Building a genome-based phylogenetic tree for mango cultivars

Background

Mangifera indica, native to India, is a popular fruit crop around the world, with over one thousand different cultivars (varieties). Its genome has been assembled quite recently—within the past 5 years—opening a plethora of new avenues for analysis at the genomic level. Currently, there appears to be 5 mango cultivars whose genomes have been sequenced, with 4 that are assembled and accessible online: Tommy Atkins, Alphonso, Hong Xiang Ya, and Amrapali.

My research approach involves building relations between these recently assembled genomes. While there exists a phylogenetic tree that compares the relationship between the mango species and its close relatives, there does not seem to exist such a tree that compares between mango cultivars themselves. Mango varieties, while all belonging to the same species, have been found to exhibit high heterozygosity (Wang et al., 2020) and can vary dramatically in size, color, and taste. Building a phylogenetic tree for mango cultivars and comparing genomic differences may provide some insight towards the evolutionary history of the mango cultivars.

Along with building a phylogenetic tree for the 4 aforementioned mango cultivars, I also looked to compare genes that correspond to prominent phenotypes, such as skin color and fruit shape, between cultivars. A consensus genetic map of the mango exists in the literature (Kuhn et al., 2017), of which I used to consolidate genetic information of prominent mango phenotypes in the mango genomes of interest. Performing this kind of analysis can be of use to future mango research by revealing genetic differences that may correlate towards phenotypic differences.

Research Question & Hypothesis

The main question my research posed was in regards to the genomic relationship between mango cultivars. How closely are mango cultivars related to each other at the genetic level, and to what extent do their differences correspond to mango-distinguishing phenotypes such as shape and color?

I hypothesized that mango cultivars of similar appearance (color, shape, size) would have more similar genomes than mango cultivars that exhibit more differences in phenotype. I predicted the phylogenetic tree to reflect these phenotypic differences and relations. For instance, perhaps the Hong Xiang Ya mango and the Alphonso mango may be more closely related to each other than to the Tommy Atkins mango, due to their similarity in color. However, there would certainly be many phenotypic traits (more than just color alone) that can altogether affect the genetic similarity between these 4 mango cultivars.

Methods

I utilized the available genomic data on mangobase.org to create the phylogenetic tree of the 4 aforementioned mango cultivars (Bally et al., 2021). An additional pistachio genome was used from NCBI to serve as an outgroup in tree construction. To locate genes that are associated with phenotypes of interest, I drew from the consensus genetic map created by Kuhn et al. (2017).

Figure 1. Images (from Google) of the four mango cultivars used for this project.

To proceed with the phylogenetic tree-building, I first selected genes based on their association with phenotype in the map, then extracted them from each genome. The paper provided only segments of each gene, so I performed a BLAST search on NCBI to obtain the original complete mRNA transcript. Afterwards, I ran a local BLAST on the 4 mango genomes as most of the genomes were not listed on NCBI. This process involved a lot of copying and pasting as the local BLAST returned positional information of the alignments, of which I had to manually extract and remove introns from using samtools (thus creating a mRNA transcript of each gene for each genome).

Once all of the genes were extracted, I ran MAFFT to generate a multiple sequence alignment, which was then processed in phyml to construct a phylogenetic tree using maximum likelihood. To visualize the tree, I used phytools in RStudio as well as FigTree.

Results

Altogether, I used 9 genes from the mango consensus genetic map, genes that showed a significant association with one of the following phenotypes: blush intensity, ground skin color, pulp color, and beak shape. This resulted in a total of 45 genes extracted to construct the phylogenetic tree. Each of these genes coded for a protein of (usually) known function; for instance, the gene labeled SSKP003C1_C682T coded for a ethylene-forming enzyme and was found to associate with the blush intensity trait in mangoes.

Gene ID	Gene description	Trait association
Mi_0135	Copper amine oxidase family protein	Ground skin color
SSKP009C1_A627T	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein	Ground skin color
Mango_c48384	Core-2/I-branching beta-1;6-N-acetylglucosaminyltransferase family	Beak shape

	protein	
Mi_0450	BAK1-interacting receptor-like kinase 1	Blush intensity
SSKP003C1_C682T	Ethylene-forming enzyme	Blush intensity
Mi_0145	MYB-like 102	Blush intensity
Mi_0277	Lumazine-binding family protein	Blush intensity
Mi_0217	Plant protein of unknown function (DUF946)	Pulp color
Mi_0029	Galactose oxidase/kelch repeat superfamily protein	Pulp color

Table 1. List of genes extracted, their description, and trait association. (Kuhn et al., 2017)

Once these genes were extracted from each of the four mango genomes and the pistachio genome, all 45 sequences were concatenated into one file to run MAFFT. This resulted in an output FASTA file that formatted in a way that could be converted into a .phy file and further interpreted by phyml in bash.

After running phyml, a txt file was produced which I then used to plot the tree in RStudio (using phytools). Additionally, I rerooted the tree using the tip label of the tree (obtained using the View method). The final tree from RStudio is shown in the figure below. Note that labels were also added for ease of identification of each gene with its associated phenotype.

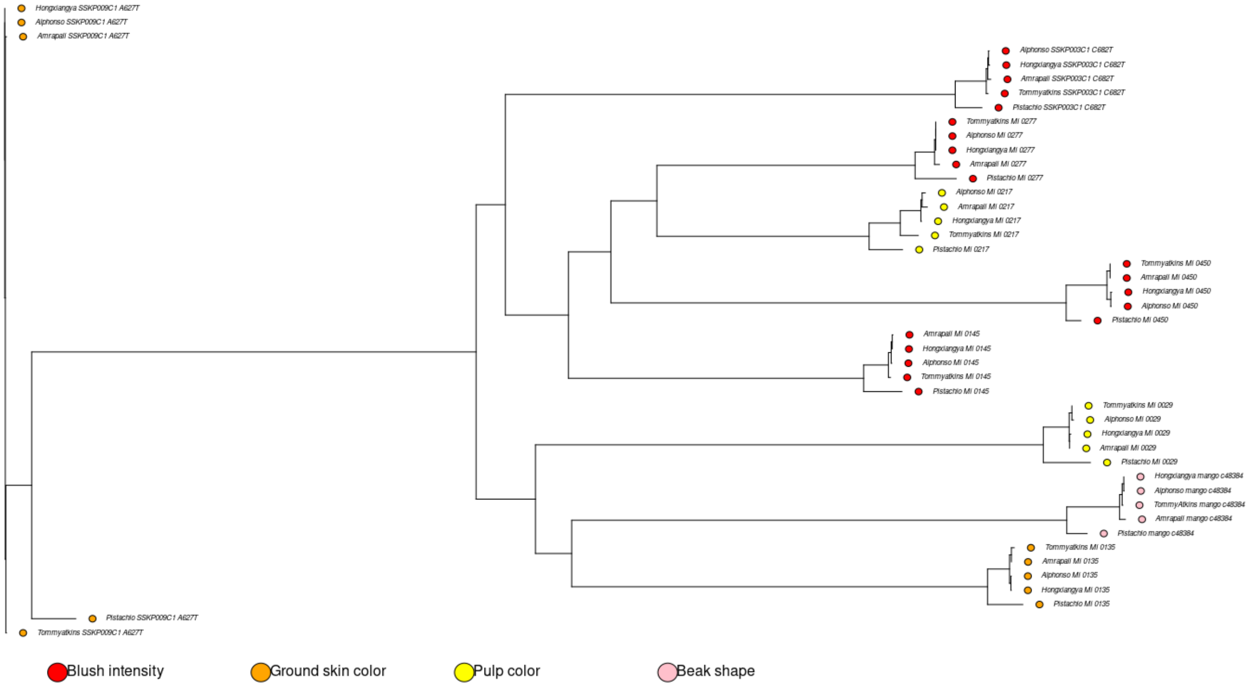

Figure 2. Rerooted annotated phylogenetic tree using phytools in RStudio.

At first glance, I noticed that the branches at the end of the tree were much shorter than the rest of the branches, suggesting that there is not much variation between genomes for the same gene. This was expected as the genomes for this project (except for the pistachio) were all cultivars of the same species. As an additional step, I opened the txt file in FigTree for a better visualization of the tree topology.

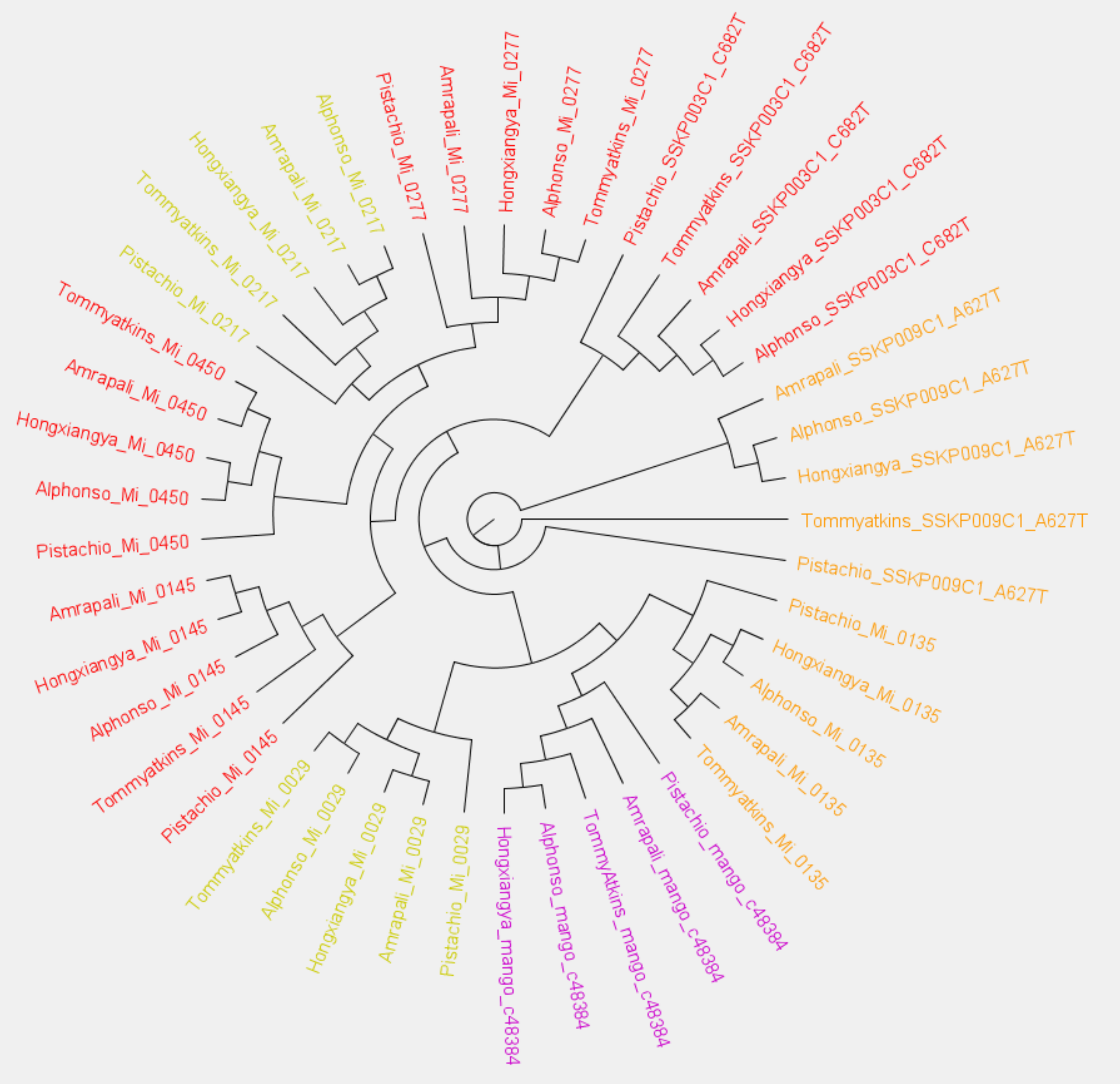

Figure 3. Annotated phylogenetic tree using FigTree. Note the branches are not proportional to the relative similarity between genes. The colors in this tree match the legend of Figure 2: red = blush intensity, orange = ground skin color, yellow = pulp color, pink = beak shape

Discussion

Regarding my research question, not many conclusions could be drawn using the above results. The Alphonso and Hong Xiang Ya cultivars were found to be phylogenetically closest to each other for genes corresponding to ground skin color, which does match my hypothesis as both of those cultivars are similar in color (yellow). However, it cannot be said with confidence that this is the case for all phenotypes, or even for all genes related to ground skin color for other mango genomes. For instance, while Hong Xiang Ya and Amrapali appear more similarly shaped (longer with some curl at the end), the phylogenetic tree indicated that Alphonso and Hong Xiang Ya were more closely related for the gene associated with beak shape. Additionally, for the various blush intensity genes, there did not appear to be a consistency for relationships between the cultivars (other than that the pistachio was the least related out of all genes). The same can be said about pulp color.

There are many reasons why the results could have turned out this way. First, I was only able to look at 1-4 genes per phenotype due to time constraints and data availability. There were more blush intensity genes that I could have looked at, and by contrast, there were only 2 pulp color genes, based on what was provided by Kuhn et al. (2017). It is also possible that my hypothesis is simply wrong and that there is no significant correlation between similarity in fruit phenotype and genetic similarity between these 9 genes. I had only looked at fruit phenotypes rather than the appearance of the entire organism, and as many of these genes code for proteins that can have a function throughout the entire plant, it is possible for genome similarity to be more strongly associated with protein function, or phenotypes of the mango tree itself. Regardless, if more genes could be found that are associated with the same phenotype, I believe that would still help with gaining more insight on the true level of association between genetic similarity and phenotypic similarity.

References

Bally, I. S. E., Bombarely, A., Chambers, A. H., Cohen, Y., Dillon, N. L., Innes, D. J., Islas-Osuna, M. A., Kuhn, D. N., Mueller, L. A., Ophir, R., Rambani, A., Sherman, A., Yan, H., & Mango Genome Consortium. (2021). The ‘Tommy Atkins’ mango genome reveals candidate genes for fruit quality. *BMC Plant Biology*, *21*(1), 108. https://doi.org/10.1186/s12870-021-02858-1

Kuhn, D. N., Bally, I. S. E., Dillon, N. L., Innes, D., Groh, A. M., Rahaman, J., Ophir, R., Cohen, Y., & Sherman, A. (2017). Genetic Map of Mango: A Tool for Mango Breeding. *Frontiers in Plant Science*, *8*(577). doi: 10.3389/fpls.2017.00577

Wang, P., Luo, Y., Huang, J., Gao, S., Zhu, G., Dang, Z., Gai, J., Yang, M., Zhu, M., Zhang, H., Ye, X., Gao, A., Tan, X., Wang, S., Wu, W., Cahoon, E.B., Bai, B., Zhao, Z., Li, Q., … Chen, Y. (2020). The genome evolution and domestication of tropical fruit mango. *Genome Biology, 21*(60). https://doi.org/10.1186/s13059-020-01959-8