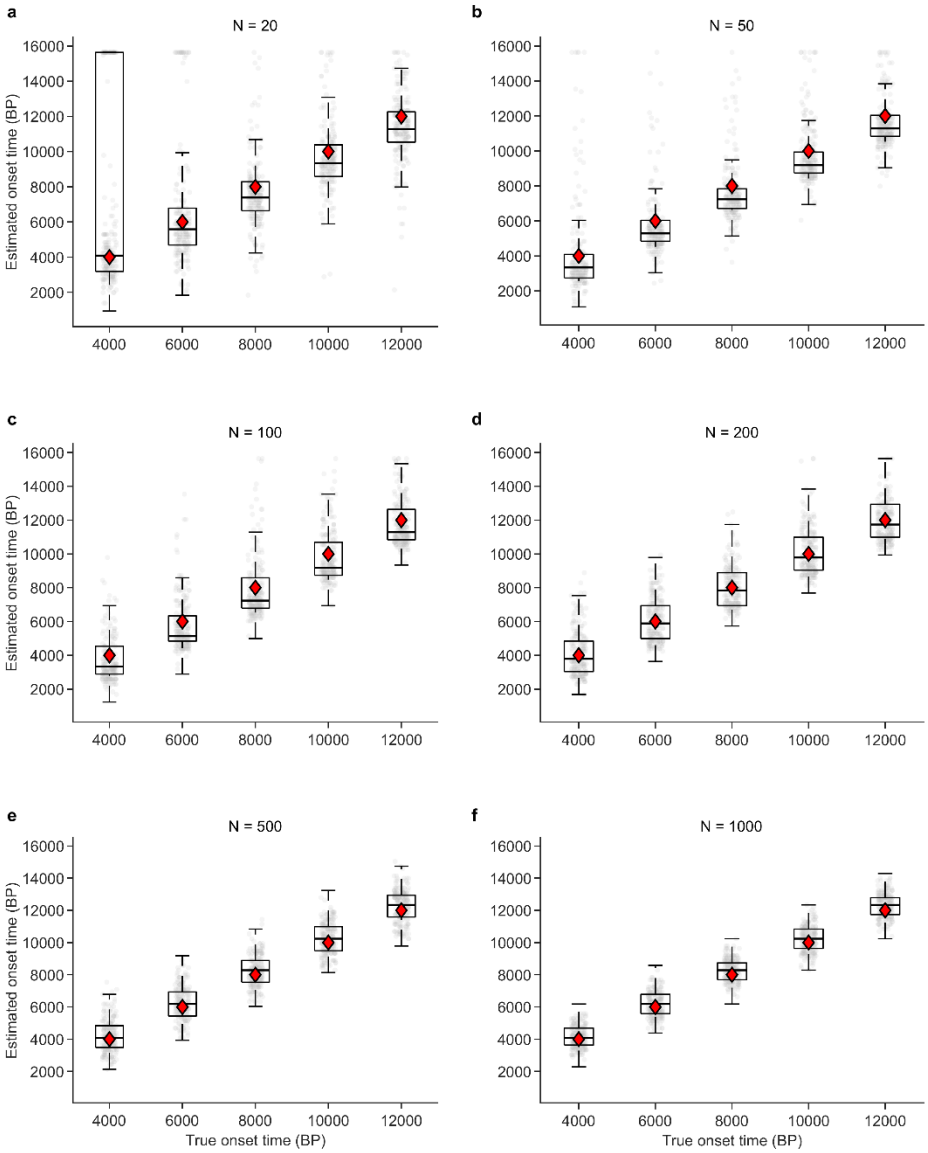
Supplementary Fig. 37 | Simulation benchmark of dropout-adjusted frequency estimation within temporal bins. Single-bin simulations were used to evaluate the performance of the dropout-adjusted frequency estimator under ancient-DNA-like detection conditions. Local sequencing depth and per-depth junction-read detection efficiency were sampled from the empirical ancient-DNA data, and insertion presence was simulated under genotype-specific dropout probabilities. Simulations were performed across sample sizes of 20, 50, 100, 200, 500,

and 1,000 and true insertion frequencies of 0.05, 0.10, 0.20, 0.50, and 0.80, with 1,000 replicates for each parameter combination. Each panel corresponds to one sample size. Boxplots summarize the full distribution of estimated insertion frequencies across all 1,000 replicates, gray points show 200 randomly sampled replicate estimates for visualization, and red diamonds indicate the true simulated frequencies.



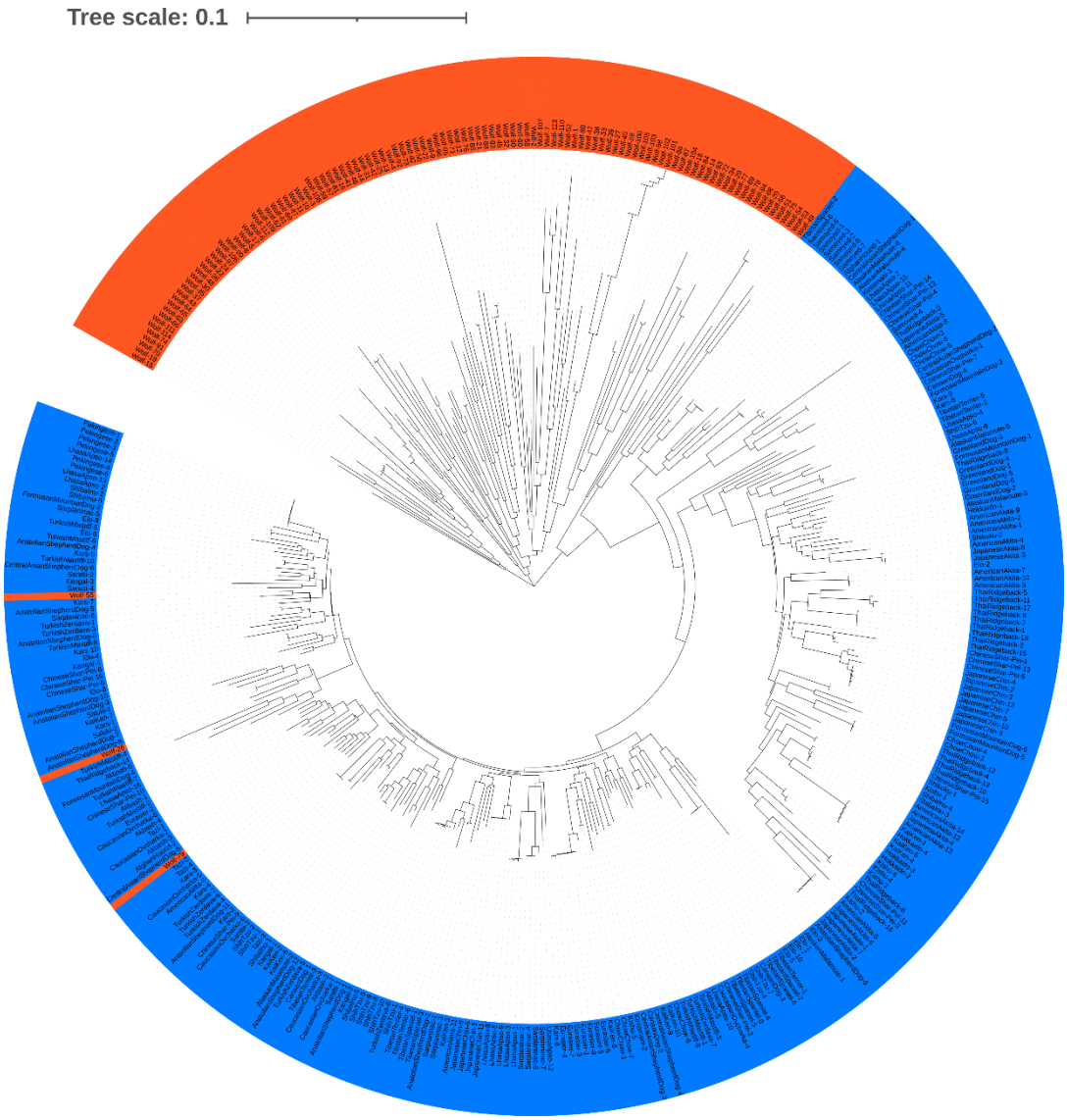
Supplementary Fig. 38 | Simulation benchmark of selection-intensity estimation in the trajectory model. Simulations were used to evaluate recovery of the shared selection coefficient under the dropout-adjusted regional trajectory model. The true selection-onset time was fixed at 8,000 BP, while the true selection intensity was varied across 0, 0.001, 0.002, 0.005, 0.01, 0.02, and 0.05. Simulated datasets were generated under ancient-DNA-like local depth and junction-read detection-efficiency distributions, with sample sizes of 20, 50, 100, 200, 500, and

1,000 and 1,000 replicates for each parameter combination. Each panel corresponds to one sample size. Boxplots summarize the full distribution of fitted selection-intensity estimates across all 1,000 replicates, gray points show 200 randomly sampled replicate estimates for visualization, and red diamonds indicate the true selection-intensity values.

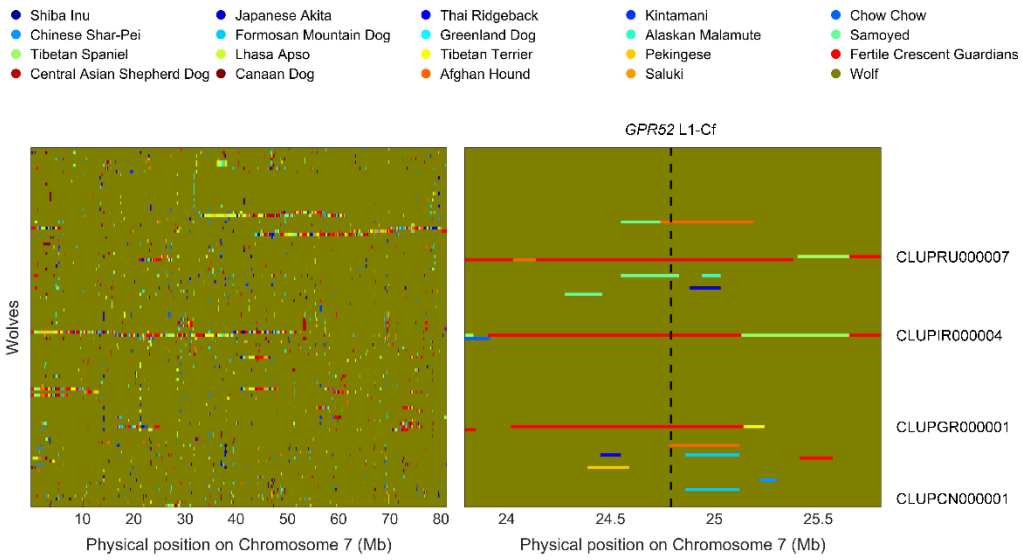


Supplementary Fig. 39 | Simulation benchmark of selection-onset time estimation in the trajectory model. Simulations were used to evaluate recovery of selection-onset time under the dropout-adjusted regional trajectory model. The true selection intensity was fixed at 0.01, while the true selection-onset time was varied across 4,000, 6,000, 8,000, 10,000, and 12,000 BP. Simulated datasets were generated using empirical distributions of local sequencing depth and per-depth junction-read detection efficiency, with sample sizes of 20, 50, 100, 200, 500, and 1,000 and 1,000 replicates for each parameter combination. Each panel corresponds to one sample size.

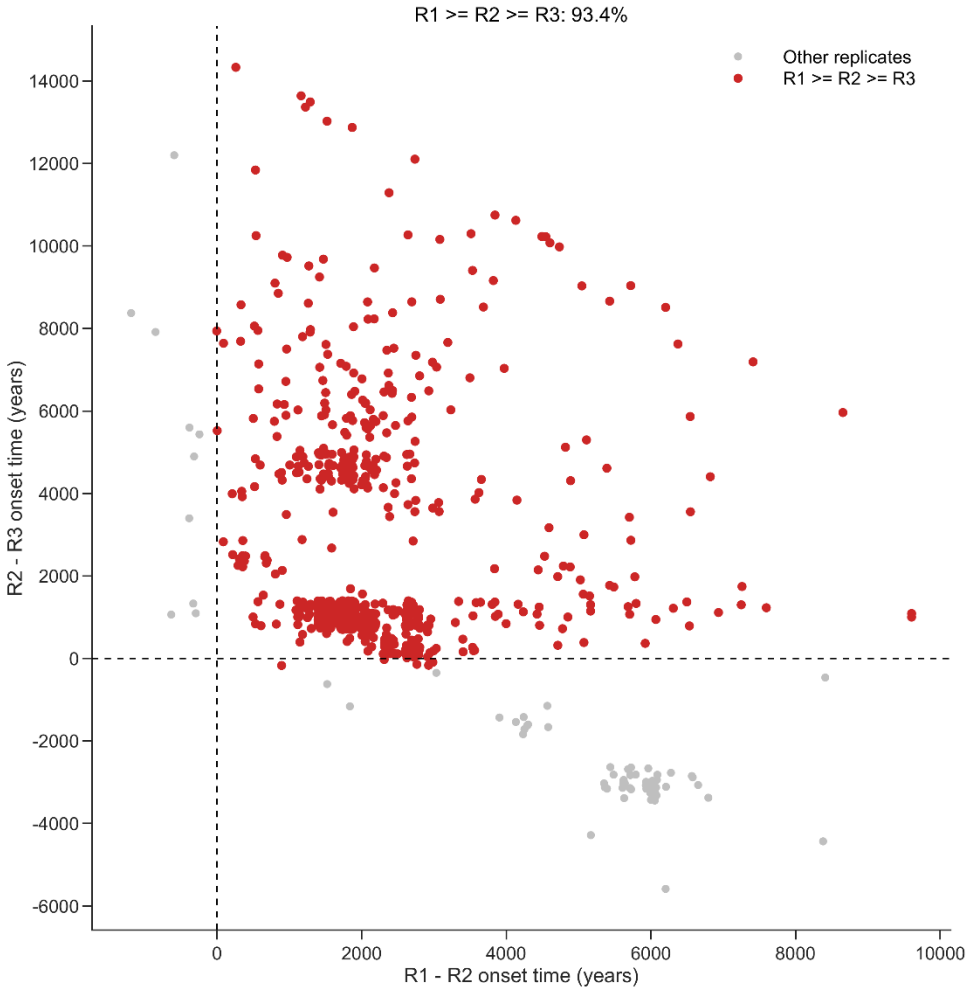
Boxplots summarize the full distribution of fitted onset-time estimates across all 1,000 replicates, gray points show 200 randomly sampled replicate estimates for visualization, and red diamonds indicate the true onset times.



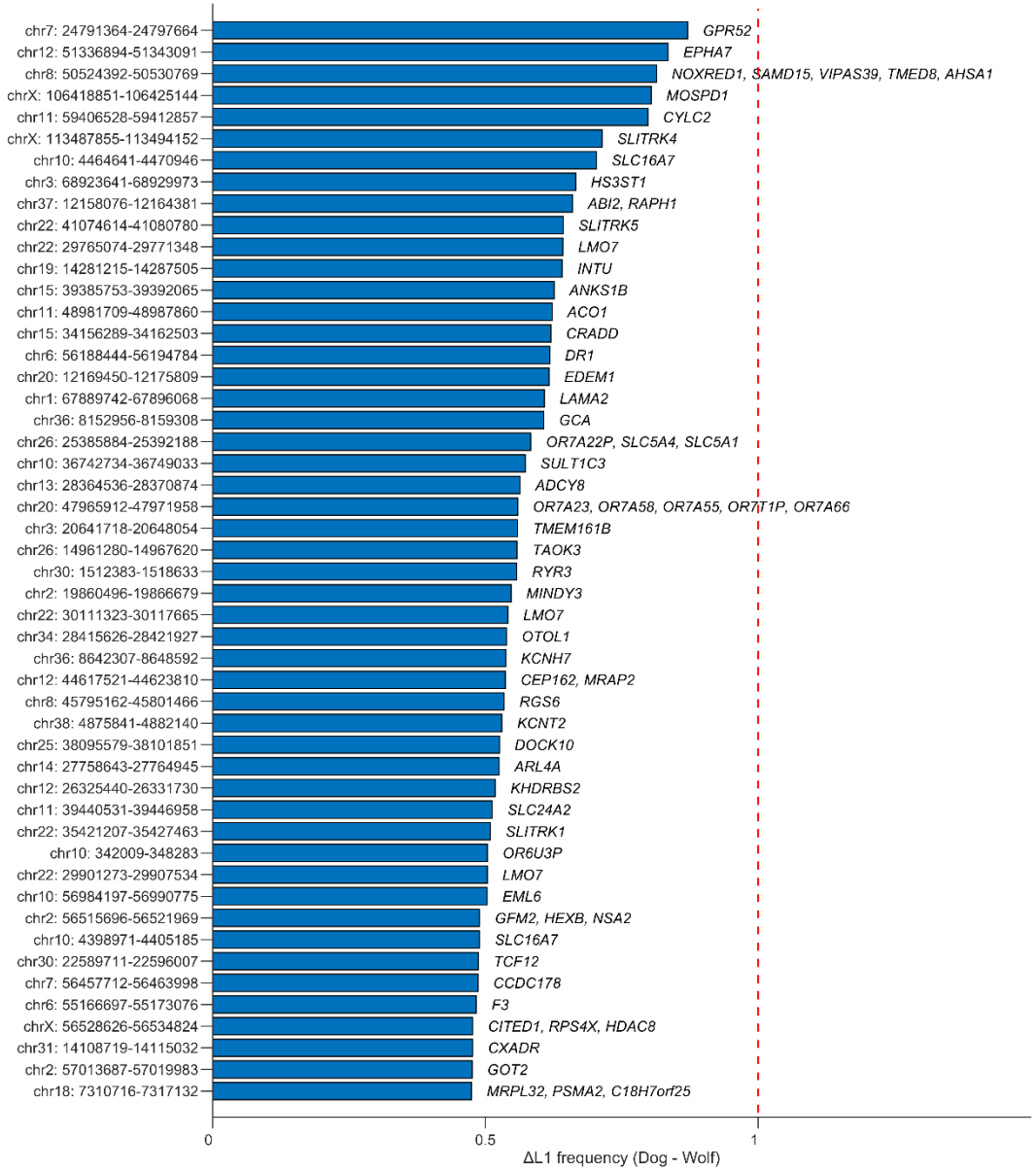
Supplementary Fig. 40 | Neighbor-joining tree of 114 wolf haplotypes and 284 dog haplotypes sampled from 37 basal domestic breeds, as reported in Supplementary Fig. 11. The tree was constructed from phased sequence across Chr7:23,950,070–25,389,886 bp, the shortest interval that captures *GPR52* upstream L1-Cf introgression in wolves. Wolf haplotypes are colored red and dog haplotypes blue. Branch lengths are proportional to genetic distance.



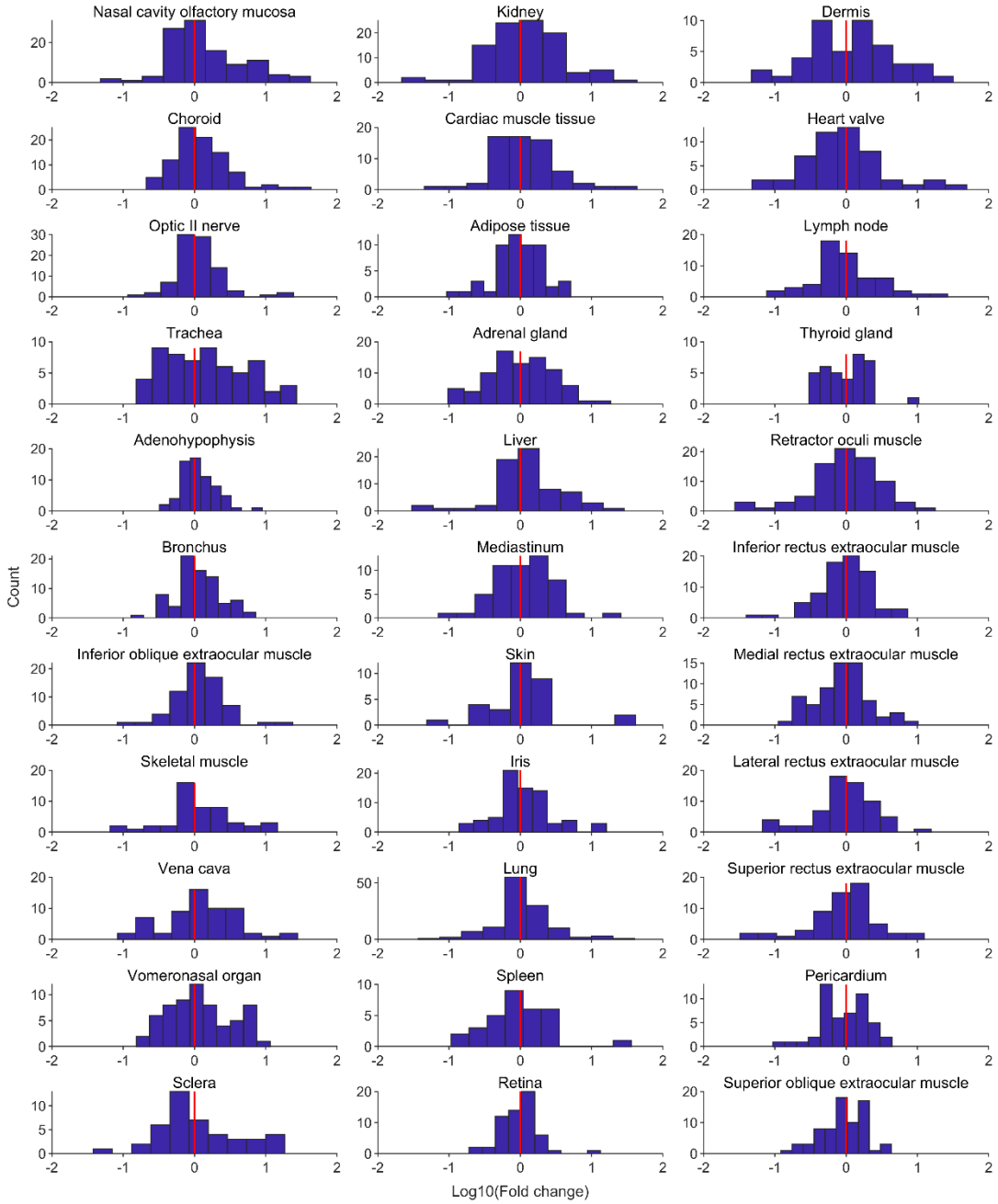
Supplementary Fig. 41 | Local-ancestry inference on chromosome 7 in wolves. The left panel shows chromosome-wide local ancestry across chromosome 7 in wolf haplotypes, and the right panel shows the zoomed *GPR52* region. Colors denote inferred proxy ancestry components, and the dashed line marks the position of the *GPR52* upstream L1-Cf insertion. Among the four wolf haplotypes carrying the insertion, three were assigned most closely to the Fertile Crescent Guardian–associated component in the surrounding region (CLUPRU000007, CLUPIR000004, and CLUPGR000001).



Supplementary Fig. 42 | Bootstrap support for regional ordering of L1-Cf selection onset under the shared-*s* model. R1 denotes Fertile Crescent–Anatolia–Balkans-related dogs, R2 denotes Western Eurasian dogs, and R3 denotes Eastern/Northern Eurasian and Arctic North American dogs. Each point represents one stratified bootstrap replicate from the shared-s selection-onset model, which jointly fitted all three regions with a shared selection coefficient. Onset-time estimates were converted to years BP, such that larger values indicate older inferred onset times. The x axis shows the onset-time difference between R1 and R2, and the y axis shows the onset-time difference between R2 and R3. Dashed lines mark zero difference; points in the upper-right quadrant therefore support the expected regional ordering R1 $\geqslant$ R2 $\geqslant$ R3. Red points indicate bootstrap replicates satisfying this ordering, whereas gray points indicate other replicates. The shared-selection model was summarized from 1,000 stratified bootstrap replicates, with the expected ordering supported in 93.4% of replicates.



Supplementary Fig. 43 | L1 frequency differences (dog − wolf) and nearby genes. Shown are the top 50 L1 insertions with the largest frequency difference between dogs and wolves. For each L1, the distance to the nearest annotated gene was defined as the minimum distance from the L1 core to either the gene's 5′ or 3′ untranslated region (UTR). Genes lacking official gene symbols and microRNAs were excluded. Genes within 50 kb of an L1 insertion are displayed; when no gene lies within 50 kb, the nearest gene is shown.



Supplementary Fig. 44 | L1-associated expression fold-change distributions in non-CNS tissues. Gene-level $\log_{10}$(fold-change) values for L1-adjacent genes are shown across 33 retained non-CNS tissues, complementing the CNS distributions in Fig. 4d. Red vertical lines mark $\log_{10}$(fold-change) = 0, corresponding to no expression shift.

Supplementary Tables

Supplementary Table 1. Classification of basal canid lineages used as proxy ancestry components.

This table classifies basal canine lineages by geographic origin, functional history, and genetic clustering, and indicates the representative lineages used as basal reference components. This table is included in the Supplementary Information PDF.

Supplementary Table 2. *GPR52*-linked L1-Cf insertion frequencies across canid breeds and populations.

This table summarizes the number of chromosomes examined, the number carrying the *GPR52*-linked L1-Cf insertion, and the corresponding insertion frequency for modern dog breeds, village-dog populations and wolf populations. Main-text group-level frequencies were calculated as means of breed- or population-level frequencies rather than by pooling chromosomes across unevenly sampled groups. Provided as an Excel worksheet.

Supplementary Table 3. Breed-level behavioral phenotypes used for exploratory L1-Cf association analyses.

This table contains breed-level behavioral trait scores used to assess exploratory associations between L1-Cf frequency and behavioral phenotypes in modern dog breeds. Provided as an Excel worksheet.

Supplementary Table 4. Ancient canid samples used for L1-Cf detection and regional trajectory analyses.

This table summarizes sample accession, species, age, geographic origin, ENA study accession, L1-Cf detection status, local breakpoint depth, and regional category assignments for ancient dog and wolf genomes analyzed in this study. Three ancient dog samples and three ancient wolf samples lacking age information were excluded from downstream analyses. Among samples with age information, those with zero

local coverage across the *GPR52*-linked L1-Cf breakpoint region were further excluded. After these filtering steps, 73 ancient dog genomes and 70 ancient wolf genomes were retained for subsequent analyses. Provided as an Excel worksheet.

Supplementary Table 5. Canid genome assemblies used for pangenome construction and L1 insertion discovery.

This table lists the domestic dog, dingo, gray wolf, and coyote genome assemblies used for canid pangenome construction and L1 insertion discovery. This table is included in the Supplementary Information PDF.

Supplementary Table 6. Genome-wide full-length L1 insertions and nearby genes identified from the canid pangenome.

This table lists full-length L1 transposable-element insertions longer than 6 kb identified from the canid pangenome, together with gene annotations located within 50 kb of each insertion and genotype information across the analyzed genomes. The 1,213 genes used for enrichment analysis were obtained by excluding LOC-prefixed predicted or unnamed gene models and collapsing the remaining annotations to unique named gene symbols. Provided as an Excel worksheet.

Supplementary Table 7. Pathway enrichment analysis of 1,213 genes located within 50 kb of full-length L1 insertions.

This table reports KOBAS pathway enrichment results for genes located within 50 kb of full-length L1 insertions, including database source, pathway identifiers, input gene counts, background gene counts, nominal P values, corrected P values, and contributing genes. Provided as an Excel worksheet.

Supplementary Table 8. Disease enrichment analysis of 1,213 genes located within 50 kb of full-length L1 insertions.

This table reports disease associated enrichment results for genes located within 50 kb of full-length L1 insertions, including enrichment terms, database sources, statistical

values, corrected P values, and contributing genes. Provided as an Excel worksheet.

Supplementary Table 9. Gene Ontology enrichment analysis of 1,213 genes located within 50 kb of full-length L1 insertions.

This table reports Gene Ontology enrichment results for genes located within 50 kb of full-length L1 insertions, including GO terms, GO identifiers, input gene counts, background gene counts, nominal P values, corrected P values, and contributing genes. Provided as an Excel worksheet.

Supplementary Table 10. Tissue-level L1-associated gene-expression shifts across 52 retained tissues.

This table summarizes tissue-level analyses of L1-associated gene-expression shifts across 52 retained tissues spanning 11 organ systems after tissue-level filtering, including organ system, tissue term, mean $\log_{10}$(fold-change), nominal P value and FDR-adjusted P value. Provided as an Excel worksheet.

Supplementary Table 11. Organ-system-level L1-associated gene-expression shifts across 12 organ systems.

This table summarizes L1-associated gene-expression shifts across 12 organ systems after organ-system-level aggregation, including mean $\log_{10}$(fold-change), nominal P value, FDR-adjusted P value and the proportion of genes with fold change greater than one. Provided as an Excel worksheet.

Supplementary Table 12. Organ-system-level comparison of L1-associated expression shifts between dog-dominant and wolf-dominant frequency groups.

This table compares mean $\log_{10}$(fold-change) values between dog-dominant and wolf-dominant L1 frequency groups across organ systems and reports the corresponding statistical comparisons. Provided as an Excel worksheet.

References

1 Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747-751, doi:10.1126/science.1243518 (2014).

2 Campbell, C. L., Bherer, C., Morrow, B. E., Boyko, A. R. & Auton, A. A Pedigree-Based Map of Recombination in the Domestic Dog Genome. *G3 (Bethesda)* **6**, 3517-3524, doi:10.1534/g3.116.034678 (2016).

3 Haller, B. C. & Messer, P. W. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Molecular Biology and Evolution* **36**, 632-637, doi:10.1093/molbev/msy228 (2019).

4 Browning, S. R., Waples, R. K. & Browning, B. L. Fast, accurate local ancestry inference with FLARE. *American Journal of Human Genetics* **110**, 326-335, doi:10.1016/j.ajhg.2022.12.010 (2023).

5 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534, doi:10.1093/molbev/msaa015 (2020).

6 Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* **52**, W78-W82, doi:10.1093/nar/gkae268 (2024).

7 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).

8 Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459-471, doi:10.1534/genetics.113.150029 (2013).

9 Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832-837, doi:10.1038/nature01140 (2002).

10 Zhao, S., Chi, L., Fu, M. & Chen, H. HaploSweep: Detecting and Distinguishing Recent Soft and Hard Selective Sweeps through Haplotype Structure. *Mol Biol Evol* **41**, doi:10.1093/molbev/msae192 (2024).

11 Chen, H., Hey, J. & Slatkin, M. A hidden Markov model for investigating recent positive selection through haplotype structure. *Theor Popul Biol* **99**, 18-30, doi:10.1016/j.tpb.2014.11.001 (2015).

12 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

13 Zhang, X. *et al.* The distinct morphological phenotypes of Southeast Asian aborigines are shaped by novel mechanisms for adaptation to tropical rainforests. *Natl Sci Rev* **9**, nwab072, doi:10.1093/nsr/nwab072 (2022).

14 Chi, L. *et al.* HapNetworkView: a tool for haplotype network exploration and visualization. *BMC Genomics* **26**, 52, doi:10.1186/s12864-025-11206-8 (2025).

15 Botigué, L. R. *et al.* Ancient European dog genomes reveal continuity since the Early Neolithic. *Nature Communications* **8**, doi:10.1038/ncomms16082 (2017).

16 Skoglund, P., Ersmark, E., Palkopoulou, E. & Dalén, L. Ancient Wolf Genome Reveals an Early Divergence of Domestic Dog Ancestors and Admixture into High-Latitude Breeds. *Current Biology* **25**, 1515-1519, doi:10.1016/j.cub.2015.04.019 (2015).

17 Sinding, M. H. S. *et al.* Arctic-adapted dogs emerged at the Pleistocene-Holocene

transitiond. *Science* **368**, 1495-+, doi:10.1126/science.aaz8599 (2020).
18 Marsh, W. A. *et al.* Dogs were widely distributed across western Eurasia during the Palaeolithic. *Nature* **651**, doi:10.1038/s41586-026-10170-x (2026).
19 Leathlobhair, M. N. *et al.* The evolutionary history of dogs in the Americas. *Science* **361**, 81-+, doi:10.1126/science.aao4776 (2018).
20 Ramos-Madrigal, J. *et al.* Genomes of Pleistocene Siberian Wolves Uncover Multiple Extinct Wolf Lineages. *Current Biology* **31**, 198-+, doi:10.1016/j.cub.2020.10.002 (2021).
21 Frantz, L. A. F. *et al.* Genomic and archaeological evidence suggests a dual origin of domestic dogs. *Science* **352**, 1228-1231, doi:10.1126/science.aaf3161 (2016).
22 Zhang, S. J. *et al.* Genomic evidence for the Holocene codispersal of dogs and humans across Eastern Eurasia. *Science* **390**, 735-740, doi:10.1126/science.adu2836 (2025).
23 Bergström, A. *et al.* Grey wolf genomic history reveals a dual ancestry of dogs. *Nature* **607**, 313-+, doi:10.1038/s41586-022-04824-9 (2022).
24 Bougiouri, K. *et al.* Imputation of ancient canid genomes reveals inbreeding history over the past 10,000 years. *Proceedings of the National Academy of Sciences of the United States of America* **122**, doi:10.1073/pnas.2416980122 (2025).
25 Feuerborn, T. R. *et al.* Modern Siberian dog ancestry was shaped by several thousand years of Eurasian-wide trade and human dispersal. *Proceedings of the National Academy of Sciences of the United States of America* **118**, doi:10.1073/pnas.2100338118 (2021).
26 Bergström, A. *et al.* Origins and genetic legacy of prehistoric dogs. *Science* **370**, 557-564, doi:10.1126/science.aba9572 (2020).
27 Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* **38**, 276-278, doi:10.1038/s41587-020-0439-x (2020).
28 Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes* **9**, 88, doi:10.1186/s13104-016-1900-2 (2016).
29 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
30 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
31 Gardner, E. J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Research* **27**, 1916-1929, doi:10.1101/gr.218032.116 (2017).
32 Hickey, G. *et al.* Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol* **42**, 663-673, doi:10.1038/s41587-023-01793-w (2024).
33 Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput Biol* **18**, e1009123, doi:10.1371/journal.pcbi.1009123 (2022).
34 Bu, D. *et al.* KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res* **49**, W317-W325, doi:10.1093/nar/gkab447 (2021).
35 Hortenhuber, M. *et al.* The DoGA consortium expression atlas of promoters and genes in

100 canine tissues. *Nat Commun* **15**, 9082, doi:10.1038/s41467-024-52798-1 (2024).
36 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222, doi:10.1093/bioinformatics/btv710 (2016).
37 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
38 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).
39 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
40 Niu, L. J. *et al.* Three-dimensional folding dynamics of the genome. *Nature Genetics* **53**, 1075-+, doi:10.1038/s41588-021-00878-z (2021).
41 Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884-890, doi:10.1093/bioinformatics/bty560 (2018).
42 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
43 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).
44 Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
45 Lopez-Delisle, L. *et al.* pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics* **37**, 422-423, doi:10.1093/bioinformatics/btaa692 (2021).
46 Poulter, S. *et al.* The Identification of GPR52 Agonist HTL0041178, a Potential Therapy for Schizophrenia and Related Psychiatric Disorders. *Acs Med Chem Lett* **14**, 499-505, doi:10.1021/acsmedchemlett.3c00052 (2023).
47 Nadler, J. J. *et al.* Automated apparatus for quantitation of social approach behaviors in mice. *Genes Brain Behav* **3**, 303-314, doi:10.1111/j.1601-183X.2004.00071.x (2004).
48 Yilmaz, M. & Meister, M. Rapid Innate Defensive Responses of Mice to Looming Visual Stimuli. *Current Biology* **23**, 2011-2015, doi:10.1016/j.cub.2013.08.015 (2013).
49 Pellow, S., Chopin, P., File, S. E. & Briley, M. Validation of open:closed arm entries in an elevated plus-maze as a measure of anxiety in the rat. *J Neurosci Methods* **14**, 149-167, doi:10.1016/0165-0270(85)90031-7 (1985).
50 Swerdlow, N. R., Geyer, M. A. & Braff, D. L. Neural circuit regulation of prepulse inhibition of startle in the rat: current knowledge and future challenges. *Psychopharmacology (Berl)* **156**, 194-215, doi:10.1007/s002130100799 (2001).
51 Poulter, S. *et al.* Preclinical pharmacokinetics, metabolism, and disposition of NXE0041178, a novel orally bioavailable agonist of the GPR52 receptor with potential for treatment of schizophrenia and related psychiatric disorders. *Xenobiotica* **55**, 151-166, doi:10.1080/00498254.2025.2501593 (2025).