FINAL PROJECT REPORT **YEAR: 3 (of 2+1yr NCE)**
WTFRC Project Number: PR-17-104

Project Title: Functional genomics of 'd'Anjou' pear fruit quality and maturity

PI: Loren Honaas
Organization: USDA-ARS
Telephone: 509.664.2280 x211
Email: loren.honaas@ars.usda.gov
Address: 1104 North Western Ave
City/State/Zip: Wenatchee, WA 98801

Cooperators: Stefano Musacchi & Sara Serra (WSU-TFREC), David Rudell & Jim Mattheis (USDA-ARS), Claude dePamphilis (PennState)

Total Project Request: Year 1: $52,707 **Year 2:** $33,488 Year 3: NA

Other funding sources: USDA-ARS technician salary and benefits - $31,734

Budget 1
Organization Name: USDA, ARS **Contract Administrator: Chuck Myers**
Telephone: 510-559-5769 **Email address: chuck.myers@ars.usda.gov**

Item	2017	2018
Wages[1]	$12,500	$12,500
Equipment[2]	$1,980	NA
Supplies	$8,407	$5,483
Miscellaneous[3]	$29,820	$15,505
Total	$52,707	$33,488

Footnotes:
[1] Cooperative Agreement to Penn State for data processing and data analysis
[2] Service contract for CLC genomics workbench support
[3] Illumina sequencing & library prep at Penn State Genomics Core via Cooperative Agreement

OBJECTIVES

Long term objective: Develop a detailed understanding of the genetics of pear fruit maturation, ripening and quality towards enhanced diagnostics, therapeutics, and production practices, focusing on postharvest technology.

Specific objectives:
1) **Identify gene activity** correlated with fruit quality and maturity as it relates to on-tree fruit position
2) **Discover genes** in 'd'Anjou' pear for comparative genomics with 'Bartlettt' pear
3) **Generate a list of potential biomarkers** for use in research and fruit production

SIGNIFICANT FINDINGS

Progress on specific objectives:

- **Objective 1) Complete –** A full analysis of gene activity that relates to on tree fruit position is complete, including analyses leveraging the new 'd'Anjou' genome
- **Objective 2) Exceeded –** Instead of building fruit specific models for genes using 'd'Anjou' pear fruit gene activity data, we sequenced the entire genome of 'd'Anjou' pear capturing virtually all of the genes in this cultivar
- **Objective 3) Complete –** We have a candidate list of genes for additional maturity work in European pear, and the 'd'Anjou' genome provides the opportunity to discover additional genes and gene forms that are related to cultivar differences between 'Bartlett' and 'd'Anjou'

Other important findings:

- Synergy with "Enhancing reference genomes for cross-cultivar functional genomics" provides a foundation to fully explore genetic differences between 'Bartlett' and 'd'Anjou' pear with fully annotated *de novo* 'd'Anjou' pear genome
- Massive gene activity changes occur during storage
- Relatively few gene activity signatures distinguish fruit from internal canopy positions vs. external canopy positions
- Genes do not act alone – co-expression and differential expression implicate groups of genes

RESULTS & DISCUSSION

Recap of project proposal justification

In European pear, tens of thousands of genes are active when fruit are harvested. The activity of these genes changes as fruit matures, in response to postharvest conditions, and is different in various fruit tissues. By examining these gene activity signatures we gain insight into poorly understood biological processes that influence pear fruit ripening and quality. Additionally, gene activity that is correlated with fruit maturation and ripening may provide a tool to finely monitor fruit. In the context of research, this could facilitate experiments that target manipulation of fruit maturity and ripening by monitoring gene signals in responses to various postharvest conditions. Also, because changes in gene activity often precede otherwise undetectable physiological changes in plant tissues these changes can be potentially used to predict future fruit quality, thereby providing tools to enhance fruit quality management for the industry. The approach for this project included functional genomics which aims to use very large gene activity data sets (100s of millions of measurements) to learn about complex

biological processes in plants. The specific method, called **RNA-Seq**, is the method of choice because it can be used to monitor the activity of *all* genes simultaneously.

A critical requirement for RNA-Seq is the availability of a reference genome. The reference genome contains the genes that are needed to interpret the massive gene activity data sets generated by RNA-Seq. When this work was proposed, the only available reference genome for European pear was for 'Bartlett.' In the past 3 years, genomes for various *Pyrus* species/cultivars (including a dwarfing rootstock and 'Bartlett' v2.0) have been published (https://www.rosaceae.org/species/pyrus/all) reflecting the increased accessibility of genomes for applied plant research. Towards exploring the effect of using genetically mismatched genomes for RNA-Seq, this project was synergistic with Honaas' Tech project "Enhancing reference genomes for cross-cultivar functional genomics (TR-17-100)." That project:

> **1)** Aimed to improve the reference genomes for cross-cultivar RNA-Seq – for instance interpreting 'd'Anjou' gene activity data with the genetically mismatched 'Bartlett' genome
> **2)** Contributed data that enhanced gene discovery in 'd'Anjou' to facilitate comparative genomics in European pear.

The synergy with Honaas' tech project directly impacted each objective in this project – the 'd'Anjou genome enhances gene activity analyses, facilitates gene discovery, and expands our potential gene candidate list in our search for genes that influence European pear fruit quality in the postharvest period.

This project also had significant cooperation with the WTFRC project "Improving Quality and Maturity Consistency of 'd'Anjou'" that was led by Stefano Musacchi. Primarily, we were given access to cryopreserved fruit samples (for gene activity analysis) and associated fruit quality data that could be used to guide the gene activity analysis. We also used data from that project to guide sample selection for gene activity analysis - **Musacchi's project showed that 'd'Anjou' pear fruit from external canopy positions ripened differently compared to fruit from internal canopy positions.** We reasoned that targeting a very fine contrast – pear fruit harvested from the outer canopy vs. the inner canopy that also fell into the same DA meter class – would provide the best opportunity to find gene activity signatures that could differentiate the fruit. This project, by drawing on other WTFRC funded work, has provided a foundation for functional genomics in 'd'Anjou' pear towards development of biomarkers to predict future fruit quality, especially with regard to ripening capacity.

RNA-Seq: the first look at global gene activity measurements in 'd'Anjou' pear

Our first step was to extract the cryopreserved samples from Musacchi's project (see above) using the protocol developed in Honaas' lab (https://doi.org/10.1186/s13104-017-2564-2) specifically for the recalcitrant tissues of long-stored pears. We chose to analyze all 5 biological replicates from Musacchi's banked tissues to maximize our statistical power for identification of significantly differential gene activity signatures between fruit from the inner vs. outer canopy. However, the 'Bartlett' v1.0 genome was not an ideal reference due to the low rate of data inclusion, most likely due to the genetic differences between the 'Bartlett' genome and 'd'Anjou' gene activity data. Our initial RNA-Seq experiment brought about 56% of the RNA-Seq data into the analysis, similar to published work (https://doi.org/10.3389/fpls.2017.00455). While we did not expect all of the data to be brought into the analysis (for a host of reasons that are beyond the scope of this report), failing to bring nearly half of the data into the analysis certainly results in many false negatives, providing at best an incomplete picture of the biology the experiment attempts to describe.

The story that can be told from the parts of the biological picture we can see must be validated. There are no hard and fast standards for validation, though the research community generally agrees that congruent gene activity estimates from another technology, like quantitative PCR, are acceptable. However, how to choose genes, and what threshold for congruence is acceptable varies widely. The data from this project was used to refine our published RNA-Seq validation protocol

(https://doi.org/10.21273/JASHS04424-18), and results in improved validation results (Figure 1) specifically by targeting genes, and regions of genes, that are highly similar between 'Bartlett' and 'd'Anjou.' This suggests that the genetic differences between the two cultivars are problematic when attempting to estimate gene activity from one using the genome of the other.

Without the 'd'Anjou' genome, we did not have genetically matched reference genes to interpret our RNA-Seq data – but the next best option was to build gene models directly from our RNA-Seq data – an approach called *de novo* transcriptome assembly (for more info see - https://doi.org/10.1371/journal.pone.0146062). This approach has limitations that are inherent to the process that result in fragmented and incomplete gene models – but the models will exactly match the gene activity data and should therefore allow more of it to be assigned than when we use the 'Bartlett' genome. Indeed this was the case as we were able to increase the data inclusion rate by about 15% when using our *de novo* gene models. While this confirmed that genetic differences may be interfering with gene activity measurements, the limitations of using these gene models do not make it a better choice than the 'Barlett' v1.0 genome, albeit for different reasons.

Concurrent with the aforementioned experiments, Honaas' WTFRC Tech project shifted from improving the 'Bartlett' v1.0 reference to generating data to build a 'd'Anjou' genome from scratch. Because the gene activity data we had showed structure that was sufficient to distinguish fruit that ripened differently in the postharvest period (inner canopy vs. outer canopy positions – Figure 2) the additional sequencing budget was used to generate several types of genome data for 'd'Anjou' pear rather than to generate more transcriptome data – this additional data substantially improved the 'd'Anjou' genome assembly (roughly 7 fold).

Gene discovery in 'd'Anjou' pear for comparative genomics

We met objective 2 by gathering a massive gene activity data set (~1.5 billion reads) from pear fruit that was then used to build gene models via *de novo* transcriptome assembly. This approach provided models for genes were active in our pear fruit experimental samples but exclude roughly 1/3 of pear genes that were not active in our samples. By pivoting to whole genome sequencing for gene discovery, we cast a broader net that included a near complete compliment of genes. The annotation, or survey of genes across the genome, showed that the number of genes in our 'd'Anjou' draft genome (45,981) was similar to the draft genome of Bartlett (45,217).

During the last months of this project a second, higher quality 'Bartlett' genome became available and was officially published in Dec 2019 (https://doi.org/10.1093/gigascience/giz138). The gene count in this genome was lower than the draft genomes for 'Bartlett' and 'd'Anjou' at 37,445 due in part to the purge of alleles from the genome - alleles can cause considerable redundancy in gene predictions. If this process were error-free, the gene complement of 'Bartlett' v2.0 would be a perfect subset of 'Bartlett' v1.0, thus we explored the gene content differences between the 3 genomes (Figure 3). We found that the genes in the 'Bartlett' v2.0 genome were not a perfect subset of 'Bartlett' v1.0, and that the cultivars 'Bartlett' and 'd'Anjou' do potentially contain a small number of genes that are cultivar specific. The reasons for the gene content differences might include real genetic differences in the artificially induced double haploid line used for 'Bartlett' v2.0 as well as the bioinformatic filtering of alleles during annotation of the new genome.

Importantly, in addition to finding cultivar specific genes we now have the evidence to survey the genome to find genes that are shared between the cultivars, but have small differences that may help explain cultivar specific traits. In our apple experiments, the analysis of 'Granny Smith' vs. 'Golden Delicious' showed that of the ~5.2 million polymorphisms, nearly 0.5 million occurred in genes. Others (Zhang et al. 2019 - https://doi.org/10.1038/s41467-019-09518-x) found, in apple as well, that >1/3 of apple genes encode altered proteins when comparing cultivars, specifically 'Hanfu' vs. 'Golden Delicious.' Recall that we also surveyed the 'Bartlett' genome for differences with 'd'Anjou' and found roughly 5.6 million small polymorphisms. We expect the pattern to be similar in pear, which would result in hundreds of thousands of polymorphisms in several thousand genes.

The gene discovery phase of this project is important for two main reasons. First, as the results of Honaas' WTFRC Tech project show, a genetically matched genome is preferred for RNA-Seq data because it can recover gene activity signal that is lost when using a genetically mismatched genome. Second, the ability to survey genomes for even small differences between cultivars of European pear offer the possibility to discover the genetic basis for cultivar differences. For 'd'Anjou' pear, of particular interest are the genetic factors that explain the different chilling requirement for proper ripening compared to 'Bartlett.'

A preliminary list of candidate genes for biomarker development

The primary deliverable of this project is a foundation to discover biomarkers that can be used to predict future fruit quality. Towards that end, we have leveraged the resources from this project, and those from the WTFRC projects "Improving Quality and Maturity Consistency of 'd'Anjou'" led by Stefano Musacchi, and Honaas' Tech project "Enhancing reference genomes for cross-cultivar functional genomics (TR-17-100)," to generate a list of candidate genes for biomarker development. This effort began with RNA-Seq using the 'Bartlett' v1.0 genome, but the low data inclusion rate indicated a need for a better reference. During the last year of the project, while we were finishing 'd'Anjou' v1.0 the next 'Bartlett' genome, v 2.0, was made publicly available prior to publication via the Genome Database for Rosaceae (https://www.rosaceae.org/). We repeated the full RNA-Seq analysis using both 'Bartlett' v2.0 and our 'd'Anjou' v1.0. The data inclusion rate improved substantially with each of the new genomes, increasing ~10% (Figure 4 – from TR-17-100 final report).

This result was reassuring, but not without surprises. The amount of data that matched to genes was similar for each new genome, 63.8% for 'Bartlett' v2.0 and 65.8% for 'd'Anjou' v1.0. However, the fraction of RNA-Seq reads that matched uniquely was lower in 'd'Anjou, and the fraction of reads that mapped in between genes was also lower for 'd'Anjou. These are likely artifacts of annotation and will be improved as these genome resources are improved. The annotation (a highly iterative exercise) for the new 'Bartlett' genome was more mature than our first pass annotation of 'd'Anjou.' Furthermore, because the data inclusion rate in both experiments was similar, we used the RNA-Seq analysis with 'Bartlett' v2.0 in subsequent steps.

Biomarker candidates - differential expression analysis and functional enrichment analysis

Perhaps the most striking shifts in gene activity were from harvest through the end of the storage period (8 months). The differences between fruit from internal canopy positions vs. external positions was dwarfed by changes during storage, often by a factor of >10 (Figure 5a). The peak for differential gene activity signatures between treatments was after 3 months of storage (Figure 5b), which is also the time point for the most divergent ripening between the internal vs. external canopy fruit. At this time point (3 months of storage) the gene activity signatures included enhanced cellular transport processes in fruit from the external canopy positions, while internal canopy positions had signatures of osmotic stress and wound response.

Biomarker candidates - Gene Co-expression Network (GCN) analysis

Using the same gene activity data set, John Hadish, a student in the lab of Dr. Stephen Ficklin (Honaas' collaborator from the WSU Horticulture department) constructed a GCN that considered gene activity in all internal tissues vs. all external tissues, we did not distinguish between peel and cortex. We expected that the fine contrast in fruit quality characteristics and differential gene activity would be reflected in a comparative network analysis, showing a relatively small number of genes. Indeed this analysis showed a small number of co-expressed genes, and several of these genes show significant co-expression in both treatments (Figure 6). The lists of differentially active genes and co-expressed genes serve as a starting point for biomarker candidate selection.

Biomarker candidates – preliminary selection
We manually searched the gene activity signatures for a subset of genes that are both differential and co-expressed to find signatures that can distinguish pear fruit that we know have different ripening characteristics in the postharvest period (see examples in Figure 7). We found a variety of patterns that were able to distinguish the fruit, including differences at harvest that persist throughout storage (Figure 7a), differences that emerge during storage (Figure 7b), and more complex changes that may require integration to predict ripening characteristics (Figure 7c & 7d).

Perspectives
This project has provided the foundation to take the next steps for biomarker discovery in European pear using functional genomics combined with experimental horticulture. In this initial search we explored a very fine contrast – fruit from the same tree that had similar, at-harvest fruit quality characteristics, but had divergent ripening characteristics in the postharvest period. We used very aggressive statistical methods to select the only the most prominent signatures in this fine contrast as a proof of concept. We readily found candidate genes that could distinguish fruit at harvest and during storage by manually curating the data. These results show biological signatures exist in the data that, with sufficient development, may be used to create postharvest tools for both research and industry. Next steps include leveraging cutting-edge bioinformatic approaches to mine the data for additional candidates, and then deploying these preliminary biomarkers for validation studies.

FIGURES

Figure 1. RNA-Seq validation with qPCR improves when validation genes are genetically identical. This indicates that genetic polymorphisms influence gene activity estimates.

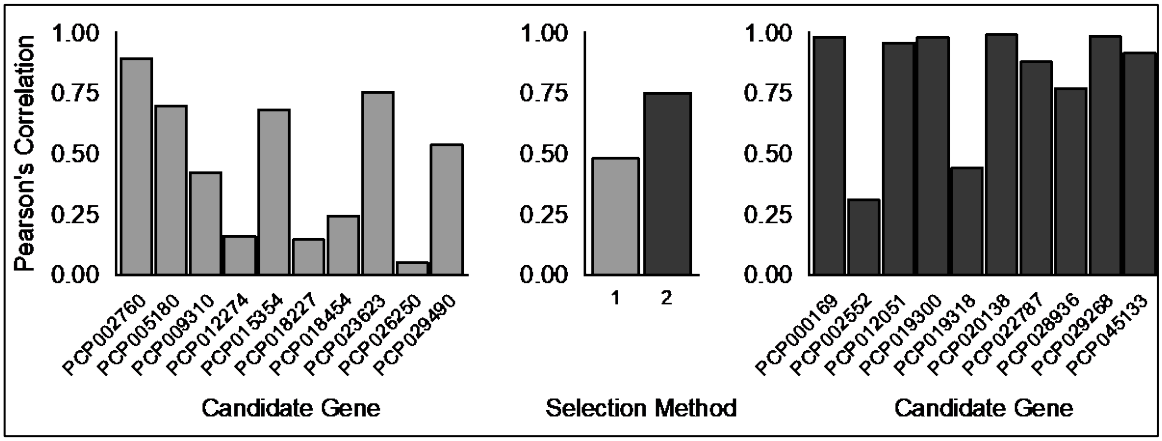


Figure 2. **A Principle Component Analysis of gene activity data shows structure that separates fruit from internal vs. external canopy positions.** This indicates that the gene activity data we gathered for the project might contain individual signatures that may be useful to distinguish fruit that have different ripening characteristics in the postharvest period.

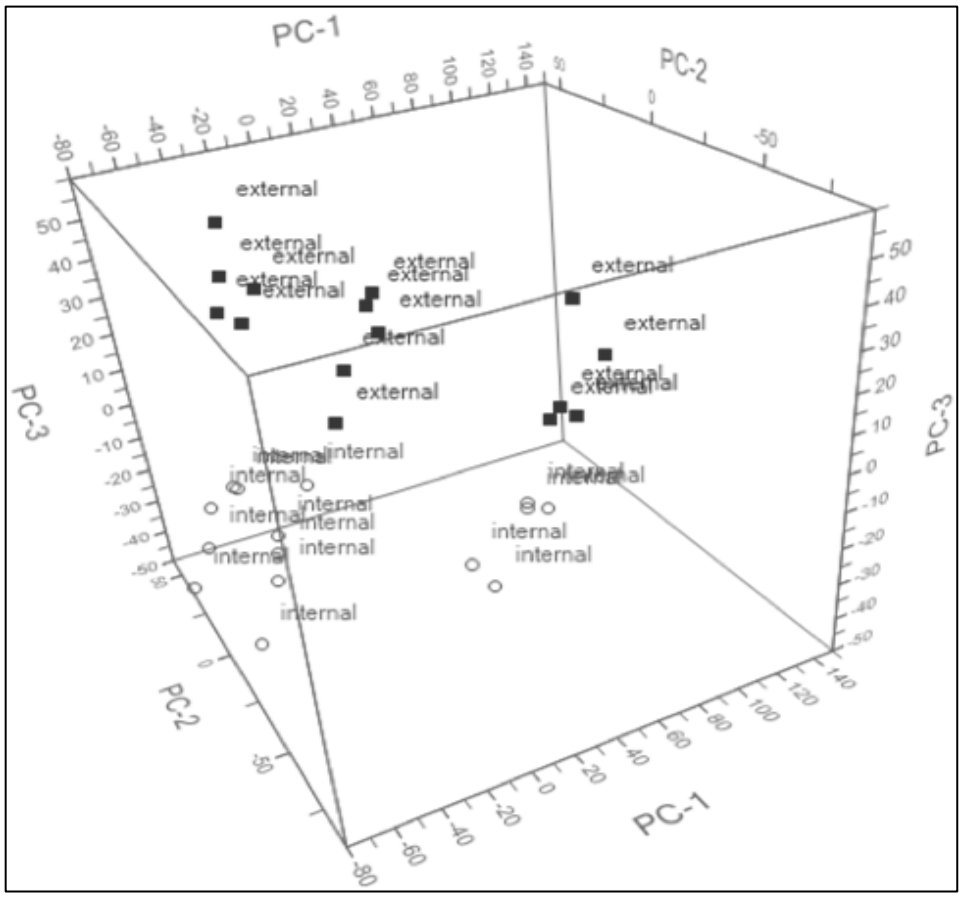


Figure 3. **Genome wide analysis of known plant gene families shows a majority of pear gene predictions from 3 different genomes overlap, but there are potentially cultivar specific gene families (plant gene families typically contain a small number of genes).** These differences in gene family predictions are probably attributable to real biological differences and also methodological differences between the approaches used to build and annotate each of the genomes.

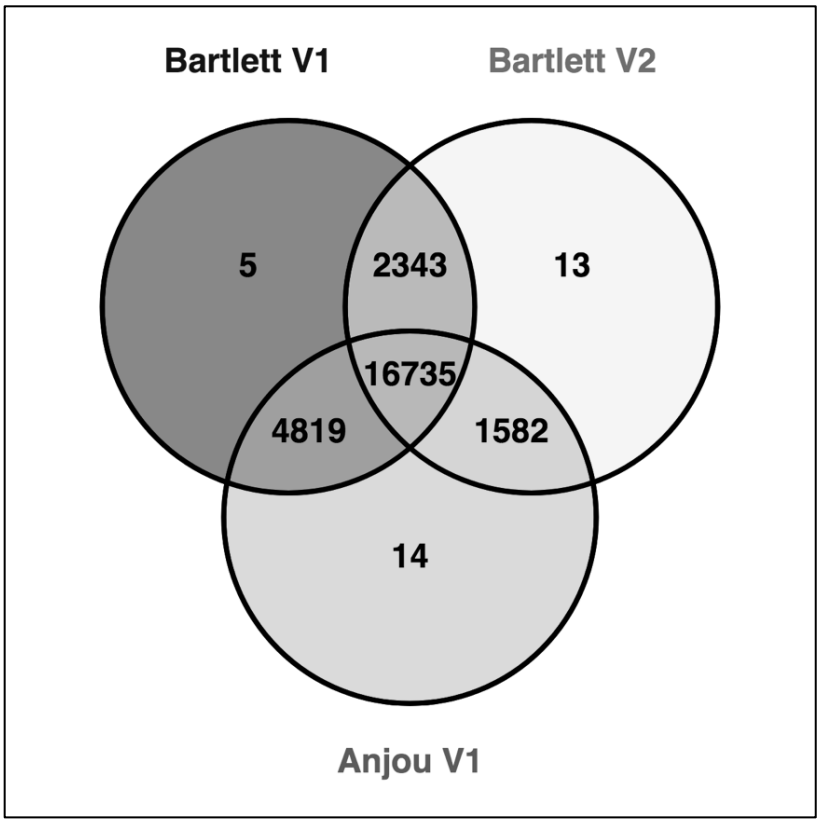


Figure 4. **The amount of 'd'Anjou' gene activity data included in RNA-Seq analyses increases in superior genomes (Bartlet v2.0) and genetically matched genomes (d'Anjou v1.0).** The 'Bartlett' v2.0 genome is superior because the genome assembly is less fragmented than 'Bartlett' v1.0. Even though the 'd'Anjou' v1.0 genome is more fragmented, the pieces are genetically matched to the RNA-Seq data, allowing higher data inclusion.

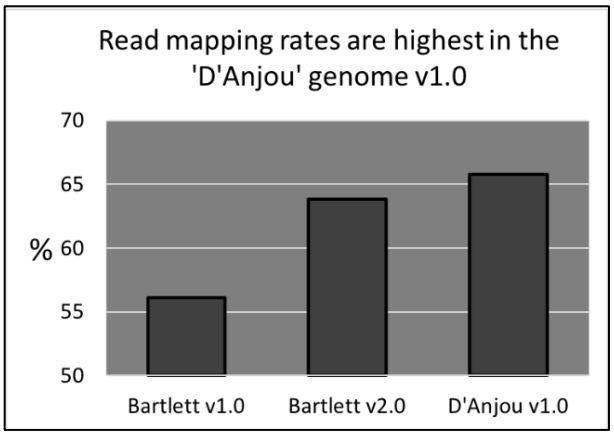


Figure 5.
A: Statistically significant (Bonferroni p<0.05) gene activity *changes* during storage dwarf gene activity *differences* between fruit that have different ripening characteristics in the postharvest period. IC = cortical tissue/internal canopy fruit, **EC** = cortical tissue/external canopy fruit, **down IC & EC =** less gene activity over time, **up IC & EC** = more gene activity over time. In external canopy (**E**) fruit cortical tissue vs internal canopy (**I**) fruit cortical tissue after 8 months of storage, **"down E vs I" =** less gene activity and **"up E vs I"** = more gene activity.
B: The maximum gene activity difference between fruit at 3 months of storage correspond to the largest differences in ripening characteristics, further suggesting the potential to identify gene activity signatures that can distinguish the fruit.

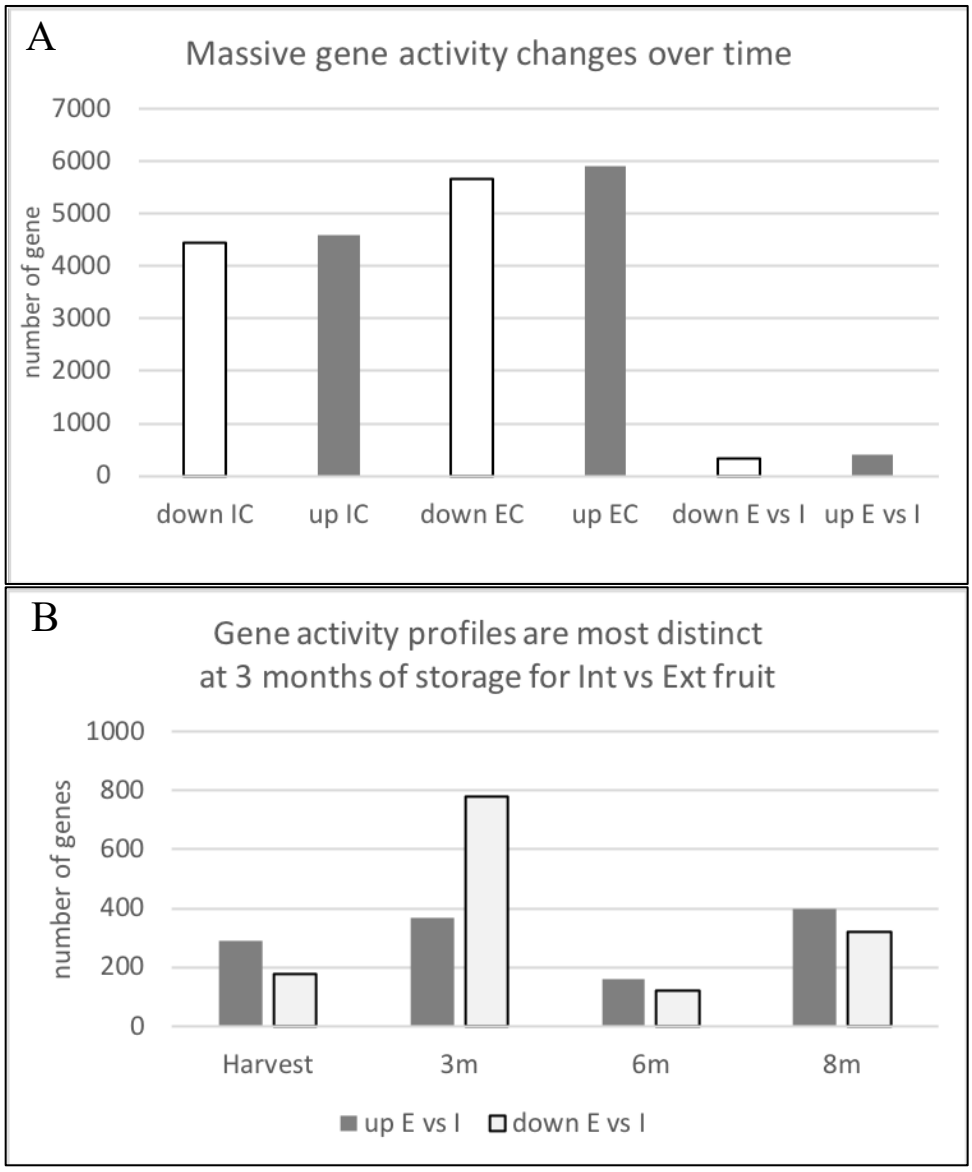


Figure 6. **The Gene Co-expression Network highlights genes with correlated expression during the postharvest period.** Dark gray edges indicate a significant relationship in fruit from internal canopy positions, light gray edges indicate a significant relationship in fruit from external canopy positions. The fruit from internal vs. external canopy positions have different ripening characteristics, and some gene activity signatures are shared (double lines) while others are distinct.

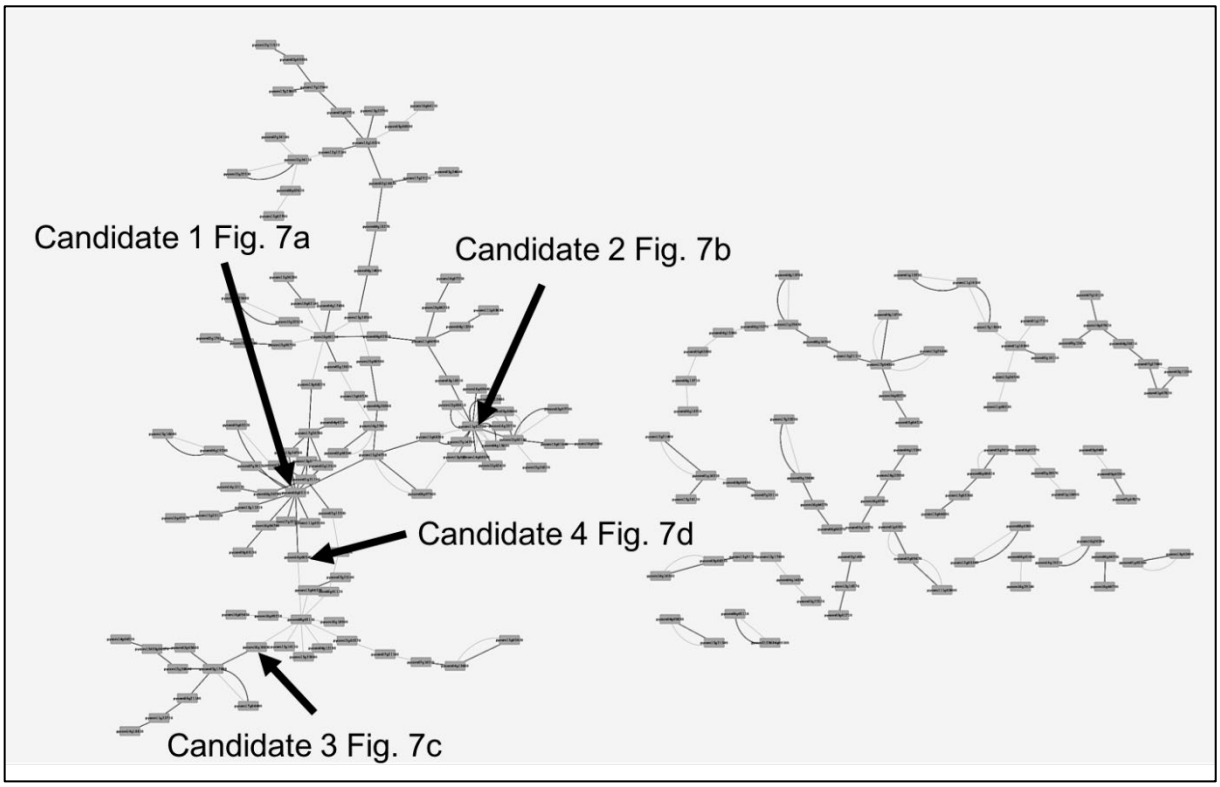
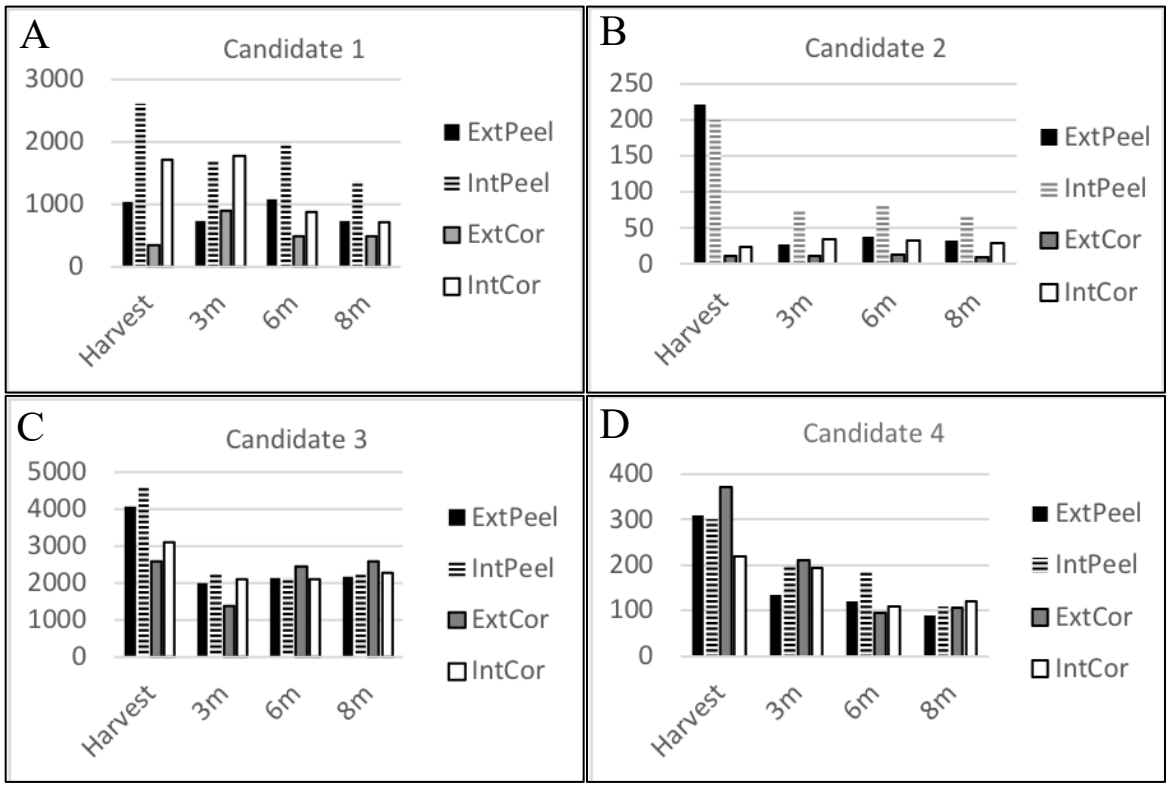


Figure 7. Gene activity signatures can be used to distinguish fruit that have different ripening characteristics – in this experiment the contrast is *internal* canopy vs. *external* canopy 'd'Anjou' pear fruit. Normalized gene activity is on the **Y-axis**, sample times are on the **X-axis**. The examples below of hand selected biomarker candidates are indicated above in Figure 6.



EXECUTIVE SUMMARY

Project title: Functional genomics of 'd'Anjou' pear fruit quality and maturity

Key words: d'Anjou, European pear, maturity, RNA-Seq, genome, gene, TR-17-100, PR14-108A

Abstract: We scanned *Pyrus communis* 'd'Anjou' fruit gene activity profiles to find genes related to postharvest quality and capacity to ripen. We now have a list of candidate biomarkers for future work aimed at understanding the genetic basis of pear fruit quality traits. We will emphasize development of tools to predict future fruit quality, especially as it relates to ripening capacity.

Summary: The aim of this project was to generate genomic resources for 'd'Anjou' pear towards the development of postharvest tools for enhanced pear fruit quality. We leveraged existing WTFRC funded research from Stefano Musacchi ("Improving Quality and Maturity Consistency of 'd'Anjou'') by using samples from their cryopreserved biobank of pear fruit tissue and the associated fruit quality data. From that project we selected fruit that ripened differently in the postharvest period – fruit from internal vs. external canopy positions. We proceeded to gather massive gene activity data sets from these samples, and began to mine the data for signatures, or hints, about future fruit quality especially with regard to estimating maturity and the capacity to ripen.

Concurrently, Honaas' WTFRC project "Enhancing reference genomes for cross-cultivar functional genomics" helped to meet our gene discovery goal for 'd'Anjou' pear by sequencing the genome of this variety. Thus we discovered virtually all of the genes in this variety, rather than just genes that were active in our fruit samples as initially proposed. This allowed us to exceed our goal for gene discovery and to build a stronger foundation for comparative genomics in European pear. During this same time period, several *Pyrus* genomes became available, including a new version of the 'Bartlett' genome. We repeated our full gene activity analysis using both versions of the 'Bartlett' genome (v1.0 and v2.0) plus our new 'd'Anjou' genome. We found that our 'd'Anjou' genome and the new 'Bartlett' genome were better than the first 'Bartlett' genome for analyzing our gene activity data sets. Because the finishing steps for building a genome were more mature for 'Bartlett' v2.0, we continued our search for potential biomarkers using that genome as a reference.

We found that the changes in gene activity (i.e. number of genes that were different) from harvest to 8 months of storage dwarfed the differences between fruit from equivalent storage time points, often by more than a factor of 10. Yet it was clear that our gene activity data set had structure that distinguished fruit with different ripening characteristics. When we combined the analysis that showed which genes were different between samples, with one that identified genes that had highly similar activity signatures within samples, we had a starting list of candidate biomarkers. By manually digging through the data we identified genes that had interesting signatures. This included patterns that allowed us to differentiate fruit from internal vs. external canopy positions at harvest and also during storage. The next steps include leveraging our new genomics resources and cutting-edge bioinformatics tools to mine the data for additional candidates, and then deploying these preliminary biomarkers in validation studies.