Gene prediction and genome functional annotation of ‘Nonpareil’

Project No.: **HORT35.Fresnedo**

Project Leader: Jonathan Fresnedo Ramirez
Mailing Address: The Ohio State University
Dept. of Horticulture and Crop Science
1680 Madison Ave. Wooster, OH 44691
Phone: 330-263-3822 Fax: 330-263-3887
E-mail: fresnedoramirez.1@osu.edu

Project Cooperators and Personnel:
Wilbur Z. Ouma & ***Tea Meulia***, *Molecular and Cellular Imaging Center – Ohio State University*
Thomas M. Gradziel, *Department of Plant Science – UC Davis*

A. Summary

Genomics is integral to many of the current solutions to problems faced by humanity and is considered when developing new solutions to our current and coming challenges. However, regardless of the power of genomics, the availability of genomic resources in almond have been lacking. In this project supported by the Almond Board of California, we have developed the genome of ‘Nonpareil’, the most important cultivar in California. This genome sequence has been annotated to represent the predicted function of the ‘Nonpareil’ genes encoding for both the tree and kernel. The ‘Nonpareil’ genome sequence annotated with its genes and the proteins these genes encode represents a relavant resource for researchers in the almond community. This key resource will enable more thorough development of solutions for almond production such as enhancing almond breeding to develop novel almond cultivars. This resource will be available and open to any person interested in examining and utilizing the almond genome. The genome is expected to be realsed within the second quarter of 2020 through the Genome Rosaceae Database.

B. Objectives

Our vision is that this tool will contribute to formulating solutions for almond production issues and ensure the sustainability of almond production and the resiliency of the California almond value supply chain. Previously, we performed gene prediction and annotation using the gene space assembly of ‘Nonpareil’ only. As a continuation of this project, we proposed refining the genome assembly using optical mapping to obtain a more accurate representation of the ‘Nonpareil’ genome and enhace the functional annotation.

The objectives for 2019 were to:

1. Use the genomic data from the genome assembly of ‘Nonpareil’ to develop the gene prediction for the genome.
2. Develop a complete transcriptome with full-length gene transcripts found in the ‘Nonpareil’ genome using MinION Oxford Nanopore technology and Illumina RNA sequencing to validate transcripts.

3. Use the gene prediction and transcriptomic data coupled with advanced computing methods to develop the gene annotation of the ‘Nonpareil’ genome and produce a list of genes and the encoded proteins.

C. Annual Results and Discussion

To provide genomic resources for almond, an interdisciplinary team including researchers from The Ohio State University (OSU) and UC Davis initiated an effort to develop the first completely assembled and annotated genome sequence for the ‘Nonpareil’ almond cultivar.

The collaborative effort has resulted in a high-quality draft assembly for almond using a combination of Illumina technology and high-throughput chromosome conformation capture (Hi-C). As a result, a high-quality, high-continuity draft assembly of the gene space of ‘Nonpareil’ was produced, enabling generation of a reliable gene prediction and functional annotation.

Objective 1: Use the genomic data from the genome assembly of ‘Nonpareil’ to develop the gene prediction for the genome.

A pipeline that employs a combination of *ab initio* (intrinsic) and evidence-based (extrinsic) gene-finding approaches was successfully implemented at the Ohio Supercomputer Center. First, repeats and low complexity regions were identified and subsequently masked in the ‘Nonpareil’ assembly. As a result, 27% of the genome was soft-masked for repeat elements, a majority of which was comprised of interspersed repeats (25%), suggestive of transposition or retrotransposition events.

Thus, the gene prediction pipeline output included 27,487 protein-coding genes in the ‘Nonpareil’ genome, which is in contrast to the 27,042 reported for ‘Mission’ (a.k.a. ‘Texas’ in Spain). This means our calculations and results thus far are within the expected range and we were able to advance to validation using transcriptomic data. The table below shows additional statistics specific to the ‘Nonpareil’ genome according to gene prediction:

Genes	**Quantification**
Number of genes	27,487
Average gene length (bp)	2,652
Total length of gene models (Mb)	72.89
Shortest gene (bp)	200
Longest gene (bp)	37,254
Exons	
Number of exons	159,889
Average number of exons per gene	5.82
Average exon length (bp)	215
Introns	
Number of introns	130,471
Average number of introns per gene	4.75
Average intron length (bp)	359

Table 1: Quantitfication and stats of genomic features in the ‘Nonpareil’ genome

Objective 2: Develop a complete transcriptome with full-length gene transcripts found in the 'Nonpareil' genome using MinION Oxford Nanopore technology and Illumina RNA sequencing to validate transcripts.

We generated 7.62 Gb of short Illumina reads, and 2.61 Gb of long-read data with an average length of 1.1 Kb using Oxford Nanopore. The gene prediction and annotation pipelines, implemented on a high-performance computing platform at the Ohio Supercomputer Center, yielded 28,637 gene models covering 48.2% of the genome assembly.

For comparison, about 27,000 gene models (~40% of the genome) were reported in the almond 'Texas' (a.k.a. Mission) and 26,873 gene models were reported in peach (~ 38% of the genome; see Table 2 for additional information). These numbers suggest that the 'Nonpareil' gene content is within the expected gene content in the *Prunus* genus. The predicted gene models were supported by expression in transcript data generated from 'Nonpareil' tissues, protein sequences for related cultivars, and/or presence of protein domains identified by a protein family database search.

Cultivar	Number of genes	Average gene length (Kb)	Length of gene models (Mb)	Percentage of genome covered by genes
'Nonpareil'	28,637	2.82	79.34	48.2
'Texas/Mission'	27,042	3.33	90.11	39.6
Peach ('Lovell')	26,873	3.24	87.10	38.3

Table 2: Comparison of the number of gene models, average gene length, total length of genome space, and percentage of the genome covered by genes in two almond cultivars and peach

Objective 3: Use the gene prediction and transcriptomic data coupled with advanced computing methods to develop the gene annotation of the 'Nonpareil' genome and produce a list of genes and the encoded proteins.

Predicted and validated protein-coding genes were subsequently functionally annotated by a combination of sequence motif-based (Pfam domain search), homology-based (NCBI and Uniprot database search), and orthology-based (KEGG database search) methods. The functional annotation of the protein sequences associated with the predicted models found ~68% were associated with at least one biological function while the rest are of unknown function.

The most abundant protein domains identified in the predicted models were associated with key biological functions such as enzyme activity (kinases, transferases, deaminases), the plant specific pentatricopeptide repeat (PPR) family of proteins, and nucleic acid-binding proteins (Figure 1).

A quantitative assessment of annotation completeness in terms of the expected gene content was performed using Benchmarking Universal Single-Copy Orthologs (BUSCO) derived from OrthoDB. The 'Nonpareil' gene prediction exhibited ~90% completeness (blue color portion in

Figure 2), slightly less than completeness observed in peach, the model plant *Arabidopsis thaliana*, and the almond ‘Mission’/’Texas’ cultivar.

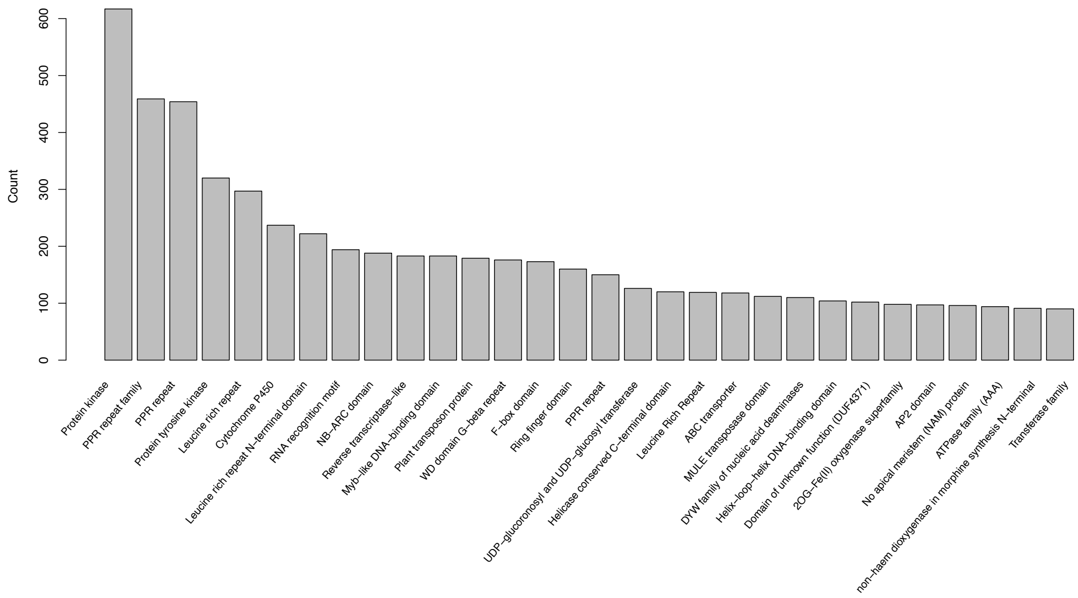

Figure 1: Frequency of the most abundant protein domains in ’Nonpareil’

The ‘Nonpareil’ completeness score suggests opportunity for refining the annotation of the genome by providing additional evidence such as RNA-Seq data derived from different almond plant tissues (e.g. flower and meristematic tissues). Increasing the number of tissues represented will greatly improve the quality of the annotation. Unfortunatley in the case of ‘Nonpareil’, it is almost impossible to obtain samples of actual root tissues given that the original seedling of ‘Nonpareil’ no longer exists.

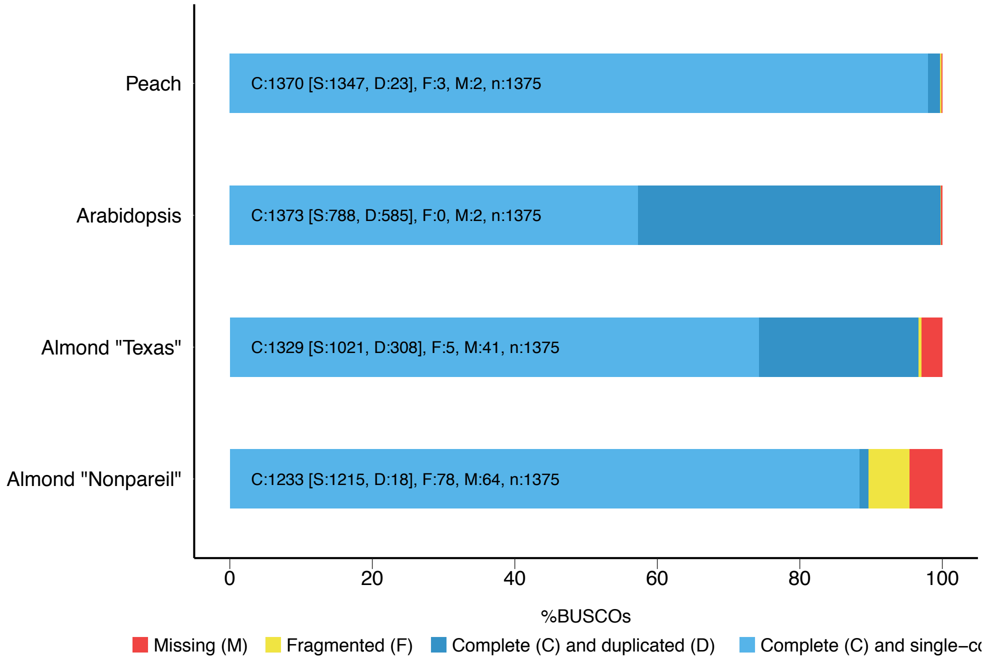

Figure 2: Benchmarking Universal Single-Copy Orthologs (BUSCO) of ‘Nonpareil’

Finally, a high proportion of ‘Nonpareil’ gene content supports the ongoing effort of improving the genome assembly using optical mapping technology to uncover additional gene regions.

D. Outreach Activities

The preliminary results of this project were presented at the Great Lakes Bioinfromatics Conference on May 19-22, 2019 at the University of Wisconsin at Madison. This conference included ~600 researchers and students in the area of bioinformatics. Results of this project were also presented at the 2019 Almond Conference in Sacramento, CA on December 10-12, 2019. The audience was composed of students, researchers, growers and stake holders of the almond community.

E. Materials and Methods:

Previously, a preliminary genome assembly of 164.55 Mb with a 13.96 Mb N90 contiguity in eight scaffolds was produced using a combination of Illumina technology and high-throughput chromosome conformation capture (Hi-C). This was the basic data used for gene prediction. Transcriptomic data was produced from both leaf and whole fruit tissues from the same ‘Nonpareil’ individual sampled for the genome sequencing.

A gene prediction and functional annotation pipeline that employs a combination of *ab initio* and evidence-based gene finding approaches was successfully implemented at the Ohio Supercomputer Center using both short (Illumina) and long read (Oxford Nanopore MinION) data (Figure 3). Repeats and low complexity regions were identified and subsequently masked in the ‘Nonpareil’ assembly. Illumina reads were assembled with Trinity software prior to training the SNAP gene finder classifier. An iterative training process resulted in the generation of a SNAP hidden markov model (HMM), which was subsequently employed—together with an AUGUSTUS gene finder—in predicting an initial list of gene models.

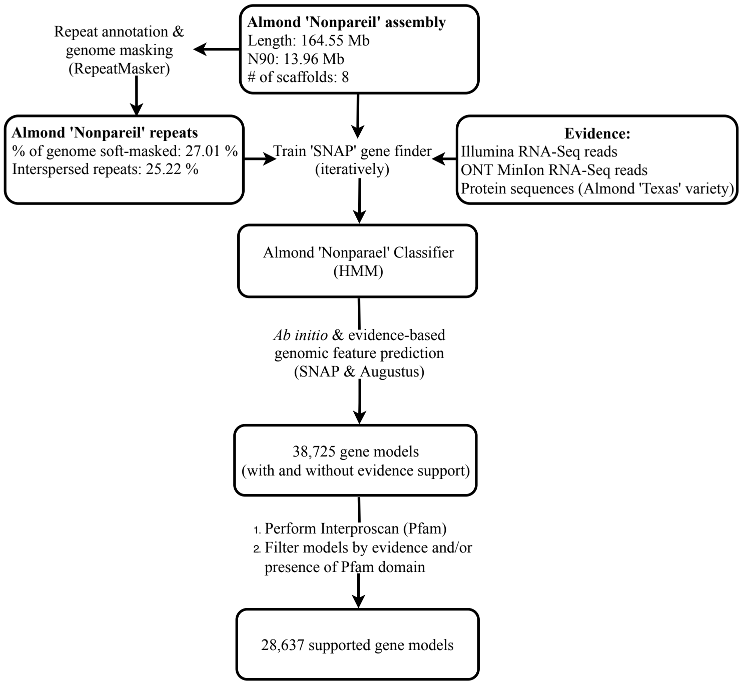
Almond 'Nonpareil' assembly
Length: 164.55 Mb
N90: 13.96 Mb
# of scaffolds: 8
Repeat annotation &
genome masking
(RepeatMasker)
Almond 'Nonpareil' repeats
% of genome soft-masked: 27.01 %
Interspersed repeats: 25.22 %
Train 'SNAP' gene finder
(iteratively)
Evidence:
Illumina RNA-Seq reads
ONT MinIon RNA-Seq reads
Protein sequences (Almond 'Texas' variety)
Almond 'Nonparael' Classifier
(HMM)
Ab initio & evidence-based
genomic feature prediction
(SNAP & Augustus)
38,725 gene models
(with and without evidence support)
1. Perform Interproscan (Pfam)
2. Filter models by evidence and/or
presence of Pfam domain
28,637 supported gene models

Figure 3: Pipeline for gene prediction and annotation of the ‘Nonpareil’ genome

For functional annotation, predicted gene models were submitted to a pipeline utilizing database searches of protein sequences in Uniprot (homology search), KEGG database (orthology search), and Pfam (protein domain search; Figure 4). The resulting descriptions of putative gene functions, Pfam domain identifiers (IDs), and gene ontology (GO) terms were included in the genome annotation feature file (GFF).

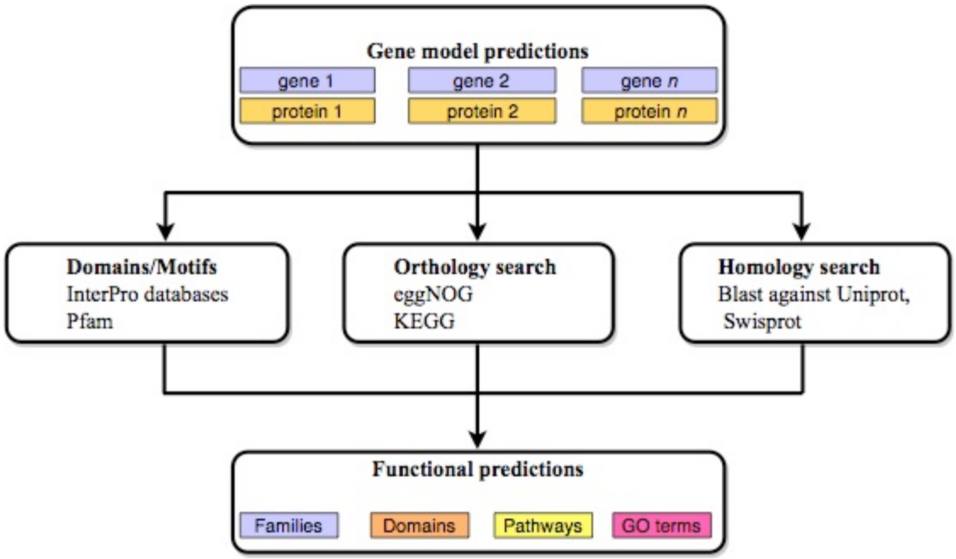

Figure 4: Process for functional annotation of ‘Nonpareil’ genes

F. Publications that emerged from this work

1. List peer review publications in preparation, accepted or published
   - *DNA-(de)methylation status is associated with the divergent exhibition of non-infectious bud failure, an age-related disorder in twin almonds (Prunus dulcis)* D’Amico-Willman, et. al (in preparation).

This manuscript is the result of a collaborative project with UC Davis, Michigan State, and Ohio State utilizing the sequenced and annotated ‘Nonpareil’ genome. The manuscript is currently under preparation for submission to The Plant Journal or Plant Science following completion in March 2020.

The public release of the ‘Nonpareil’ genome and its annotation is expected within the second quarter of 2020 through the Genome Rosaceae Database and in the National Center for Biotechnology Information websites and repositories. The manuscript mentioned above will work as reference when the use of the genome will be cited.