This analysis uses the dataset and figures generated by Huang et. al. in “Rewiring of the Fruit Metabolome in Tomato Breeding” (2018). This paper demonstrates how domestication of the tomato, and selection for high weight drastically shifted the metabolomic profile of the tomato, which affects the flavor and nutritional content. They were able to recapitulate some of those domestication effects by altering the genome of an otherwise undomesticated variety. While the later data was fascinating, I chose to focus on two of the supplemental figures, both based on the initial metabolomic profiling of the three tomato groups: S. lycopersicum var. cerasiforme group (CER)[1], S. lycopersicum group (BIG)[2], and S. pimpinellifolium (PIM)[3]. As a lover of tomatoes, and a gardener who misses the heirlooms of her summer tomato garden, I was interested in learning more about the evolution of tomato flavor and subgroups, and the metabolic and genetic underpinnings.

Methods:

The first figure shows the coefficient of for each metabolite in each group (BIG, CER, and PIM). Assuming the variants within a group are closely related, the metabolite concentrations should be similar. Figure 1 (Figure S1E in the paper) is a box and whisker plot where each point is the coefficient of variation for the measurements of single metabolite across the variants in a single group. According to the specified statistical methods, the coefficient of variation was calculated using σ/u, where σ and u are the standard deviation and mean of each metabolite in the population, respectively. I calculated this by creating a list of the three groups (BIG, CER, and PIM) and calculating means and stdev for rows where the Group column matched the item in the list. I then plotted the points as a box and whisker plot and set the y limit to match the figure in the paper.

Next I sought to recreate Figure S1B, a PCA plot of tomato accessions, by metabolite. The dataset provided was labeled “the raw dataset of metabolites”. The Metabolomic profiling method indicated that this table had already been normalized by an internal reference control and then further normalized by a log2 transformation, but the scale of the data (maximum values reaching the scale of 10^8) suggested that this data table had not been log2 transformed. In support of this was Table M5a, which includes the average of each significantly changed metabolite between the groups. These averages were within the range of 10-20, and when I took the log2 of TableM1, the results generally fell within that range.

Unfortunately, the paper was unhelpful in the subsequent data processing steps, as the authors used a software for processing the data. I decided to consult the literature and online tutorials in order to build, if not the same figure as the authors, then a statistically robust and equally informative figure. The general analysis pipeline is as follows: (1) normalize the data to an internal control, (2) log2 transform the data for further normalization, (3) address the issue of missing data, (4) perform the PCA analysis and (5) visualize the data. One important note: because the raw metabolite table provided data for two replicate experiments, and the paper indicated that those two sets were averaged to result in the final dataset for use in their

[1] Otherwise known as cherry tomato
[2] Expectedly, the larger tomato varieties
[3] An undomesticated tomato variety, found in Ecuador and Peru. Important for understanding the tomato pre-domestication and for use in later genetic manipulation experiments.

analyses, an averaging step must take place at some point between steps 2 and 4. Initial normalization of the data had already been completed. My analysis begins with step 2.

2. Log2 transformation of the data:

The dataframe was log2-transformed by removing any metadata or string-type columns and then using numpy's log2() function.

3. Missing Data

Missing data was addressed in two different ways: the first was a k-nearest neighbors imputation, using sklearn's KNNImputer class to fill in missing values by the value of their nearest neighbor. As a different approach, missing values were set to 0. The two datasets (replicate 1 and replicate 2) were averaged after either approach. The reasoning behind these two approaches is discussed in the Discussion. Evaluations of patterns of missing data across dataframes were performed using Seaborn's heatmap function.

4. PCA & visualization

PCA was performed on both the knn-imputed dataframe and the zero-filled dataframe. Scikit-learn packages PCA, SparsePCA, and t-SNE were used and the resulting plot was colored by group membership of that tomato sample.

Discussion

Processing the only semi-processed metabolomics data was a good lesson in the difficulties of processing high-dimension datasets. The raw metabolite measurements demonstrated some classic issues of biological data: batch effects (differences between the measurements of samples across replicate experiments), proper normalization (log transformation, scaling) missing data.

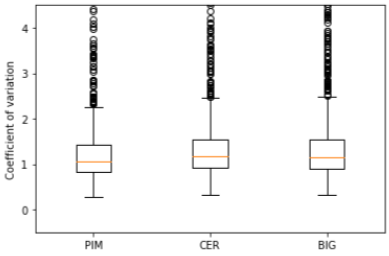

Figure 1: Boxplot of coefficient of variation of metabolites for three populations

	Huang et. al.	My Calculation
PIM	1.13	1.22
CER	1.22	1.36
BIG	1.24	1.42

Table 1. Coefficient of variation of metabolites, by group

Calculations of the coefficient of variation of the metabolites by group yielded results that were close to those of the paper. The same is true for the box and whisker plot (Figure 1). The distribution of coefficients of variation is unsurprising, given the inherent variation in both metabolomics data and in actual metabolomic variation between hundreds of tomato varieties. The authors noted their surprise at the low coefficient of variation in PIM, given the high genetic variability. I believe that there may be two explanations for this. First, the PIM group contains only 62 samples, which is far less than the 248 CER and 574 BIG varieties, meaning there is a decreased possibility of outliers that could skew the data. Second, given that PIM varieties tend to occupy a smaller niche and perhaps convergently adapted a similar metabolic phenotype to survive in the wild, it makes sense that they exhibit less variety. BIG and CER are farmed and certain traits are selected for in an artificial way, which may have created a more diverse dataset.

Log2 normalization of the data corrected for batch effects between the two replicate experiments. Figure 2a and 2b show a random sample metabolite and the correlation between the results pre- and post-log2 transformation, respectively. Figure 3a and 3b show the same for

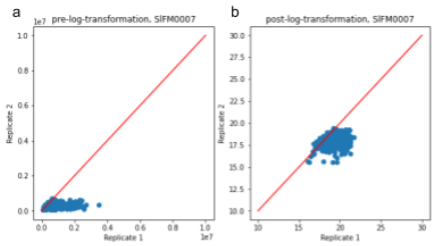

Figure 2

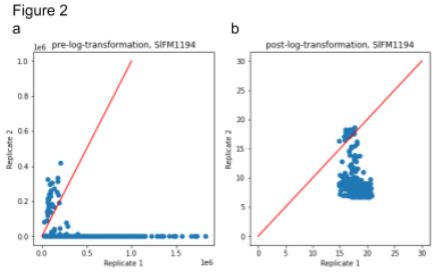

Figure 3

another random metabolite. (Log2 transformation can leave each metabolite on a very different scale from others, so it is more helpful to visualize the transformation for each metabolite separately. The data is not completely linear, but it is improved, and the averaging step will help as well.

The missing data points proved to be the most challenging to resolve, and I believe may be the reason for the difference between my PCA plots and the authors'. The approach to dealing with missing data depends on the reason it is missing. At first, I assumed that it was missing due to random instrumental failure, or random error, and therefore the k-nearest neighbors imputation would have been appropriate.

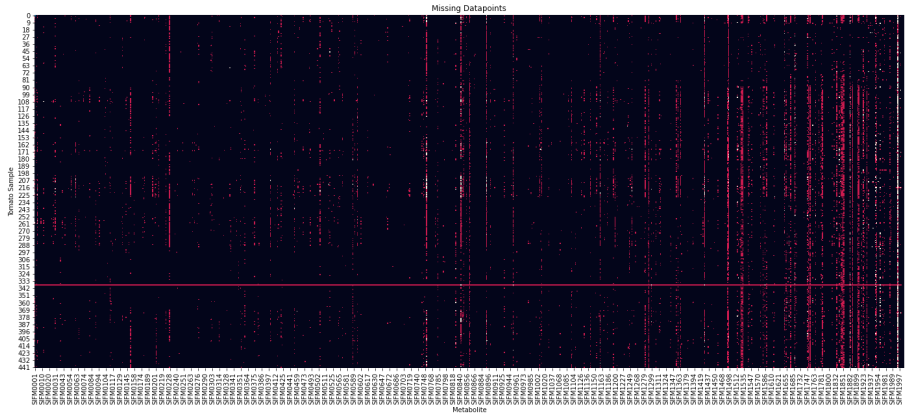

Figure 4. Missing metabolites across two replicate datasets. 2 = missing in both datasets, 1 = missing in one, 0 = missing in neither.

While the k-nearest neighbors imputation should likely have filled in missing values with the values of the most similar samples, which presumably would be from the same group, it is possible that because of the pattern of missing data, this algorithm was not appropriate, or diluted the important effects of missing metabolites. After creating a heatmap of the missing data, it became clear that there were trends across groups and metabolites, and that the pattern of missing data could be considered as part of the metabolomic profile of a sample or group, and should not be filled in with data. Attempting to fill the missing values with 0 made sense biologically, assuming that non-measured metabolites were simply not present (and that absence is important for the metabolic profile). These zeroes, however, seemed to skew the data, and generated a plot that looked similar to that created with the imputed dataframe, but with one outlier (Figure 5b).

Ultimately, no PCA plots I generated showed the clustering and separation demonstrated in the PCA plot. Because the parameters and algorithm used to generate this plot are not supplied, it is possible that this is in fact not a PCA, but a different type of dimensionality reduction algorithm/plot. Because I was able to quite accurately generate the other figure (measuring the

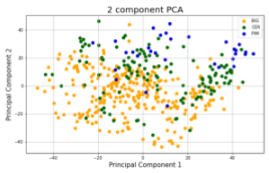

Figure 5a PCA with NaN values imputed by KNN

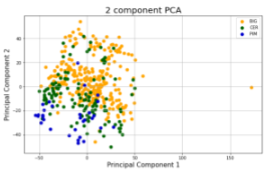

Figure 5b. PCA with NaN values filled with 0

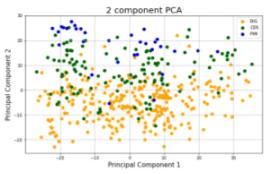

Figure 5c. PCA with scaled imputed dataset

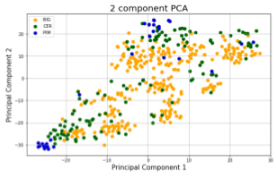

Figure 5d. t-SNE plot

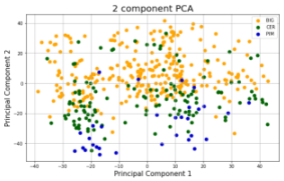

Figure 5e. Sparse PCA plot

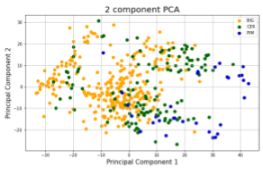

Figure 5f. PCA plot with Nan values imputed, filtered to include only significantly different metabolites (PIM vs. CER)

coefficient of variation across samples in each group), I believe that my initial processing of the data was correct, and the software used by the authors must have clustered, filtered, or normalized the data further before plotting. There are likely more sophisticated algorithms that may produce better results, or those more similar to the paper, but that must be left to future work.

However, I was able to generate plots with some degree of clustering between the three groups (Figures 5a-f). Considering that the authors may have filtered out less significant metabolites and built their plot only using a subset, I filtered my data to include only those metabolites determined significantly different between PIM and CER (Table M5a from paper). This, Figure 5f, did not resemble the PCA from the paper, but did show greater separation between those two groups than the other PCA plots. Ultimately, in all of my PCA plots, CER and BIG data points tended to cluster closer together, and PIM was more separated. This makes sense evolutionarily, as CER and BIG are modern, heavily bred varieties and PIM is more closely related to the older, wild ancestor. Additionally CER and PIM cluster closer together than BIG and PIM, which makes sense given the evolutionary and domestication timeline of tomatoes.

References:

CIMCB Metabolomics Workflow Tutorial, https://cimcb.github.io/MetabWorkflowTutorial/

Kirpich AS, Ibarra M, Moskalenko O, et al. SECIMTools: a suite of metabolomics data analysis tools. *BMC Bioinformatics*. 2018;19(1):151. Published 2018 Apr 20. doi:10.1186/s12859-018-2134-1

" Principal Component Analysis in Python." DataCamp, www.datacamp.com/community/tutorials/principal-component-analysis-in-python.

van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7:142. Published 2006 Jun 8. doi:10.1186/1471-2164-7-142

Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C, Lin T, Qin M, Peng M, Yang C, Cao X, Han X, Wang X, van der Knaap E, Zhang Z, Cui X, Klee H, Fernie AR, Luo J, Huang S. Rewiring of the Fruit Metabolome in Tomato Breeding. Cell. 2018 Jan 11;172(1-2):249-261.e12. doi: 10.1016/j.cell.2017.12.019. PMID: 29328914.

Data provided:

Mmc5: original source of supplementary data, including table M1, the raw metabolite matrix
Mmc1: Description of each tomato accession, including group
tomato_metabolites.xlsx : table M1 copied into a new spreadsheet for ease of analysis
tomato_varieties.xlsx : mmc1 table copied and headers removed
significant_metabolites.csv : list extracted from mmc5 table M5a of all metabolites where there was a significant difference between two groups