DECADES OF MISCOMPUTATION IN GENOMIC CLADES AND DISTANCES

Richard B. Frost

Frost Concepts, Vista CA, USA

ABSTRACT

Hardly a week seems to go by without encountering a new genetics study that contains a diagram of specimen genetic similarities and clades. For these diagrams, biologists have long relied on university-based and/or commercial computational packages which are not only prone to pilot errors but also contain "analysis" methods which should never be used for genetic distance or clustering. Not that all the software is poor – it appears there is a mixture of good and bad in each package. The troublesome methods, however, have enjoyed acceptable use for so long that serious errors are published on a frequent basis. What follows is a list of concerns that will hopefully be useful to authors and reviewers alike. The report concludes with a graph-theoretical alternative to the current status quo in genomics.

KEYWORDS

Bayesian clustering, Graph partitioning, Missing values, Pair joining, Pseudo-metrics.

1. ITEMS OF CONCERN

1.1. Use of pseudo-metrics

A portion of the genomic literature utilizes pseudo-metrics to compute similarity or dissimilarity among genetic profiles. But to be used as a distance, the values are not valid for comparison unless the measure is a qualified metric [1]. This matter was contested 38 years ago by Felsenstein [2] who insisted biologists were only making adjacency comparisons in nodes of topological ancestry trees. However, the construction of those trees involves all-to-all comparisons of dissimilarities[3].

The well-known 1979 coefficient of Nei & Li [4] (eqns. 8 and 26) is an example of a pseudo-metric:

$$Nei1979(x,y) \equiv -\log \frac{\sum(x_i \cdot y_i)}{\sqrt{\sum(x_i \cdot x_i) \cdot \sum(y_i \cdot y_i)}}$$

Consider the 8 specimens with 8 random 0,1 markers listed in Table 1. In the spatial domain Nei & Li's 1979 coefficient produces 2 infinite distances due to zero denominators. Ignoring these, the remaining subset violates the metric triangle axiom 16 times out of 270. In the marker frequency domain (Nei's intended use) the results are equally poor with the $Nei1979$ measure producing 49 triangle axiom errors out of 336 tests plus 1 zero distance value. A list of non-metric measures offered as "distances" in a selection of commonly used software packages is given in Supplemental Table S1. It is also worth noting that only one biostatistical software package is careful with the term "metric" and refers to all genetic "distances" as dissimilarities [5].

Table 1.

Part 1: Eight example specimens with randomly generated marker values.
Part 2: Marker value frequencies.
Part 3: Jaccard metric spatial distances, scaled to integers so that values represent number of marker mismatches. In Part 3, nearest neighbours have the smallest value in any row or column. For example, row #6 indicates that specimen 6 has nearest neighbours 1, 2, and 8.

	marker spatial values								marker value frequencies								specimen spatial distances							
#	A	B	C	D	E	F	G	H	A	B	C	D	E	F	G	H	1	2	3	4	5	6	7	8
1	1	1	1	0	1	1	0	1	3/8	7/8	1/2	5/8	1/4	1/2	3/8	5/8		4	5	5	4	3	6	4
2	0	1	1	0	0	0	1	1	5/8	7/8	1/2	5/8	3/4	1/2	5/8	5/8	4		5	1	2	3	4	4
3	0	1	0	1	0	1	0	0	5/8	7/8	1/2	3/8	3/4	1/2	3/8	3/8	5	5		4	3	4	5	3
4	0	1	0	0	0	0	1	1	5/8	7/8	1/2	5/8	3/4	1/2	5/8	5/8	5	1	4		3	4	3	3
5	0	1	1	1	0	0	0	1	5/8	7/8	1/2	3/8	3/4	1/2	3/8	5/8	4	2	3	3		5	6	4
6	1	1	1	0	0	1	1	0	3/8	7/8	1/2	5/8	3/4	1/2	5/8	3/8	3	3	4	4	5		5	3
7	0	0	0	0	1	0	1	0	5/8	1/8	1/2	5/8	1/4	1/2	5/8	3/8	6	4	5	3	6	5		6
8	1	1	0	1	0	1	1	1	3/8	7/8	1/2	3/8	3/4	1/2	5/8	5/8	4	4	3	3	4	3	6	

1.2. Use of synonyms in cluster analysis

When a metric produces a zero distance between two or more profiles, they are termed synonyms under that metric. Investigators should make a note of such occurrences and then pick one as an ambassador to represent the synonymous group going forward. Synonyms are a violation of the metric positive definite axiom and should not be present when dissimilarity values are being compared since they skew the analysis of connectivity among profiles (see Figure 1).

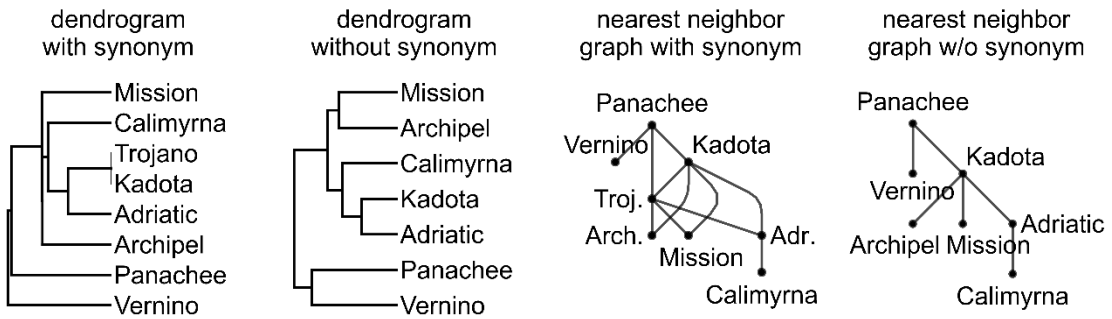

Figure 1. The effects of synonymy on cluster analysis of genetic distances. In the above graphs the marker data for specimen Trojano is identical to the data for specimen Kadota and thus there is zero distance between them. The presence and absence of Trojano produces dendrograms that are structurally different in both clustering and depth of branch points. Structural differences are also apparent in the nearest neighbour graphs on the right. Vertical hierarchy does not imply ancestry. See Supplemental Table S2 for marker data and computed distances.

1.3. Flattening multi-dimensional data into vectors

Except for pattern-matching metrics, most distance measures available in software packages are vector-based e.g. `List` in Mathematica® [6], `pdist` and `clustergram` in MATLAB® [7, 8], and single spreadsheet rows in SPSS [9]. As such investigators and software packages often "flatten" their tensor marker data (multiple primers per marker) into vectors by the following procedure:

$$X \rightarrow \left[X_{1,1} \cdots X_{1,K}\right] \cup \cdots \cup \left[X_{L,1} \cdots X_{L,K}\right] \rightarrow x$$

where X is tensor data, x is the resulting vector, L is the number of markers and K the number of primers per marker. This is sometimes done silently without the users knowledge e.g. function`mat_gen_dist` in R [10]. Doing so is equivalent to assuming all primers per marker are independent. It also changes the problem under study when a metric with a non-trivial normative expression (e.g. Euclidean) is used, since

$$cos\ \theta(X,Y) \equiv \frac{\|X \cdot Y\|_2}{\|X\|_2 \cdot \|Y\|_2} \neq cos\ \theta(x,y) \equiv \frac{\|x \cdot y\|_2}{\|x\|_2 \cdot \|y\|_2}$$

except in rare instances of X, Y. Further, any distance $\delta(x,y)$ computed by such metrics cannot be viably projected back into the original problem space because of the nature of the transform. In particular, the scalars of the normed $nm \times 1$ vector space would have to be inverted to scalars of the normed $n \times m$ tensor space which is generally infeasible when $n, m > 1$ for non-trivial norms due to the loss of dimensionality. Investigators desiring a normative metric for multi-allele data should consider the spectral radius

$$\rho(X-Y) \equiv \max_{i,j} \left| \sqrt{\lambda_i\left((X-Y)\cdot(X-Y)^T\right)} \right|_j, \left(X, Y \in \mathbb{Q}^{+0}_{p\times q}\right) \wedge (p \leq q)$$

Where $\lambda_i(M)$ is the ith eigenvalue of square matrix M and $|\cdot|_j$ denotes the modulus of the jth root of λ_i.

1.4. Misuse of metrics on data with errors and omissions

Some investigators will use a pattern matching metric and incorporate all their markers into an analysis of genetic distances – including those with missing values due to recording errors. This malpractice has been previously discussed by Schlueter and Harris[11]. Investigators plus authors of clustering and genetic distance software (e.g. [5, 8, 9, 12, 13]) should consider what happens when profiles $A = \{1,1,1,1,0\}$, $B = \{1,1,1,1\}$, and $C = \{1,1,1,1,1\}$ where B is missing a final value due to recording error. Jaccard's metric [14] will produce $J_{AB} = J_{BC}$, but without the recording error the result would be $J_{AB} \neq J_{BC}$. Likewise if any metric is used to compare only the primers with all values intact the result will be the same. Hence both approaches introduce at least as many errors as are removed.

1.5. Use of marker frequencies to distinguish individuals

Investigators who choose non-pattern-matching metrics on data composed of amino letter sequences are forced to use marker frequencies due to the lack of numeric values. To do so, the distribution of letters within a specimen marker profile is considered a population sample whose positional frequencies (Table 1, Part 2) are compared by distance metric or pseudo-metric to the letter distribution of another specimen profile. The first problem that occurs here is with multi-dimensional data: investigators and software packages are flattening tensors as discussed above instead of comparing multi-dimensional distribution samples. The second, more general problem is that distances computed between specimen marker distributions cannot be considered valid when the correlation matrix (or tensor) of all profile frequencies is singular – demonstrating that a valid encompassing distribution has not been established for the positional frequencies. Singular frequency correlation objects are typically the case with data from genetic profiles e.g., the correlation matrices and tensors of Supplemental Tables S2, S3, S4 are singular as discussed in the Table legends. Investigators desiring a dissimilarity measure for single-value marker frequencies should consider Mahalanobis' metric [15]

$$\text{Mahalanobis}(f_i, f_j) = \sqrt{(f_i - f_j)^T \cdot C^{-1} \cdot (f_i - f_j)}$$

where $F = \{f_1, \cdots, f_n\}$ contains frequency vectors of the n specimen spatial vectors and C is the correlation matrix of F. Be sure to check the numerical condition [16] of C^{-1} before computing distances.

1.6. Structures derived from profile enrichment

To classify genetic profiles into clades or distance clusters, some software packages use the sample distribution of primer values per marker to generate additional profiles for an "enriched" population, sometimes referred to as "burn-in", "bootstrap", or "Bayesian clustering". Partitions of the enriched set are then used to decide cluster memberships for the original sample e.g., [17-22]. Some implementations provide a questionable confidence interval calculated from within the enriched profile set.

It is unclear whether the enriched sets are relevant to the original data [23] or whether the resulting partition should be considered as anything more than one of several possibilities [24]. The enrichments are commonly of magnitude 1k to 100k. However, a non-trivial set of biased profiles with L markers and K primers per marker will imply a full population of at least magnitude 10^L to 10^{LK}, requiring a statistical sample size $S \cong 10^{L-4}$ to 10^{LK-4} [25]. This is an intractable situation in terms of computing order $S^2/2$ distances or order $S^3/3$ metric tests for nontrivial size L. Implementers of enrichment methods claim that smaller magnitudes are sufficient. To verify this claim they need to provide an analytic function of the distance metric and the marker distributions of all specimen profiles that for a given profile and radial displacement yields an unbiased measurable set of enriched profiles that fill (topologically cover) the enclosed hypersphere. This will enable computation of an accurate confidence interval for the profile enrichment results.

1.7. False nearest neighbours

Nearest neighbour analysis is preferred for cluster determination among genetic distances due to the complex topology of multi-dimensional genetic distance spaces (the alternative is to establish an eigenbasis with a Lie algebra). It is important for investigators to realize that a genetic profile can have multiple nearest neighbours (n.n.) of the same distance – a condition termed multiplicity (Table 1, Part 3). Unfortunately, some of the n.n. algorithms used internally in biostatistical software only pick the first element returned by sorting instead of the entire multiplicity group – a behaviour inherited from graph traversals (see Figure 2). When multiplicity exists and is ignored by the n.n. algorithm the distance to the selected neighbour is effectively shortened and the problem under study is changed. Consequently, many published sets of genetic distance clusters are erroneous.

1.8. Misuse of pair group analysis

The concordance correlation of pair group analysis forces elements into pairs by design [3]. As a result, pair group analysis cannot accurately partition distance sets containing odd-ranked multiplicity – a condition due to the nature of genetic profiles (see Figure 2). Another concern is the interpretation of branching points in pair group dendrograms as mutation or procreation points. This might be true in a system with only binary branching and no re-entrant breeding, but statements of these qualifications are not found in the genomic literature even though ample evidence is available for both.

1.9. Misapplication of graph partitioning software

Among the many graph partitioning algorithms available from computer science and graph theory, only a subset is applicable to distances, and only those that do not ignore multiplicity are viable for genetic profiles with significant number of markers. Further, just because an investigator picks a good algorithm does not mean the results will have any relevance as population subgroups or clades. To check this, examine a distance-limited nearest neighbour graph using either the component maximal or an empirically known upper bound of distance separation for n generations. The result will be one of 3 outcomes: viable clusters, lack of cohesion, or too much cohesion (Figure 3). These conditions can be due to the choice of metric, choice of markers, or the reality of the specimens.

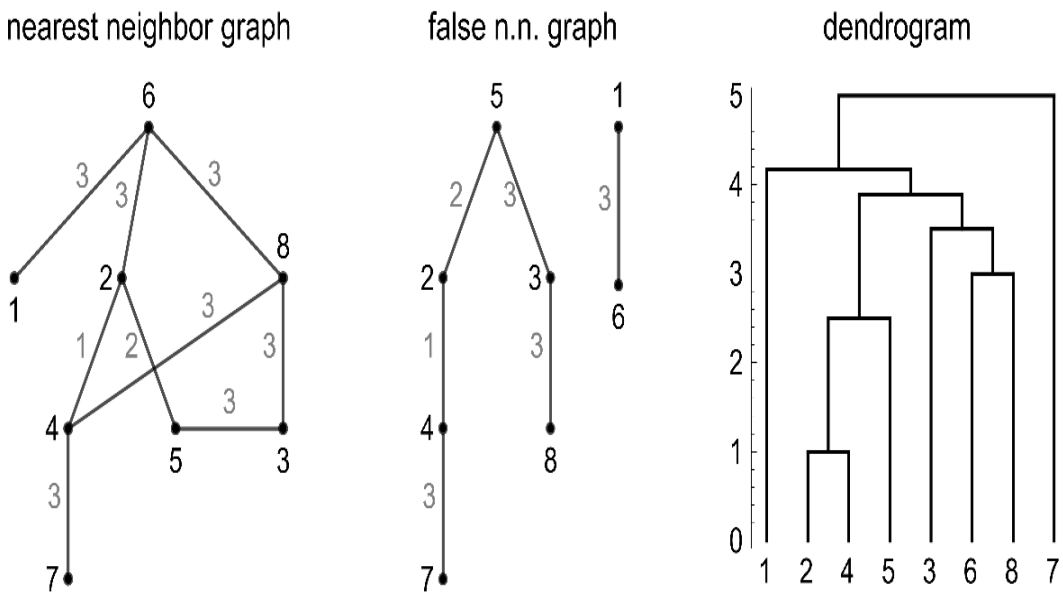

Figure 2. Three renderings of the example specimen data from Table 1 using Jaccard's metric on spatial values. In the first graph, notice the nearest-neighbor multiplicities of length 3 for specimens 6, 8, and 3, while specimen 4's n.n. is at length 1. The second graph is from an algorithm that ignores multiplicity and leads to false structural conclusions. The dendrogram on the right illustrates the inability of Sokal's pair-joining algorithm to parse odd-ranked multiplicity. Vertical hierarchy does not imply ancestry.

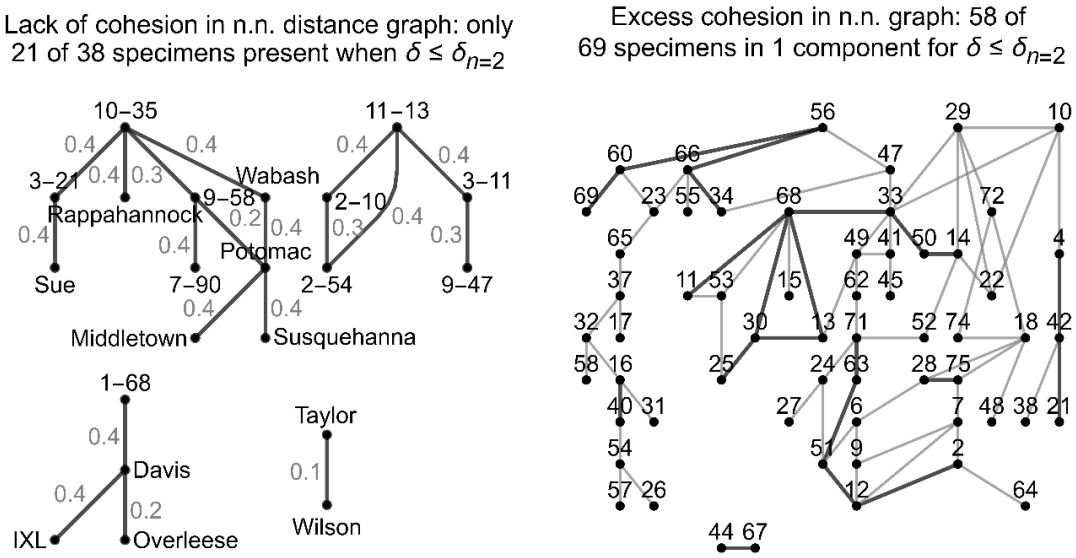

Figure 3. Cohesion extremes in distance limited nearest neighbour graphs. Distance limits were empirically determined by comparing specimen ancestry records to 2nd generation distances computed with Jaccard's metric. Vertical hierarchy does not imply ancestry. The left graph illustrates lack of cohesion in SSR data (see Supplemental Table S3. The right graph shows excess cohesion in SSR data (see Supplemental Table S5). Thick black edges are length 1/9, thin grey edges are length 2/9. Note that distance partitioning cannot improve the right-hand graph because to cut any grey line one must cut them all.

2. A TOPOLOGICAL APPROACH

If G is a collection of genetic profiles plus a set of some or all the distances between them and these qualify as a metric, then G is termed a distance graph with profiles for vertices and distances for edges. The traditional nearest neighbor graph G_N is the subset of G containing only nearest neighbor edges and their vertices. (Note: some computer programs ignore multiplicity so it is recommended to check automated results.) A least bridges graph G_{LB} of G will have overlap with G_N but offers more insight to components. The construction method is hierarchical. Vertices are first added as disconnected components. The shortest available edge connections are then added incrementally. Edges are only added between disconnected components and thus termed "bridges" [26]. A new component is created each time an edge is added, replacing the prior two. If there are multiple edges of the same distance that qualify then the entire set is added, possibly engulfing multiple components. The distances among components must be re-evaluated after an edge or edge set is added. Inter-component distances are determined by selecting the shortest vertex-to-vertex distance between them. This is often but not always a nearest neighbor edge. The process is continued incrementally until a prescribed limit is reached (e.g. a maximum distance) or a connected graph is achieved.

When G_N or G_{LB} is distance-limited by a non-trivial amount δ_i, the number of edges will be reduced and hence the number of components can increase. In any such graph, consider the function f of the product of the # of components c with the # of vertices v: $f(\delta_i) = c \cdot v$. The value δ_{opt} which produces a graph that maximizes f is termed the component maximal (Figure 4). Since this value maximizes the number of graph components with respect to vertices, it produces elemental clusters of original graph G for its given distance metric. An example is shown in Figure 5.

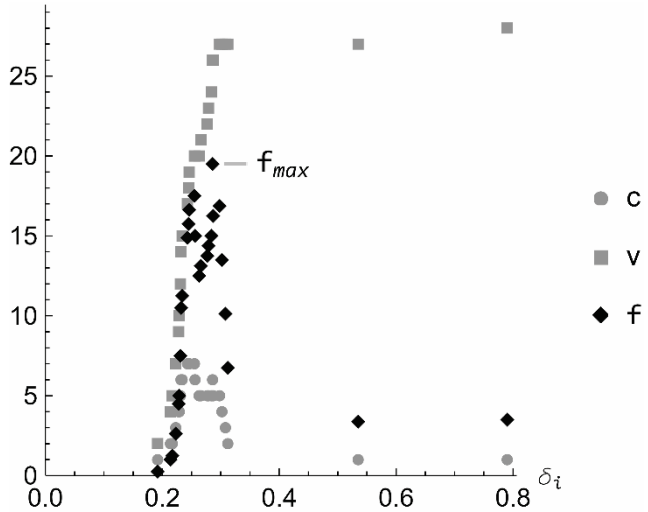

Figure 4. Variation of c = # of components (isolated clusters), v = # of vertices (specimens), and f ∝ c · v in least bridges graphs of 28 *Moraceae* specimens limited to genetic distance δ_i. The component maximal occurs at $(\delta_{opt}, f_{max}) = (0.286, \alpha \cdot 6 \cdot 26)$ where α is used for illustration purposes (see Figure 5).

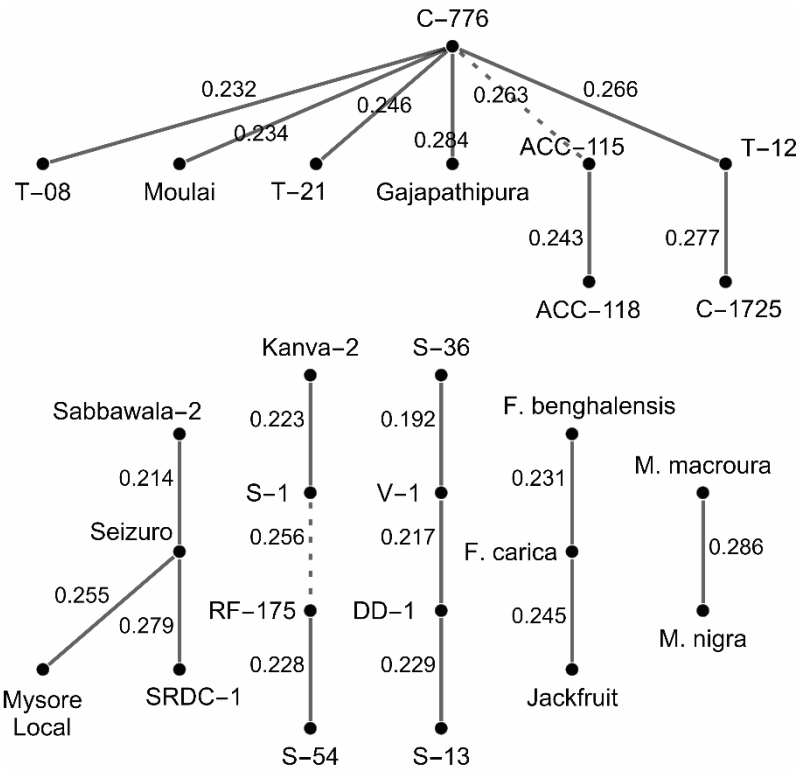

Figure 5. Distance-limited least bridges graph showing elemental clusters of 28 *Moraceae* specimens [27]. Solid lines denote nearest neighbors and dashed lines are least bridges. 26 specimens are present in the graph with 2 remaining cladeless. In the bottom left component Seizuro and Sabbawala-2 are mutual nearest neighbors, while Seizuro is the single nearest neighbor of Mysore Local and SRDC-1. A complete set of distance values are available in Supplemental Table S6 and the referenced publication. Distance limit is the component maximal $\delta_{opt} = 0.286$. Vertical hierarchy does not imply ancestry.

REFERENCES

[1] J. K. Hunter and B. Nachtergaele, *Applied analysis*. World Scientific Publishing Company, 2001, p. 438, doi: https://doi.org/10.1142/4319.

[2] J. Felsenstein, "Distance methods for inferring phylogenies: a justification," *Evolution*, pp. 16-24, 1984. https://www.jstor.org/stable/2408542.

[3] J. H. Camin and R. R. Sokal, "A method for deducing branching sequences in phylogeny," *Evolution*, pp. 311-326, 1965. https://www.jstor.org/stable/2406441.

[4] M. Nei and W.-H. Li, "Mathematical model for studying genetic variation in terms of restriction endonucleases," *Proceedings of the National Academy of Sciences*, vol. 76, no. 10, pp. 5269-5273, 1979, doi: https://doi.org/10.1073/pnas.76.10.5269.

[5] X. J.-C. Perrier, Jean-Pierre. "DARwin - Dissimilarity Analysis and Representation for Windows." CIRAD. https://darwin.cirad.fr/.

[6] W. Research. "Mathematica." https://www.wolfram.com/mathematica.

[7] MATLAB. "Pairwise distance between pairs of observations - MATLAB pdist." MathWorks. https://www.mathworks.com/help/stats/pdist.html.

[8] MATLAB. "Object containing hierarchical clustering analysis data - MATLAB." MathWorks. https://www.mathworks.com/help/bioinfo/ref/clustergram.html.

[9] IBM. "SPSS Statistics | IBM." IBM. https://www.ibm.com/products/spss-statistics.

[10] P. Savary. "Landscape and genetic data processing with graph4lg." The R Project. https://cran.r-project.org/web/packages/graph4lg/vignettes/input_data_processing_1.html.

[11] P. M. Schlueter and S. A. Harris, "Analysis of multilocus fingerprinting data sets containing missing data," *Molecular Ecology Notes*, vol. 6, no. 2, pp. 569-572, 2006, doi: https://doi.org/10.1111/j.1471-8286.2006.01225.x.

[12] Biostat. "NTSYSpc." Applied Biostat LLC. http://www.appliedbiostat.com/ntsyspc/ntsyspc.html.

[13] R. "The R Project for Statistical Computing." The R Foundation. https://www.r-project.org/.

[14] S. Kosub, "A note on the triangle inequality for the Jaccard distance," *Pattern Recognition Letters*, vol. 120, pp. 36-38, 2019, doi: https://doi.org/10.1016/j.patrec.2018.12.007.

[15] P. C. Mahalanobis, "On the generalized distance in statistics," 1936. http://library.isical.ac.in:8080/jspui/bitstream/10263/6765/1/Vol02_1936_1_Art05-pcm.pdf.

[16] G. W. Stewart, *Afternotes on numerical analysis*. SIAM, 1996. https://doi.org/10.1137/1.9781611971491.

[17] M. J. Hubisz, D. Falush, M. Stephens, and J. K. Pritchard, "Inferring weak population structure with the assistance of sample group information," *Molecular ecology resources*, vol. 9, no. 5, pp. 1322-1332, 2009, doi: https://doi.org/10.1111/j.1755-0998.2009.02591.x.

[18] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation PLINK: rising to the challenge of larger and richer datasets," *Gigascience*, vol. 4, no. 1, pp. s13742-015-0047-8, 2015, doi: https://doi.org/10.1186/s13742-015-0047-8.

[19] G. Guillot, S. Renaud, R. Ledevin, J. Michaux, and J. Claude, "A unifying model for the analysis of phenotypic, genetic, and geographic data," *Systematic biology*, vol. 61, no. 6, pp. 897-911, 2012, doi: https://doi.org/10.1093/sysbio/sys038.

[20] L. Excoffier and H. E. Lischer, "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows," *Molecular ecology resources*, vol. 10, no. 3, pp. 564-567, 2010, doi: https://doi.org/10.1111/j.1755-0998.2010.02847.x.

[21] O. François, S. Ancelet, and G. Guillot, "Bayesian clustering using hidden Markov random fields in spatial population genetics," *Genetics*, vol. 174, no. 2, pp. 805-816, 2006, doi: https://doi.org/10.1534/genetics.106.059923.

[22] C. Chen, E. Durand, F. Forbes, and O. François, "Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study," *Molecular Ecology Notes*, vol. 7, no. 5, pp. 747-756, 2007, doi: https://doi.org/10.1111/j.1471-8286.2007.01769.x.

[23] D. J. Witherspoon *et al.*, "Genetic similarities within and between human populations," *Genetics*, vol. 176, no. 1, pp. 351-359, 2007, doi: https://doi.org/10.1534/genetics.106.067355.

[24] J. Novembre, "Pritchard, Stephens, and Donnelly on population structure," *Genetics*, vol. 204, no. 2, pp. 391-393, 2016, doi: https://doi.org/10.1534/genetics.116.195164.

[25] M. F. Triola, *Elementary Statistics*, 8th ed. Addison-Wesley, 2001. https://books.google.com/books?id=G6u8PwAACAAJ.

[26] C. Godsil and G. F. Royle, *Algebraic graph theory*. Springer Science & Business Media, 2013. https://link.springer.com/book/10.1007/978-1-4613-0163-9.
[27] B. Mathi Thumilan, R. Sajeevan, J. Biradar, T. Madhuri, K. N. Nataraja, and S. M. Sreeman, "Development and characterization of genic SSR markers from Indian mulberry transcriptome and their transferability to related species of Moraceae," *PloS ONE*, vol. 11, no. 9, p. e0162909, 2016, doi: https://doi.org/10.1371/journal.pone.0162909.
[28] MATLAB. "Pairwise distance between pairs of observations - MATLAB pdist - Distance metric." MathWorks. https://www.mathworks.com/help/stats/pdist.html#mw_39296772-30a1-45f3-a296-653c38875df7.
[29] Wolfram. "Distance and Similarity Measures - Wolfram Language Documentation." Wolfram Research, Inc. https://reference.wolfram.com/language/guide/DistanceAndSimilarityMeasures.html.
[30] IBM. "Distances - IBM Documentation." IBM Corporation. https://www.ibm.com/docs/en/spss-statistics/28.0.0?topic=features-distances.
[31] USDA. "Ficus carica L. GRIN-Global." USDA ARS. https://npgsweb.ars-grin.gov/gringlobal/taxon/taxonomydetail?id=16801.
[32] K. W. Pomper *et al.*, "Characterization and identification of pawpaw cultivars and advanced selections by simple sequence repeat markers," *Journal of the American Society for Horticultural Science*, vol. 135, no. 2, pp. 143-149, 2010, doi: https://doi.org/10.21273/JASHS.135.2.143.
[33] K. Vinod, "Structured association mapping using STRUCTURE and TASSEL," *Bioinformatics Tools for Genomics Research*, p. 103, 2011. https://www.academia.edu/706699/Structured_Association_Mapping_using_STRUCTURE_and_TASSEL.
[34] A. Wünsch and J. Hormaza, "Molecular characterisation of sweet cherry (Prunus avium L.) genotypes using peach [Prunus persica (L.) Batsch] SSR sequences," *Heredity*, vol. 89, no. 1, pp. 56-63, 2002, doi: https://doi.org/10.1038/sj.hdy.6800101.

AUTHOR

R.B. Frost is an old-school numerical analyst with academic and vocational experience in applied mathematics, computer science, and horticulture. He is currently pursuing research in the genomics of lesser-studied fruits.

SUPPLEMENTARY INFORMATION

Supplemental Table S1. Non-metric dissimilarity measures erroneously presented as distances in software packages commonly used for biostatistics research [12, 28-30]. Symmetric pairs of zero distances were removed prior to analysis.

					Metric Tests		
Software	**Measure Name**	**Data Type**	**Domain**	**Dataset**	**Reflexive**	**Com-mutative**	**Triangle Inequality**
MATLAB R2022a	Correlation	Numerical	Spatial	Table x	Passed	Passed	Failed
MATLAB R2022a	Correlation	Numerical	Frequency	Table x	Passed	Passed	Failed
MATLAB R2022a	Cosine	Numerical	Spatial	Table x	Passed	Passed	Failed
MATLAB R2022a	Cosine	Numerical	Frequency	Table x	Passed	Passed	Failed
Mathematica v13.0	Bray-Curtis	Numerical	Spatial	Table x	Passed	Passed	Failed
Mathematica v13.0	Bray-Curtis	Numerical	Frequency	Table x	Passed	Passed	Failed
Mathematica v13.0	Correlation	Numerical	Spatial	Table x	Passed	Passed	Failed
Mathematica v13.0	Correlation	Numerical	Frequency	Table x	Passed	Passed	Failed
Mathematica v13.0	Cosine	Numerical	Spatial	Table x	Passed	Passed	Failed
Mathematica v13.0	Cosine	Numerical	Frequency	Table x	Passed	Passed	Failed

NTSYSpc v2.21w	Hillis	Multi-allele Counts	Spatial	Table S2	Passed	Passed	Failed
NTSYSpc v2.21w	Hillis	Multi-allele Counts	Frequency	Table S2	Passed	Passed	Failed
NTSYSpc v2.21w	Nei 1972	Multi-allele Counts	Spatial	Table S2	Passed	Failed	Failed
NTSYSpc v2.21w	Nei 1972	Multi-allele Counts	Frequency	Table S2	Passed	Failed	Failed
NTSYSpc v2.21w	Nei 1978	Multi-allele Counts	Spatial	Table S2	Passed	Passed	Failed
NTSYSpc v2.21w	Nei 1978	Multi-allele Counts	Frequency	Table S2	Passed	Passed	Failed
SPSS v28.0.1.1	Size Difference	0,1 Binary	Spatial	Table 1	Passed	Passed	Failed
SPSS v28.0.1.1	Pattern Difference	0,1 Binary	Spatial	Table 1	Passed	Passed	Failed
SPSS v28.0.1.1	Binary Shape	0,1 Binary	Spatial	Table 1	Passed	Passed	Failed
SPSS v28.0.1.1	Lance and Williams	0,1 Binary	Spatial	Table 1	Passed	Passed	Failed

Supplemental Table S2. Multi-dimensional SSR and genetic distance data from 8 *Ficus carica* specimens at NCGR Davis [31]. Distances computed with spectral radius of $X - Y$, where X, Y are the 2×15 tensors below each specimen name. The correlation tensor of frequencies for this data is singular due to marker frequencies of M8N1.2 all having value 1/2, and several other markers having no variation e.g., C22F1.1.

		specimens							
		Adriatic	Archipel	Calimyrna	Kadota	Mission	Panachee	Trojano	Vernino
SSR markers	C22F1	283	283	283	283	283	283	283	283
		283	283	283	283	285	283	283	283
	C24H1	272	270	272	272	270	272	272	270
		272	272	272	272	272	272	272	272
	C26N1	234	234	234	234	234	234	234	234
		234	236	234	234	236	234	234	234
	C31F1	224	224	239	224	224	224	224	224
		239	224	239	239	239	239	239	239
	C35H1	252	254	254	254	254	252	254	254
		254	254	256	254	254	254	254	254
	C37N1	204	204	204	204	204	204	204	204
		204	208	204	208	204	204	208	204
	LM12H1	214	214	214	214	214	233	214	233
		243	243	243	243	243	233	243	243
	LM14H1	200	200	200	200	200	200	200	198
		200	200	200	200	200	200	200	200
	LM30N1	245	245	243	243	245	237	243	231
		251	251	245	245	247	251	245	251
	LM36N1	248	248	248	248	248	248	248	248
		248	248	250	248	250	248	248	250
	M1F1	172	189	172	172	189	155	172	172
		184	189	184	189	189	184	189	188
	M2H1	155	161	153	153	161	153	153	155
		161	167	161	167	167	153	167	155
	M3N1	124	120	132	120	122	124	120	132
		132	132	132	132	122	132	132	132
	M4F1	194	194	194	194	194	194	194	214

		214	214	214	218	218	214	218	214
	M8N1	171	171	171	171	171	171	171	171
		171	175	171	175	175	171	175	171
distances	Adriatic		21.383	17.378	12.042	20.772	27.715	12.042	32.14
	Archipel	21.383		30.989	19.046	19.546	40.948	19.046	37.917
	Calimyrna	17.378	30.989		19.209	27.995	32.223	19.209	33.892
	Kadota	12.042	19.046	19.209		19.134	27.599	0.	33.119
	Mission	20.772	19.546	27.995	19.134		40.773	19.134	37.264
	Panachee	27.715	40.948	32.223	27.599	40.773		27.599	28.505
	Trojano	12.042	19.046	19.209	0.	19.134	27.599		33.119
	Vernino	32.14	37.917	33.892	33.119	37.264	28.505	33.119	
		Adriatic	Archipel	Calimyrna	Kadota	Mission	Panachee	Trojano	Vernino

Supplemental Table S3. SSR data values for 5 loci from 2010 *Asimina triloba* (Pawpaw) study of Pomper et al [32]. Suffix _.F refers to forward SSR primer, _.R to reverse. Note that the zeros are missing values which skew the analysis as discussed in the main text. The correlation tensor of frequencies for this data is singular due to the interdependence of primers C104.F and C104.R.

Genotype	B3.F	B3.R	B103.F	B103.R	B129.F	B129.R	C104.F	C104.R	G119.F	G119.R
10-35	183	191	266	339	166	172	184	0	158	164
11-13	191	0	264	305	166	172	184	0	158	0
1-23	185	189	290	310	158	0	184	0	158	176
1-68	185	187	268	341	158	179	175	184	158	164
2-10	191	0	264	270	170	172	184	0	161	164
2-54	191	0	264	270	162	166	184	0	161	0
3-11	191	0	272	288	158	172	184	0	158	161
3-21	189	191	266	305	166	170	184	0	161	164
5-5	183	189	270	305	166	168	184	0	161	0
7-90	185	191	305	342	170	176	184	0	161	164
8-20	189	191	264	270	162	0	184	0	158	167
9-47	183	0	272	274	158	166	184	0	158	161
9-58	183	191	264	339	170	176	184	0	158	164
BH10	189	0	319	321	162	170	184	0	144	161
Cales Creek	175	183	266	274	156	158	184	0	158	164
Davis	185	189	264	268	158	164	175	184	158	164
Greenriver Belle	183	189	264	266	162	172	184	0	158	161
IXL	187	189	274	309	158	162	175	184	158	164
M. Gordon	185	195	270	312	164	170	184	0	161	164
Middletown	183	193	270	321	170	0	184	0	158	161
Mitchell	0	0	266	321	158	172	184	0	158	167
NC-1	185	193	266	0	158	162	184	0	158	161
Overleese	185	189	264	0	158	164	175	184	158	0
PA-Golden#1	191	193	336	343	172	176	184	0	158	164
PA-Golden#3	189	191	336	343	158	172	184	0	161	170
PA-Golden#4	175	183	319	326	164	0	184	0	158	161
Potomac	183	191	264	324	170	0	184	0	158	164
Prolific	189	191	309	323	158	162	184	0	158	0
Rappahannock	183	191	266	0	166	0	184	0	164	170
Rebeccas Gold	185	193	266	0	158	162	184	0	158	161
Shenandoah	185	187	264	274	162	164	184	0	158	164
Sue	175	189	266	329	166	180	184	0	161	164
Sunflower	187	0	274	341	162	180	175	184	164	0
Susquehanna	189	191	264	270	162	0	184	0	158	167
Sweet Alice	175	183	260	324	166	182	184	0	144	164
Taylor	183	185	268	322	173	193	184	0	167	170
Taytwo	175	185	252	290	158	0	184	0	164	176
Wabash	183	0	266	324	170	172	184	0	158	170
Wells	175	191	276	290	177	0	184	0	161	164
Wilson	183	185	268	321	173	193	184	0	167	170
Zimmerman	191	195	303	324	164	177	184	0	158	0

Supplemental Table S4. Example SSR data table from Vinod 2011 [33]. Row "GENO6" has been deleted due to a missing data value in marker SSR2. The correlation matrix of frequencies for this data is singular due to the lack of variation in columns 4, 6, and 8.

	SSR1	SSR2	SSR3	SSR4	SSR5	SSR6	SSR7	SSR8	SSR9
GENO1	110	330	190	140	220	140	240	160	200
GENO2	110	330	190	140	230	140	240	160	190
GENO3	110	320	190	140	220	140	240	160	200
GENO4	110	320	190	140	220	140	240	160	200
GENO5	110	330	180	140	220	140	240	160	200
GENO7	110	330	190	140	220	140	240	160	200
GENO8	110	320	180	140	220	140	240	160	200
GENO9	110	330	190	140	220	140	250	160	200
GENO10	120	320	180	140	220	140	240	160	200

Supplemental Table S5. SSR data values for 9 loci from 2002 *Prunus avium* (Sweet Cherry) study of Wünsch & Hormaza [34]. Dual values are treated as rational numbers. Note that single values are instances of missing scores which skew the analysis as discussed in the main text. The correlation matrix of frequencies for this data is not singular, with determinant ≅0.280843. This would not be the case if the missing scores were recorded as zeros (see Table S3).

Cultivar	Index	Pchcms1	Pchcms3	Pchcms5	UDP96-005	UDP98-409	UPD98-021	UPD98-022	UPD97-402	UPD98-412
Ambrunes	1	140	180/160	290	150/125	160/130	110/100	110/90	125	130
Arcina	2	140	180	260	150/130	130	110/100	100/90	140/125	130
Beige	3	190/140	180	260	150/120	130	110/100	110/90	130	130
Bing	4	190/140	180	260	150/120	130	110/100	110/90	130	130
Blanca de Provenza	5	140	180/160	290/260	120	130	100	105/90	145/125	130/100
Brooks	6	140	180	290	150/120	130	110/100	90	135/125	130
Burlat	7	140	180	290	150/130	130	110/100	100/90	130/120	130
Burlat C-1	8	140	180	290	150/130	130	110/100	100/90	130/120	130
Celeste	9	140	180	290	150/130	130	110/100	100	135/125	130
Chinook	10	190/140	180	260	150/120	130	110/100	110	145/130	130
Compact Stella	11	190/140	180	290/260	150/120	130	110/100	110/100	145/125	130
Coralise	12	140	180	260	150/130	130	110/100	100/90	135/125	130
Corum	13	190/140	180	290/260	150/120	130	110/100	110/100	135/125	140/130
Cristalina	14	190/140	180/160	260	150/120	130	110	110/100	135/125	130
Cristo-balina	15	190/140	180	290/260	150/115	130	110/100	110/90	135/125	130
Duroni 3	16	140	180	290/260	150/115	130	110/100	100/90	125	130
Earlise	17	140	180	290/260	150/130	130	100	90	130/120	130
Earlystar	18	190/140	180	290	150/120	130	100	100/90	145/125	130
Early Van Compact	19	190/140	180	290/260	150/120	130	110	100/90	135/125	130/120
Ferrovia	20	140	180	290/260	150/120	130	110	100	145/130	140/130
Garnet	21	140	180	260	120	130	110/100	110/90	130	130
Gil Peck	22	190/140	180/160	260	150/120	130	110	110	145/130	130
Giorgia	23	140	180	290/260	150/120	130	110	100/90	125	130/120
Hartland	24	140	180	260	150/120	130	110	100/90	140/130	130
Hedel-finger	25	140	180	290/260	150/120	130	110/100	110/100	135/125	130/100
Lambert	26	190/140	180/160	290/260	150/120	130	110/100	110	125	130
Lamida	27	140	180	260	150/120	130	110	110	145/125	130

Lapins	28	190/140	180	290	150/120	130	110/100	100/90	135/125	130
Larian	29	190/140	180	260	150/120	130	110	110/100	145/130	130
Marmotte	30	190/140	180	290/260	150/120	130	110/100	110/100	135/125	130/100
Marvin	31	140	180	290/260	120	130	110/100	100/90	130/125	130
Moreau	32	140	180	290/260	150/115	130	100	100/90	145/125	130
Napoleon	33	190/140	180	260	150/120	130	110/100	110/100	135/125	130
Newstar	34	140	180	290/260	150/120	130	110/100	100	135/125	130/120
Pico Colorado	35	140	180/160	290/260	150/130	160/130	110/100	110	140/125	130
Pico Negro	36	140	180/160	290/260	150/115	160/130	100	110/90	140/120	130
Precoce Bernard	37	140	180	290/260	150/130	130	100	100/90	145/130	130
Rainier	38	190/140	180	290/260	120	130	110/100	90	130	130
Ramon Oliva	39	140	180/160	290	150/130	130	100	90	120	130
Reverchon	40	140	180	290/260	150/115	130	110/100	100	125	130
Royalton	41	190/140	180	260	150/130	130	100	110/100	135/125	130
Ruby	42	190/140	180	260	120	130	110/100	110/90	130	130
Sam	43	140	180/160	290/260	150/120	130	100	110/100	145/125	130
Samba	44	190/140	180	290/260	150/120	130	110	110/90	135/130	130
Santina	45	190/140	180/160	290/260	150/130	130	100	110/100	135/125	130
Skeena	46	190	180	290/260	150/120	130	100	90	145/125	130
Somerset	47	190/140	180	260	150/120	130	110/100	100	135/125	130/120
Sonata	48	190	180	290	150/120	130	100	100/90	145/125	130/120
Spalding	49	190/140	180	260	150/130	130	110/100	110/100	135/125	140/130
Star	50	190/140	180/160	260	150/120	130	110/100	110/100	135/125	130
Starky Hardy Giant	51	140	180	260	150/120	130	110/100	100/90	135/125	130
Sue	52	140	180/160	260	150/120	130	110	110/100	135/125	140/130
Sumesi	53	140	180	290/260	150/120	130	110/100	110/100	145/130	130
Summit	54	140	180/160	290/260	150/120	130	110/100	100	125	130
Sunburst	55	140	180	290	150/120	130	110/100	100/90	145/130	130/120
Sweetheart	56	190/140	180	290/260	150/120	130	110/100	100/90	135/125	130/120
Sylvia	57	140	180/160	290	150/120	130	110/100	100	145/125	130
Taleguera Brilliante	58	140	180/160	290/260	150/115	130	100	100/90	135	130
Tigre	59	140	180	290	150/115	130	110	110/90	120	130
Van	60	190/140	180	290/260	150/120	130	110	100/90	135/125	130/120
Van Spur	61	190/140	180	290/260	150/120	130	110	100/90	135/125	130/120
Vega	62	190/140	180	260	150/130	130	110	100/90	135/125	140/130
Vic	63	140	180	260	150/120	130	110/100	100/90	135/125	140/130
Vignola	64	140	180	260	150/130	130	110	100	140/125	130
Vittoria	65	140	180	290/260	150/130	130	110	100/90	145/130	130/120
13N.7.19	66	140	180	290/260	150/120	130	110/100	100/90	135/125	130/120
13S.17.20	67	190/140	180	290/260	150/120	130	110	110/90	130	130
13S.18.10	68	190/140	180	290/260	150/120	130	110/100	110/100	135/125	130
13S.18.15	69	190/140	180	290/260	150/120	130	110	110/100	135/125	130/120
13S.21.7	70	140	180	290	150/130	130	110/100	100	135/125	130
13S.27.17	71	140	180	260	150/120	130	110	100/90	135/125	140/130
13S.3.13	72	190/140	180	290	150/120	130	100	110/100	145/130	130
44W.11.8	73	190/140	180	260	150/120	130	110	110/100	145/130	130
83703007	74	190/140	180	290	150/120	130	100	100	145/120	130
84703002	75	190/140	180	290	150/120	130	110/100	100/90	130/120	130
84704006	76	140	180	290	150/130	130	110/100	100/90	130/120	130

Supplemental Table S6. Distance matrix from 2016 Moraceae SSR study by Mathi Thumilan et al [27]. The investigators computed distances with one of the metrics in the Darwin v.5.0 program [5]. It is unknown whether the original SSR data was vector-valued or flattened from a tensor.

Genotype	M. macrourа	M. nigra	ME-107	T-12	C-1725	T-21	Moulai	T-08	C-776	ACC-118	ACC-115	Gajapathipura	SRDC-1	
M. nigra	0.286													
ME-107	0.302	0.298												
T-12	0.322	0.319	0.308											
C-1725	0.353	0.349	0.339	0.277										
T-21	0.36	0.357	0.346	0.285	0.313									
Moulai	0.355	0.351	0.341	0.279	0.307	0.259								
T-08	0.354	0.35	0.339	0.278	0.306	0.258	0.235							
C-776	0.341	0.338	0.327	0.266	0.294	0.246	0.234	0.232						
ACC-118	0.364	0.36	0.35	0.288	0.316	0.288	0.283	0.282	0.27					
ACC-115	0.357	0.354	0.343	0.282	0.31	0.282	0.276	0.275	0.263	0.243				
Gajapathipura	0.36	0.356	0.346	0.285	0.313	0.303	0.297	0.296	0.284	0.306	0.3			
SRDC-1	0.394	0.39	0.38	0.318	0.346	0.336	0.331	0.33	0.318	0.34	0.333	0.319		
Sabbawala-2	0.397	0.393	0.383	0.321	0.349	0.339	0.334	0.333	0.32	0.343	0.336	0.322	0.288	
Seizuro	0.388	0.384	0.374	0.313	0.34	0.331	0.325	0.324	0.312	0.334	0.328	0.313	0.279	
Mysore Local	0.404	0.4	0.39	0.328	0.356	0.346	0.341	0.34	0.327	0.35	0.343	0.329	0.295	
Kanva-2	0.398	0.394	0.384	0.323	0.35	0.341	0.335	0.334	0.322	0.344	0.338	0.323	0.309	
S-1	0.397	0.393	0.383	0.322	0.349	0.34	0.334	0.333	0.321	0.343	0.337	0.322	0.308	
RF-175	0.407	0.403	0.393	0.331	0.359	0.349	0.344	0.343	0.33	0.353	0.346	0.332	0.317	
S-54	0.417	0.414	0.403	0.342	0.37	0.36	0.354	0.353	0.341	0.363	0.357	0.342	0.328	
S-13	0.399	0.395	0.385	0.324	0.352	0.342	0.336	0.335	0.323	0.345	0.339	0.324	0.31	
DD-1	0.398	0.394	0.384	0.322	0.35	0.34	0.335	0.334	0.321	0.344	0.337	0.323	0.308	
V-1	0.406	0.403	0.392	0.331	0.359	0.349	0.343	0.342	0.33	0.353	0.346	0.332	0.317	
S-36	0.413	0.41	0.399	0.338	0.366	0.356	0.35	0.349	0.337	0.36	0.353	0.339	0.324	
F. benghalensis	0.613	0.609	0.599	0.538	0.565	0.556	0.55	0.549	0.537	0.559	0.553	0.538	0.547	
F. carica	0.612	0.608	0.598	0.536	0.564	0.555	0.549	0.548	0.536	0.558	0.551	0.537	0.546	
Jackfruit	0.612	0.608	0.597	0.536	0.564	0.554	0.549	0.547	0.535	0.558	0.551	0.537	0.546	
Dudia white	0.866	0.863	0.852	0.791	0.819	0.809	0.803	0.802	0.79	0.813	0.806	0.792	0.801	
Genotype	Sabbawala-2	Seizuro	Mysore Local	Kanva-2	S-1	RF-175	S-54	S-13	DD-1	V-1	S-36	F. benghalensis	F. carica	Jackfruit
Seizuro	0.214													

Mysore Local	0.264	0.255												
Kanva-2	0.312	0.303	0.318											
S-1	0.311	0.302	0.318	0.223										
RF-175	0.32	0.312	0.327	0.257	0.256									
S-54	0.331	0.322	0.338	0.268	0.267	0.228								
S-13	0.313	0.304	0.32	0.29	0.289	0.298	0.309							
DD-1	0.311	0.303	0.318	0.288	0.287	0.297	0.307	0.229						
V-1	0.32	0.311	0.327	0.297	0.296	0.305	0.316	0.238	0.217					
S-36	0.327	0.318	0.334	0.304	0.303	0.312	0.323	0.245	0.224	0.192				
F. benghalensis	0.55	0.542	0.557	0.552	0.551	0.56	0.571	0.553	0.551	0.56	0.567			
F. carica	0.549	0.541	0.556	0.55	0.549	0.559	0.57	0.552	0.55	0.559	0.566	0.231		
Jackfruit	0.549	0.54	0.556	0.55	0.549	0.559	0.569	0.551	0.55	0.559	0.566	0.246	0.245	
Dudia white	0.804	0.795	0.811	0.805	0.804	0.814	0.824	0.806	0.805	0.813	0.821	0.861	0.859	0.859