**S3: Description of Data Processing in R**

By Brandon Jordon-Thaden and Ingrid Jordon-Thaden

**R Files:**

Link.R

Cluster.R

Call.R

Calc-parents.R

Calc-inheritance.R

Analyze-drift.R

**Link.R**

This script takes the data from the Typer4 software, links them to the identifiers for each sample, and creates a data matrix that is used in the next steps.

**Cluster.R**

This script does the Gaussian clustering. This is done in four main steps: (1) Removes spectra of poor quality both from the Typer4 designation “Bad Spectra” and our own call of too much uncertainty in spectral quality based on the ratio of the area of peak 1 to area of peak 2 (A1/A2); (2) Converts the area of the spectral area peaks from Cartesian coordinates to polar coordinates to reduce the experiment to the two most important parameters, namely the ratio of the A1/A2 and the sum of the amplitude of A1/A2 (see Figure S3-1); (3) Carries out the Gaussian clustering of the samples, finding the center of mass of the cluster, calculating the standard deviation of the cluster, and numbering each cluster (see Figure S3-2); (4) Removes the clusters that did not give a well-localized cluster or had a poor deviation; (5) Removes the clusters that did not represent enough individuals for the populations and synthetic lines. This script is responsible for the greatest number of spectra removed from the dataset.

**Call.R**

This script calls (or names) each sample data point by the cluster center; all samples in the cluster are given the same generic call depending on their high or low mass (diploids: HH or LL, and tetraploids HHLL). This allowed us to assign designations to both diploid and tetraploid individuals, circumventing the limitation of Typer4 not being able to handle polyploids. It also incorporated the actual SNP call (A,T,G, or C) from Typer4. This resulted in a new column in the data matrix labeled “new call” resulting in e.g. AATT or AT.

**Calc-parents.R**

This script calculates the parent call percentages of samples across each contig per DNA type (cDNA and gDNA separately). It ultimately identifies if parental values for a contig can be distinguished and are useful. A contig was considered useful if there was an identifiable cluster of low mass and high mass for each parent. If both were low mass or both were high mass, then they would not be useful for interpreting the tetraploid pattern in the next step. This step removes contigs that did not meet this criterion.

**Calc-inheritance.R**

After all of the cleaning of the data, we are now ready to confirm the contigs that not only give distinct mass clusters for the two parents, but also give clusters for tetraploid individuals, resulting in an identifiable parent-tetraploid ‘triangle.’ This script uses the high and low mass we identified for the parents for a contig and DNA type to call the tetraploid correctly. The choices are based on the percent ‘dubius-like’ calls because *T. dubius* is the common parent in both tetraploid triangles of relationship. Inheritance patterns are assigned based on the cluster position in the experiment triangle at 50/50, 25/75, or 75/25% of the dubius homeologs (see Figure S3-2). This step is where the final experimental triangles were reduced to 493 contigs for the *T. mirus* dataset, and 462 contigs for the *T. miscellus* dataset. After this step was complete, we were able to read from the data matrix which samples had loss of one or two parental homeolog copies. At the end of this script is an extensive graphing script that resulted in the color-coded four-panel graphs for each experimental set and the samples (parent 2-parent2-tetraploid-parent mix) for each contig for the gDNA and cDNA (see Figure S3-2 and S4). These graphs were exported into a large pdf and are found in S4\_Clustering plots all.pdf.

The **parent mixes** (manually mixed extracted DNA and RNA of the parents and ran through the experiment) were plotted at this step to make a visual review of the contigs in the graph format. If the parent mix samples did not appear as a center cluster (i.e. an artificial tetraploid) for that contig, then it was concluded that the contig had failed. At this point, this check was almost redundant but helped us confirm that our previous screening methods were successful. We created parent mixes of *T. dubius* and *T. porrifolius* for two different populations, and of *T. dubius* and *T. pratensis* for four different populations. They were our control samples.

**Analysis-drift.R**

This final script ultimately combines the experiment triangles that provide data for a sample across both the above data for the cDNA and gDNA for the same contig. It looked at each sample per triangle, per contig and records how the cDNA varies (drifts) from the gDNA call, i.e., if the homeologs are both expressed, if one homeolog or the other silenced, or if neither homeolog is expressed. The latter case could refer to complete silencing of both homeologs or could indicate that the contig in question is not expressed in leaf tissue in the tetraploids, even if it is expressed in leaf tissue in one or both parents. Complete silencing was marked with blue triangles in the heat maps. In order to determine successful contigs for this analysis the contig had to work not only in both the cDNA and gDNA for that sample, but also for the complete experiment triangle (i.e., triangle: *T. mirus*, *T. porrifolius*, and *T. dubius*, or triangle: *T. miscellus*, *T. pratensis*, and *T. dubius*). Those contigs that did not fit these criteria were eliminated from the drift analysis. Filtering to remove these contigs in which gDNA and cDNA could not be compared across the experimental triangles reduced the number of successful contigs to 250 for the *T. mirus*, *T. porrifolius*, and *T. dubius* species set and to 233 contigs for the *T. miscellus*, *T. pratensis*, and *T. dubius* set. Which is about half of the data determined in the inheritance step above (493 and 462 respectively).

**Final Remarks:**

At each step along the way of applying the above scripts, spreadsheets were exported, and data were observed and summarized in Excel (Microsoft Excel, 2010) with bar graphs of parental contig percentages across samples and contigs. Over a period of three years, these data were analyzed until we felt our human bias had been completely removed from the homeolog calling process. The final called genotypic/expression and drift patterns can be found in S6, and the raw data with the identifiers can be found in S7.

Ultimately, we were able to ensure we could confidently identify true homeolog loss and silencing based on the following rigorous statistical steps taken with the raw data. First: we set a tight threshold on the measurement and experimental uncertainty. Second: we only included well localized and statistically valid Gaussian clusters. Third: we only included on a per-contig and per-data type (gDNA or cDNA) basis those samples that had a complete experimental triangle. This allowed us to feel confident that the final grey boxes in the heat maps are true failed samples, and not poor methodology on the homeolog base calling. Ultimately allowing us to be confident that the total silencing cases, marked by blue boxes, are biologically true, and not failed samples. We successfully found contigs with SNPs that were found across all players in the experimental triangle, across many populations, and across many synthetic generations and lines in order to make the conclusions we have within this project, all without using human bias of SNP calling and graph peak interpretation by eye.

**Figure S3-1:** Conversion of data from Cartesian to Polar Coordinates in order to reduce the experiment to the two most important parameters, namely the ratio of A1/A2 and the sum of the amplitude of A1/A2.

A picture containing diagram

Description automatically generated

**Figure S3-2:** Model-based clustering optimizes the number of clusters and fits clusters to Gaussians in 2D space. The mathematically determined center of the clusters is marked by a “+” just for visualization in this figure. We illustrate how we used the position of the center of the clusters in conjunction with results of the parents from the same contig to assign new allele calls to each sample. The ovals are drawn on this image for visualization purposes of this figure. If the cluster had a poor localized center and/or a poor standard deviation, it was eliminated from the dataset.

Left panel: Data for the gDNA samples are shown for contig 2818-68, both natural and synthetic individual tetraploids and the parents. The *T. miscellus* samples created 5 clusters, indicated by the circles around the clusters. Genetic theory states that the majority of the tetraploid samples will equally combine the parental contributions, as seen by the green cluster in the center and therefore would be called LLHH (50/50%). Note that in this example, the green cluster is slightly shifted to the left (sitting at ~40 degrees). It is does not make biological sense to have a 40% inheritance being the main cluster. This makes it more difficult to automatically call it 50/50 without first identifying the centroid of the cluster. We concluded this is an experimental flaw of the Sequenom method and software and led us to develop these steps to accurately interpret the data. This is ultimately the reason we needed to analyze the data in 2D Gaussian space. The other clusters can easily be identified as alternative patterns and set within a degree range compared to the “50/50” inheritance cluster. The identity of the parental allele that represents each cluster is extrapolated from the corresponding spectra from the same contig for each of the parents. The behavior of this contig in parent *T. pratensis* samples is found at 90 degrees and called HH (blue cluster in the *T. pratensis* graph), thus identifying the two tetraploid samples in *T. miscellus* (small pink cluster) as containing 100% of the pratensis homeolog for this allele and called HHHH. Note that an allotetraploid should combine the parental copies; the fact that it has only this single parental genotype indicates loss of both copies of the *T. dubius* homeolog, for an actual genotype of HH, however, for clarity and automation ease in the data analysis we kept the genotypes for the tetraploids as HHHH or LLLL. Conversely, the parent *T. dubius* samples are found at 0 degrees and called LL (yellow cluster), thus identifying the two tetraploid samples in *T. miscellus* (small blue cluster) as containing 100% of the *T. dubius* homeolog and called LLLL. Here, because an allotetraploid should combine the parental copies, the fact that it has only this single parental genotype indicates loss of both copies of the *T. pratensis* homeolog, for an actual genotype of LL. This leaves the two clusters remaining, burgundy and red, which would represent allele dosage of 25% of one parent and 75% of the other and would be LLLH for the burgundy cluster and LHHH for the red cluster. Note that these dosages are idealized to the presence of all four alleles in an allotetraploid; in reality, true loss of both copies of one parental homeolog would yield two copies (e.g., resulting in LL or HH rather than LLLL or HHHH), and true loss of one copy of one parental homeolog would result in one copy of parent 1 (33%) and two copies of parent 2 (67%), with genotypes LLH or LHH, despite the idealized designations in the figure and graphs in S4.

Right panel: Here we show the same samples for cDNA for the same contig. The expression of the homeologs for each sample was called in the exact same manner as the gDNA. However, with the cDNA, we are able to distinguish between homeolog loss (not present in either cDNA or gDNA) or silencing (present in gDNA but not in cDNA).

Graphical user interface

Description automatically generated