COUPLED ANALYSIS OF PAWPAW (*ASIMINA TRILOBA*) GENETIC MARKERS AND ANCESTRY RECORDS

Richard Frost

Frost Concepts, Vista CA, USA
tangent.vectors@gmail.com

ABSTRACT

Subsets of 49 RAPD markers for 36 Asimina triloba specimens from U.S. NCGR repository sites are examined for matches to ancestry records. Several known parent-progeny and sibling relationships are verified, but a few specimens are also determined mislabeled due to excessive dissimilarities. An insight to the debate of cultivar Overleese vs NC-1 is also presented.

KEYWORDS

Cultivar ancestry, Graph theory, RAPD markers

1. INTRODUCTION

The Pawpaw is a deciduous tree native to eastern North America. It produces a potato-size fruit which has been cultivated by native peoples since antiquity and more recently in home orchards and small farms in the U.S. [1, 2]. Through the efforts of USDA agro-economist Neal Peterson [3] the use of Pawpaw has increased in the past few decades due to his breeding of advanced cultivars (see Figure 1) and establishment of USDA satellite repositories for Asimina specimens [4]. The fruit is also being considered as a crop in other parts of the world [5].

There have been 4 genomic studies of the specimens assembled by Peterson. The first 3 were by Hongwen Huang, also known for his studies with chestnuts. Of them, the 1st in 2000 was a preliminary study to determine appropriate single-loci RAPD markers [6]. The 2nd in 2003 applied 71 of these markers to 37 specimens [7]. One pair of the specimens has synonymous marker values, thus bringing the usable total to 36. Also, 22 of the markers returned by the lab contained missing values and unfortunately the measurements could not be repeated. Regardless, Huang processed the data with the NTSYS-pc biostatistical package and employed two questionable practices: use of markers with missing values [8] and dissimilarity measurements with a pseudo-metric [9]. On a positive note, Huang published all the marker data which provides an opportunity to revisit the study. A few years later Huang published his 3rd study of the specimens, this time using AFLP markers [10]. This study also used pseudo-metric analysis and only published the resulting dissimilarity values. The fourth genomic study of the Pawpaws was by Pomper et al in 2010 [11]. Only 6 SSR markers were used leading to a grossly underdetermined data matrix. The data also contains many missing values and is thus of no use for further investigation.

The present effort involves rectifying the useable data from Huang's 2nd study with known ancestry data (Figure 2). Huang's original 71 markers are considered a balanced set which were then arbitrarily reduced to 49. As such, dissimilarity relations among the specimens are examined using subsets of the markers with the goal of identifying one or more marker groups that are meaningful with respect to ancestry records at an acceptable genetic distance resolution.

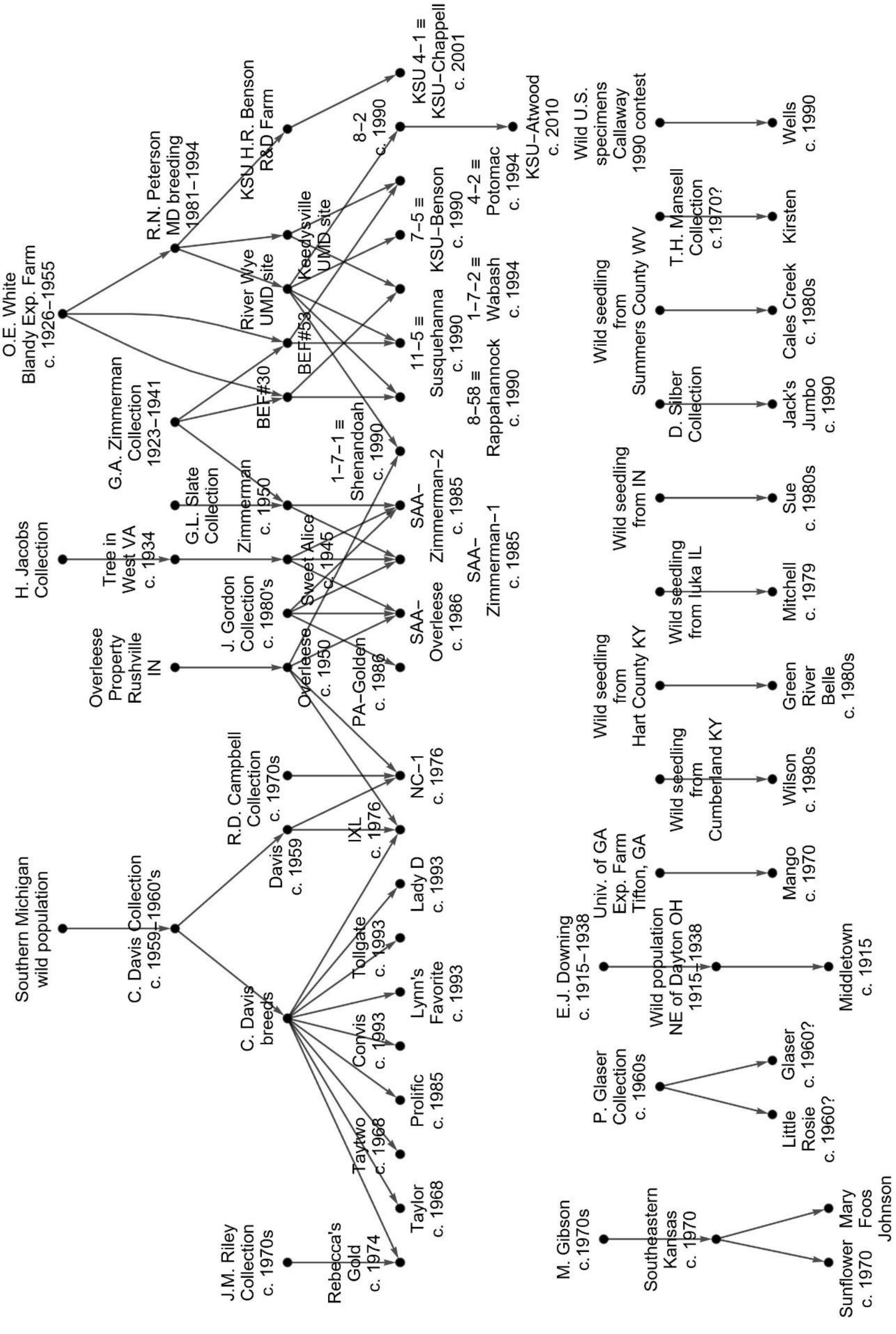

Figure 1. Known ancestry of U.S. Pawpaw cultivars currently in circulation [3, 11-13].

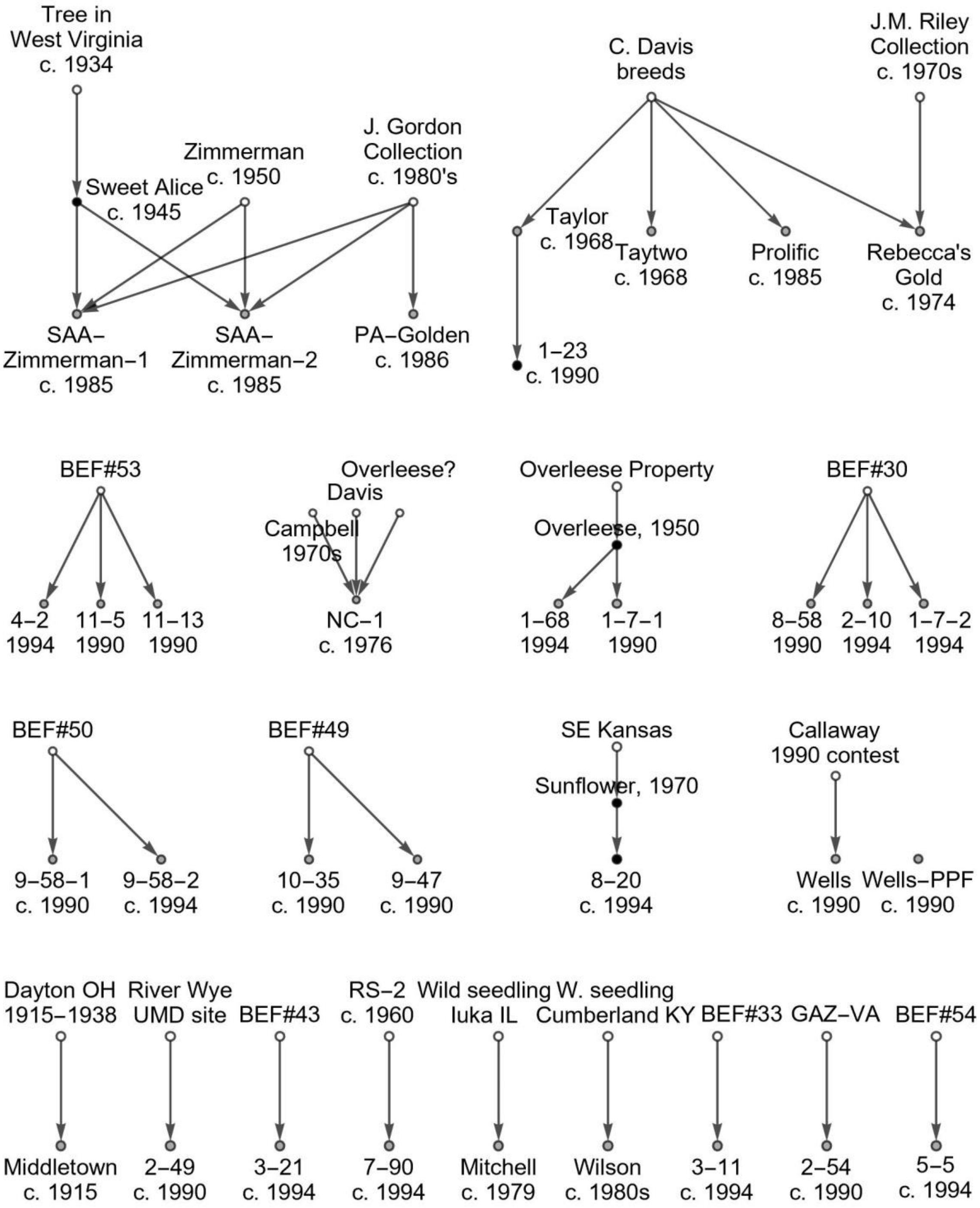

Figure 2. Ancestry and origins of specimens in H. Huang's RAPD study [3, 11]. The origin of Wells-PPF is unknown. BEF = Blandy Experimental Farm.

2. RESULTS

A candidate group of markers was identified after an exhaustive, automated search of 2,121,017 topological graphs produced by subsets of size 44 through 49 of Huang's error free markers – sans 17,393 sets which produced one or more zero distances. A complete distance graph G determined by the selected marker set was constructed along with a connected least bridges graph G_{LB} [14]. Four known parent-progeny pairs appeared as nearest neighbours in G_{LB}. The distribution of mismatches is shown in Figure 3 and minimum and maximum distances are exhibited in Figure 4. Loci mismatches of ancestry relations are given in Tables 1, 2.

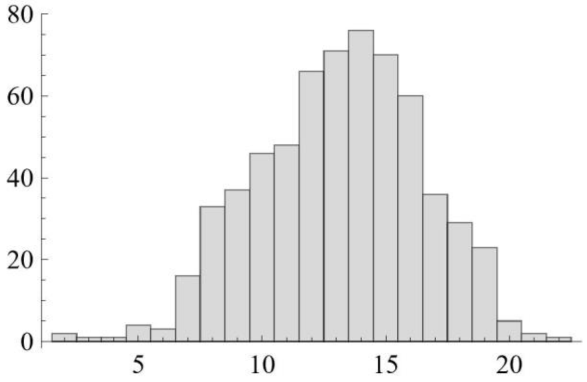

Figure 3. Distribution of loci mismatches in complete graph of selected marker set.

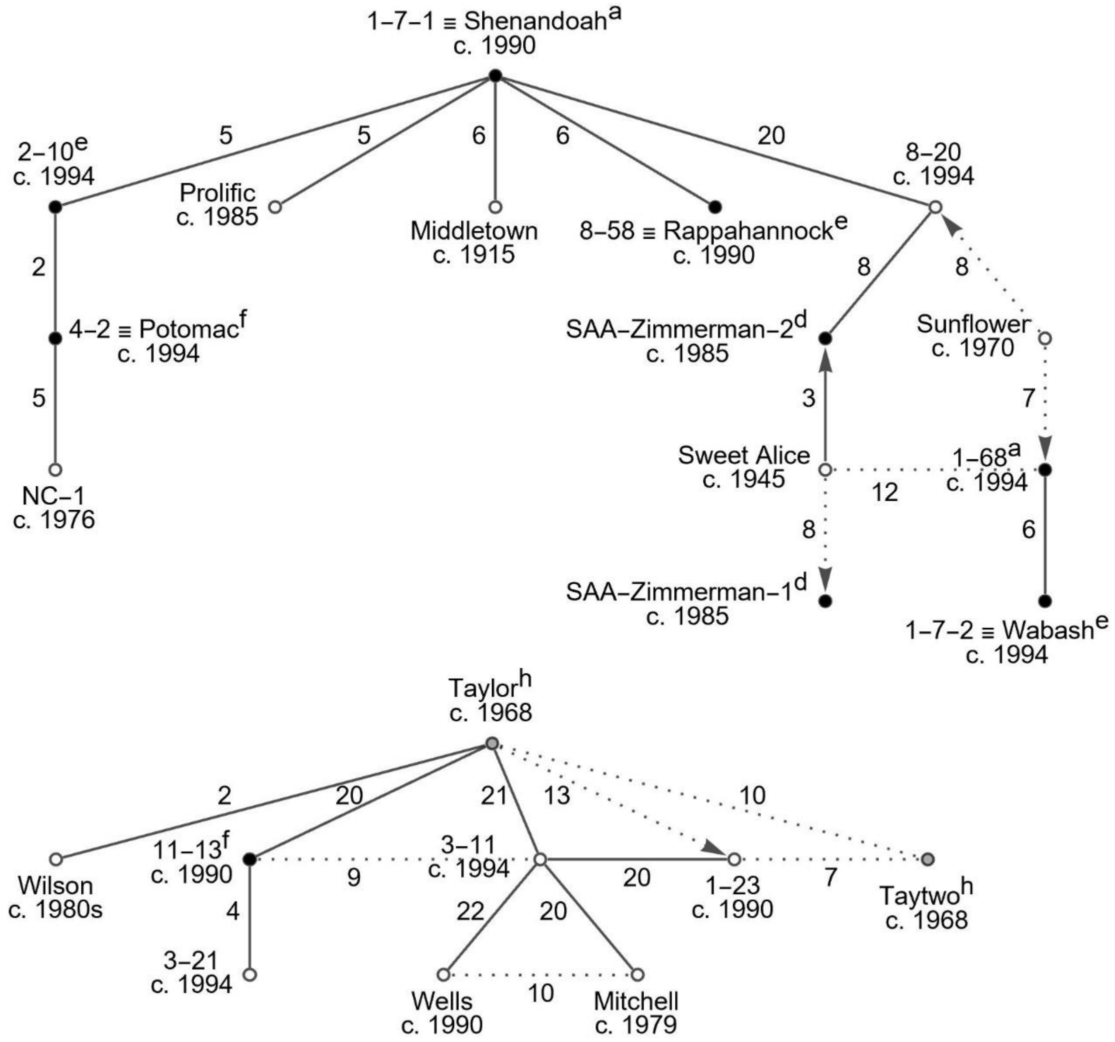

Figure 4. Combined graph of distance extrema shown with solid lines, plus selected neighbouring specimens for orientation denoted by dotted lines. Black vertices denote members of known sibling sets a-f while grey vertices are members of suspected sibling sets g and h (see Table 2). Arrowheads specify parent-progeny relations, otherwise spatial orientation is arbitrary.

Table 1. Known and suspected parent-progeny relations. Distance units are loci mismatches.

Parent	Progeny	Distance	Status
Overleese	1-68	8	Found
Overleese	1-7-1 ≡ Shenandoah	11	not nearest neighbour
Overleese	NC-1 (suspected)	7	Found
Sunflower	8-20	8	Found
Sweet Alice	SAA-Zimmerman-1	8	Found
Sweet Alice	SAA-Zimmerman-2	3	Found
Taylor	1-23	13	not nearest neighbour

Table 2. Known and suspected sibling relations.

Set #	Specimens	Distances
a	1-68, 1-7-1 ≡ Shenandoah	15
b	9-47, 10-35	13
c	9-58-1, 9-58-2	9
d	SAA-Zimmerman-1, SAA-Zimmerman-2	9
e	1-7-2 ≡ Wabash, 2-10, 8-58 ≡ Rappahannock	14, 15, 11
f	4-2 ≡ Potomac, 11-5 ≡ Susquehanna, 11-13	8, 10, 10
g (suspected)	PA-Golden, SAA-Zimmerman-1 and SAA-Zimmerman-2	10, 9
h (suspected)	Taylor, Taytwo	10

3. DISCUSSION

From the top of Figure 4, one observes the tight cluster of specimens NC-1, Potomac, Prolific, Middletown, Shenandoah, and Rappahannock. The cultivars 2-10 and Potomac are close enough to imply at least a sibling relationship. Off to the right note the long distance to the parent-sibling group of Sweet Alice and the SAA-Zimmermans, plus the adjacent group containing Sunflower and Wabash. Two of C. Davis' first cultivars Taylor and Taytwo are found below along with Wilson - a possibly undocumented offspring of Taylor. The numbered cultivar 11-13 – a sibling of Potomac appears there at great distance from Taylor indicating the large dissimilarity of these Davis breeds from the specimens above. Wells and Mitchell are found at the bottom – also dissimilar from those above and the Davis breeds. The specimen Wells-PPF is displaced by 5 mismatches from Wells, indicating one could be the progeny of the other. Peterson apparently believes the latter is the original.

The distance from Overleese to its progeny 1-7-1 appears excessive and from Taylor to its progeny 1-23 even more so. Since both parents have suitable distances between other relations, this calls into question the validity of the labels on 1-7-1 and 1-23. In the case of 1-7-1, the problem is further emphasized by its relatively large distance to sibling 1-68. For specimen 1-23, the discrepancy appeared in all marker subsets during the selection process.

In the long-debated case of Overleese vs its suspected sibling NC-1, a distance of 7 was found which is in the range of 3-8 found for other parent-progeny relations. However, it is also close to the range of 8-15 found in sibling relations. Consequently the debate appears unresolved by any genomic measurements performed to date.

The results of the present study are limited by relevance of the original marker set and the process of selection by ancestry records. Given the range of dissimilarities produced for known relations, the measurements here should be considered a coarse approximation to the actual displacements among specimens. Even so, the high correspondence (75%) between measurements and the known relations of Tables 1, 2 indicate that H. Huang's markers have merit. Therefore the author believes a retesting of the specimens using all 71 markers at a lab capable of producing error-free results would be beneficial.

4. METHODS

The data from H. Huang's paper was extracted using Adobe Acrobat® and placed in CSV files. The markers with missing data values were entirely deleted. Specimens 11-13-KSU and 11-13-PPF were found to have identical marker values and thus replaced by the single label 11-13. This vetted set contains 36 specimens with 49 markers each.

A software program was then constructed to iterate through progressively smaller subsets of the original size L = 49. For each subset, basic statistics such as distances in known relations was extracted, along with parameters of the least bridges graph [14] produced by the markers including the component maximal and a list of component vertices. Marker sets producing one or more zero distances were discarded for poor resolution. The number of zero producing marker sets increased from 0.085% at L = 47 to 0.85% at L = 44. Also at this latter size the resulting graphs suffered from too much cohesion and thus no smaller sizes were pursued. Elapsed execution times for this software program ranged from 0.2 seconds for L = 49 to 3.6 days for L = 44, including I/O.

A second program was written to rate the results. For each subset, a specimen pair from a known relation was considered "present" if both members of the pair occurred in the same component of the least bridges graph limited by δopt. Two vectors were formed from this data: numbers of known relations and numbers of suspected relations, with each value in the last columns of Tables 1, 2 representing a vector component. The 2-norm of the outer product of these vectors was then used as a score. From the scores a group of 291 candidates was produced. A high degree of duplication was noticed among the topological structures. The candidates were examined for cohesion properties and a best-of-class with L = 45 and δopt = 8 was selected. A connected graph of the selection is shown in Figure 5.

All computation and visualizations for this study were performed with Mathematica® versions 12 and 13. The hardware platform was a deskside Intel® i9-10900KF PC with 32GB RAM and 1TB SSD running Windows® 11. No compatibility issues were detected within this environment.

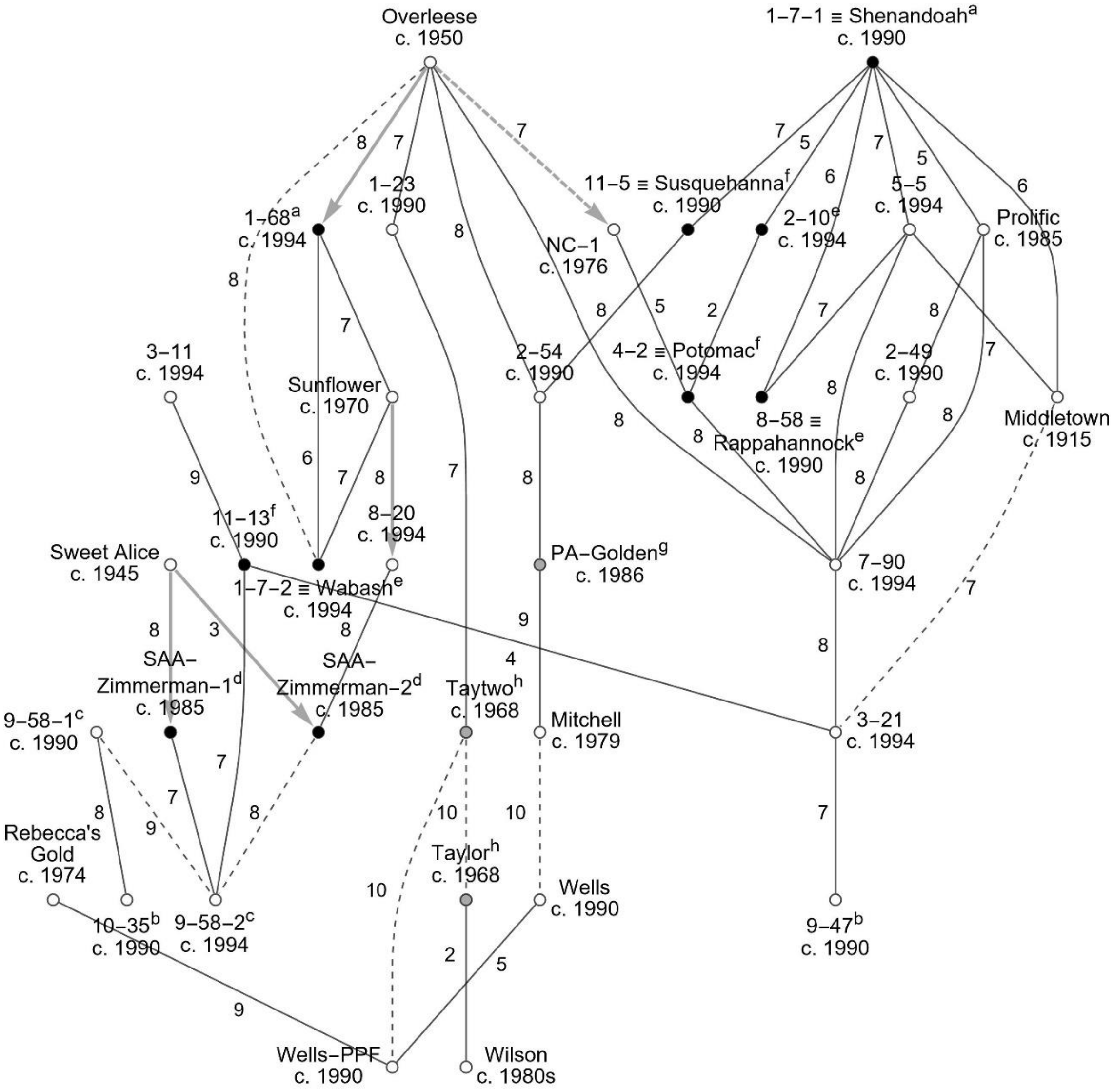

Figure 5. Least genetic distances between 36 Pawpaw cultivars tested by Huang et al [7]. Distances represent # of loci mismatches between a rectified set of 45 markers from Huang's original error-free set of 49. Orientation of Pawpaws is arbitrary except for solid arrows indicating parent-progeny relations. Dashed arrow indicates suspected parent-progeny relation. Solid lines (not arrows) are nearest-neighbour relations and dashed lines are least bridges. Names assigned upon release of a breed to the nursery industry are specified by “≡”. Labels with superscripts a-f are sets of known siblings, while g-h are sets of possible siblings.

References

[1] R. N. Peterson, "PAWPAW (ASIMINA)," *Genetic Resources of Temperate Fruit and Nut Crops 290*, pp. 569-602, 1991. [Online]. Available: https://www.actahort.org/books/290/290_13.htm.

[2] C. Ferrer-Blanco, J. Hormaza, and J. Lora, "Phenological growth stages of “pawpaw”[Asimina triloba (L.) Dunal, Annonaceae] according to the BBCH scale," *Scientia Horticulturae,* vol. 295, p. 110853, 2022, doi: https://doi.org/10.1016/j.scienta.2021.110853.

[3] R. N. Peterson, "Pawpaw variety development: a history and future prospects," *HortTechnology,* vol. 13, no. 3, pp. 449-454, 2003, doi: https://doi.org/10.21273/HORTTECH.13.3.0449.

[4] "NCGR Corvallis - Asimina Germplasm." USDA ARS. https://www.ars.usda.gov/pacific-west-area/corvallis-or/national-clonal-germplasm-repository/docs/ncgr-corvallis-asimina-germplasm/ (accessed 2022).

[5] R. G. Brannan and M. N. Coyle, "Worldwide Introduction of North American Pawpaw (Asimina triloba): Evidence Based on Scientific Reports," *Sustainable Agriculture Research*, vol. 10, no. 3, pp. 1-19, 2021, doi: https://doi.org/10.5539/sar.v10n3p19.

[6] H. Huang, D. R. Layne, and T. L. Kubisiak, "RAPD inheritance and diversity in pawpaw (Asimina triloba)," *Journal of the American Society for Horticultural Science*, vol. 125, no. 4, pp. 454-459, 2000, doi: https://doi.org/10.21273/JASHS.125.4.454.

[7] H. Huang, D. R. Layne, and T. L. Kubisiak, "Molecular characterization of cultivated pawpaw (Asimina triloba) using RAPD markers," *Journal of the American Society for Horticultural Science*, vol. 128, no. 1, pp. 85-93, 2003, doi: https://doi.org/10.21273/JASHS.128.1.0085.

[8] 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.

[9] R. Frost, "Re-evaluation of NCGR Davis *Ficus carica* and *palmata* SSR profiles," *PLoS ONE*, vol. 17, no. 2, p. e0263715, 2022, doi: https://doi.org/10.1371/journal.pone.0263715.

[10] Y. Wang, G. L. Reighard, D. R. Layne, A. G. Abbott, and H. Huang, "Inheritance of AFLP markers and their use for genetic diversity analysis in wild and domesticated pawpaw [Asimina triloba (L.) Dunal]," *Journal of the American Society for Horticultural Science*, vol. 130, no. 4, pp. 561-568, 2005, doi: https://doi.org/10.21273/JASHS.130.4.561.

[11] 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.

[12] K. W. Pomper, S. B. Crabtree, and J. D. Lowe, "The North American Pawpaw Variety:'KSU-Atwood (TM)'," *Journal of the American Pomological Society*, vol. 65, no. 4, pp. 218-221, 2011. [Online]. Available: https://www.pubhort.org/aps/65/v65_n4_a6.htm.

[13] K. Gasic, J. E. Preece, and D. Karp, "Register of new fruit and nut cultivars list 50," *HortScience*, vol. 55, no. 7, pp. 1164-1201, 2020, doi: https://doi.org/10.21273/HORTSCI50register-20.

[14] R. Frost. "Least Bridges Graphs." https://frostconcepts.org/LeastBridgesGraphs/ (accessed 2022).

Author

Richard 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.