FINAL PROJECT REPORT **YEAR: 3 (of 2+1yr NCE)**
WTFRC Project Number: TR-17-100

Project Title: Enhancing reference genomes for cross-cultivar functional genomics

PI: Loren Honaas
Organization: USDA, ARS
Telephone: (509)664-2280
Email: loren.honaas@ars.usda.gov
Address: 1104 N. Western Ave.
City/State/Zip: Wenatchee, WA 98801

Co-PI: Joshua Der
Organization: Cal State University, Fullerton
Telephone: (657)278-4115
Email: jder@fullerton.edu
Address: 800 North State College Blvd.
City/State/Zip: Fullerton, CA 92831

Cooperators: Stefano Musacchi & Sara Serra (WSU-TFREC), Claude dePamphilis (PennState)

Total Project Request: **Year 1:** $48,832 **Year 2:** $35,207

Percentage time per crop: Apple: 60% Pear: 40% Cherry: NA Stone Fruit: NA

Other funding sources: None

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	$3,750	$750
Supplies	$3,000	$3,722
Miscellaneous[2]	$11,664	NA
Total	$30,914	$16,972

Footnotes:
[1]Data analysis including Research Support Agreements to cooperators
[2]Cooperative Agreement for PacBio library prep + sequencing to Penn State Group

Budget 2
Organization Name: CSU Fullerton **Contract Administrator: Alison Nguyen**
Telephone: 657-278-7621 **Email address: allisonnguyen@fullerton.edu**

Item	2017	2018
Salaries[1]	$8,922	$9,234
Benefits[1]	$129	$134
Wages[2]	$8,526	$8,526
Benefits[2]	$341	$341
Total	$17,918	$18,235

Footnotes:
[1]Salary and benefits for Joshua Der – 1 month
[2]Salary and benefits for Der lab student – 2 semesters

OBJECTIVES

Enhance discovery of genetic factors associated with fruit quality differences using existing and in-progress RNA-seq data, along with publicly available genomic resources:

Step 1) **identify genetic differences** between reference genomes and genomes of interest ('Golden Delicious' vs 'Granny Smith' & 'Bartlett' vs. 'D'Anjou').

Step 2) use bioinformatic approaches to **update the reference genomes** to reflect these differences creating custom, polished references for analysis of gene expression in each of the genomes of interest.

Step 3) **compare gene expression results** from the original and polished versions to calculate changes in read mapping rates focusing on total reads matched and changes in uniquely matched reads (both indicating changes in sensitivity for measuring gene activity) **to evaluate the efficacy of the genome polishing strategy.**

SIGNIFICANT FINDINGS

Progress on specific objectives:

- **Step 1) Exceeded -** We identified genetic differences between reference cultivars and cultivars of interest, finding millions of polymorphisms across the whole genome, not just in gene sequences
- **Step 2) Exceeded** - By sequencing the full genome of 'Granny Smith' and 'D'Anjou', we can account for 10s of thousands of gene sequence polymorphisms captured in recent experiments, and we discover virtually all 'Granny Smith' and 'D'Anjou' genes, not just those shared with reference cultivars 'Golden Delicious' and 'Bartlett'
- **Step 3) Exceeded –** We can bring ~10-15% more gene activity data into an analysis by mapping to our cultivar specific genomes than to the reference genomes for 'Golden Delicious' and 'Bartlett'. Additionally we characterized the effect of polymorphisms on specific genes

Other important findings:

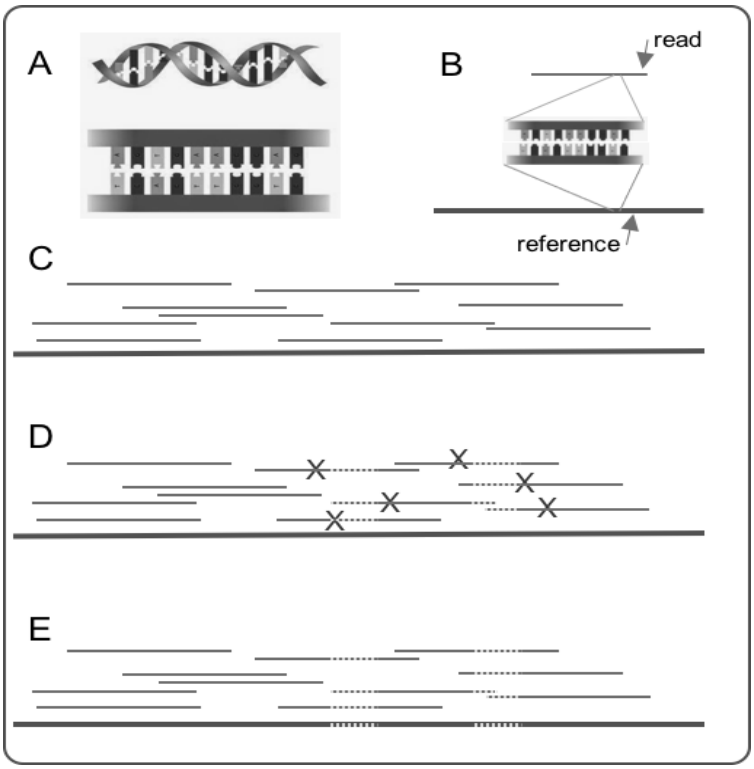
- Differences among the genome sequences of *Malus* and *Pyrus* cultivars are problematic for gene activity measurements
- Affected genes may be important for understanding important fruit quality traits
- Apple and pear genomes provide research opportunities beyond gene activity measurements that will help us understand pome fruit traits
- Genomes also facilitate:
  - access to more cost-effective gene activity measurements
  - deeper comparative genomics in apple and European pear
  - access to emerging technologies for gene activity analysis

RESULTS & DISCUSSION

Excerpt from the project proposal justification

Recent technological advances have allowed researchers to measure the activity of all genes in a sample simultaneously using a method called RNA-Seq. This approach uses Second Generation Sequencing to create hundreds of millions of measurements called "reads" – these are the RNA-Seq basic unit of measurement. The reads from each sample are counted with sophisticated software by matching reads to the DNA sequence of known genes. This critical step in analyzing RNA-Seq data relies on existing knowledge of the genes being measured and assumes that the RNA-Seq data perfectly matches the existing genes. These known genes come from sequenced **genomes** – genomes are the collection of all genetic material in an organism, a small fraction (~5%) of which are the genes. A reference genome exists only for one apple cultivar, 'Golden Delicious', and one European pear, 'Bartlett'. However, commercially produced apples and pears contain substantial genetic diversity within and among cultivars. When doing RNA-Seq in other cultivars, these genetic differences (polymorphisms) may present substantial challenges when obtaining gene activity measurements in postharvest fruit quality experiments.

For example, genetic differences between 'Golden Delicious' and 'Granny Smith' may lead to loss of signal in RNA Seq experiments because 'Granny Smith' reads may fail to be assigned to the correct gene (or at all) in the reference 'Golden Delicious' genome, due to a *mismatch* (see Box 1).



Box 1. RNA-Seq data analysis relies on digitally matching reads, letter-by-letter, to the DNA sequences of genes. A) *rendering of the DNA double helix (top) and flattened version (bottom) more typical for visualizing DNA sequences (image credit -http://biology4alevel.blogspot.com/),* ***B)*** *zoomed schematic showing how reads are matched, letter-by-letter, to reference genes where the position of the read indicates where it matches to the reference,* ***C)*** *'Golden Delicious' example where reads match the reference and are mapped successfully,* ***D)*** *'Granny Smith' hypothetical example where polymorphisms (dashed regions of the reads) fail to map to the correct position,* ***E)*** *example of a polished apple reference corrected for use with 'Granny Smith' that allows reads to map successfully.*

The question that led to this project was straightforward: "What is the effect of genetic mismatches between genomes in RNA-Seq experiments?" This led to the following hypothesis.

> ***If the genes that explain differences in cultivars tend to be different, and if these differences are related to loss of signal in gene activity measurements, then the genes that may explain important cultivar differences are the most likely to be missing from gene activity analyses.***

Simply put, the genes that make 'Granny Smith' fruit different from 'Golden Delicious' fruit may be dropping out of our gene activity experiments because they are genetically different. If those missing genes are involved with important fruit quality traits (like Superficial Scald susceptibility), this creates a critical blind spot when we use genomics tools like RNA-Seq to learn about those important fruit quality traits.

Analogy of the problem

The issue described above can be illustrated in a jigsaw puzzle analogy. The RNA-Seq data units, called reads, can be thought of as jigsaw puzzle pieces. The reference genome is best thought of as the picture that comes printed on the jigsaw puzzle box that can be consulted by the user to guide where to put puzzle pieces. The issue with mismatched genomes is that the puzzle pieces do not perfectly match the picture on the box (Box 2). A skilled user could probably finish the puzzle if it consisted of a few thousand pieces. However, an RNA-Seq puzzle consists of hundreds of millions of pieces, and the picture is ~100 times higher resolution than a 4k Ultra HD television. So, necessarily, sophisticated software on high performance computers are needed to complete the task.

In fact, large portions of the puzzle *can* be put together using software that does not need a reference "picture" – this is how we built gene models early in the project. Yet those pictures were not fully complete due to inherent limitations of the process. Assembling a genome is the only way to get the whole picture.

Box 2. *In the puzzle analogy, when the pieces are assembled correctly, they would create the picture in "A." But when using the picture in "B" to guide assembly of the puzzle, the result would be an incomplete image in portions of the picture where the two were most different.*

Genetic differences between cultivar specific genomes and reference genomes – Step 1

Step 1 was to scan the genomes of 'Granny Smith' and 'D'Anjou' to find differences compared to the reference cultivars 'Golden Delicious' and 'Bartlett,' respectively. We eventually found millions of small differences - Single Nucleotide Polymorphisms (SNPs) and small insertions or deletions (InDels) in 'Granny Smith' apple (5.18 million) and 'D'Anjou' pear (5.60 million) (Figure 1). While many of these differences were located in between genes, some of these differences did occur in the genes themselves (e.g. in one experiment we observed 490,275 polymorphisms across 'Granny Smith' apple genes). This is critically important, because genes, while they only represent about ~5% of the apple and pear genomes, are the target for interpreting gene activity measurements in RNA-Seq experiments. *Malus* and *Pyrus* genomes over all are known to be polymorphic, but the extent to which ***genes*** are polymorphic had not been thoroughly examined, nor had there been a detailed exploration of the effect on gene expression analyses using a genetically mismatched genome reference.

Initially, using our raw 'Granny Smith' gene activity data we built cultivar-specific gene models and searched for matches to known 'Golden Delicious' genes. Then for a small set of genes we explored the effect of polymorphisms by comparing gene activity measurements between a gene-by-gene approach (1 gene at a time - qPCR) vs. the global approach (all ~40,000 genes at once – RNA-Seq). We published a paper (Hargarten et al. 2018 - https://doi.org/10.21273/JASHS04424-18) that included a protocol for improved RNA-Seq data validation, and we also found some evidence that polymorphisms can interfere with RNA-Seq gene activity measurements. However, that was based on a small number of genes, so we proceeded to the next step which was to modify or polish the reference genomes to account for polymorphisms. This would provide a large gene sample size to compare experimental results the reference genomes and the polished versions.

Building a better reference for RNA-Seq – Step 2

The original strategy for Step 2 was to use the raw gene activity data to identify and then modify or polish polymorphisms in the genes of reference cultivar genomes. However, during the first year of the project, advancements in 3rd generation genome sequencing technology reduced the cost of obtaining high-quality genomes. This created the possibility of building new cultivar-specific genomes, instead of just polishing the publicly available reference genomes. This new approach was preferred for several reasons: 1) a full genome contains all the genes, whereas the raw gene activity data from fruit could only be used to polish genes that were active in fruit, 2) a full genome can show us large scale differences (not just SNPs and small InDels) that can help explain cultivar trait differences, 3) the full genome will include genes that are unique to 'Granny Smith' apple and 'D'Anjou' pear, showing us gene content differences that may help explain cultivar trait differences.

We therefore shifted our focus to obtaining data to build 'Granny Smith' and 'D'Anjou' genomes. The price decrease in 3rd generation genome sequencing technology (PacBio) allowed us to access the technology, but not enough to allow us to build a genome solely from long read data. Thus, we generated a hybrid data set consisting ultimately of 4 kinds of genome data; Illumina, PacBio, Oxford Nanopore, and 10x. Except for Illumina, which is a second-generation technology, these 3rd generation technologies take advantage of high molecular weight DNA (very long pieces) in either the sample preparation or the sequencing step. The resulting larger pieces of data effectively reduce the complexity of the genome assembly task. Following the puzzle analogy above (Box 2), this data type creates puzzles with larger pieces that are easier to assemble.

These multiple data types were assembled and postprocessed via a proprietary method that is in development in the dePamphilis lab at Penn State. We predicted based on the volume of data that the assemblies would be at least good enough to recover the complex portions of the genome that included genes. In fact, the genomes are highly complete, with recovery of fragments that account for virtually all of the *Malus* and *Pyrus* chromosomes (Figure 2). Further, when assessing the completeness of complex plant genomes like these, searching for genes that are shared widely across all plants is indicative of completeness. This method is called Benchmarking Universal Single Copy Orthologs (BUSCO) – the BUSCO completeness score for the 'Granny Smith' genome was 96.2% and for

'D'Anjou' was 97.2%. These metrics indicate that the gene content for the cultivar specific genomes are highly complete. This is a substantial increase over the BUSCO scores of ~65% in the gene model collections we initially built from the raw gene activity data in 'D'Anjou.' These results indicate that the 'Granny Smith' and 'D'Anjou' genomes will contain all the genes that are represented in RNA-Seq experiments.

With good genome assemblies in hand, we selected one to test the next step of annotation which involves scanning the ~600 million letters of DNA code (consisting of A,T,G, & C) to find genes, and then identifying them by comparison to known genes. We chose pear because our results suggested that *Pyrus* would benefit more than apple from cultivar specific genomes, and because of synergy with Honaas' pear project (WTFRC Project PR-17-104 "Functional Genomics of 'D'Anjou' Pear Fruit Quality and Maturity") that had a substantial RNA-Seq component that would likely benefit from the 'D'Anjou' genome v1.0. Briefly, the annotation process involves using the gene models we built from the raw 'D'Anjou' gene activity data in the initial phases of the project, combined with all the available resources for pear genes at the Genome Database for Rosaceae (https://www.rosaceae.org/). Luckily, this included 3 new pear genomes made public in 2019 - *Pyrus betulifolia* v1.0*, Pyrus ussuriensis* x *communis* v1.0*,* and *Pyrus communis* 'Bartlett' v2.0 (https://www.rosaceae.org/). The annotation results were concordant with results from the published Bartlett genomes v1.0 (Table 1), further indication of the high quality of the 'D'Anjou' genome v1.0. With updated RNA-Seq references and a catalog of differences between the cultivar-specific genomes and reference genomes, we moved on to Step 3.

Polymorphisms between pome fruit genomes interfere with RNA-Seq data measurements – Step 3A

As stated above, we had data from a small number of genes that suggested polymorphisms might be interfering with RNA-Seq based gene activity measurements. Concurrent with the effort to sequence and assemble the genomes of 'Granny Smith' and 'D'Anjou,' we developed a bioinformatics test that would help reveal the effects of polymorphisms on RNA-Seq measurements. Counterintuitively, this involved making the RNA-Seq reads even smaller. The idea was that if a 'Granny Smith' RNA-Seq read matched to a 'Golden Delicious' gene but had enough small differences that the computer program could not confidently match it, then by fragmenting the reads we might be able to separate out the pieces that matched perfectly from the ones that contained differences. Then the parts that matched perfectly would be successfully assigned, causing a recovery of the gene activity signal. We ran this test using a subset of RNA-Seq data from our published 'Granny Smith' work (Honaas et al. 2019 - https://doi.org/10.1016/j.postharvbio.2018.09.016).

We found that genes that contained differences showed a recovery in gene activity signals, and that the signal recovery was flanked by known genetic differences between 'Granny Smith' and 'Golden Delicious' (Figure 3A). Unfortunately, this test is not a solution to our problem. This is because while fragmentation of the reads helps recover the gene activity signal for genes with differences, for other genes the signal becomes ambiguous and therefore reduced (Figure 3B). We then examined the pattern across all genes for multiple samples, finding that there is a significant positive correlation (R^2=0.36±0.02, p=2.2e-16) between recovery of signal and polymorphisms when mapping fragmented reads (Figure 3C).

With further development this test may be useful to estimate the extent to which gene activity measurements are affected by genetic mismatches, especially with regard to tuning the parameters during the step where reads are matched to genes. For our purposes, it illustrates that there are pitfalls associated with mapping RNA-Seq data across cultivars. These pitfalls can be potentially avoided by providing a closer match between the gene activity data and the reference genome.

RNA-Seq mapping improves in a genetically matched genome – Step 3B

The final tests involve examining how the genomes perform as references for RNA-Seq experiments. These tests are straight forward, and our predictions about the increases in the proportion of RNA-Seq read data that could be brought into an experiment (15%) were surprisingly accurate. In the initial

annotation tests, we simply lifted over the gene location and structure information from the reference genomes to the new cultivar-specific genomes. This quick and dirty approach allowed us to annotate about 80% of the genes in the new genome. Starting with the apple genes that we examined in the fragmented read test above, we searched the new 'Granny Smith' genome to see if any of those genes were annotated. We did not find many genes that met this citeria, which is expected – the genes that saw signal recovery are among the most polymorphic in fragmented read analysis, and gene differences would cause the lift-over annotation strategy to fail. In Figure 4 there is an example of a gene that was different enough to show a signal recovery in the fragmented read experiment, but not too different such that it was excluded from the fast, reference lift-over annotation. When we mapped the 'Granny Smith' gene activity data at high stringency to both the 'Golden Delicious' genome and 'Granny Smith' genome, we saw the signal in the 'Granny Smith' genome increase. This result indicated that for at least some genes, a matched genome would allow signal recovery compared to a mismatched genome when used for RNA-Seq.

The last test was to run a full RNA-Seq experiment with the new fully annotated 'D'Anjou' genome. We ran the analysis 3 times with identical parameters: once with the 'Bartlett' v1.0 genome, once with the recently released 'Bartlett' v2.0 genome, and last with our 'D'Anjou' genome v1.0 (Figure 5). This test shows us that we can assign more RNA-Seq data to pear genes when we used a genetically matched reference genome. Furthermore, we observed that the mismatch rate, deletion rate and insertion rate were roughly twice as high when mapping to the 'Bartlett v2.0' genome, indicating that we could increase the stringency of the mapping parameters and enhance the confidence of our gene activity results when using the 'D'Anjou' genome.

Perspectives

The hypothesis that led to the proposal for this work started with the question "What is the effect of genetic mismatches between genomes in RNA-Seq experiments?" The answer we now have is that a genetically matched genome reference is better for RNA-Seq because we bring more data into the analysis at higher stringency, giving us an overall increase in the confidence of our gene activity measurements, leading to higher confidence in the biological stories these data tell.

By opting to sequence and assemble the whole genomes of 'Granny Smith' and 'D'Anjou' pear, we substantially increased our genomics tool kit for apple and European pear. It is important to note that these genomes are not endpoints but starting points for comparative genomics in these species. These are called draft genomes because they are works in progress. The area that needs the most improvement is the annotation of the genomes. This is a complex and iterative process, yet the first pass annotation showed us that there may be genes in the 'D'Anjou genome that are missing from 'Bartlett' versions 1 and 2. If this is proven to be the case, these genes would be totally absent from an 'D'Anjou' experiment using either of the 'Bartlett' genomes. The genomes also provide opportunities to leverage higher efficiency RNA-Seq data types that target small portions of genes, but which require the cultivar specific genomes we now have. We can now potentially access cutting edge data analysis tools, like Salmon (https://combine-lab.github.io/salmon/) which are super-efficient, yet require very high accuracy gene models which we now have with the 'Granny Smith' and 'D'Anjou genomes. Last, we can now more easily identify genes that are most different between 'Bartlett' and 'D'Anjou' providing exciting new opportunities to discover the genetic basis of the differences between these cultivars.

FIGURES AND TABLES

Figure 1. Cultivar specific genes are polymorphic.
A) An example of a **polymorphic** 'Granny Smith' gene showing both Single Nucleotide Polymorphisms (SNPs) and an Insertion/Deletion (InDel).
B) CIRCOS plots summarize a scan of 'Granny Smith' apple and
C) 'D'Anjou' pear genomes showing polymorphisms (outer trace) and genome data coverage (inner trace). Each block represents a large genomic fragment and the entire genomes of apple and pear are represented.

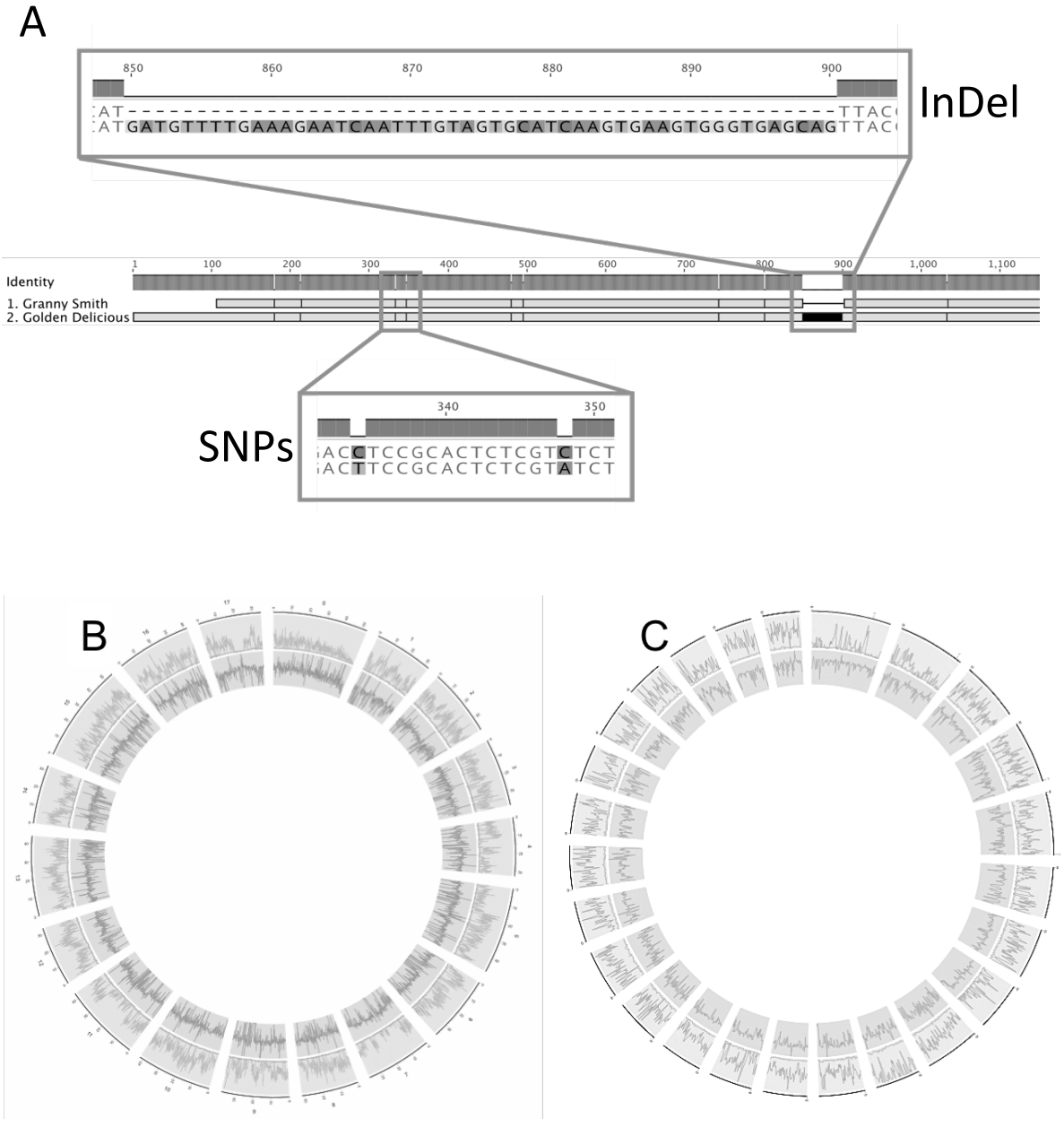


Figure 2. The *Malus* and *Pyrus* genomes we sequenced for this project are highly complete.
A) 'Granny Smith' and 'B) 'D'Anjou,' were lined up against the known reference genomes to estimate completeness. A straight and unbroken line from the lower left corner to the upper right corner indicates that for every portion of the reference genome, there is a corresponding piece of the cultivar specific genome that is in the correct orientation.

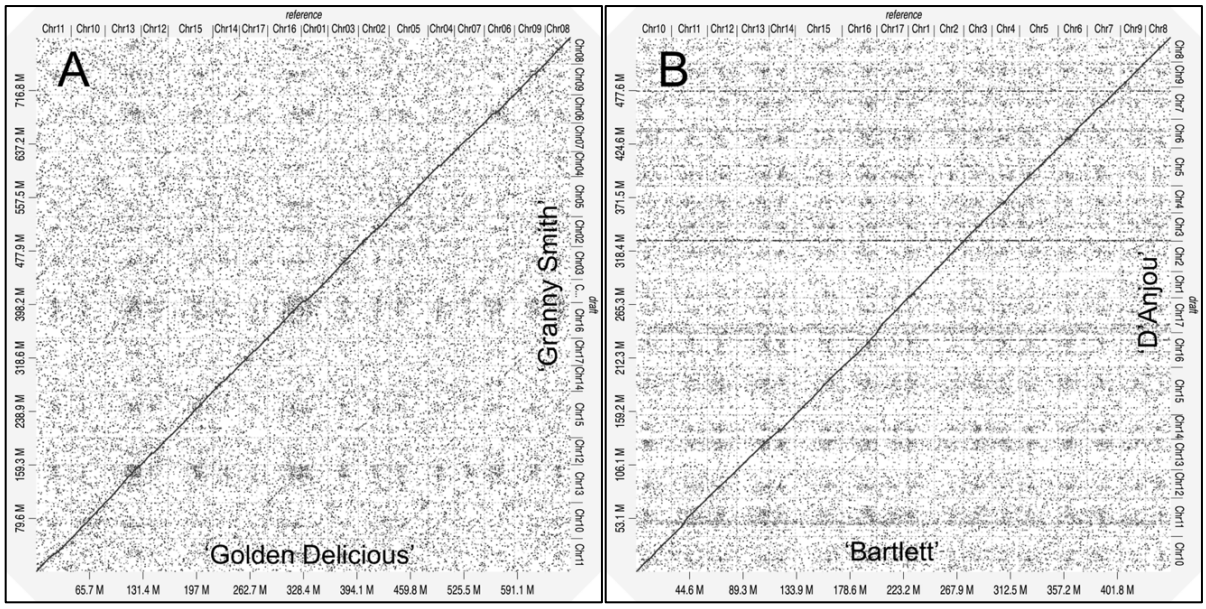


Figure 3. Polymorphisms can prevent reads from being matched to a reference gene. In panel A, Fragmented 'Granny Smith' reads match to the reference gene in between polymorphisms (arrow heads) that prevented the full length reads from matching to the reference gene. In panel B, signal is lost for a different gene when the reads are fragmented because the match becomes ambiguous – the reads match multiple similar genes in the 'Golden Delicious' genome. In panel C the pattern across all genes detected in the experiment shows that the number of differences (e.g. polymorphisms) is proportional to signal recovery in the fragmented read experiment. The genes that were most different showed the highest signal recovery when we matched read fragments in between polymorphisms.

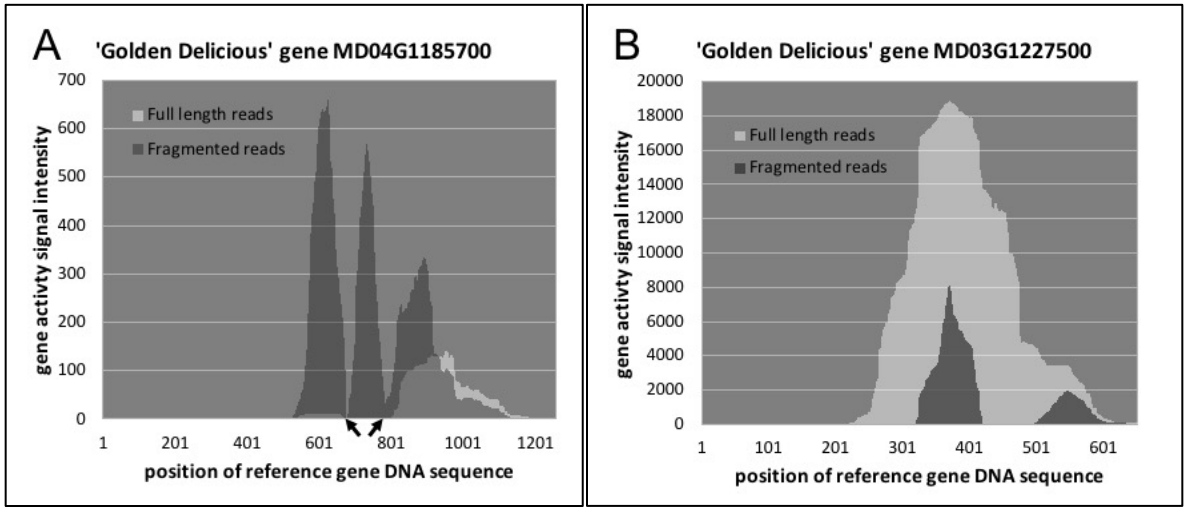


Figure 3C.

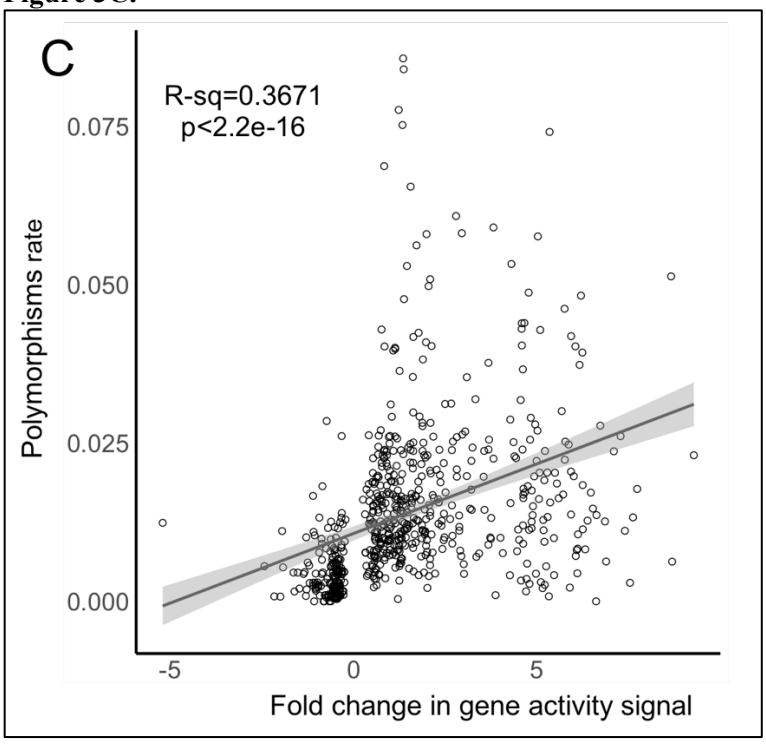


Figure 4. RNA-Seq signal recovery between to different experiments shows the same pattern. When we changed the structure of the read data to fit in between polymorphisms, we saw RNA-Seq signal recovery. In a different experiment, we mapped 'Granny Smith' RNA-Seq data to a partially annotated 'Granny Smith' genome as well as the 'Golden Delicious' genome. For this same gene, we saw a similar signal recovery when using the matched genome. This suggests that genetic polymorphisms can attenuate RNA-Seq signals.

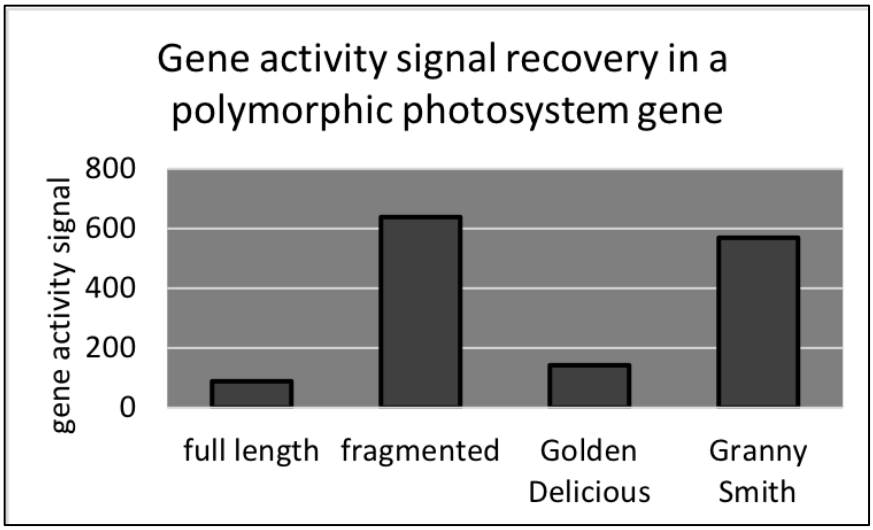


Figure 5. RNA-Seq data mapping was highest when the experimental data and genome matched. This result supports the hypothesis that genetic polymorphisms can create artifacts in RNA-Seq data, that include loss of signal. A matched genome can correct some of these artifacts.

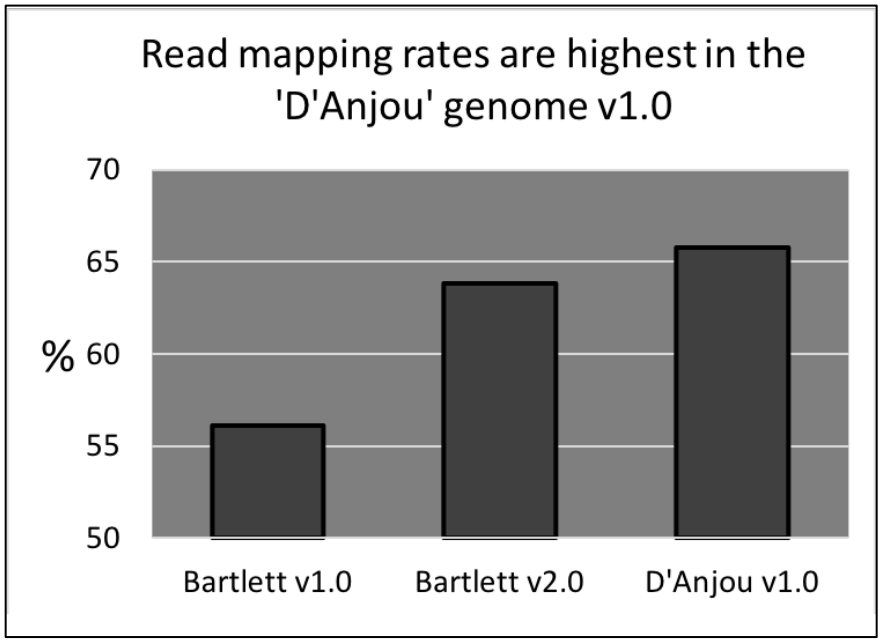


Table 1. Annotation summary for *Pyrus communis* genomes.

	Bartlett v1.0	Bartlett v2.0	D'Anjou v1.0
Number of genes	45,217	37,445	45,981
Number of genes in known plant gene families	34,444	28,547	36,976
% Genes in known plant gene families	76.2%	76.2%	80.4%
% Gene families detected	93.7%	81%	90.7%
Number of unassigned genes	10,773	8,898	9,005
% Unassigned genes	23.8%	23.8%	19.6%
Number of unique plant gene families	5	13	14
Number of genes in unique plant gene families	16	73	43

EXECUTIVE SUMMARY

Project title: Enhancing reference genomes for cross-cultivar functional genomics

Key words: Postharvest, genome, RNA-Seq, biomarker, gene expression, polymorphism, apple, pear

Abstract: The genomes among apple and pear cultivars differ, and these differences likely impact gene activity measurements. We explored methods to account for these differences, finding that cost effective genomes are a viable strategy. The new cultivar-specific reference genomes are a valuable part of efforts to characterize fruit quality traits using function genomics.

Summary: In the age of plant genomics, there are rapidly expanding opportunities to understand important plant traits. Researchers can associate genes with plant traits by scanning for gene activity signatures in experiments that highlight important plant traits. These experiments can improve our understanding of plant gene function, leading to a deeper understanding of plant traits. For instance, a better understanding of pome fruit traits may help us refine and/or develop technology to minimize postharvest losses.

Such experiments can efficiently target genes that are active in a sample. Combined with high performance sequencing machines that can generate 100s of millions of measurements, this approach can measure the activity of practically ***all*** of the ~40,000 apple (or pear genes) in a single experiment. However, like putting together a jigsaw puzzle, the researcher needs an accurate reference picture to know where to put each piece. This problem becomes enormous when trying to assemble a super high-resolution picture from a puzzle with hundreds of millions of pieces. Necessarily, researchers use cutting edge software on high performance computers to automate the task. This puzzle assembly step assumes a near perfect match between the picture and the puzzle pieces. If there are places where the puzzle doesn't match the picture, this can result in incomplete puzzle assembly. **Thus, this project aimed to understand the impact of the genetic differences between the reference genome data from the sequenced cultivars ('Golden Delicious' apple and 'Bartlett' pear) and the cultivars used in our experiments.** While the automated puzzle assembly step can be tuned to account for some small differences, we hypothesized that differences in specific genomic regions can cause the gene activity "picture" to be incomplete – the signal for certain genes would be repressed or lost. Simply, the genes that were the most different might be more likely to drop out of the analysis. Because the trait differences between cultivars are directly related to the genetic differences between cultivars, this could create a critical blind spot for polymorphic genes that help explain the unique fruit quality traits of a given cultivar.

Our results indicate that genetic differences do impact gene activity measurements, and that genes with more differences are more likely to be negatively affected. We explored ways to account for these differences, ultimately settling on genome sequencing due to rapid advances in the field of plant genomics during this project. This approach improved our ability to create a complete gene activity "picture," which now includes more of the highly distinct genes that may explain important cultivar trait differences. Our project initially focused narrowly on improving our ability to detect versions of genes that were still shared between apple and pear cultivars. With full genomes, we can discover additional large scale-genetic differences that can help us understand plant traits. This includes the discovery of genes that are unique to one cultivar or another.

The important investment in cultivar specific genomes allows us to get better gene activity measurements (for less money) and provides a foundational resource for functional genomics in pome fruit. This project has helped pave the way to obtaining this valuable resource for all the cultivars important in the Washington tree fruit market.