SEQUENCING AND ASSEMBLY OF THE PRUNUS DOMESTICA CV. IMPROVED FRENCH GENOME.

Chris Dardick, USDA-ARS, Appalachian Fruit Research Station, Kearneysville, WV 25442

OBJECTIVES

The objectives of this project are to enhance genome assembly of the Improved French genome in order to develop molecular markers useful for germplasm characterization and breeding.

PROGRESS REPORT

Over the past two years, the USDA-AFRS Genetic Improvement Research Unit has been developing genome sequences for *Prunus domestica* to facilitate molecular characterization of prune genomes and enable marker assisted breeding. To date, extensive sequence coverage has been obtained for two prune cultivars; HoneySweet (a transgenic 'eastern' type prune) and 'Improved French'. Initial genome assemblies have been performed using the closely related peach genome as a template. These assemblies revealed that 87% of the plum genome is shared with peach. Over the shared regions, prune shows approximately 83.2% nucleotide identity to peach. By analyzing mismatched and broken paired-reads we have found that there are numerous local chromosomal rearrangements relative to peach resulting from large numbers of small insertions, deletions, and inversions. These differences suggest that many of the molecular markers developed for peach may not be transferrable to prune.

Using new assembly algorithms, we have made significant progress in assembling the 'Improved French' genome. The largest gains have come from the ability to "scaffold" *de novo* assembled contigs. Scaffolding has enabled assembly of approximately 50% of the prune genome into contigs >750 base pairs (vs. the prior 30%) and has had a significant improvement on average contig length.

We performed additional sequencing of Improved French using a strategy called Mate-Pair sequencing. Mate pair data provides information that can link together smaller contigs to former longer, continuous genomic sequences. Addition of the mate pair data has allowed us to improve the contig length from an average of 1,500bp to over 3,000bp. Maximum contig size was increased from 60,000bp to nearly 200,000bp. These data have significantly approved the plum genome assembly and significantly improves our ability to design molecular markers, especially in regions that are not syntenic with the peach genome.

We analyzed Single Nucleotide Polymorphisms (SNPs) across the entire prune genome. The relative level of sequence polymorphism in prune was found to be high. Prune was found to contain approximately 10 times the number of SNPs as peach (3.3 million vs 0.3 million). Increased polymorphism is expected given the hexaploid nature of prune. However, the high level of polymorphism suggests that, unlike peach, prune cultivars have not been subject to a high degree of inbreeding.

A set of 30 prune-specific molecular markers were designed based on the prune genome. PCR based markers were designed to flank 3-10 base pair indels (insertion-deletion polymorphisms) at genomic intervals of 5 million base pairs for all 8 chromosomes (Figure 1).

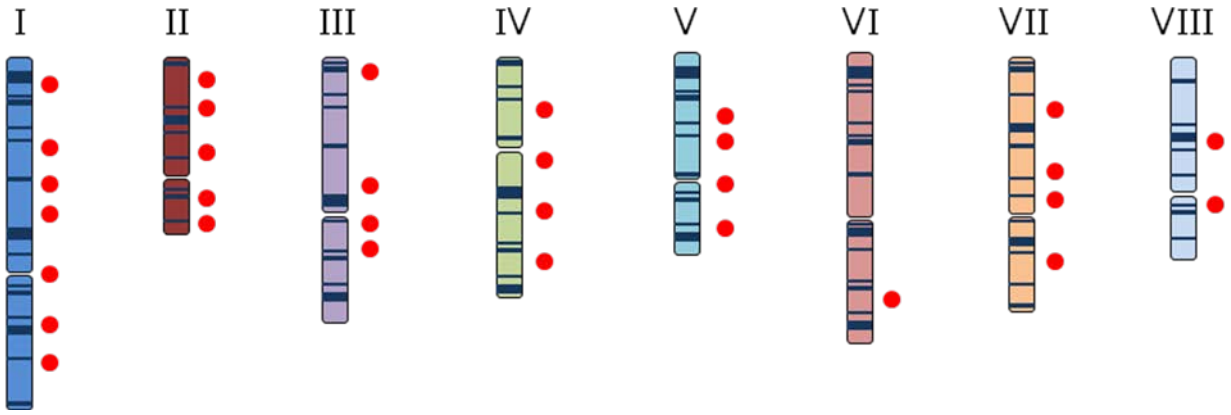

Figure 1. Distribution of molecular markers (red dots) across prune chromosomes.

The markers were chosen based on their ability to distinguish the 'HoneySweet' and 'Improved French' genomes. PCR fragments were distinguished using the High Resolution Melting technique. Resulting melt curves were categorized based on their shape. Marker data obtained from nearly 70 cultivars were analyzed to assess the ability of the markers to distinguish known prune pedigrees.

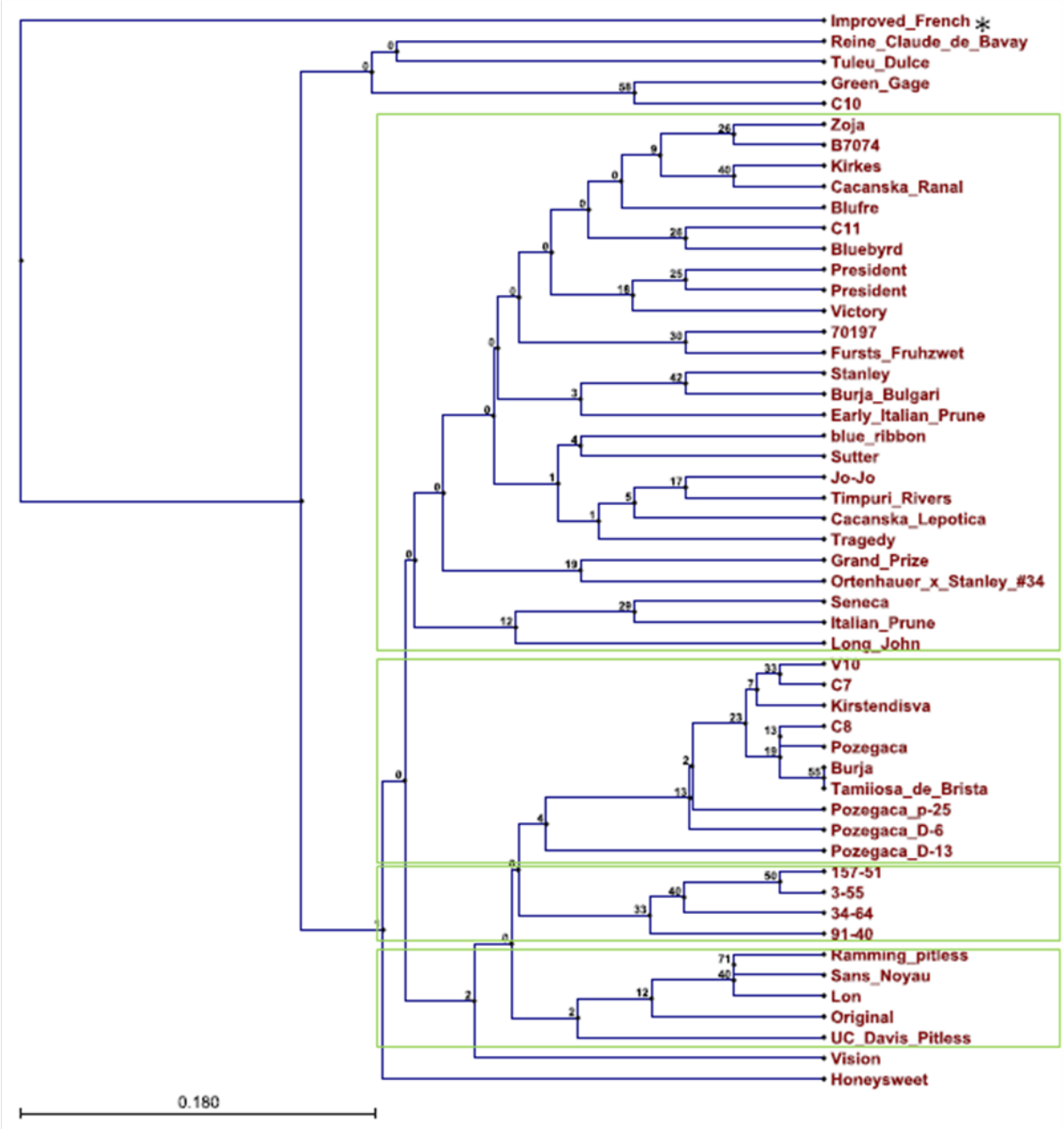

Figure 2. Phylogenetic tree derived from molecular marker data. Cultivar names are indicated to the right.

Results indicated that the markers successfully distinguished cultivars of known origins. For example, Posegaca types and related germplasm formed a distinct cluster. Stoneless types also produced a distinct clade. Surprisingly, 'Improved French' was found to be unrelated to the other cultivars examined. To confirm this finding, the experiment was repeated using a new set of cultivars that included all available 'French' types (Figure 3.)

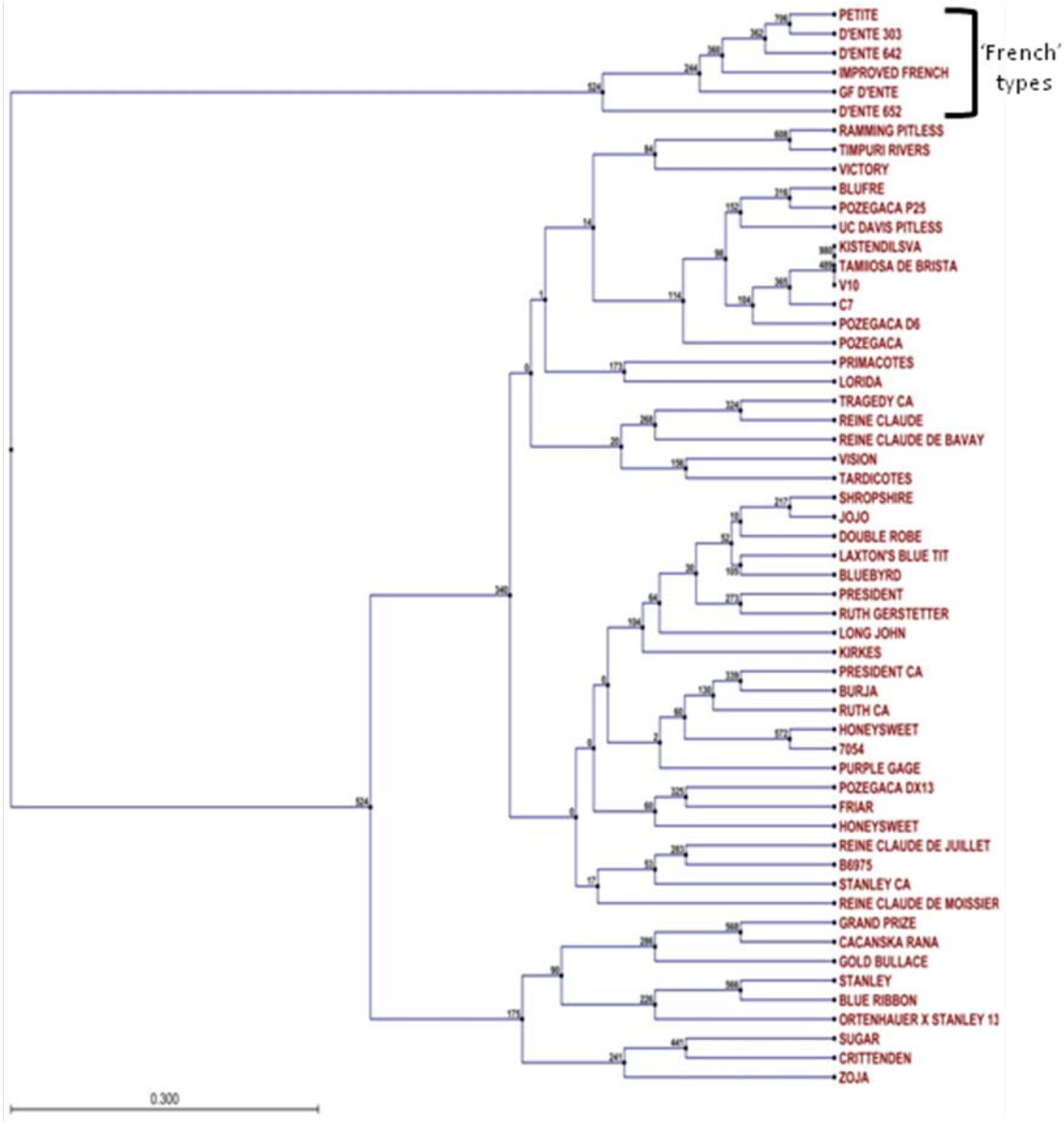

Figure 3. Phylogenetic tree of molecular marker data that includes six known 'French' cultivars (indicated).

The results confirmed our previous finding and suggest that 'French' germplasm is genetically distinct from other European and Eastern US cultivars.

We have also begun using the 'Improved French' genome data to determine the mode of inheritance within this hexaploid species. Prune is thought to have originated from a cross between an unknown tetraploid and a diploid *prunus* species resulting in a hexaploid. How the six chromosomal copies segregate is currently unknown- although genetic segregation data suggests it behaves as a diploid (4x2 or 3x3). We have identified genomic sites where all 6 chromosome copies contain unique sequences; allowing us to track the segregation of all six copies in the progeny. A molecular marker was tested that spanned a short repeat region within a

gene we previously identified as being associated with control of branch angle (TAC1). This SSR (simple sequence repeat) marker consisted of unique PCR primers that flanked the repeat region resulting in amplification of 6 different sized DNA products (92, 115, 118, 121, 124, and 128 base pairs). The resulting DNA fragments were separated using capillary gel electrophoresis and quantified. The expected sized fragments based on whole genome sequencing were largely consistent with the SSR marker, although one of the expected fragments could not be accurately scored (124 base pairs).

To determine the mechanism of segregation in *Prunus domestica*, this marker was evaluated on a set of 28 'Improved French' seedlings derived from a tree that had been self-pollinated along with 44 P. domestica cultivars and 7 other *Prunus* species. The results showed that two of the 5 alleles tracked (121 and 128bp) were present in approximately 75% of the Improved French seedlings; as would be expected for diploid segregation. However, the other 3 alleles were found at significantly higher frequencies in the offspring (>92%), and were consistent with a tetraploid mode of segregation. Likewise, the allele frequencies among all *P. domestica* cultivars tested also supported these data as the 121 and 128 fragments were present in 45% and 36% of all cultivars respectively, while the 92, 115, and 118bp fragments were found in 91%, 82%, and 75% of all cultivars, respectively. Collectively these data suggest that the *P. domestica* hexaploid genome consists of two sub-genomes; a diploid genome and a separate tetraploid genome that behaves independently. Among the 5 plum alleles, only two of them were found in other prunus species. The 92bp fragment was present in *Prunus cerasifera* while the 115 bp fragment was found in *Prunus avium* (cherry). None of the 5 plum alleles were found in *P. persica* (peach), Mariana 2624 rootstock (a P. cerasifera X P. munsonia hybrid), *P. simonii*, or *P. salacina* (Japanese plum). Prior reports suggested that *P. domestica* was the result of an interspecific hybrid between *P. cerasifera* and *P. spinosa*. Our data support the hypothesis that *P. cerasifera* may have been one of the progenitors.

ONGOING STUDIES

The molecular markers that we developed and validated are currently being analyzed within our FasTrack breeding program to confirm crosses between 'HoneySweet' and 'Improved French'. These proto-typical studies will serve as the basis for future marker assisted breeding efforts that target additional prune traits of interest to the US prune industry.

Prunus domestica genome segregation is still being studied. We are currently testing our existing polymorphic SSR marker on additional *Prunus* germplasm including a larger 'Improved French' seedling population as well as additional *Prunus* species, notably *P. spinosa* which is thought to be one of the progenitors to P. domestica. In addition, we are designing additional polymorphic markers to confirm the observed genetic segregation. Collectively, these data will establish the genetic behavior of *P. domestica* and help determine the origins of the diploid and tetraploid genomes that comprise *P. domestica*.