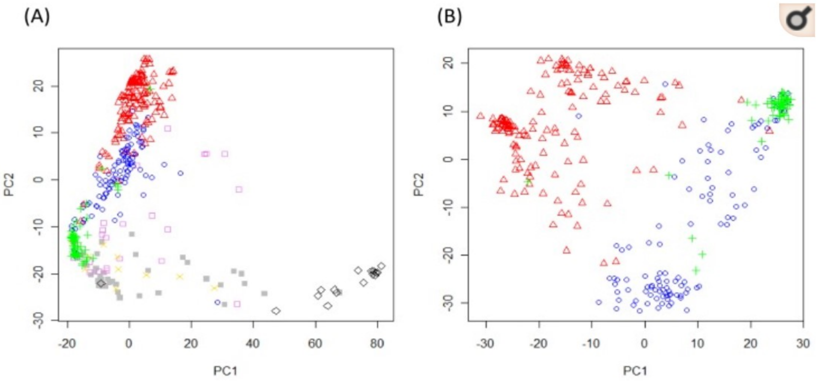
Chinmay Raut
(93345289)

Replicating the PCA of a published study (2 page write-up version)

High-Density SNP Genotyping of Tomato (Solanum lycopersicum L.) Reveals Patterns of Genetic Variation Due to Breeding (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3447764/)

by: Sung-Chur Sim, Allen Van Deynze, Kevin Stoffel, et al.

Figure 1



Principal component analysis (PCA) based on 4,393 SNP markers.

The PCA was conducted separately using data for all sub-populations of the SolCAP germplasm (A) and data for only the three large-fruited cultivated sub-populations consisting of the processing, fresh market, and vintage accessions (B). The processing accessions are indicated Δ (red); fresh market, ○ (blue); vintage, + (green); cultivated cherry, □ (violet); landrace, × (gold); wild cherry, ▪ (gray); and *S. pimpinellifolium,* ◊ (black).

About the paper:

The study took a collection of 426 tomatoes (410 inbred & 16 hybrids) and had each of their SNP's genotyped. This meant that they sequenced the tomato's genome, and then identified which nucleotide was present at each of the 7,000+ Single Nucleotide Polymorphism sites further marking which nucleotide came from which parent. Then on this data they performed Principal Component analysis using the SNP sites to verify that the tomatoes could be segregated into their corresponding subpopulations (originally defined from their phenotypes) through this SNP data of each tomato. Then using the variances and contributions they had from each of the SNP sites in their PCA they were able to determine which SNP's were the most influential in segregating the tomatoes into their subpopulation which in turn highlighted the chromosomes and sites of interest as the contributors of the phenotypic differences between the plants. These phenotypic differences were deemed important as these tomatoes have been bred by people from different cultures to achieve these certain traits and understanding the genetic background and function could prove useful while engineering new behaviors in such plants.

The study found that in even inbred tomatoes (tomatoes which when mating with themselves produce perfect clones due to histories of inbreeding) there exists high levels of genetic diversity among the SNP's. They also learned that chromosomes 2, 4, 5, 6, and 11 in particular host the SNP's which explained much of the differences observed in the subpopulations from the PCA's. Through the use of F-tests and studying the Minor Allele frequency patterns they confirmed these regions as influential.

The Data Tables Used & contents:

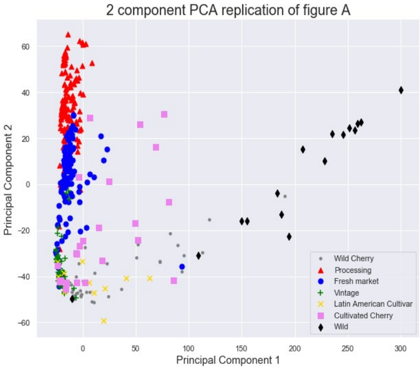
TableS1 = Table of all subjects and metadata associated with each subject. Each subject is on its own row. Used primarily to map subjects to subspecies of tomato.

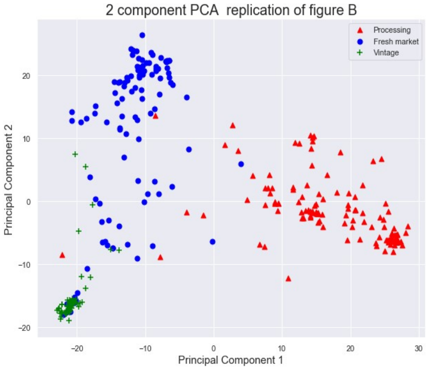
TableS2 = Main SNP data table. Consisted of SNP's as rows and 1 column for each subject. The cells contained the nucleotides that each subject had at that chromosomal position (1 from mom & 1 from dad). This was the main table referred to for the PCA. The nucleotides were translated to numeric arbitrarily trying different orderings were not observed to change the PCA in any way.

TableS4(MoneyMaker) = Table of 4,000+ SNP’s that the title of the figure mentions from the paper. This data came from a different paper from S.C. Sim and was used to recreate Figure 1A only.

TableS6,7,8 = Table of loci that were selected by the paper to be significant in differentiating the subpopulations of processing & fresh-market & vintage. The SNP’s of interest were the rows and were aligned with their chromosome position and Probability that the F-statistic of the simulation mapping would be less than that of the F-statistic of its physical mapping. Basically I used these 3 tables to identify the SNP’s that were significant in the PCA of figure 1B.

Results of my attempt of the PCA (explanations & thoughts in Jupyter Notebook):


2 component PCA replication of figure A


2 component PCA replication of figure B

Why the PCA is important & what it shows:

The PCA is important as it allows us to make sense of the vast amount of genomic data that each organism contains and find connections between the genome of organisms in a common species to further understanding the genome of that specific species and identify regions of interest within the species' genome which have phenotypic significance. To put the amount of data that the PCA is able to sift through in a matter of minutes, this SNP table was 400+ organisms with 7000+ dimensions of information which needed to be condensed to 2 dimensions to even visualize. A person cannot begin to fathom processing such data by hand, nor without linear algebra would they know where to start. Overall the PCA is important as it provides a way to interpret this vast amount of genomic and SNP data regarding the species under study.

Moving away from the importance of the PCA, the analysis identifies a few interesting things. First, that there is a clear correlation between subpopulation of the tomato species and the SNP data obtained from a single genome assembly of the tomato. The groups can even be circled in the graphs produced either by hand or k-means clustering. This distinction might sound obvious on paper, different tomato subspecies have differences in their genetic code and tomatoes from the same subspecies would have fewer differences, but the PCA goes further and says instead of looking at the entire genome we only need to care about ~7000 Single Nucleotide Differences (Polymorphisms). Figure 1B goes further and says that we need to care about fewer SNP's, only about 1100. This means that identifying and targeting these differences in tomato genomes could potentially begin with these specific nucleotides. The paper took the results of the PCA a bit further & identified the exact the SNP's which contributed to the PCA the most & determined which chromosomes that they originated from. Also the group mentioned how the SNP's and chromosomes identified subverted their expectations specifically in the case that the largest chromosome didn't contain the most significant SNP's whereas some smaller ones did, provoking more questions regarding the uniformity of SNP significance across chromosomes.

Why I chose this paper:

I chose this paper because it ran PCA on a fairly large dataset and had much to say about the results of the analysis. Also because initially I thought that SNP data would be easy to figure out and analyze due to my Hubris, but once I realized I was super wrong about everything I stuck to it due to my pride and my desire to complete things properly.