Evaluating kinship estimation methods for reduced-representation SNP data in non-model species: supplemental document

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. POPULATION GENETICS OF TEST SPECIES

For each species we conducted STRUCTURE ancestry analysis followed by StructureHarvester Evanno assessment to determine the optimal number of ancestral populations. We also conducted PCA and F_{ST} analyses to get an overall picture of the population dynamics of the test species.

A. *Acacia terminalis*
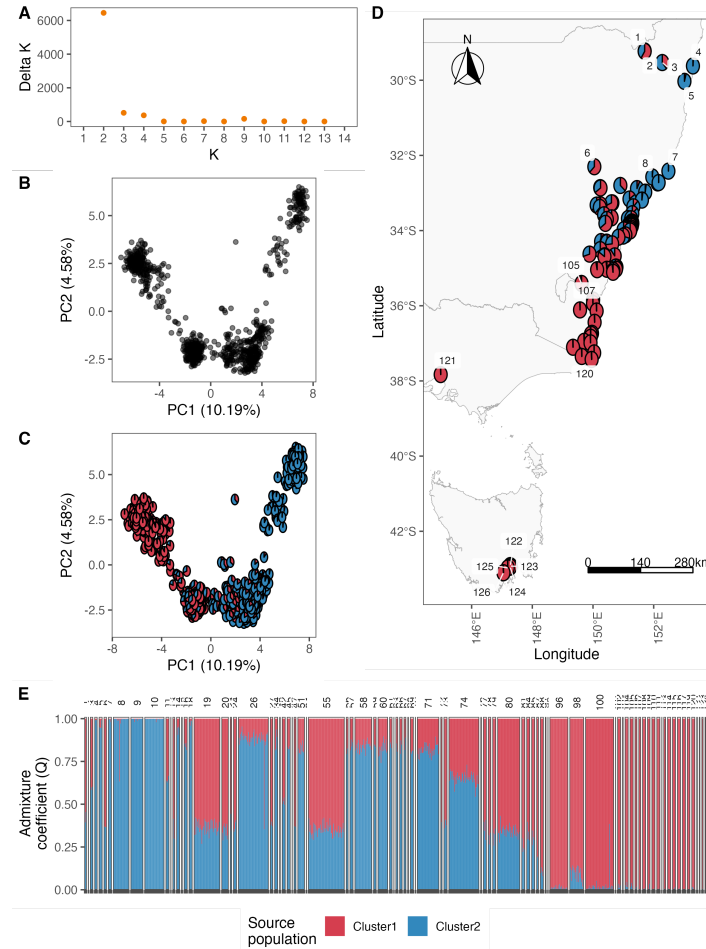
Figure S1. A) Delta K values, derived through the Evanno method, are presented for each evaluated STRUCTURE k value. B) Principal Component Analysis (PCA) of genotypes. C) PCA of genotypes with data points representing the relative proportions of ancestral populations as determined by STRUCTURE. D) Geographical representation of the average ancestry for each site, with select sites labelled. E) Ancestry proportions for each individual, determined by STRUCTURE for the optimal K with individuals grouped by their respective sites.
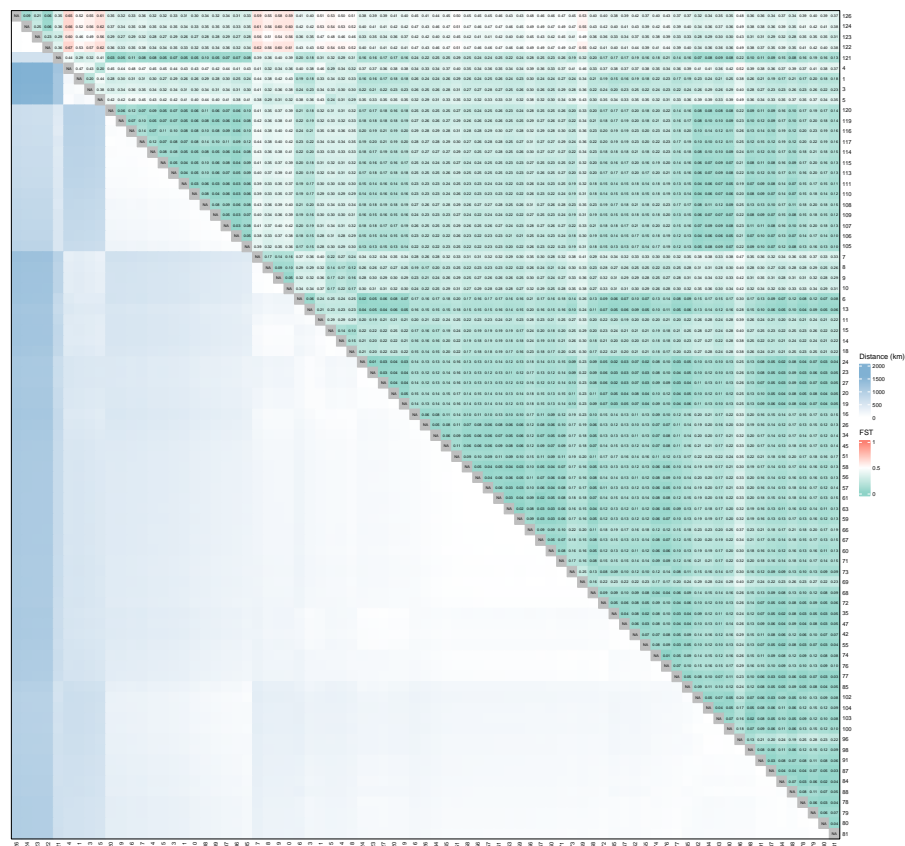
Figure S2. Combined heatmap illustrating geographic distances (in km, lower left half) and F_{ST} values (upper right half) among all sites. The Site ID numbers are positioned adjacent to each column and row.

Figure S3. Geographic distances (in km) and F_{ST} values between site pairs, with a delineating dotted line at $F_{ST} = 0.3$. The Mantel statistic, computed using Pearson correlation, is provided above the plot.

4

B. *Acacia suaveolens*

Figure S4. A) Delta K values, derived through the Evanno method, are presented for each evaluated STRUCTURE k value. B) Principal Component Analysis (PCA) of genotypes. C) PCA of genotypes with data points representing the relative proportions of ancestral populations as determined by STRUCTURE. D) Geographical representation of the average ancestry for each site, with select sites labelled. E) Ancestry proportions for each individual, determined by STRUCTURE for the optimal K with individuals grouped by their respective sites.

Figure S5. Combined heatmap illustrating geographic distances (in km, lower left half) and F_{ST} values (upper right half) among all sites. The Site ID numbers are positioned adjacent to each column and row.

Figure S6. Geographic distances (in km) and F_{ST} values between site pairs, with a delineating dotted line at $F_{ST} = 0.3$. The Mantel statistic, computed using Pearson correlation, is provided above the plot.

C. *Banksia aemula*
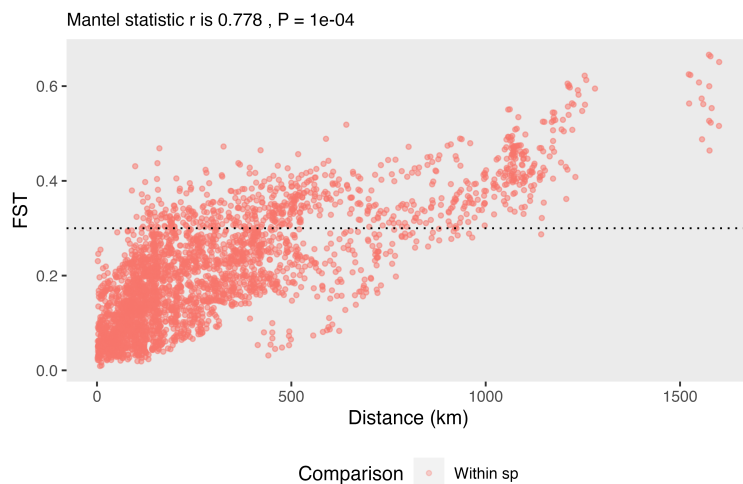

Figure S7. A) Delta K values, derived through the Evanno method, are presented for each evaluated STRUCTURE k value. B) Principal Component Analysis (PCA) of genotypes. C) PCA of genotypes with data points representing the relative proportions of ancestral populations as determined by STRUCTURE. D) Geographical representation of the average ancestry for each site, with select sites labelled. E) Ancestry proportions for each individual, determined by STRUCTURE for the optimal K with individuals grouped by their respective sites.
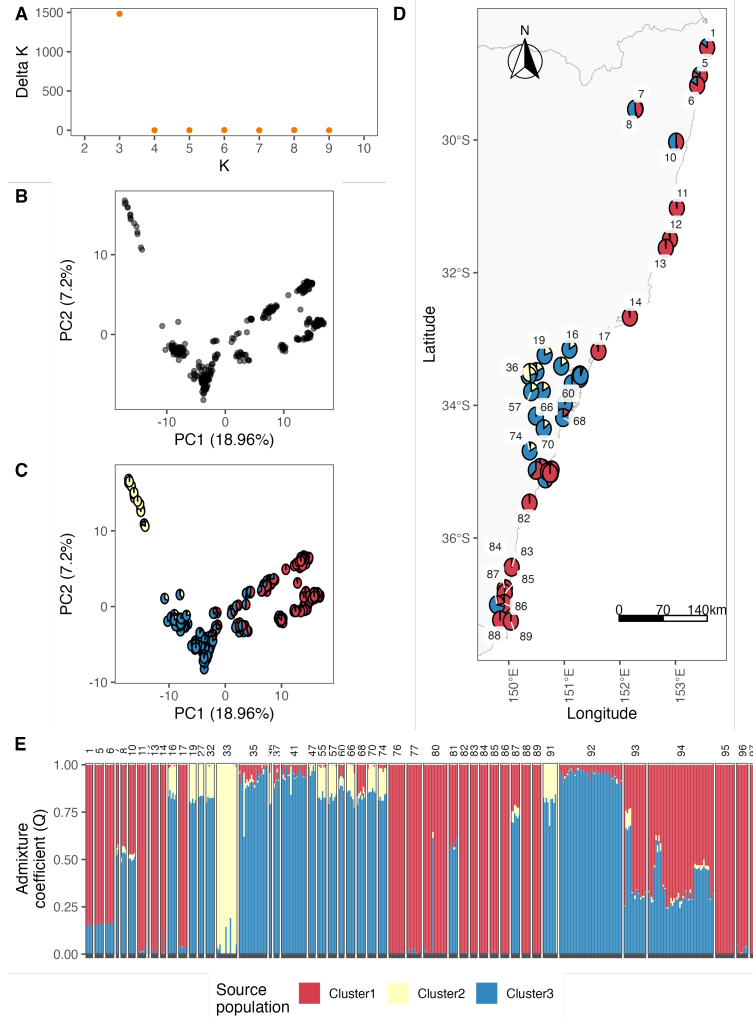
Figure S8. Combined heatmap illustrating geographic distances (in km, lower left half) and F_{ST} values (upper right half) among all sites. The Site ID numbers are positioned adjacent to each column and row.
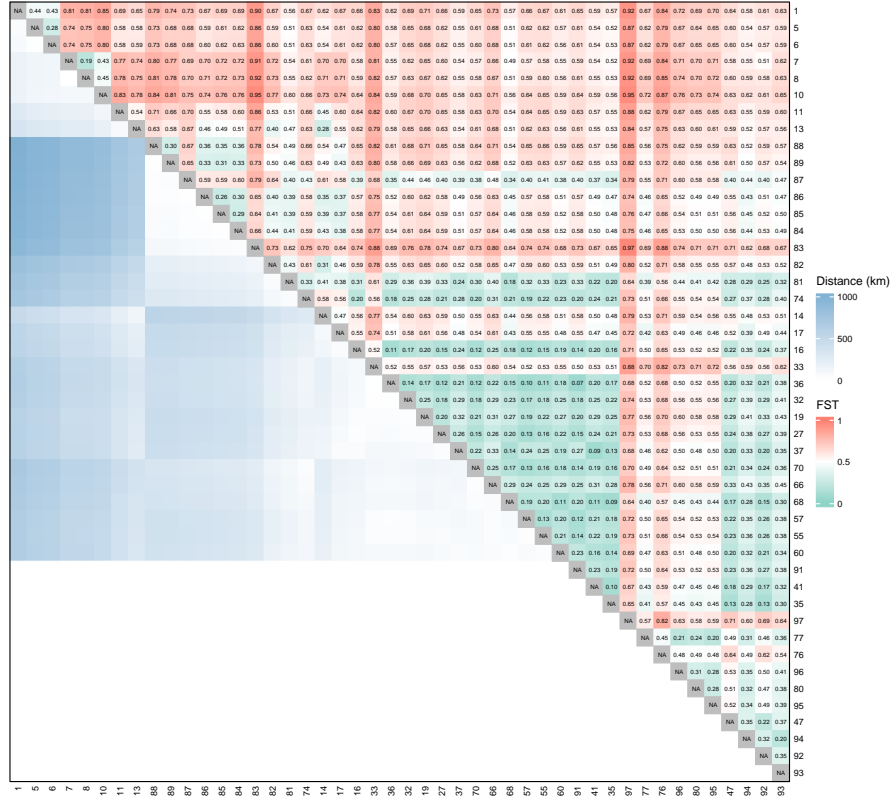

Figure S9. Geographic distances (in km) and F_{ST} values between site pairs, with a delineating dotted line at $F_{ST} = 0.3$. The Mantel statistic, computed using Pearson correlation, is provided above the plot.

D. *Banksia serrata*

Figure S10. A) Delta K values, derived through the Evanno method, are presented for each evaluated STRUCTURE k value. B) Principal Component Analysis (PCA) of genotypes. C) PCA of genotypes with data points representing the relative proportions of ancestral populations as determined by STRUCTURE. D) Geographical representation of the average ancestry for each site, with select sites labelled. E) Ancestry proportions for each individual, determined by STRUCTURE for the optimal K with individuals grouped by their respective sites.

9

Figure S11. Combined heatmap illustrating geographic distances (in km, lower left half) and F_{ST} values (upper right half) among all sites. The Site ID numbers are positioned adjacent to each column and row.
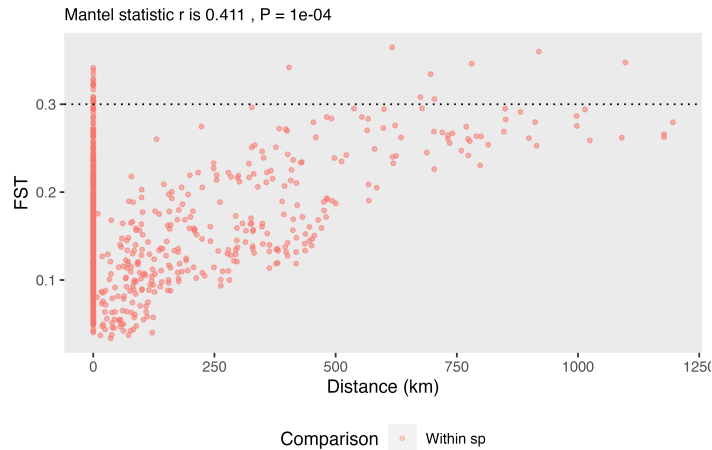
Figure S12. Geographic distances (in km) and F_{ST} values between site pairs, with a delineating dotted line at $F_{ST} = 0.3$. The Mantel statistic, computed using Pearson correlation, is provided above the plot.
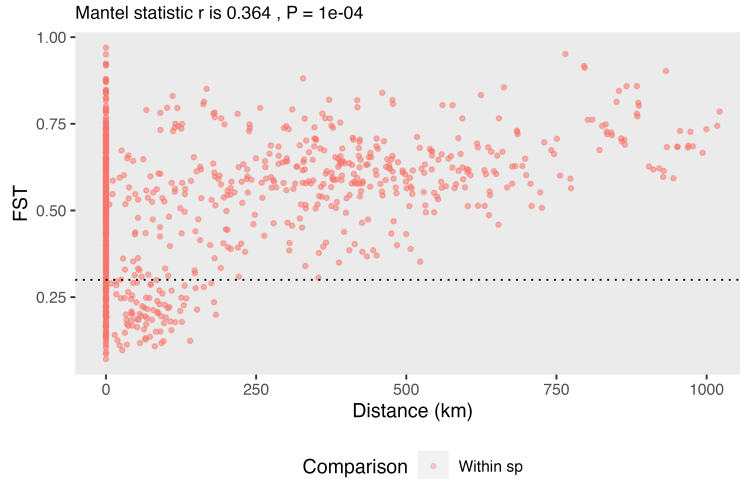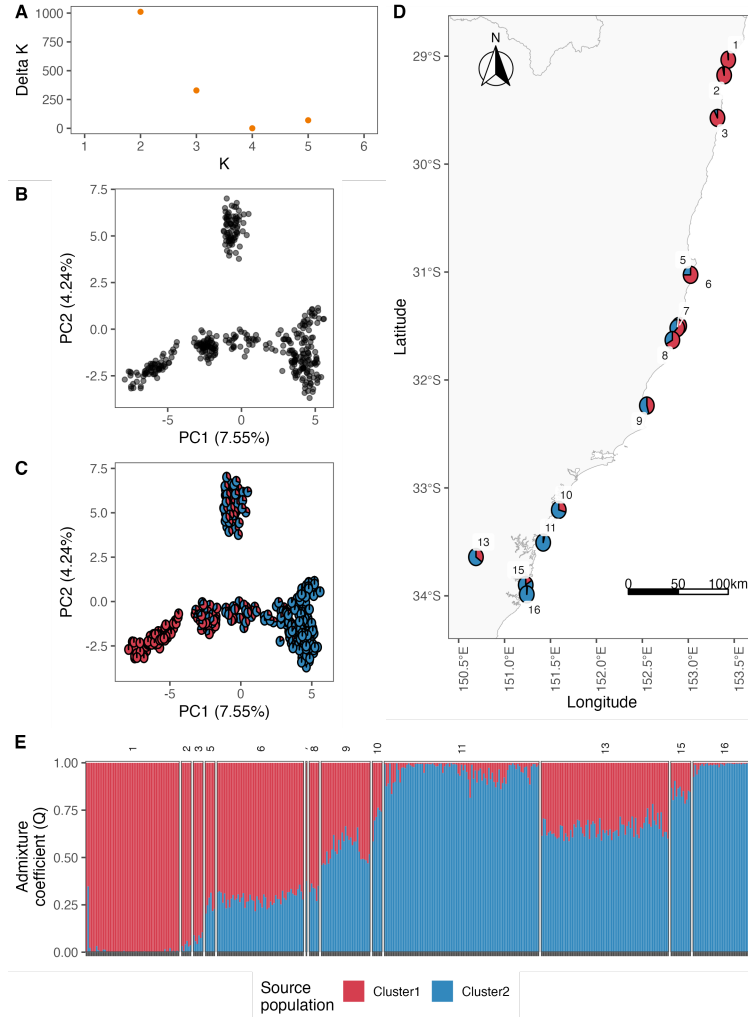
E. *Hakea sericea*
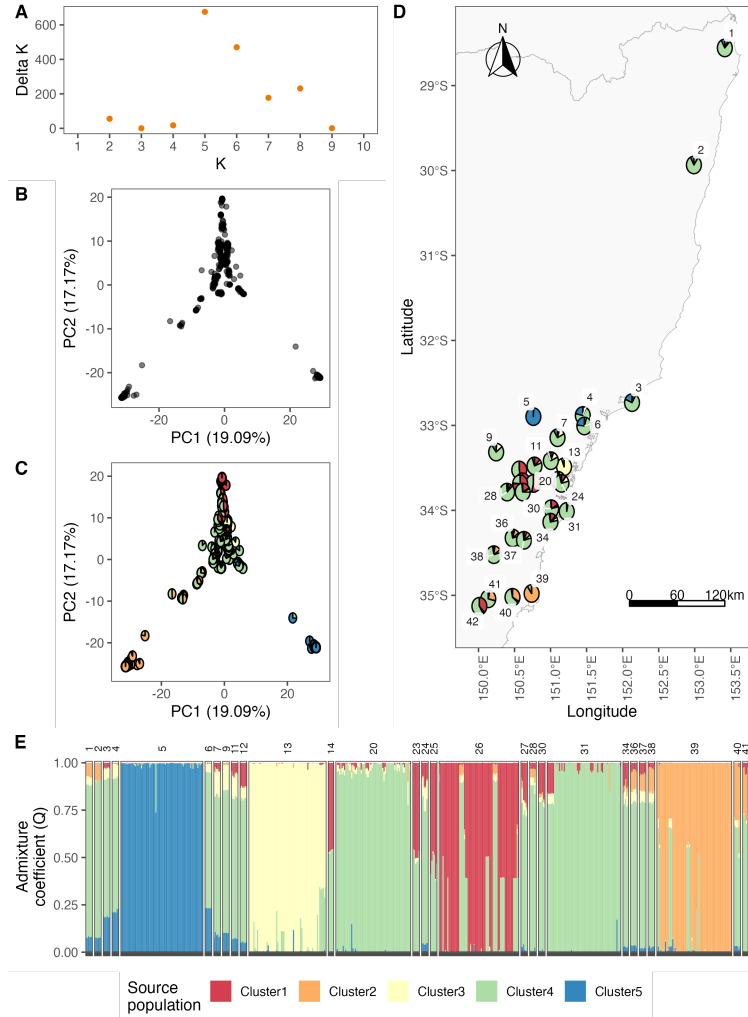
Figure S13. A) Delta K values, derived through the Evanno method, are presented for each evaluated STRUCTURE k value. B) Principal Component Analysis (PCA) of genotypes. C) PCA of genotypes with data points representing the relative proportions of ancestral populations as determined by STRUCTURE. D) Geographical representation of the average ancestry for each site, with select sites labelled. E) Ancestry proportions for each individual, determined by STRUCTURE for the optimal K with individuals grouped by their respective sites.
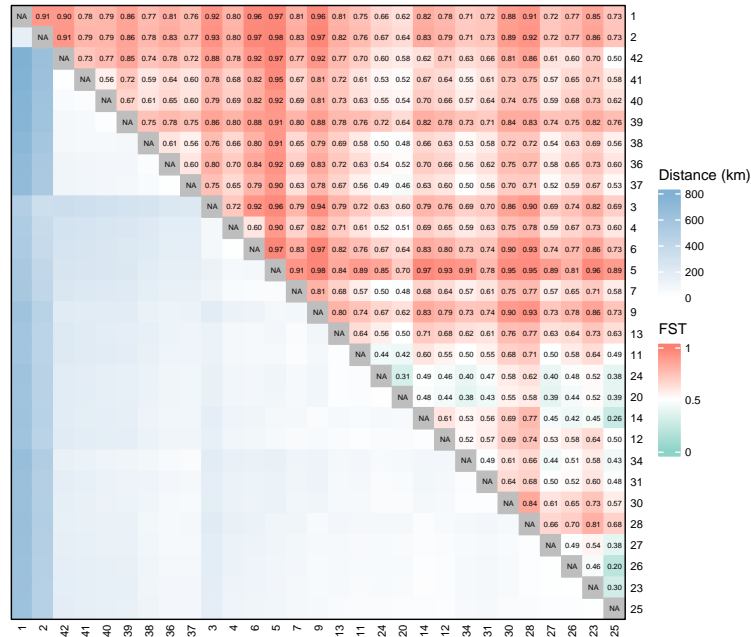
Figure S14. Combined heatmap illustrating geographic distances (in km, lower left half) and F_{ST} values (upper right half) among all sites. The Site ID numbers are positioned adjacent to each column and row.
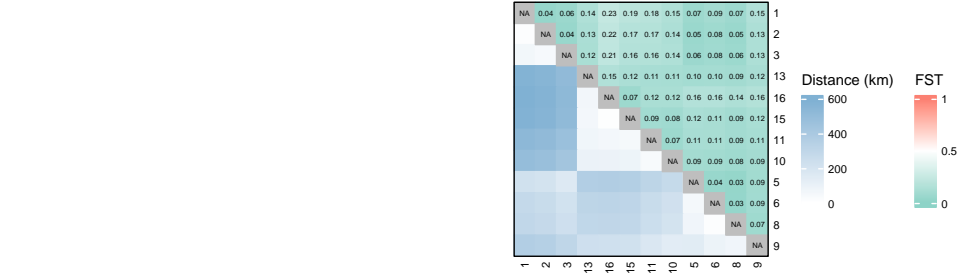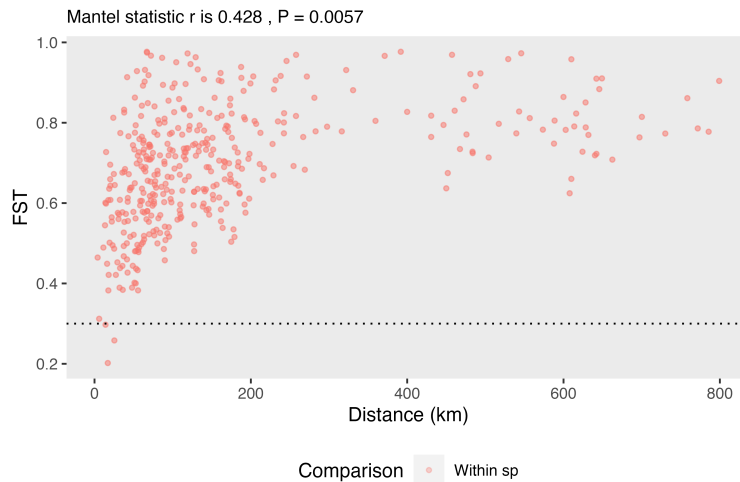

Figure S15. Geographic distances (in km) and F_{ST} values between site pairs, with a delineating dotted line at $F_{ST} = 0.3$. The Mantel statistic, computed using Pearson correlation, is provided above the plot.
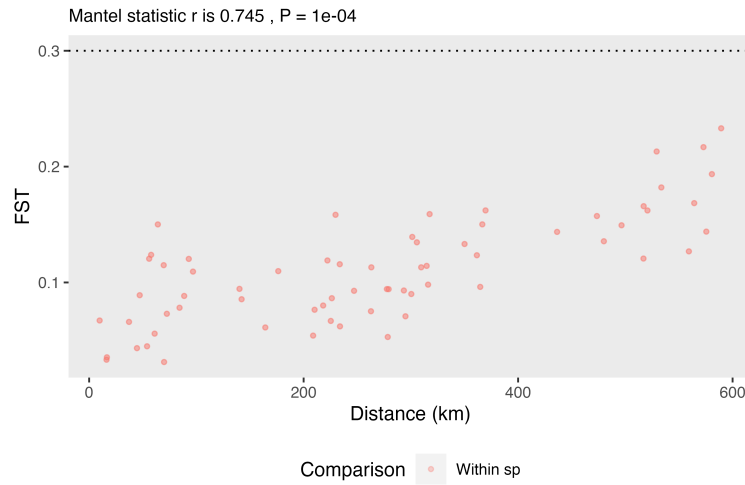
F. *Hakea teretifolia*
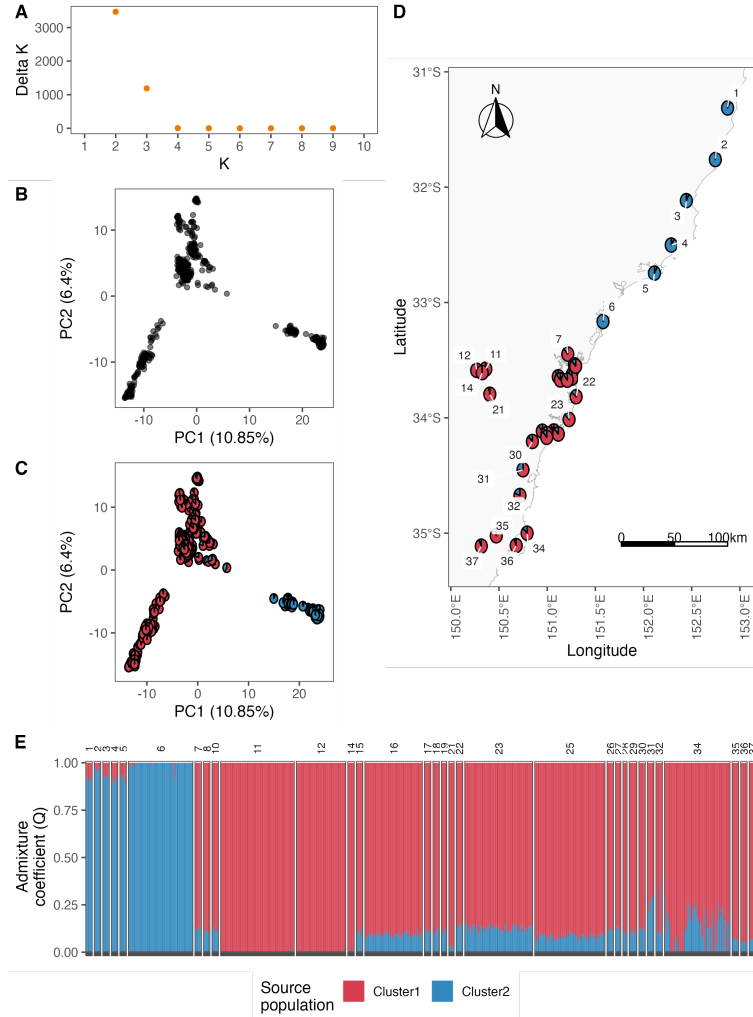
Figure S16. A) Delta K values, derived through the Evanno method, are presented for each evaluated STRUCTURE k value. B) Principal Component Analysis (PCA) of genotypes. C) PCA of genotypes with data points representing the relative proportions of ancestral populations as determined by STRUCTURE. D) Geographical representation of the average ancestry for each site, with select sites labelled. E) Ancestry proportions for each individual, determined by STRUCTURE for the optimal K with individuals grouped by their respective sites.
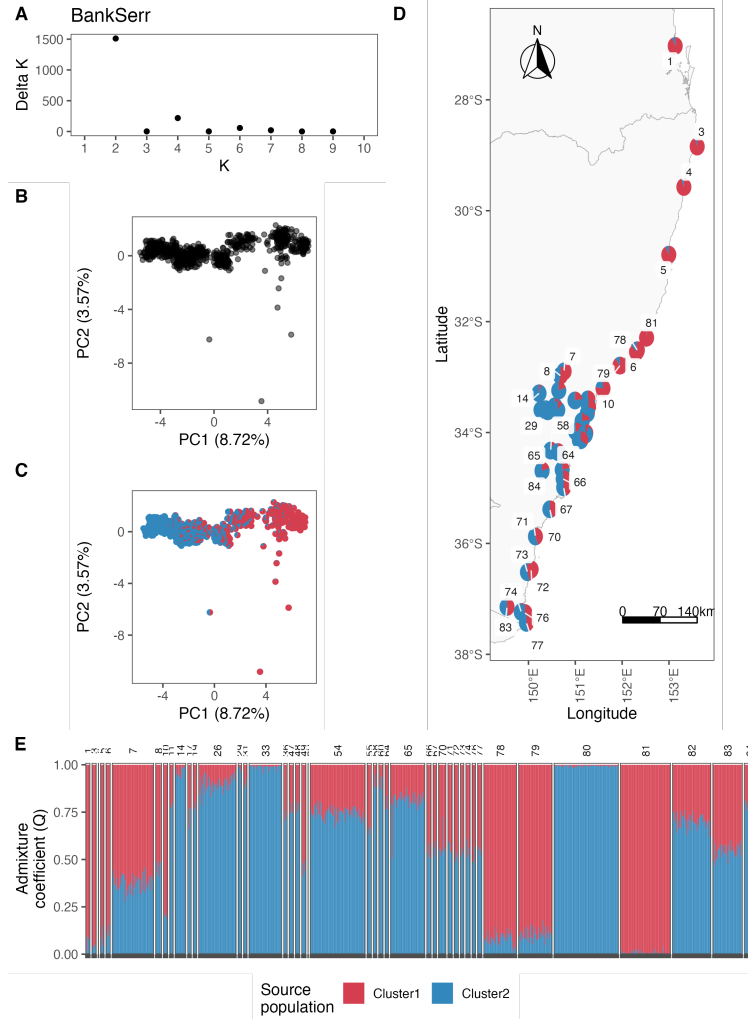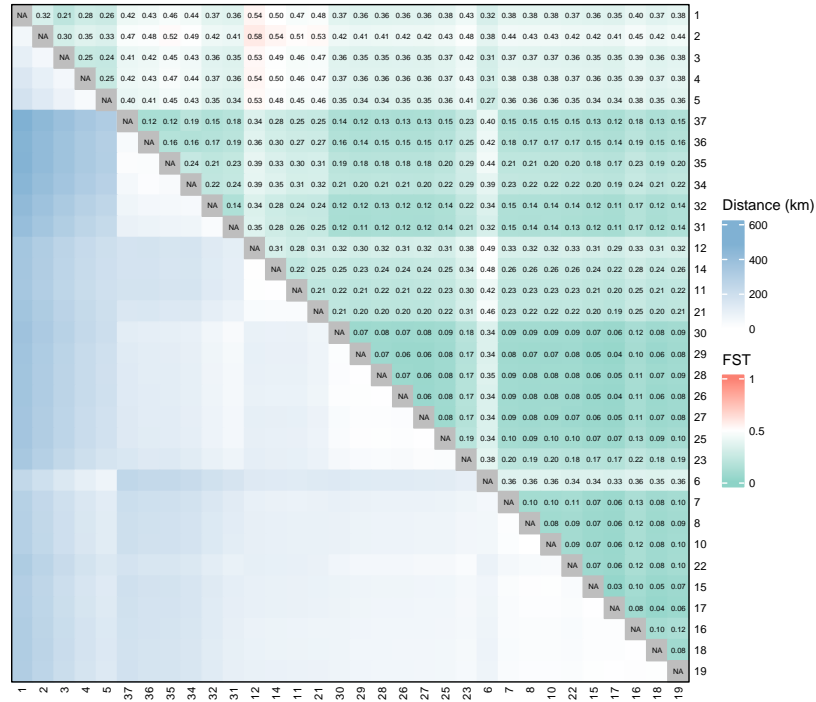
Figure S17. Combined heatmap illustrating geographic distances (in km, lower left half) and F_{ST} values (upper right half) among all sites. The Site ID numbers are positioned adjacent to each column and row.
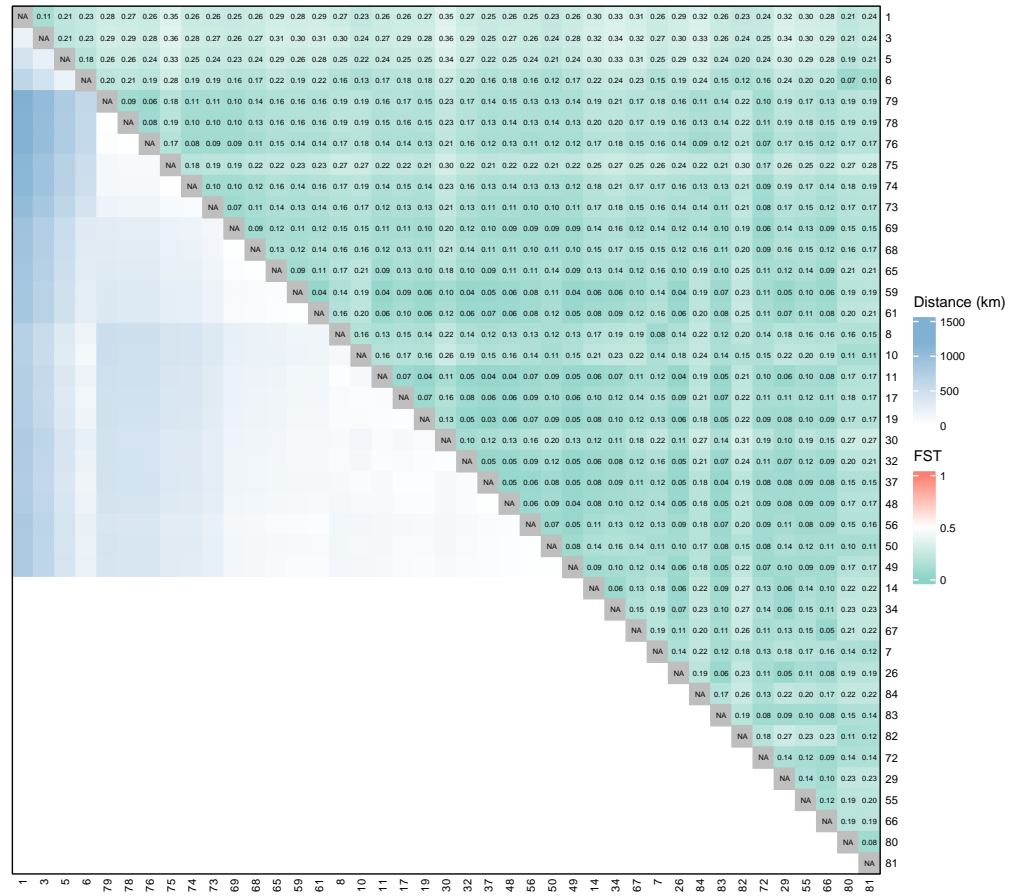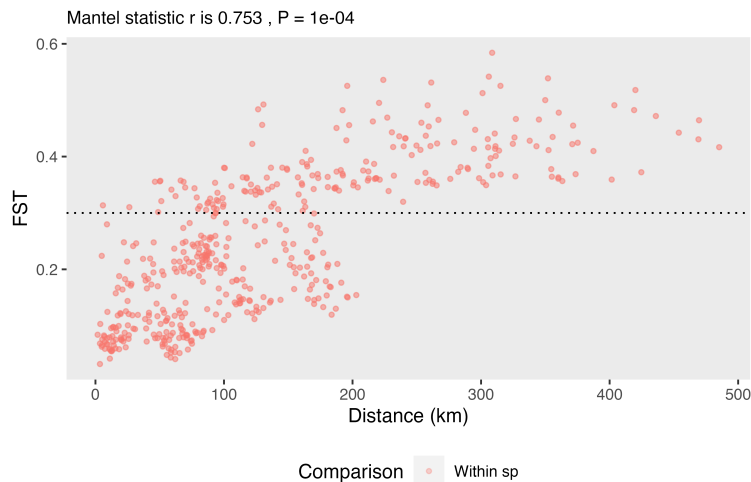

Figure S18. Geographic distances (in km) and F_{ST} values between site pairs, with a delineating dotted line at $F_{ST} = 0.3$. The Mantel statistic, computed using Pearson correlation, is provided above the plot.

14

2. KINSHIP CRITERIA

Table S1. Threshold kinship values used to classify zero-, first-, and second-degree relatives. These are based on the theoretical thresholds presented by (Manichaikul et al., 2010), which were also used by (Ramstetter et al., 2017). We modified the thresholds to work with kinship estimators where kinship could exceed the bounds of 0–0.5 (i.e., KING, PC-Relate, and Goudet).

Degree	Min kin	Max kin
0	$1/2^{\frac{3}{2}}$	0.5 (modified to be inf)
1	$1/2^{\frac{5}{2}}$	$1/2^{\frac{3}{2}}$
2	$1/2^{\frac{7}{2}}$	$1/2^{\frac{5}{2}}$
3	$1/2^{\frac{9}{2}}$	$1/2^{\frac{7}{2}}$
4	$1/2^{\frac{11}{2}}$	$1/2^{\frac{9}{2}}$
Unrelated	0 (modified to be -inf)	$1/2^{\frac{11}{2}}$

3. FILTERING PARAMETERS FOR DIFFERENT METHODS

For each method, we selected a specific set of filtering parameters (including maximum locus missingness, maximum sample missingness, and minimum minor allele frequency) to scrutinize the results more closely. These values were chosen based on their ability to yield consistently high precision and sensitivity across all species for each method. The following presentation offers a detailed examination of the data, which is also depicted in Figure 2, aiming to provide a more focused view of where the optimal filtering parameters are positioned within the outcome distribution.
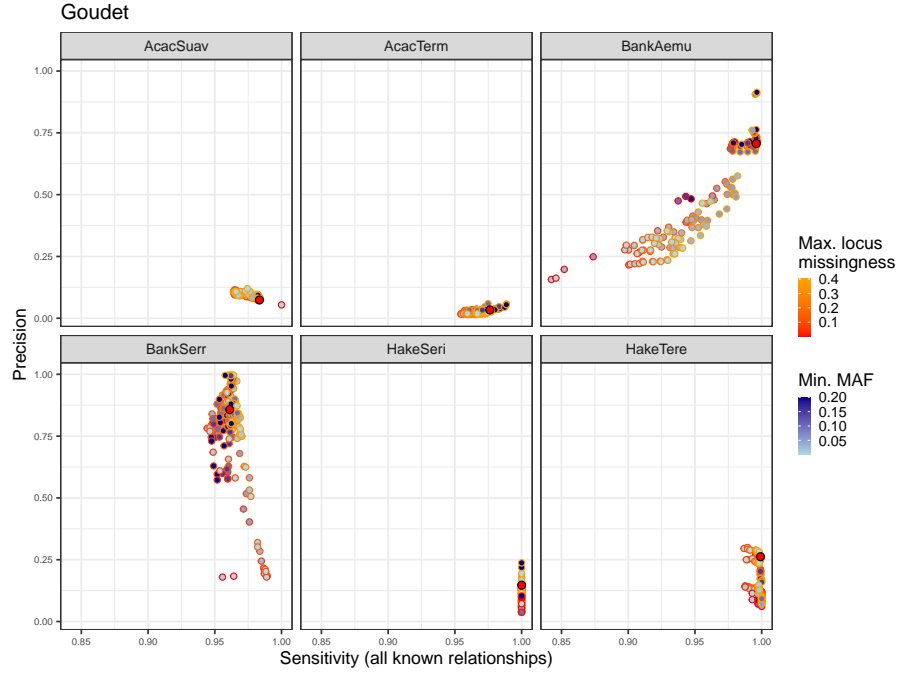
Figure S19. Scatter plot of overall precision vs overall sensitivity of Goudet's beta dosage kinship method for data filtered using different parameters. The outline colour of points scales by maximum locus missingness and the fill colour indicates minimum MAF.
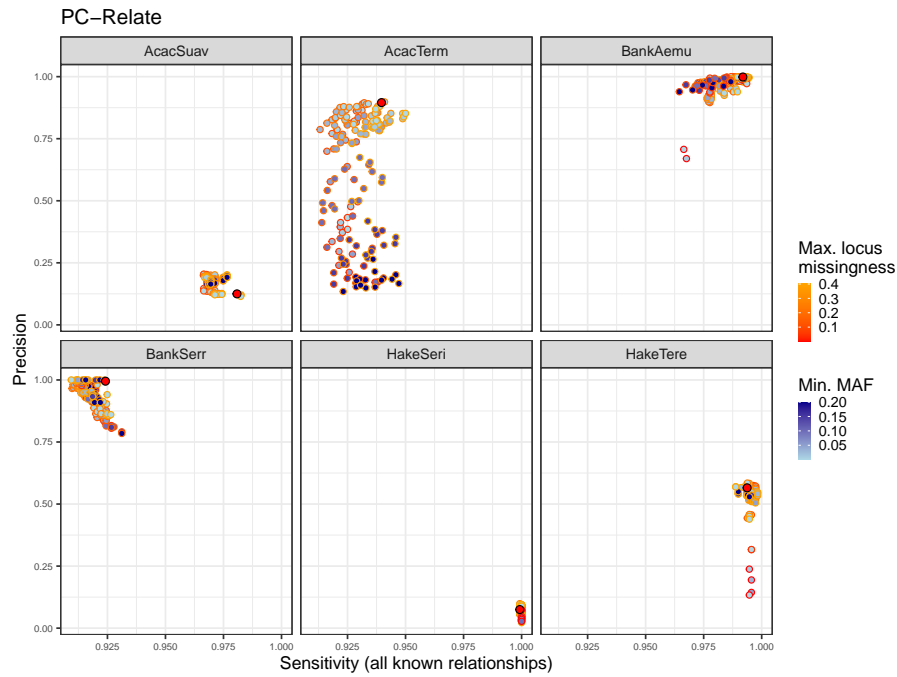
Figure S20. Scatter plot of overall precision vs overall sensitivity of PC-Relate kinship method for data filtered using different parameters. The outline colour of points scales by maximum locus missingness and the fill colour indicates minimum MAF.
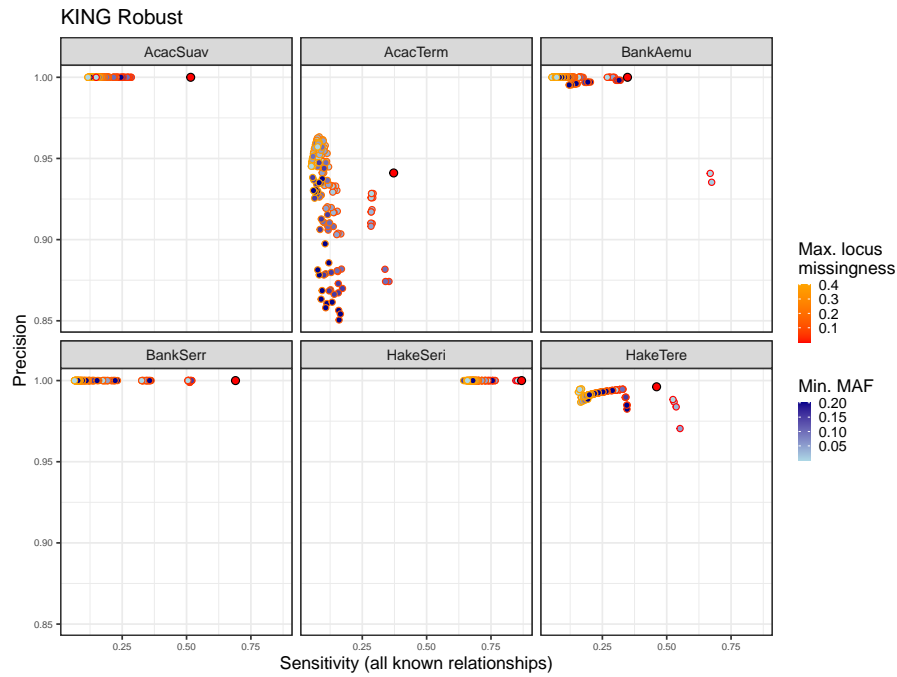
Figure S21. Scatter plot of overall precision vs overall sensitivity of KING-Homo kinship method for data filtered using different parameters. The outline colour of points scales by maximum locus missingness and the fill colour indicates minimum MAF.
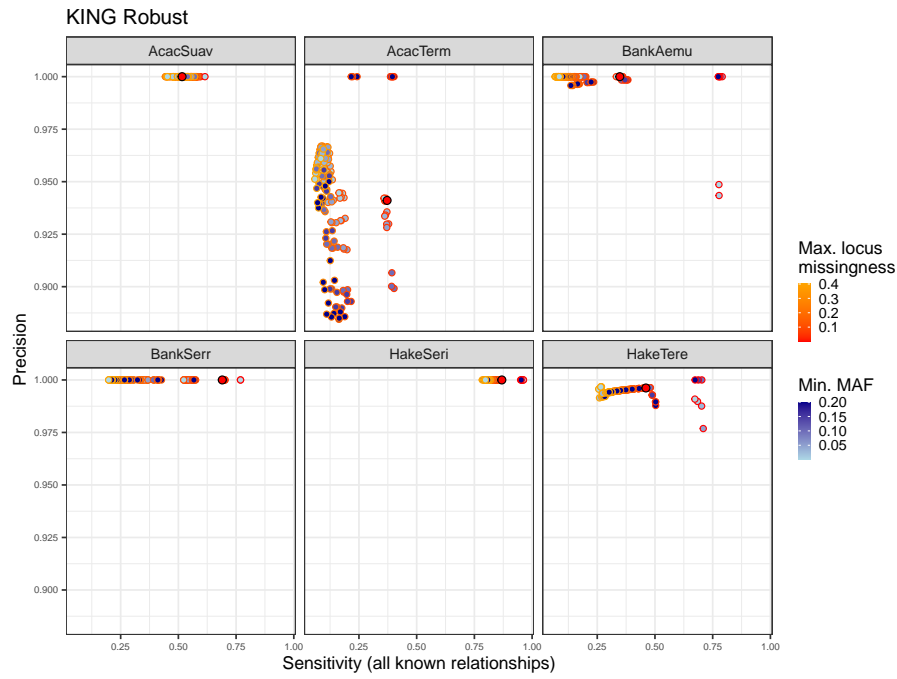
Figure S22. Scatter plot of overall precision vs overall sensitivity of KING-Robust kinship method for data filtered using different parameters. The outline colour of points scales by maximum locus missingness and the fill colour indicates minimum MAF.
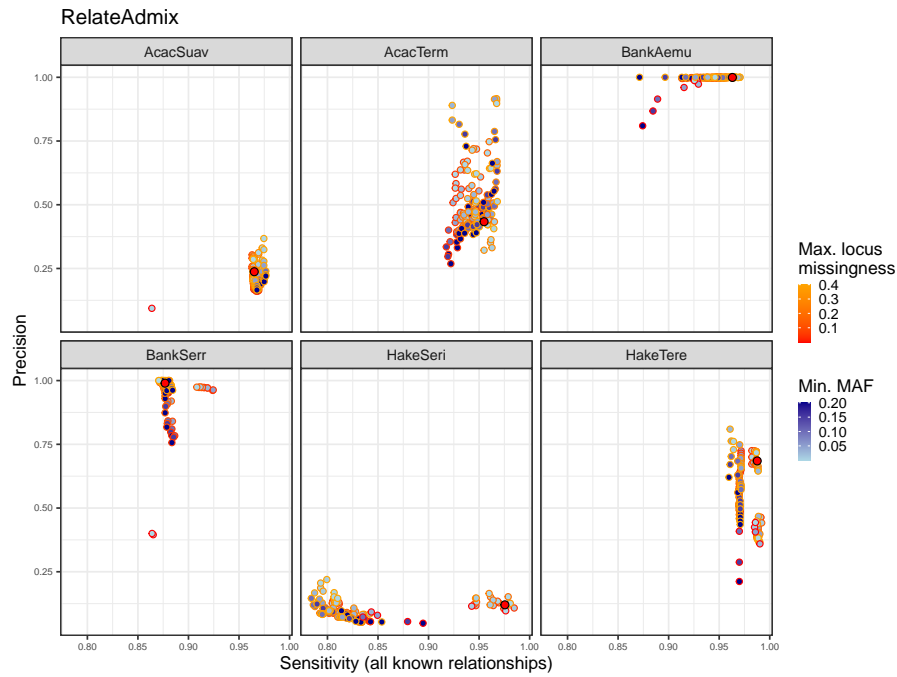
Figure S23. Scatter plot of overall precision vs overall sensitivity of RelateAdmix kinship method for data filtered using different parameters. The outline colour of points scales by maximum locus missingness and the fill colour indicates minimum MAF.
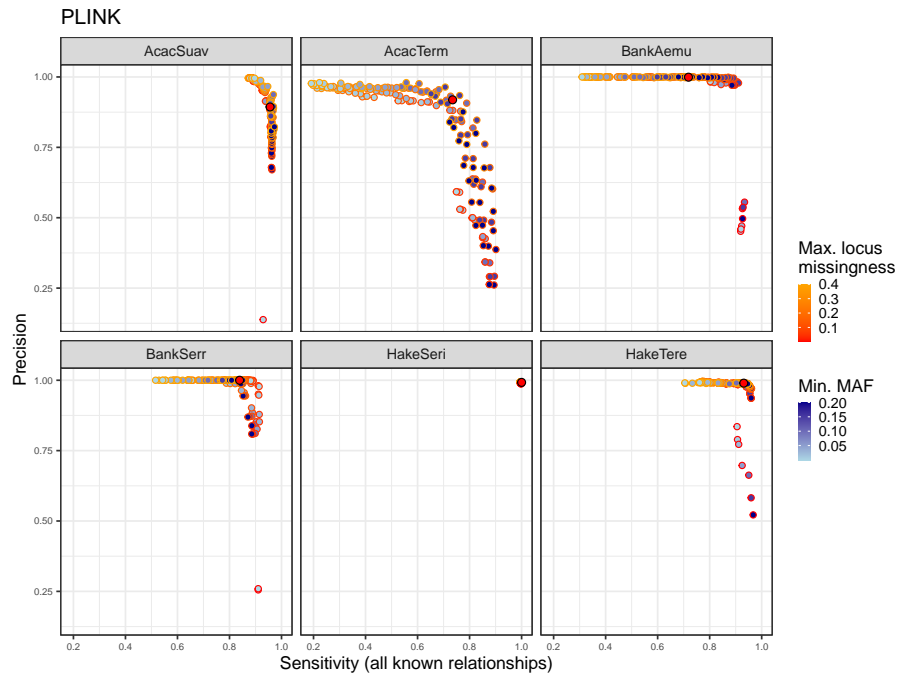
Figure S24. Scatter plot of overall precision vs overall sensitivity of PLINK kinship method for data filtered using different parameters. The outline colour of points scales by maximum locus missingness and the fill colour indicates minimum MAF.

Figure S25. Scatter plot of overall precision vs overall sensitivity of PLINK kinship method for *Acacia terminalis* data filtered using different parameters. The outline colour of points scales by maximum locus missingness and the fill colour indicates minimum MAF.

Table S2. Results of optimal filtering scenarios including sensititivity (sens) and precision (PPV) for all known relationships, parent off-spring relationships (PO), and sibship (Sib). The number of loci and samples (N) after filtering are also included.

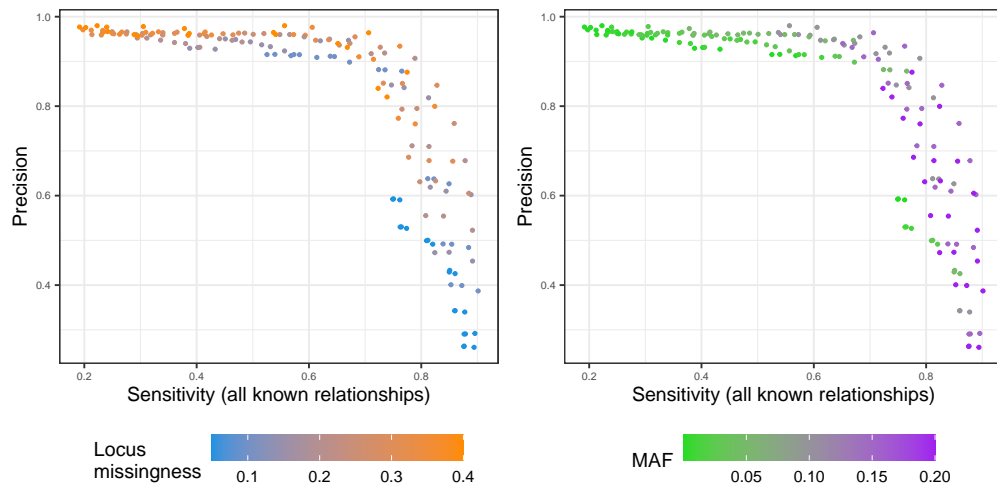| Method	Species	Sample miss	Locus miss	MAF	Sens (All)	Sens (PO)	Sens (Sib)	PPV (All)	PPV (PO)	PPV (Sib)	Loci	N
Goudet	AcacSuav	0.2	0.3	0.2	0.98	1.00	0.98	0.07	0.01	0.06	2149	392
Goudet	AcacTerm	0.2	0.3	0.2	0.98	0.96	0.98	0.03	0.01	0.03	780	761
Goudet	BankAemu	0.2	0.3	0.2	1.00	1.00	1.00	0.71	0.11	0.60	1182	386
Goudet	BankSerr	0.2	0.3	0.2	0.96	0.92	0.98	0.86	0.21	0.64	1078	713
Goudet	HakeSeri	0.2	0.3	0.2	1.00	1.00	1.00	0.15	0.03	0.11	4257	520
Goudet	HakeTere	0.2	0.3	0.2	1.00	1.00	1.00	0.26	0.07	0.19	3770	533
KING Homo	AcacSuav	0.1	0.05	0.01	0.24	0.31	0.23	1.00	0.20	0.80	5907	393
KING Homo	AcacTerm	0.1	0.05	0.01	0.29	0.23	0.31	0.93	0.17	0.76	2122	805
KING Homo	BankAemu	0.1	0.05	0.01	0.28	0.49	0.24	1.00	0.27	0.73	3049	386
KING Homo	BankSerr	0.1	0.05	0.01	0.51	0.81	0.41	1.00	0.42	0.58	1640	751
KING Homo	HakeSeri	0.1	0.05	0.01	0.75	0.78	0.74	1.00	0.24	0.76	11058	519
KING Homo	HakeTere	0.1	0.05	0.01	0.31	0.45	0.25	0.99	0.39	0.60	15465	533
KING Robust	AcacSuav	0.1	0.05	0.01	0.52	0.68	0.47	1.00	0.26	0.74	5907	393
KING Robust	AcacTerm	0.1	0.05	0.01	0.37	0.29	0.40	0.94	0.17	0.77	2122	805
KING Robust	BankAemu	0.1	0.05	0.01	0.35	0.66	0.29	1.00	0.30	0.70	3049	386
KING Robust	BankSerr	0.1	0.05	0.01	0.69	0.94	0.59	1.00	0.39	0.61	1640	751
KING Robust	HakeSeri	0.1	0.05	0.01	0.87	0.91	0.86	1.00	0.24	0.76	11058	519
KING Robust	HakeTere	0.1	0.05	0.01	0.46	0.69	0.38	1.00	0.40	0.60	15465	533
PC-Relate	AcacSuav	0.1	0.3	0	0.98	0.91	0.99	0.12	0.01	0.11	25134	368
PC-Relate	AcacTerm	0.1	0.3	0	0.94	0.99	0.93	0.90	0.09	0.80	15889	652
PC-Relate	BankAemu	0.1	0.3	0	0.99	1.00	0.99	1.00	0.15	0.85	6651	359
PC-Relate	BankSerr	0.1	0.3	0	0.92	0.91	0.93	1.00	0.25	0.74	5773	635
PC-Relate	HakeSeri	0.1	0.3	0	1.00	1.00	1.00	0.07	0.02	0.06	25057	470
PC-Relate	HakeTere	0.1	0.3	0	0.99	0.98	1.00	0.57	0.15	0.42	40251	521
PLINK	AcacSuav	0.2	0.2	0.1	0.96	0.92	0.96	0.89	0.13	0.76	3246	394
PLINK	AcacTerm	0.2	0.2	0.1	0.73	0.78	0.72	0.92	0.20	0.72	1340	786
PLINK	BankAemu	0.2	0.2	0.1	0.72	0.95	0.68	1.00	0.21	0.79	1727	387
PLINK	BankSerr	0.2	0.2	0.1	0.84	0.90	0.82	1.00	0.28	0.72	1467	739
PLINK	HakeSeri	0.2	0.2	0.1	1.00	1.00	1.00	0.99	0.24	0.76	7666	522
PLINK	HakeTere	0.2	0.2	0.1	0.93	0.98	0.91	0.99	0.28	0.71	7010	533
RelateAdmix	AcacSuav	0.2	0.3	0.01	0.97	0.92	0.97	0.24	0.04	0.20	13677	392
RelateAdmix	AcacTerm	0.2	0.3	0.01	0.95	0.93	0.96	0.43	0.08	0.36	6663	761
RelateAdmix	BankAemu	0.2	0.3	0.01	0.96	0.99	0.96	1.00	0.16	0.84	5360	386
RelateAdmix	BankSerr	0.2	0.3	0.01	0.88	0.90	0.87	0.99	0.26	0.73	3796	713
RelateAdmix	HakeSeri	0.2	0.3	0.01	0.98	0.94	0.99	0.12	0.03	0.09	18602	520
RelateAdmix	HakeTere	0.2	0.3	0.01	0.99	0.96	1.00	0.68	0.18	0.51	25695	533

4. RANDOM FOREST VARIABLE IMPORTANCE
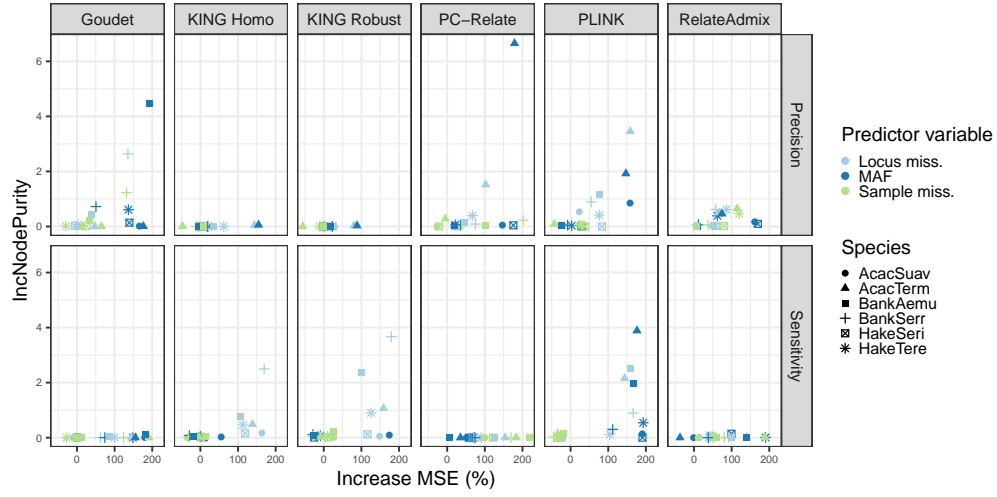
Figure S26. Increase in mean squared error (%) and increase in node purity for each predictor variable for Random Forest models. These measures indicate the importance of the variable to the predictive power of the model i.e. have the greatest impact on the response variable (precision or sensitivity).

5. RANDOM FOREST VARIANCE EXPLAINED

Table S3. Variance explained (%) by random forest models for each method, species, and response variable (overall sensitivity or overall precision). Variance explained represents the proportion of variance in the target variable captured by the model, quantifying its goodness of fit by comparing predicted values with observed values. Higher values indicate better predictive performance. Empty cells occur when a model could not be fitted (i.e., all outcomes were the same).

|		Variance explained (%)	
Method	**Species**	**Sensitivity model**	**Precision model**
Goudet	*Hakea teretifolia*	81.44	78.25
	Hakea sericea		69.61
	Acacia suaveolens	79.51	77.73
	Acacia terminalis	84.64	84.32
	Banksia serrata	65.61	76.65
	Banksia aemula	86.75	89.96
KING Homo	*Hakea teretifolia*	78.91	31.35
	Hakea sericea	80.19	
	Acacia suaveolens	83.31	
	Acacia terminalis	87.41	78.88
	Banksia serrata	89.57	7.69
	Banksia aemula	74.74	54.58
KING Robust	*Hakea teretifolia*	82.87	10.05
	Hakea sericea	81.9	
	Acacia suaveolens	79.09	
	Acacia terminalis	89.87	43.34
	Banksia serrata	89.94	
	Banksia aemula	84.06	23.47
PC-Relate	*Hakea teretifolia*	61.62	72.27
	Hakea sericea	59.78	83.53
	Acacia suaveolens	66.08	76.4
	Acacia terminalis	76.16	89.38
	Banksia serrata	52.46	66.74
	Banksia aemula	78.77	64.18
PLINK	*Hakea teretifolia*	89.31	76.78
	Hakea sericea	91.31	73.31
	Acacia suaveolens	90.38	56.79
	Acacia terminalis	87.78	85.94
	Banksia serrata	84.6	58.43
	Banksia aemula	88.38	84.89
RelateAdmix	*Hakea teretifolia*	77.3	47.57
	Hakea sericea	43.79	78.85
	Acacia suaveolens	9.68	69.86
	Acacia terminalis	74.51	53.7
	Banksia serrata	60.08	63.03
	Banksia aemula	67.14	43.99

6. ADMIXTURE PERFORMANCE ASSESSMENT

A. Sites used in admixture performance assessment

Table S4. Summary of the sites included in the qualitative assessment of kinship method performance in more complicated subsets of the data. The individuals included were either from highly admixed (individuals with the largest ancestral population proportion below 75%), from a large genetic structure group (over 90% ancestry from a majority structure group in the species), or a small structure group (over 90% ancestry from a minority structure group in the species). Only *Acacia suaveolens* and *Hakea sericea* individuals were used because these species had strong population structure.

Species	Site	Number of Families	Type
AcacSuav	93	2	Admixed
AcacSuav	94	8	Admixed
AcacSuav	92	6	Large structure group 1
AcacSuav	76	1	Large structure group 2
AcacSuav	77	1	Large structure group 2
AcacSuav	95	2	Large structure group 2
AcacSuav	96	2	Large structure group 2
AcacSuav	97	1	Large structure group 2
AcacSuav	33	2	Small structure group 1
HakeSeri	20	10	Large structure group 1
HakeSeri	31	9	Large structure group 1
HakeSeri	5	10	Small structure group 1
HakeSeri	39	10	Small structure group 2

**B. Estimated kinship in admixed or minority groups**
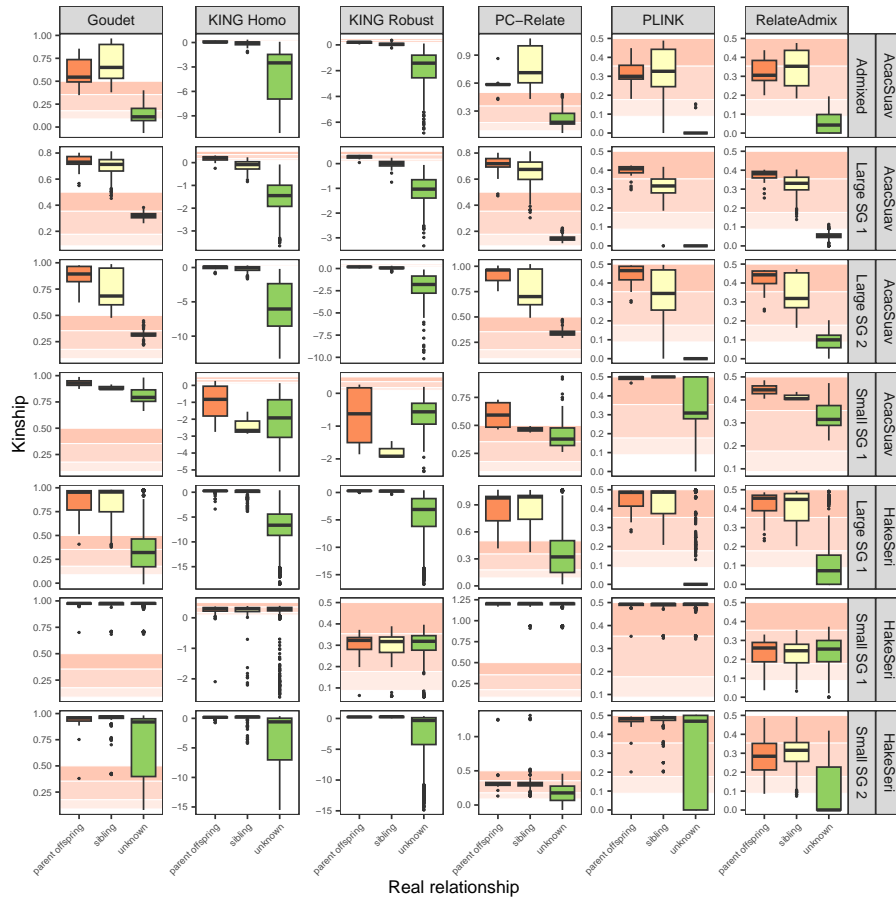
Figure S27. Boxplots of estimated kinship values among individuals within the same site applied to specific subsets of the *Acacia suaveolens* and *Hakea sericea* data. Rows represent distinct population subsets, categorized as highly admixed (individuals with the largest ancestral population proportion below 75%), from a large genetic structure group (SG; has over 90% ancestry from a majority structure group in the species), or a small structure group (has over 90% ancestry from a minority structure group in the species). Each panel presents estimated kinship values for known parent-offspring, known sibling, or unknown relationships within the same site. Background shading shows theoretical kinship levels for zero-degree relatives (dark pink, 0.354–0.5), first-degree relatives (medium pink, 0.177–0.354), and second-degree relatives (light pink, 0.088–0.177).

7. IMPACT OF LARGE FAMILIES ON KINSHIP OUTCOMES

Theoretically, utilizing a base population containing numerous related individuals might adversely affect prediction accuracy by distorting observed allele frequencies. However, Ramstetter et al. found that restricting the data to only one individual in each related pair did not have a negative impact (less than a 3% difference). To investigate further, we selected specific filtering parameters (locusmiss=0.2, samplemiss=0.2, MAF=0.1) and, for each species, performed ten random subsets, reducing families to three individuals each time. Subsequently, we applied various kinship methods to these subsetted datasets and compared sensitivity and precision to the full dataset under the same filtering conditions.

Our findings revealed that filtering did affect sensitivity and precision, but this effect was primarily dataset-specific rather than method dependent. Interestingly, using datasets with a reduced number of related individuals led to decreased precision in some instances, an unexpected outcome. Nevertheless, we opted to continue using our complete datasets (not subsetted by family) for comparisons. This decision was based on the observation that there was no consistent impact attributable to the kinship method, making them still comparable.
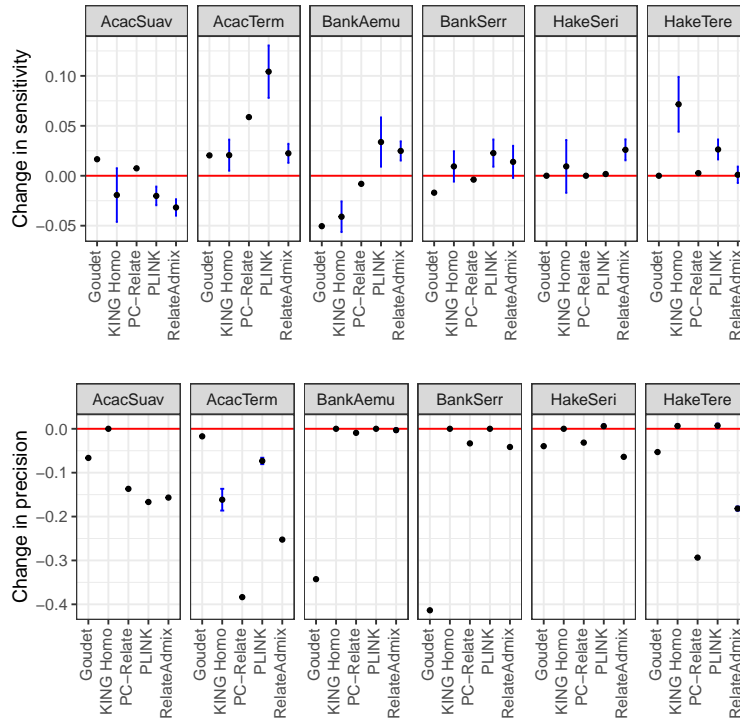
Figure S28. Flat change in overall sensitivity and precision when subsetting data from the full dataset down to 3 individuals per family. The red line indicates 0 change from the full dataset results. the black points are the average change for the 10 subsetting repetitions, and the blue lines are standard error for the repetitions.
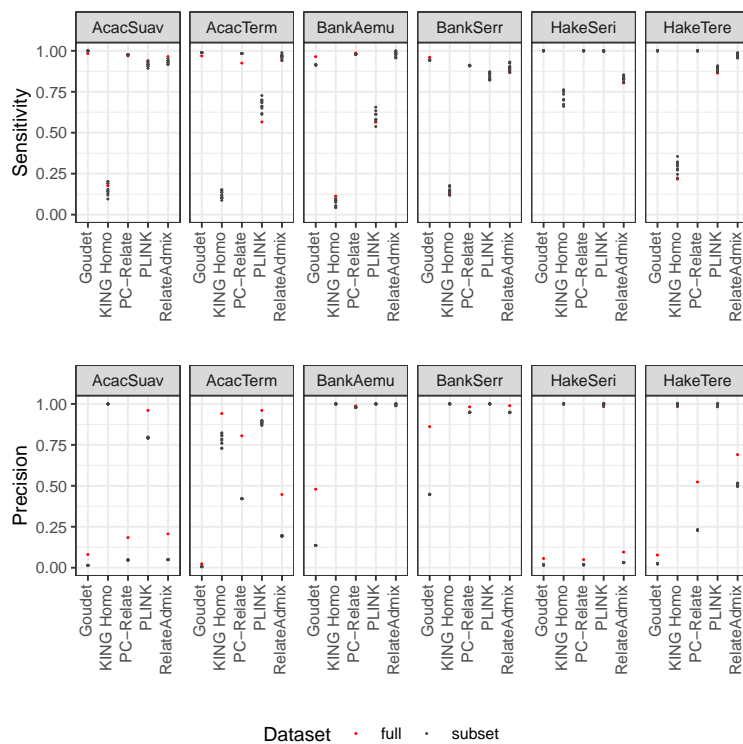
Figure S29. Raw values of overall sensitivity and precision from the full dataset (red point) and when subsetted datasets (black points).

8. KINSHIP AND IBD PROBABILITY

The theoretical thresholds for different degrees of relationship presented by Manichaikul et al. (2010) and Ramstetter et al. (2017) included kinship and probability of 0 IBD (k0) values. In this study and also Ramstetter et al., we only used kinship values to designate degree of relationship, but it is potentially very beneficial to also be able to use k0. The main benefit is that in theory, parent-offspring and full sibling relationships will both have kinship values between $1/2^{5/2}$ and $1/2^{3/2}$, but parent-offspring relationships will have higher k0 values. To assess whether this approach would be suitable for our data, we plotted kinship versus k0 values from the results of the ideal method runs (3).

Evidently, although the parent-offspring values form clusters within the data for most methods, they do not typically occur in the parent-offspring (1PO) theoretical bounds. Additionally, many of the known relationships occur outside of expected bounds entirely, which is likely a result of inbreeding. This means that although the use of k0 to predict relationship type is tempting in theory, in practice it is not suitable for RRS data of non-model plant species, which do not always conform to expectations.
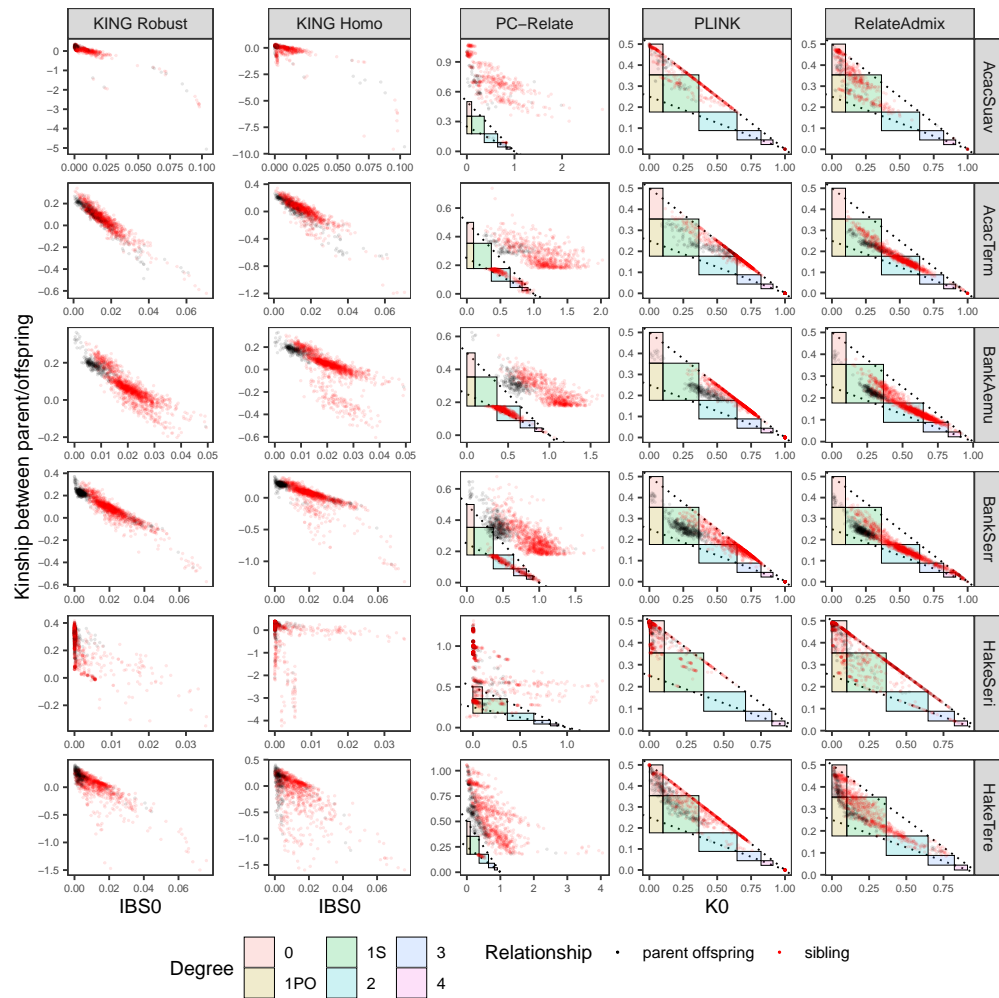
Figure S30. Probability of 0 IBD alleles (k0) versus kinship for optimal runs. Parent-offspring relationships are black points and sibships are red points. KING does not produce IBD probabilities, so IBS0 is shown instead. Goudet does not produce IBS or IBD values so is not shown. PC-Relate, PLINK, and RelateAdmix, which produce k0, are plotted here with the k0 and kinship theoretical ranges.

9. METHOD RUN TIME

We recorded method run time for for a single set of filtering parameters across all species and methods to get an idea of method efficiency. All executions were performed on datasets filtered for maximum locus missingness of 0.2, maximum sample missingness of 0.2, and minimum MAF of 0.1. Overall, PLINK and KING were significantly faster than the other methods, although Goudet was surprisingly close given it is implemented through R. PC-relate and RelateAdmix were much slower, but this is because we also included the required PC-Air and ADMIXTURE processes in the timed runs for these respective tools.
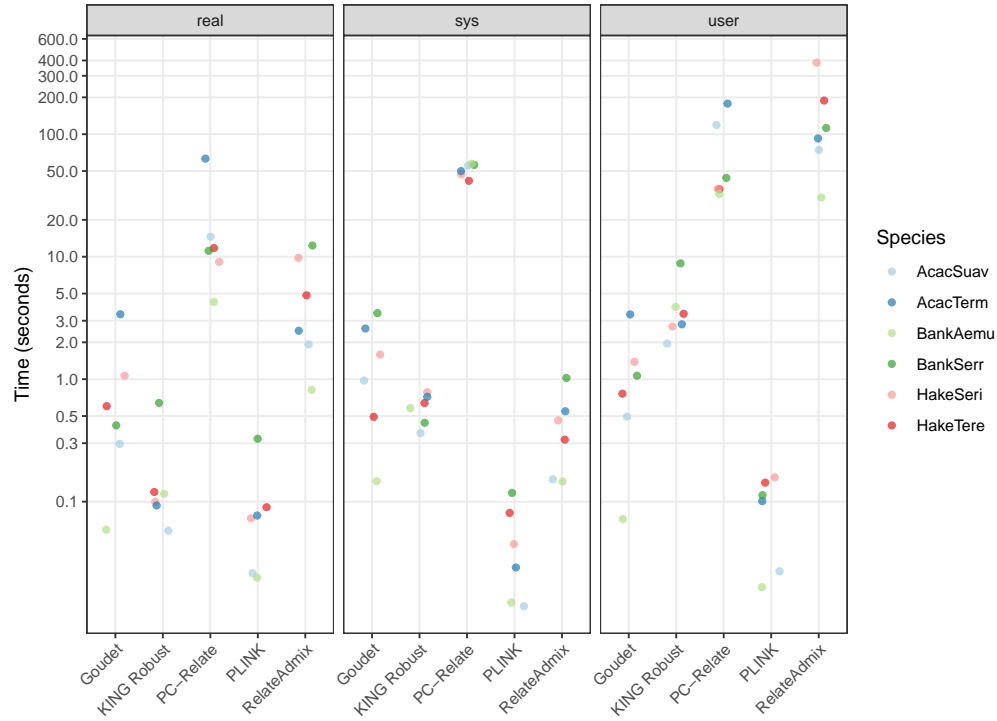
Figure S31. Real run times, system run times, and user run times of kinship methods across various species using a single set of filtering parameters across all species.

10. *DODONAEA VISCOSA* RELATIONSHIPS SUMMARISED

Table S5. Summary of relationships among individuals of *Dodonaea viscosa* within different Seed Production Areas. For each Seed Production Area, the table includes the mean and standard deviation of relationship counts observed across individuals with various degrees of relatedness. The columns contain information about the average number of relationships per individual (Mean) and the standard deviation of relationship counts (SD).

Seed Production Area	Degree	Mean	SD
SPA-A	0	0.000	0.000
SPA-A	1	0.245	0.434
SPA-A	2	0.925	1.328
SPA-A	3	0.075	0.331
SPA-A	4	0.019	0.137
SPA-B	0	0.000	0.000
SPA-B	1	2.844	3.167
SPA-B	2	1.675	2.663
SPA-B	3	0.130	0.522
SPA-B	4	0.013	0.114
SPA-C	0	0.000	0.000
SPA-C	1	0.392	0.603
SPA-C	2	0.745	1.129
SPA-C	3	0.216	0.673
SPA-C	4	0.039	0.280
Wild	0	0.007	0.081
Wild	1	0.479	1.091
Wild	2	0.400	0.865
Wild	3	0.010	0.099
Wild	4	0.007	0.081