Pilot Project Final Report: Establishing Genetic Testing Methods for Dingo Ancestry Estimation

For: National Wild Dog Action Plan

By: Dr Yassine Souilmi and Shyamsundar Ravishankar, Australian Centre for Ancient DNA, The University of Adelaide

Date: 20 March 2026

Australian Centre for Ancient DNA
205D Darling Building, Adelaide University,
North Terrace Campus, Adelaide, SA 5005

Tel: +61 8 8313 5565
Email: Yassine.Souilmi@adelaide.edu.au
Web: adelaide.edu.au

Executive Summary

Across Australia, state governments, conservation groups, and management agencies require accurate information about free-roaming canine populations to inform management strategies, legislation and regulatory instruments. Determining whether a wild canine is a dingo, a feral domestic dog, or a hybrid has been technically challenging, frequently producing conflicting test results. This pilot project was commissioned to resolve that confusion by developing and validating a highly accurate, cost-effective test for dingo ancestry, statistically compared against an ancient pre-colonial dingo DNA reference and newly developed statistical methods.

The legacy of past testing

Recent Single Nucleotide Polymorphism (SNP)-array testing combined with clustering algorithms (e.g., `FastStructure` as used in Cairns *et al.*, 2023) offered a genome-wide approach. However, these methods share a limitation with microsatellite's (STR) as they both rely on contemporary populations to define what a "pure" dingo is. If a reference population already carries historical admixture, the algorithm may absorb that admixture into the "dingo" cluster, systematically underreporting the true extent of European dog ancestry. Additionally, these methods are highly sensitive to arbitrarily defined number of clusters, completely skewing their results (Ravishankar *et al.*, in press).

Furthermore, while population-level geographic patterns derived from legacy STR data remain broadly consistent with estimates from whole-genomes, the individual STR estimates should not be treated as interchangeable with whole-genome estimates.

The new approach

To address these limitations, we developed an analytical framework based on two advances:

- **Established a pre-colonial baseline.** DNA from ancient dingo specimens (from the Nullarbor and Curracurrang sites, dating back 1,000 to 2,000 years respectively) provides an absolute reference that predates any European dog contact (Souilmi *et al.*, 2024).
- **High-resolution genomic scanning.** We analysed hundreds of thousands of specific markers across the entire genome using a statistical model (`qpAdm`) that mathematically separates the ancient dingo and European dog contributions in each animal.

A second, independent whole-genome method (MOSAIC) — which reads ancestry along the chromosomes in a different way (using haplotypes) — was used as a cross-check. The two approaches agree closely ($R^2 \approx 0.97$ across 46 specimens), which increases confidence that the reference result is robust.

Key findings

1. The qpAdm framework, anchored by pre-colonial ancient dingo genomes, provides the most robust estimates of European dog versus dingo ancestry available.
2. MOSAIC and qpAdm agree closely ($R^2 \approx 0.97$), validating the reference approach with an independent method.

3. Benchmarking confirms that low-pass whole-genome sequencing (WGS), retaining approximately 10,000 informative transversion markers, yields results comparable to high-coverage whole-genome sequencing–used in this report, presenting a cost-effective pathway for large-scale application (Ravishankar *et al.*, in press).
4. The `MOSAIC` method is, however, very sensitive to data missingness and therefore unsuitable for low-cost testing.
5. Low-pass sequencing approach coupled with our testing method lowers the testing cost by 10-fold, with further cost reductions possible at scale.
6. STR testing, while consistent overall at the broader regional level, shows high variance at the individual level and consistently diverges from whole-genome estimates ($R^2 \approx 0.25$, n = 36). For example, sample HF093 (VIC) returned 63% dingo ancestry using STR-based testing but 80.1% using `qpAdm`. The bulk of this variance is attributable to the low number of markers, making STR vulnerable to genotyping errors.

Application

This project delivers a scientifically robust and economically viable diagnostic framework. It is designed to be used by states and jurisdictions as a reliable foundational tool, allowing them to apply highly accurate genomic data in whatever manner best aligns with their specific regulatory, legislative, and management requirements.

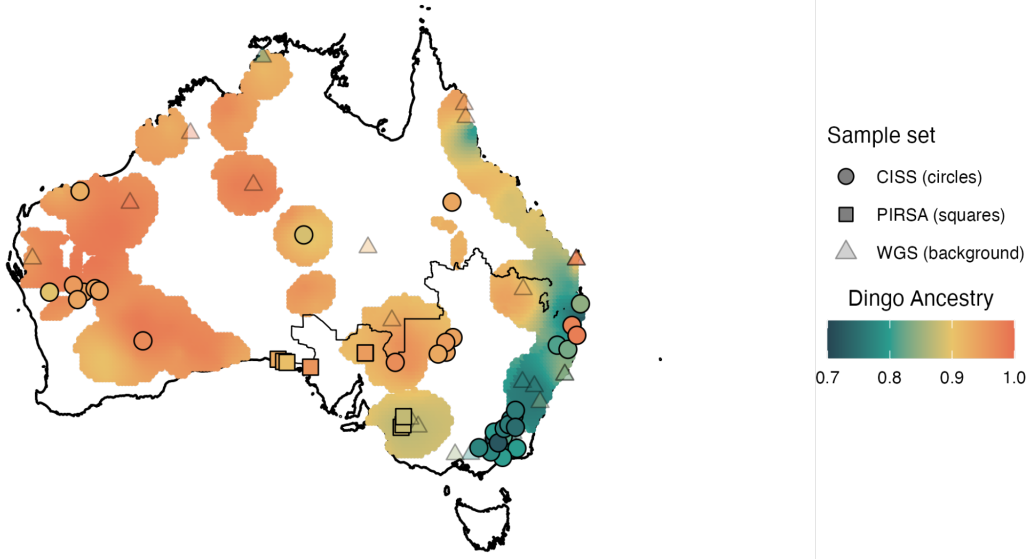

Figure 1. Continental panels display interpolated genetic ancestry distributions estimated using the best-practice method introduced in this report. This approach utilises the `qpAdm` method, anchored by ancient DNA to represent dingo ancestry. Background data points represent SNP-array samples (Cairns et al. 2023), while faded symbols indicate background WGS data from the literature (Zhang et al. 2020; Souilmi et al. 2024). CISS pilot samples are shown as circles, and PIRSA samples as squares. Point fill colour corresponds to `qpAdm` dingo ancestry (ancient-anchored model). Western Australian samples cluster at high dingo ancestry (warm tones), whereas southeastern samples show lower values (cool tones).

Project Background and Objectives

Policy and research programs increasingly need reliable dingo ancestry estimates. When tests assume today's dingoes are an unchanged baseline, results can be misleading if those populations already carry older mixing with domestic dogs.

This report details the outcomes of a pilot project designed to benchmark and establish new genomic testing methods. While the original CISS agreement provisioned for 12 biospecimens, the project scope was significantly expanded. We successfully processed 37 samples provided by CISS. Concurrently, 11 samples were processed for PIRSA under a separate agreement. Given the mutual benefit of expanded data pooling, both agencies agreed to a joint presentation of the results. Of the 48 total samples, 47 successfully passed strict quality control (QC) and sequencing thresholds.

This project aimed to:

1. Generate a high-quality genomic dataset from contemporary biospecimens across CISS and PIRSA cohorts.
2. Establish an unbiased ancestry baseline using pre-colonial ancient dingo genomes.
3. Benchmark the accuracy and reliability of STR testing and clustering methodologies against advanced genomic approaches (`qpAdm` and `MOSAIC`).
4. Provide actionable recommendations for standardising future ancestry testing methodologies.

Methodology

A comprehensive breakdown of the bioinformatics and statistical procedures is available in Appendix A.

Review of legacy testing methods

Prior to this project, ancestry testing relied on two primary approaches:

- SNP-arrays and clustering (e.g., Cairns *et al.*, 2023): Uses commercial canine arrays to examine thousands of markers, then groups animals using algorithms such as `FastStructure` or `ADMIXTURE`. Without a pre-colonial reference, these algorithms assign contemporary animals to clusters. If a population shares widespread historical admixture with European dogs, the algorithm may define that admixed state as "pure dingo," systematically underreporting the true extent of European dog ancestry. Additionally, these algorithms, if not fit properly, are highly sensitive to arbitrary pre-defined number of clusters (K value).
- Microsatellites (STRs): Examines a small number of markers (typically 23 loci). While foundational for early research and successful at identifying broad geographic trends, STR panels lack the resolution to accurately quantify individual admixture percentages. With so few markers, a single mis-called or poorly visualised locus can shift an individual's result by several percentage points.

Current project methodology

To overcome these limitations, 48 tissue samples were extracted and sequenced using Illumina shotgun sequencing (37 CISS, 11 PIRSA). One PIRSA sample failed QC, resulting in 47 successful whole-genome datasets.

We mapped to the `CanFam3.1` reference genome — the standard coordinate system for the domestic dog genome, used here as the common framework for aligning and comparing DNA sequences across all samples. To ensure maximum accuracy when comparing modern DNA to ancient DNA, the analysis was restricted to transversion single nucleotide polymorphisms (SNPs). Ancient DNA degrades over time in predictable ways, creating specific types of errors called "damage biases" (C-to-T and G-to-A changes). Transversions are a class of DNA variants that is unaffected by this degradation, so restricting the analysis to transversions effectively filters out any potential errors.

Ancestry proportions were calculated using the `qpAdm` framework, contrasting target samples against high-coverage German Shepherd genomes (as a proxy for European dog ancestry) and pre-colonial ancient dingo lineages. Results were cross-validated using `MOSAIC` for local ancestry inference and compared against legacy STR test results where available. By algorithmically masking identified European dog segments, we then performed unbiased Principal Component Analysis (PCA) and unsupervised clustering to assess the underlying ancestral dingo population structure.

Results

Geographic distribution and admixture landscape

The geographic patterns of admixture (Figure 1) place the pilot samples in the context of three independent ancestry estimates. Samples from Central and Western Australia show the highest dingo ancestry in this pilot and in the broader surfaces, with very limited European dog ancestry detected; Victoria and New South Wales samples tend toward higher European dog ancestry.

However, these remain fairly limited across animals in this pilot, estimated European domestic dog ancestry under the reference method spans roughly 1.7% – 28.8% (Figure 2 and Appendix C). These results are consistent with earlier research (Stephens *et al.*, 2015; Souilmi *et al.*, 2024; Ravishankar *et al.*, in press). Additionally, we observe a positive correlation between the level of European dog admixture and human population density (Figure 1; Ravishankar *et al.*, in press). Higher human activity correlates with increased exposure to domestic dogs, driving the elevated admixture rates observed along the eastern seaboard.

Crucially, `MOSAIC` local ancestry inference indicates this gene flow in the southeast is largely historical: initial admixture dates to approximately 50–150 years before sampling (Ravishankar *et al.*, in press; Scarsbrook *et al.*, 2025). Given a dingo generation time of approximately 3–5 years, this represents 10–50 generations of ancestral mixing, with the 1950s–1960s constituting a peak in intensity rather than the onset of hybridisation. The eastern Australian dingo population has therefore carried domestic dog ancestry for well over a century, rather than representing ongoing active hybridisation.

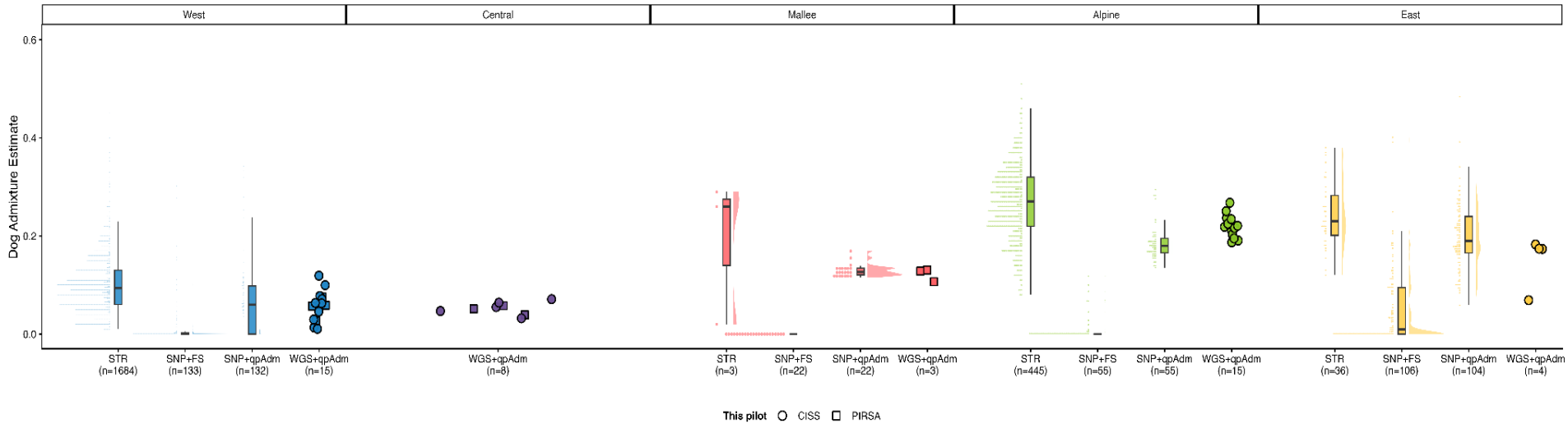

Figure 2. Grouped comparison of estimated dog admixture across four testing approaches — STR (Stephens *et al.*, 2015), SNP-array with `FastStructure` clustering (Cairns *et al.*, 2023; labelled SNP+FS), SNP-array with `qpAdm` (broader study; labelled SNP+qpAdm), and whole-genome sequencing with `qpAdm` (this pilot; labelled WGS+qpAdm) — stratified by regional population. Each panel shows the distribution of dog admixture estimates within a population, with individual data points, boxplots, and density curves. Sample sizes (n) are annotated for each method–population combination. CISS pilot samples are shown as filled circles and PIRSA samples as filled squares in the WGS+qpAdm column. The Central population panel appears only in the WGS+qpAdm column because earlier studies did not sample that region with comparable methods.

Limitations of legacy methods

The most significant methodological advance prior to the present study was the SNP-array approach of Cairns *et al.* (2023), which moved from a handful of microsatellite loci to genome-wide markers and substantially improved population-level resolution. However, even that approach relied on contemporary reference populations in the absence of a pre-colonial ancestral baseline. The critical advance of the present study is the integration of ancient pre-colonial dingo genomes as the reference, which removes the circularity inherent in using potentially admixed modern dingoes to define “dingo purity”.

Comparisons between legacy STR results and whole-genome benchmarks for matched samples reveal significant unreliability in the STR methodology at the individual animal level, although population-level geographic patterns can remain broadly similar (Figure 2; Figure 3). There is a poor correlation ($R^2 \approx 0.25$) between STR estimates and qpAdm estimates for the 36 specimens for which we have both estimates. STR testing frequently underestimates true dingo ancestry: for example, sample HF093 (VIC) returned an STR result of 63% dingo ancestry, whereas `qpAdm` analysis confirmed 80.1%.

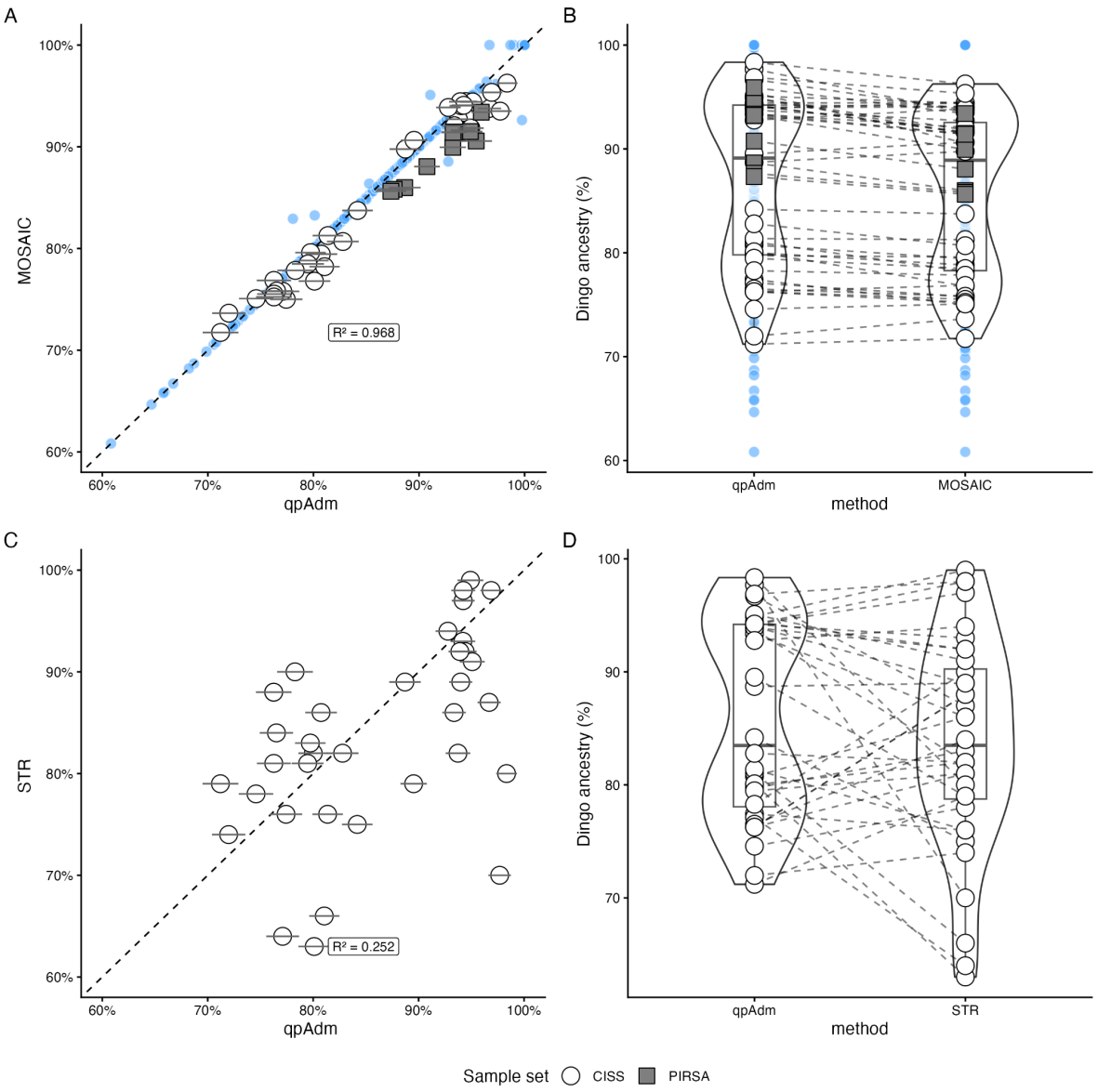

Figure 3. Comparison of ancestry estimation methods. Panels A and B: concordance between MOSAIC and qpAdm for whole-genome dingo samples. Panels C and D: comparison of STR dingo percentage with qpAdm dingo ancestry. White circles: CISS; grey squares: PIRSA. R^2 annotations refer to the Appendix C pilot subset plotted. Background cloud in panel A shows additional WGS samples for context.

Robustness of genomic methods (qpAdm versus MOSAIC)

We found high levels of concordance ($R^2 \approx 0.97$) between haplotype-based tests using the MOSAIC method and qpAdm dingo proportions estimated using pre-colonial ancient DNA dingoes as a reference (n = 46, Figure 3; Appendix C).

Furthermore, subsampling (Appendix B; Ravishankar et al., in press) reports qpAdm stability of test results down to roughly 10,000 haploid transversions per-sample, supporting the use of highly cost effective low-pass whole-genome sequencing designs, subject to quality control. This means the analytical rigour of the reference approach can be maintained at a cost profile competitive with current testing methods. However, while MOSAIC showed resilience to subsampling, the data needed to support haplotype-based methods exceed that of ancient DNA based testing coupled with qpAdm limiting their scalability and affordability.

Population structure of free-roaming dingoes

After masking European dog segments (removed using MOSAIC outputs) the underlying biogeographic population structure of the free-roaming dingoes becomes visible. Conversely, clustering algorithms (such as ADMIXTURE or FastStructure), when unanchored, i.e. used without a representative population of all potential source ancestries, can confounded European dog admixture as internal dingo population structure, thus causing them to underestimate dog ancestry when used alone as a diagnostic tool. Masking bypasses this limitation.

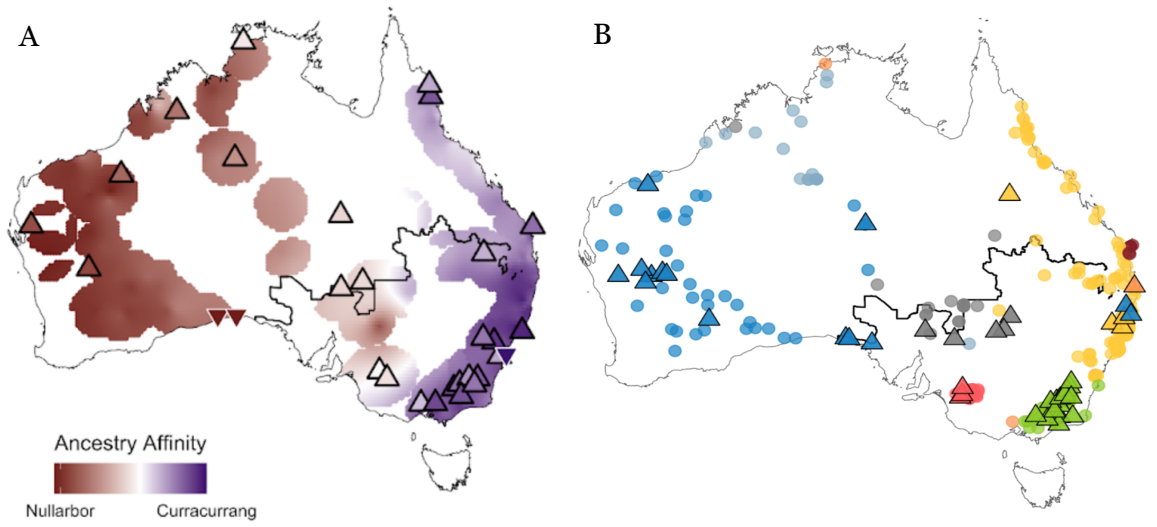

Figure 4. Ancestral dingo population structure. (A) Broad affinities of free-living dingoes to the two primary ancient dingo lineages as represented by the Nullarbor dingoes and Curracurrang dingoes. Reproduced from Ravishankar *et al.* (in press). Not all dingoes represented. (B) Geographic spatial map displaying the distribution of identified ancestral components.

We observe deep ancestral lineages, with contemporary dingoes broadly divided into two major groupings corresponding to deep historical splits. Western populations share affinity with the ~1,000-year-old Nullarbor ancient genomes, while eastern populations align with the ~2,000-year-old Curracurrang lineage (Souilmi *et al.*, 2024).

Further exploration reveals distinct population units. Dimensionality reduction (UMAP on EMU PCA of masked genomes) resolves modern populations into regional clusters (West, East, North, Central, Alpine, Mallee, and K'gari in this analysis). These are descriptive genetic population units; they are not management categories. Additionally, these population units, especially in the Southeastern seaboard appear at first glance partially shaped by human activity, such as densely populated areas and farmlands, rather than an historic structure.

Central Australian populations represent an intersection zone with high ancestral genetic diversity. Southeastern groups (Mallee, Alpine, East) display reduced dingo-specific genetic diversity, indicative of historical bottlenecks. The exact mechanisms driving these bottlenecks remain unconfirmed and require further study.

Discussion

The critical limitation of current testing methods is plagued by two key limitations. The first is the absence of an ancestral baseline. The incorporation of ancient pre-colonial dingo genomes in the present study eliminates this circularity and is the key methodological advance over Cairns et al. (2023).

The second flaw lies in the use of clustering algorithms without the appropriate reference baseline. While historical admixture is present and detectable in descendants, clustering algorithms without pre-colonial baselines can misinterpret it. If the contemporary "reference" populations themselves carry fixed historical admixture, the algorithm may inadvertently define this admixed state as the "pure dingo" baseline, systematically underreporting the true extent of European dog ancestry.

By anchoring qpAdm models with pre-colonial ancient genomes (Souilmi *et al.*, 2024), we eliminate this circularity. The resulting data provides high-resolution clarity beneath the broad management umbrella term of "wild dog" — a functional term denoting free-living canine of unknown specific genetic status, encompassing dingoes, hybrids, and the free-living domestic dog. The data demonstrates that many of these canines carry historically fixed admixture of varying degrees.

While overall continent-wide admixture is generally low, Eastern Australian populations exhibit considerable admixture (20–28% European dog ancestry). This level of admixture aligns closely with human population density and access to domestic dogs, consistent with earlier findings (Stephens *et al.*, 2015), and at odds with Cairns *et al.* (2023). Ravishankar et al. (in press) and Scarsbrook et al. (2025) discuss the timing and dynamics of admixture

in southeastern populations; those publications should be consulted for detailed interpretation.

The primary value of this new methodology lies in providing an unbiased, high-resolution tool for population-level genetic monitoring. When jurisdictions require accurate taxonomic differentiation for specific legislative, regulatory, or planning purposes, they now have access to a method that can deliver it with documented precision.

Recommendations

Based on the empirical evidence generated from the 47 successful CISS and PIRSA samples, we recommend the following methodological shifts for testing frameworks:

1. Transition away from SNP-array data used with unanchored clustering, and legacy STR testing as primary diagnostic tools due to their demonstrated statistical variance and inability to leverage pre-colonial baselines.
2. Adopt low-pass whole-genome sequencing. Agencies should transition to low-pass WGS analysed via `qpAdm` using ancient baselines. Benchmarking demonstrates that the ~10,000 transversion threshold delivers results virtually identical to high-coverage sequencing, at a cost profile competitive with current testing (Appendix B; Ravishankar *et al.*, in press).
3. Standardise testing methodologies across jurisdictions. To facilitate cross-state collaboration and data sharing, we recommend adopting this testing framework as a standardized scientific baseline, if genetic testing is part of management decisions.
4. Standardise metadata exchange. Coordinates, batch IDs, and sequencing parameters should be recorded in a consistent format to enable future bridging studies and cross-jurisdictional comparisons.

Agencies decide how any test result is used under law; this report does not prescribe management outcomes.

References

1. Cairns, K. M., Crowther, M. S., Parker, H. G., Ostrander, E. A., & Letnic, M. (2023). Genome-wide variant analyses reveal new patterns of admixture and population structure in Australian dingoes. *Molecular Ecology*, 32, 4133–4150.
2. Harney, É., Patterson, N., Reich, D., & Wakeley, J. (2021). Assessing the performance of qpAdm: a statistical tool for studying population admixture. *Genetics*, 217(4), iyaa045.

3. Meisner, J., Liu, Y., Huang, Y., & Albrechtsen, A. (2021). Large-scale inference of population structure in the presence of missingness using PCA. *Bioinformatics*, 37(13), 1868–1875.
4. Ravishankar, S., Nguyen, N. C., Taufik, L., Michielsen, N. M., Bergström, A., Tobler, R., Fordham, D., Brüniche-Olsen, A., Rahbek, C., Llamas, B., & Souilmi, Y. (in press). Palaeogenomics-informed inferences of European dog admixture enables scalable dingo conservation. bioRxiv 2026.04.08.717357. https://doi.org/10.64898/2026.04.08.717357. Also *in Press*.
5. Salter-Townshend, M., & Myers, S. (2019). Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics*, 212(3), 869–889.
6. Scarsbrook, L., et al. (2025). The impacts of European arrival on Australian dingoes. *Proceedings of the National Academy of Sciences*, 122, e2421749122.
7. Souilmi, Y., et al. (2024). Ancient genomes reveal over two thousand years of dingo population structure. *Proceedings of the National Academy of Sciences*, 121(30), e2407584121.
8. Stephens, D., et al. (2015). Death by sex in an Australian icon: a continent-wide survey reveals extensive hybridization between dingoes and domestic dogs. *Molecular Ecology*, 24, 5643–5656.

Appendix A: Detailed Methodology

A.1 Data generation and bioinformatic processing

Tissue samples provided by CISS and PIRSA were subjected to DNA extraction and prepared for double-stranded library construction. Shotgun sequencing was performed on the Illumina NovaSeq 6000 platform. The resulting contemporary dingo genomes were processed using the `nf-core/eager` (v2.4.5) pipeline with parameters optimised for modern DNA. Poly-G tails were trimmed at read termini using `fastp` (v0.20.1), and Illumina adapters were removed using `AdapterRemoval` (v2.3.2). Sequence reads were mapped to the `CanFam3.1` reference genome using `BWA-MEM` (v0.7.17), eliminating alignments with a mapping quality below 20. PCR duplicates were subsequently removed using Picard `MarkDuplicates` (v2.26.0).

A.2 Genotyping and variant calling

To ensure consistency with ancient genome models and to avoid transition biases associated with ancient DNA damage (e.g., C→T and G→A deamination), genotyping was restricted to transversion loci. `GATK HaplotypeCaller` (v4.6.1.0) was run with `EMIT_VARIANTS_ONLY`, focusing on a predefined set of biallelic transversions identified in modern canids. Genotypes were set to be missing if depth was below half or more than three times the sample mean. Pseudohaploid genotypes were generated with `PileupCaller` (`sequenceTools` 1.5.2).

A.3 Admixture modelling (`qpAdm`)

`qpAdm` (`ADMIXTOOLS2` v2.0.0 R package; Harney *et al.*, 2021; Ravishankar *et al.*, in press) was applied with German Shepherd as a European dog proxy and pre-colonial ancient dingoes (Nullarbor ~1k BP; Curracurrang-related sources as specified in the analysis workflow) as dingo-related sources. Outgroups included CoyoteCalifornia and deep-time Eurasian dog/wolf-related genomes (e.g., Zhokhov9500BP, Germany_7k, Ireland_Neolithic, Russia_Baikal_7k). These ancient reference outgroups possess ancestral diversity to modern domestic dogs but are unconfounded by recent European dog breed admixture. A rotating `qpAdm` approach was implemented with `allsnps` enabled. Models were accepted when goodness-of-fit $p \geq 0.01$ and admixture coefficients were non-negative and interpretable.

A.4 Local ancestry inference (`MOSAIC`)

`MOSAIC` (v1.5.0; Salter-Townshend & Myers, 2019) was run on imputed/phased WGS using European breed dogs (excluding Australian-derived breeds) and pre-contact ancient dingoes as references. This method provided a secondary, haplotype-aware confirmation of the `qpAdm` transversion-based estimates.

A.5 Population structure and clustering analyses

Sites assigned to European dog ancestry (`MOSAIC`) were masked to missing, then `EMU` PCA (Meisner *et al.*, 2021), `UMAP` embedding, and `ADMIXTURE` K-selection (with cross-

validation) were performed on the masked matrix. `EMU` is specifically optimised to handle high rates of random and non-random missingness. Dimensionality reduction was conducted on the top 10 eigenvectors using `UMAP` to visualise distinct genetic clusters. Unsupervised clustering was executed via `ADMIXTURE` (v1.3.0) on the masked dataset, calculating cross-validation errors to determine the optimal number of ancestral components (K).

A.6 Review of legacy testing methods

For benchmarking purposes, legacy STR results were obtained where available for CISS pilot specimens. STR panels typically examine 23 microsatellite loci and assign ancestry using clustering against modern reference panels. SNP-array data and `FastStructure`/`ADMIXTURE` q-values from Cairns et al. (2023) were compared against `qpAdm` estimates for the same individuals where overlap existed. Neither legacy method uses pre-colonial reference material.

Appendix B: Downsampling Validation and Cost-Benefit Context

A primary objective of this pilot project was to identify a testing methodology that is not only highly accurate but also economically viable for large-scale application by wildlife management agencies like CISS and PIRSA.

B.1 Computational subsampling experiment

To evaluate the minimum data requirements necessary for robust `qpAdm` ancestry modelling, we conducted a rigorous downsampling analysis (Ravishankar *et al.*, in press). For a baseline set of unrelated WGS dingo individuals, we generated independent pseudohaploid replicates. Within each replicate, the total number of whole-genome SNPs was randomly mathematically subsampled to progressively lower thresholds:

- 100,000 SNPs
- 50,000 SNPs
- 10,000 SNPs
- 5,000 SNPs (approximating DArTSeq density)
- 1,000 SNPs

B.2 Results of downsampling

The analysis demonstrated that `qpAdm` models retained high accuracy and stable standard errors down to 10,000 haploid transversions. At this threshold, the algorithm reliably differentiated between multi-source models (admixed) and single-source models (unadmixed). However, below this threshold (e.g., at the ~5,000 SNP mark common to reduced-representation techniques like DArTSeq), the statistical resolution collapsed, causing the model to erroneously fail to detect known low-level dog admixture.

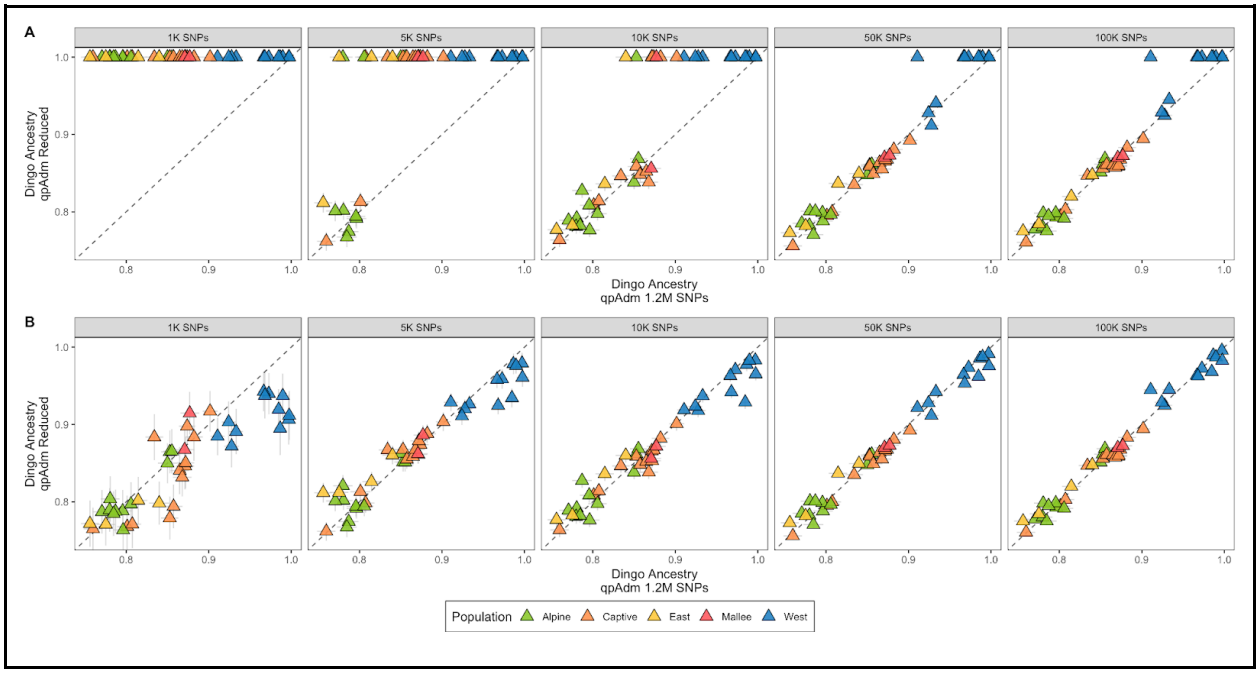

Figure B1. Reproduced with permission from Ravishankar *et al.* (in press). Comparing the effect of data quality. `qpAdm` estimates comparing dingo ancestry estimates using 1.2 million transversions heterozygous in "CoyoteCalifornia" individuals compared to reduced loci, as represented in the facets ranging from 1,000 to 100,000 pseudo-haplodised SNPs. Ten replicates were generated for each sample, varying both the pseudohaploid calls and the subset of SNPs used. Samples are coloured by population. (A) Both single- and two-source models were considered, with preference given to models with fewer sources when multiple models were accepted. (B) Only two-source models were considered.

B.3 Robustness to missing data: `qpAdm` vs local ancestry inference (`MOSAIC`)

While local ancestry inference methods like `MOSAIC` are highly effective for locating specific admixture tracts on high-quality, deep-coverage genomes, they exhibit severe vulnerabilities when applied to fragmented or low-coverage sequence data.

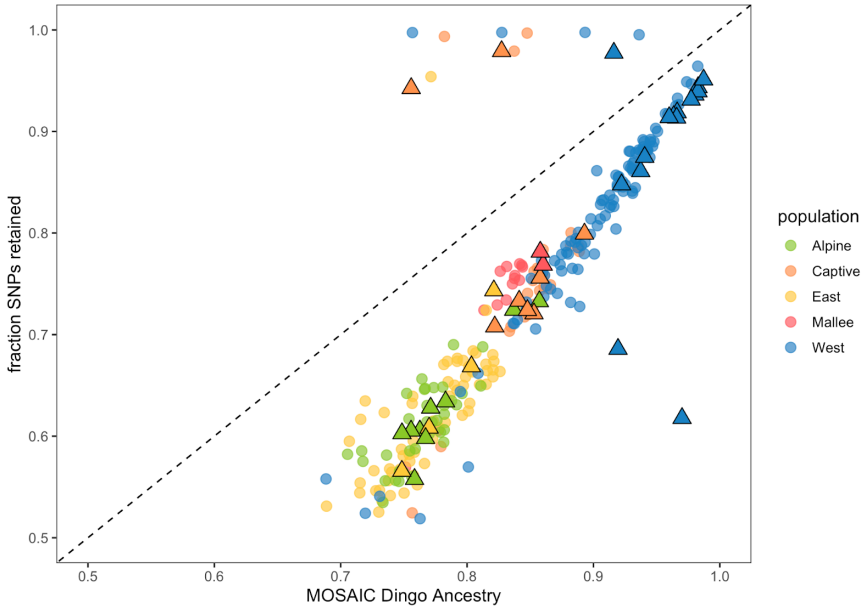

Figure B2. Relationship between `MOSAIC`-estimated dingo ancestry and the fraction of SNPs retained after masking. The dashed diagonal line represents the expected 1:1 relationship if masking is perfectly correlated with global ancestry estimates. Points falling significantly below the line indicate that a disproportionately high fraction of the genome is being discarded during the masking process.

As demonstrated in Figure B2, when mapping `MOSAIC`-estimated ancestry (x-axis) against the fraction of SNPs successfully retained after masking out flagged dog loci (y-axis), a significant systematic failure occurs. Most samples — particularly from the Alpine, East,

and Mallee clusters — fall drastically below the expected 1:1 parity line. This downward deviation indicates that `MOSAIC` struggles with lower coverage or unphased data, resulting in over-aggressive masking and the discarding of a disproportionately large fraction of valid genomic information.

By contrast, `qpAdm` evaluates aggregate allele frequencies (f4-statistics) across the genome and requires neither contiguous physical linkage nor haplotype phasing. Consequently, `qpAdm` is fundamentally immune to this specific mode of algorithmic failure, providing highly robust global admixture proportions even from sparse, pseudohaploid, or low-coverage inputs.

B.4 Cost-benefit implications

Current high-coverage whole-genome sequencing (median depth ~7X) entails significant outsourcing costs, typically ranging from $700 to $900 per sample (excluding extraction and library preparation labour). Conversely, commercial SNP-arrays, while cheaper, introduce ascertainment bias and, as shown in this report, lack the required unadmixed baseline integration.

Because our benchmarking validates that only ~10,000 informative transversion markers are required, agencies can now confidently transition to low-pass whole-genome sequencing (skimming). Low-pass sequencing provides the exact genomic coverage necessary to reach the 10,000 SNP threshold without the prohibitive costs of deep-sequencing.

- Estimated high-coverage WGS cost: ~$1,000 per sample (all-inclusive)
- Estimated low-pass WGS cost: ~$200 per sample

This transition facilitates an up to 10-fold cost saving per sample while maintaining the analytical rigour of the pre-colonial reference approach, making population-level genomic monitoring financially sustainable for state and federal programs.

Appendix C: Sample-Level Results

Individual-level results presented in attachment spreadsheet CISS-PIRSA Joint Final Report - Appendix C.xlsx