How 'Bout 'Dem Apples: Has Modern Agriculture Made Apple Cultivars More Related And Less Genetically Diverse?

Michael Jordan

Abstract

This paper uses a data set of single nucleotide polymorphisms (SNPs) from Canada's Apple Biodiversity Collection (ABC) to assess whether apple cultivars have become more related and less genetically diverse.

Introduction

The domestic apple (*Malus domestica*) is one of the oldest, most commonly cultivated, and most valuable fruit crops in the world, ranking second only to the banana in gross production value (O'Rourke, 2021). Although today *M. domestica* is produced by many countries such as America, Chile, and New Zealand, it originated in Central Asia and evolved from other *Malus* species, primarily *M. sieversii* (Volk et al., 2021). Following domestication, *M. domestica* spread from Asia along the Silk Road to Europe and beyond (Volk et al., 2021).

As *M. domestica* spread, trees with desirable traits were developed as cultivars through vegetative propagation (Volk et al., 2021). Asexual reproduction was necessary because apples are typically self-incompatible and highly heterozygous, meaning that cultivars must be cross-bred and offspring will exhibit traits not expressed in either parent tree (Volk et al., 2021). Therefore, asexual reproduction was required to maintain desirable traits.

Prior to the industrial era apple production was regional, with small farms choosing new cultivars to develop largely through trial and error or chance pollination events (O'Rourke, 2021). After industrialization the industry became increasingly structured, complex, and globally integrated as new technology allowed apples to be sold over greater distances, leading to the large vertically integrated firms of today (O'Rourke, 2021). As industry organization became more sophisticated, so to did apple breeding strategies. Controlled breeding programs were established by the early 1800s (Khan et al., 2021), during the 20th century the apple industry embraced genomics, and today it uses modern technologies like SNP arrays to develop new cultivars (Volk et al., 2021).

Although these structural and technological developments facilitated a large globally connected industry, this may have come at the cost of genetic diversity. Research has shown that many accessions in collections are clonally related to just a few commercially dominant cultivars (Migicovsky et al., 2021). I hypothesize that this is a recent phenomenon, with apple cultivars becoming more related and less genetically diverse over time as the industry has consolidated and used more sophisticated techniques to select for traits desired by the global market. If so, this would support the continued maintenance of heritage cultivars as an important reservoir of potentially valuable genotypes.

Methods

The data and code behind this report can be found at https://github.com/Skirnir3141/AppleProject. To evaluate my hypothesis, I used a data set of 278,231 SNPs from 1,175 accessions in Canada's Apple Biodiversity Collection (ABC) (Migicovsky et al., 2022). To describe these accessions, I used a data set from a phenomic review of the ABC (Watts et al., 2022). I filtered accessions using R version 4.3.2 to include only *M. domestica* with either a cultivation or a release year (earliest was used). Because SNP data did not account for non-diploids, I dropped triploids by using the het function in Plink v1.90 to estimate heterozygosity and filtering out high values (see PloidismExploration.R). This resulted in 446 remaining accessions, which I split into three time periods – 1800-1899, 1900-1959, and 1960-Present – reflecting stages in apple industry development.

SNP quality control was conducted in Plink. SNPs were pruned for linkage disequilibrium (indep-pairwise 10 3 .05). SNPs with a minor allele frequency of less than .01 and a missingness of greater than .05 were removed. Accessions with greater than .1 missingness were removed. This resulted in 154,167 SNPS and 131 accessions for P1, 148,692 SNPs and 158 accessions for P2, and 149,354 SNPs and 148 accessions for P3.

To evaluate relatedness, I estimated identity by descent (IBD) using Plink. Accessions with a $\hat{\pi}$ (proportion of inheritance) greater than .125 (a 3rd degree relationships or closer) were considered related. Related accessions were visualized using the tidygraph and ggraph R packages with a Fruchterman-Reingold (FR) layout algorithm using $\hat{\pi}$ as a weight. To evaluate genetic diversity, I calculated the genetic hamming distance (i.e., the percentage of SNPs that differ) using Plink and conducted PCoA with the cmdscale function in R.

Results

Per Figure 1a, the percentage of accessions with at least one 3rd degree or closer relationship increased from 44% in P1 to 59% in P2 and to 73% in P3. The relatedness of accessions with at least one relationship also increased. This can be seen visually in Figures 1b-d. The FR algorithm will place more related accessions closer together and increased node density is evident across periods. Extracting global efficiency scores – which are inversely related to the distance between nodes – confirms this. Global efficiency increased from 12% in P1 to 86% in P2 and to 115% in P3, indicating shorter distances between nodes. Figures 1b-d also exhibit higher connectivity over time. Average node degree increased from 1.9 in P1 to 4.7 in P2 and to 7.9 in P3.

In contrast, genetic diversity did not decrease. Per Figure 1e, a PCoA of genetic distance shows no notable difference across periods. In this PCoA, accessions with a similar genetic distance will be grouped closer together. The graph clearly shows that accessions in all three periods occupy a roughly contiguous area of graph space, with no changes in density or different groupings evident across periods.

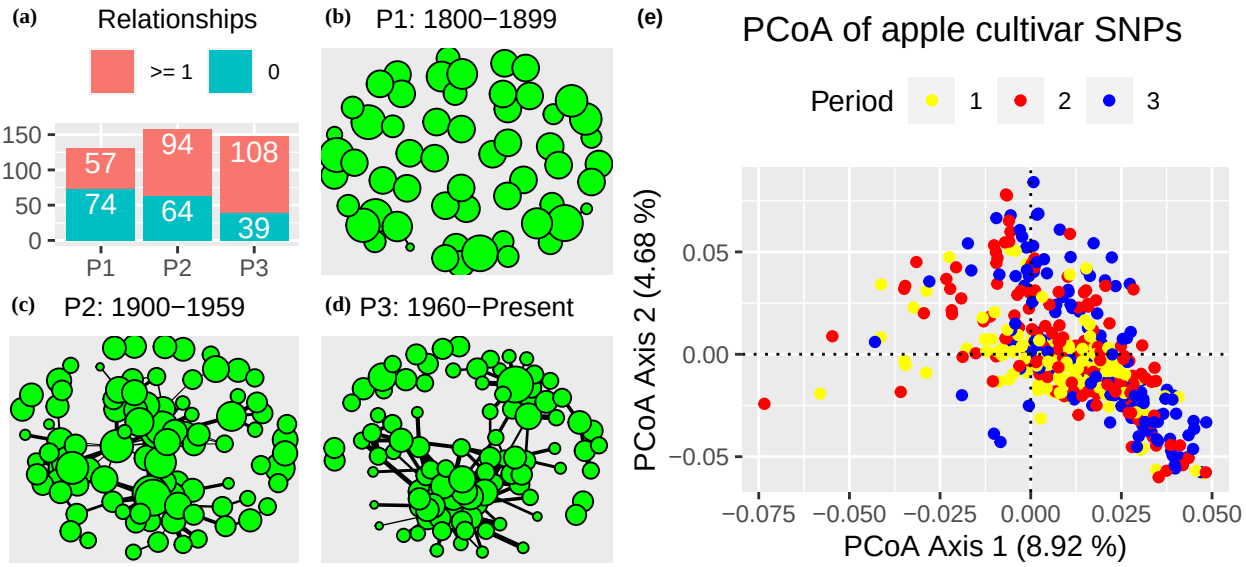

Figure 1: (a) Bar plot of accessions with 0 and >= 1 relationship by period, (b-d) Network graphs of related accessions for each period, (e) PCoA of genetic distance.

Discussion

There are several caveats to this analysis. First, the time periods used were defined semi-arbitrarily. Ideally I would have defined time periods based on precisely when new breeding strategies were applied, but this

information did not arise in my literature review. Second, my data set was not comprised of a random sample of cultivars from each period but instead of cultivars that happened to be preserved in the ABC. If preservation was non-random, then my results may be a biased measure of relatedness and diversity (in fact the ABC does have a geographic bias toward Canada and the US). Third, Plink's estimation of IBD is imperfect. Researchers have compared Plink to other methods and found significant differences (Stevens et al., 2011). That said, spot checking confirms that Plink's results are reasonable. For example, Plink identified Sir Prize and Clear Gold as having a $\hat{\pi}$ of 56% indicating a 1st degree relationship. This makes sense, since Clear Gold is a mutation of a Golden Delicious and Sir Prize is a cross between Golden Delicious and another cultivar. Finally, excluding triploid cultivars does limit the scope of these findings, since many commercially successful cultivars are triploid.

Caveats aside, I think these results are cause for both concern and some peace of mind regarding apple cultivar genetic diversity. On the one hand, newly developed apple cultivars are becoming increasingly related to each other over time as indicated by an increase in both the proportion of cultivars with at least one third degree relationship or closer and an increase in the $\hat{\pi}$ of such cultivars. On the other hand, genetic diversity as measured by the genetic distance between cultivars does not seem to have changed over time. My speculative explanation of these two seemingly disparate results is that the high heterozygosity of apples may have prevented a decrease in genetic variation up to now despite the increasing relatedness of new cultivars. However, it's worth observing that the hamming distance of SNPs just measures the percentage of overall difference between accessions. Not all of these differences will be coding or pertain to commercially relevant phenotypes. I may have reached a different conclusion if I had focused only on SNPs that code for relevant traits.

Conclusion

While it is comforting to know that we are not at a stage of apple domestication where breeding has significantly constrained genetic diversity, I would still advocate for continued preservation of heritage apple cultivars. Although modern breeding strategies have not yet significantly constrained apple genetic diversity, they may in the future and in that case being able to select for new traits from heritage cultivars will be important. Also, I'd argue that heritage cultivars represent a window into what phenotypes people found valuable in the past, which is in a sense a type of cultural heritage that in my opinion is worth preserving.

Reference List

Khan, A., Gutierrez, B., Chao, C.T., & Singh, J. (2021) Origin of the Domesticated Apples. In: Podwyszyńska, M., & Marasek-Ciołakowska, A. (eds.) (2021) *The Apple Genome*. Springer Cham, pp. 383-394.

Migicovsky, Z., Douglas, G.M., & Myles, S. (2022) Genotyping-by-sequencing of Canada's apple biodiversity collection. *Genomics of Plants and the Phytoecosystem*. 13.

Migicovsky, Z., Gardner, K.M., Richards, C., Chao, C.T., Schwaninger, H.R., Fazio, G., Zhong, G., & Myles, S. (2021) Genomic consequences of apple improvement. *Horticultural Research*. 8.

O'Rourke, D. (2021) Economic Importance of the World Apple Industry. In: Podwyszyńska, M., & Marasek-Ciołakowska, A. (eds.) (2021) *The Apple Genome*. Springer Cham, pp. 1-18.

Stevens, E.L.,Heckenberg, G., Roberson, E.D.O., Baugher, J.D., Downey, T.J., & Pevsner, J. (2011) Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. Plos Genetics.

Volk, G., Cornille, A., Durel, C., & Gutierrez, B. (2021) Botany, Taxonomy, and Origins of the Apple. In: Podwyszyńska, M., & Marasek-Ciołakowska, A. (eds.) (2021) *The Apple Genome*. Springer Cham, pp. 19-32.

Watts, S., Migicovsky, Z., McClure, K., Yu, C., Amyotte, B., Baker, T., et al. (2021) Quantifying apple diversity: A phenomic characterization of Canada's Apple Biodiversity Collection. *Plants People Planet*. 3 (6), 747-760.

Working Style Assessment

I enjoyed working on this mini project quite a bit. Overall it confirmed a lot of what I know about the pros and cons of my working style.

On the pro side, I found it deeply satisfying to learn about so many different subjects, whether that was a specific research topic like apples, a broad field like genomics, or technical details like genomic analysis software or ordination. Far from being a chore, each time my analysis hit a snag and I needed to learn about a new subject, I was excited to tackle that challenge. I was also able to pivot effectively and make judicious decisions about when to cut my losses rather than keep investing time in a losing approach. For example, after developing some background in apples as a subject area and finding the SNP data, my initial plan was to use it to conduct a Genome Wide Association Study. However, by the end of a full day of work I had run into a lot of technical and interpretive snags that I was worried I wouldn't be able to solve in the limited time available to me. Although I was quite frustrated, I had the presence of mind to not push my luck and work down to the wire hoping to resolve these issues. I also didn't throw away all of the background I had developed. Instead, I pivoted to conducting an analysis that still relied on SNP data and still answered an interesting question, but used simpler functions and relied on more easily interpreted approaches. This decision allowed me to complete the bulk of the work by the end of the first week and spend the next week polishing my draft and writing the presentation.

On the con side, I did at times exhibit downsides of the curiosity I have for the subjects I was learning: analysis paralysis and chasing after shiny objects. I'm well acquainted with both dynamics from my years working in analytical roles in tech and I and pretty much any other analyst will occasionally be caught up in them. Analysis paralysis describes a situation in which the analyst continues to assess a problem from different angles far past the point of diminishing returns when it would be preferable to simply pick an approach and move on. Chasing after shiny objects involves an analyst getting excited by some interesting, but ultimately inconsequential piece of information or technique. Overall I didn't get too sunk into either, but certainly there were moments when I found myself doing one or the other (e.g., I caught myself reading a textbook on Network Analysis chapter-by-chapter when really all that was necessary and practicable given the scope of this project and the time limit was skimming for relevant content). In my experience, this kind of thing usually happens when someone is a bit stressed or tired and the best solution is to simply close your laptop and take a walk.

This leaves me feeling ready to tackle my full research project. Certainly making good decisions about time usage and being willing and able to pivot if I hit snags will come in handy in the full research project. Overall I'm really enjoying putting together the technical and subject area expertise I've developed so far in the program and applying them to interesting problems. I think I'm going into it with the right attitude and the skills to succeed.