Project Title: Apple genomes for postharvest fruit quality biomarkers

Report Type: Final Project Report, Year 4 of 4 (Including 1 year NCE) for AP-19-103

Primary PI: Dr. Loren Honaas
Organization: USDA ARS
Telephone: 509.664.2280
Email: loren.honaas@ars.usda.gov
Address: 1104 North Western Ave
Address 2:
City/State/Zip: Wenatchee, WA 98801

Co-PI 2: Dr. Stephen Ficklin
Organization: WSU Dep. of Hort.
Telephone: 509.335.4295
Email: stephen.ficklin@wsu.edu
Address: PO Box 646414
Address 2:
City/State/Zip: Pullman, WA 99164

Co-PI 3: Dr. Jim Mattheis
Organization: USDA ARS
Telephone: 509.664.2280
Email: james.mattheis@ars.usda.gov
Address: 1104 North Western Ave
Address 2:
City/State/Zip: Wenatchee, WA 98801

Cooperators: Dr. Claude dePamphilis (Penn State Dep. of Biology), Dr. Dave Rudell (USDA ARS), Dr. Alex Harkess (HudsonAlpha Institute for Biotechnology)

Project duration: 3 Year (+1 year NCE)

Total Project Request Year 1: $87,142
Total Project Request Year 2: $96,692
Total Project Request Year 3: $97,991

Other related/associated funding sources:

Funding Duration:
Funding Duration: annual congressional appropriation (USDA TFRL Base funds)
Amount: $85,000
Agency Name: USDA ARS
Notes: 3-year total $220,000: Personnel $100,000, RNA-Seq $90,000, Consumables $30,000,

Funding Duration: N/A
Amount: $86,000
Agency Name: WSU Ficklin Start-Up Funds
Notes: These funds were used to purchase high-performance computing resources on WSU's Kamiak computing cluster. These resources will be used to perform data analysis for this project.

Funding Duration: 2017-2022
Agency Name: US National Science Foundation (NSF) Award #1659300
Amt. awarded: $150,000
Notes: A portion of this award was used to fund almost 1 Petabyte of storage for execution of scientific workflows and storage of results. We will use that infrastructure for this project.

Budget 1
Primary PI: Dr. Loren Honaas
Organization Name: USDA ARS TFRL
Contract Administrator: Chuck Meyers & Sharon Blanchard
Telephone: 510.559.5769 (CM), 509.664.2280 (SB)
Contract administrator email address: chuck.myers@ars.usda.gov, sharon.blanchard@ars.usda.gov

Item	**2019**	**2020**	**2021**
Salaries	33,000		
Benefits			
Wages			
Benefits			
RCA Room Rental			
Shipping			
Supplies	5,000	5,000	5,000
Travel			
Plot Fees			
Miscellaneous[1]	49,142		
Total	87,142	5,000	5,000

Footnotes: [1]Miscellaneous expenses category is genome sequencing for 3 apple varieties

Budget 2
Co PI 2: Dr. Stephen Ficklin
Organization Name: WSU Department of Horticulture
Contract Administrator: Anastasia Mondy
Telephone: 509.335.6885
Contract administrator email address: anastasia.mondy@wsu.edu

Item	**2019**	**2020**	**2021**
Salaries[1]		70,326	71,339
Benefits[1]		20,121	20,357
Wages[1]		1,245	1,295
Benefits			
RCA Room Rental			
Shipping			
Supplies			
Travel			
Plot Fees			
Miscellaneous[1]			
Total		91,692	92,991

Footnotes: [1]Salaries, wages, and benefits will support a fulltime postdoc for 2 years and will provide partial support to a graduate student in Co-PI Ficklin's lab

Budget 3: Co-PI Mattheis requested no budget

Objectives:

1. **Exceeded:** Sequence genomes to build variety-specific genomes for 'Honeycrisp,' 'WA 38' (Cosmic Crisp®), and 'Gala'

   NOTE: The 'Gala' genome was published by another group, so we diverted resources from the 'Gala' genome to the 'Granny Smith' genome.
2. **Exceeded:** Refine biomarker discovery pipeline using machine learning algorithms, comparative network analyses, and comparative genomics
3. **Complete:** Begin validation of biomarkers via PCR gene tests in multi-lot, multi-year surveys

Significant findings:

1. Assembled top quality apple genomes, posted to GDR for public access, published 'Honeycrisp'
2. Prototype biomarker models perform well
3. Insights into molecular response of 'Gala' apple fruit to CA - updated molecular model
4. Validation studies generally show expected results in other cultivars/orchards/years
5. Year 4 validation fruit samples obtained, ready for new project AP-22-101
6. New methods to quality check genomes enhance gene studies

Results and Discussion

New apple genomes *(Significant findings 1 & 6)*

The genomes for 'Honeycrisp,' 'WA 38,' and 'Granny Smith' are diploid assemblies, which means they are essentially ***two perfect*** apple genomes, containing the haplomes inherited from each parent - 1 each from pollen and ovule (see 'Honeycrisp' haplomes in Fig. 1). The field of genome assembly and annotation is rapidly evolving, and our team was particularly well positioned to leverage the state-of-the-art technology in genome sequencing. Improved genome resources will promote the identification of genes that drive important traits. In addition to leaving less data on the table during the analysis phase (as Honaas has previously reported, see reports for AP-19-103, PR-17-104), the exceptional quality of the genomes from this project opens new doors for genome scale analyses in apple. This effectively increases the number of genome features (e.g. genes, gene arrangements, etc.) we can use to build models that aim to predict maturity and future fruit quality. For example, we detected a structural rearrangement of a chromosome in 'Honeycrisp' that contains >100 genes; it was apparently inherited by 'WA 38' (Fig. 2). These kinds of structural changes can have massive impacts on genes that are in or near these regions, potentially explaining traits that are unique to a cultivar. Another example relates to the activity of alleles; each gene in the apple genome has two versions called alleles (one in each haplome). Until we built our genomes, this kind of allele-specific analysis was not possible for each project cultivar. Moreover, work in 'Gala' has shown that 1 in 5 genes shows allele activity differences during fruit development (Sun et al. 2020). This is important because if only one of two alleles plays a strong role in a trait (think "dominant" vs "recessive") we would not be able to detect this without our diploid genomes. Last, our sophisticated gene family analysis approach (recently published - (Zhang et al. 2022; Khan et al. 2022) has shown that there are potentially ~100 unique genes in the 'Honeycrisp' genome that are not detected in the other 6 apple genomes. We were able to detect this by carefully classifying all available apple genes into plant gene families (using our software PlantTribes2 - Wafula et al. In Press). These examples are important because unique gene and genome features might help explain unique cultivar traits. All of these examples illustrate new opportunities, as well as key resources that help us avoid pitfalls, as we search for important genes to monitor for risk assessment and maturity prediction tools.

Models for textural changes in 'Gala' apple fruit during storage *(Significant findings 2 & 4)*

There were two main experiments aimed at the development of technology for new risk assessment tools. The first was focused on textural changes in 'Gala' apple fruit during storage. In this experiment we stored 'Gala' apple fruit in various conditions (that include commercially relevant schemes) and tracked changes in fruit quality. There are a few main lessons from this work. The first is that while we can identify genes that are relatable to fruit texture changes, the wide range of possible storage conditions poses substantial challenges to biosignature development. This means that potential future tests may 1) need to be developed with a much larger training data set if biomarkers are to be deployed across ***all possible storage conditions***, or 2) need to be developed in a ***condition-limited manner***, such as for CA vs. air storage. Following the first experiments, we conducted a validation experiment where we stored 'Gala' fruit from a different orchard/year in similar conditions as the first experiment. We tested genes from our models using qPCR to see if the patterns were consistent across orchard/years. The results of this validation show that our top genes show very similar patterns of activity in fruit from a different orchard in a different year (75% agreement among all genes and storage conditions/treatments - Fig. 3).

Important also are practical considerations for biosignature tests beyond model performance, such as good signal to noise ratio (high vs. low levels of gene activity), lack of highly similar genes that can dilute or confound the signal (apple genomes are full of duplicate genes and large, complex gene families), and large scale changes through time (making tests more sensitive). We applied these and other criteria to select genes from the model, and also randomly selected a similar number of genes from the top genes in the model. Both of these subsets had similar performance, R^2 of 0.754 vs. 0.705 respectively. This indicates that applying additional criteria that are meant to enhance performance of PCR tests do not substantially reduce the predictive value of the model. This is important because we can choose genes that are likely to be easier to measure without sacrificing predictive value. All-in-all, while the models for textural changes seem to require many genes for maximum performance, it is reassuring that most of the genes we validated show consistent patterns across orchard/years.

Insights into fruit responses to low oxygen environments *(Significant finding 3)*

Another strategy to enhance postharvest fruit quality revolves around understanding how fruit respond to postharvest environments. This can offer clues about how fruit respond at a molecular level to, for example, 1-MCP, low temperature, and/or low oxygen (i.e. CA - controlled atmosphere). Our 'Gala' storage experiment provided excellent opportunities to examine how molecular models (that were elucidated over decades of work in model plants like rice, Arabidopsis, and others) operate in pome fruit species. In these fruit tree species a necessary first step is a careful classification of genes because the genes are not present in clean 1:1 ratios across plants - especially across distantly related plants like rice and apple. Our team leveraged our evolutionary expertise to classify all known apple genes into gene families, and then by looking at gene family trees identify apple genes that belong in molecular models from model plants. Doing this, we identified apples genes that respond in unexpected ways to environmental stresses in the postharvest period (Fig. 4), providing clues about the role of ethylene in losses of quality in long term fruit storage. We are in the process now of updating the molecular models for apple, and will continue to pursue this new line of inquiry towards optimized storage conditions for apple fruit. This is important because we might be able to identify windows of opportunity to apply certain types of crop protectants or plant growth regulators (or even new combinations thereof) that could be useful to maintain fruit quality in the postharvest period.

Prototype biomarkers - Next Generation Maturity Indices (NGMIs) *(Significant findings 2, 4, & 5)*

Our NGMI prototype models can use gene activity alone to predict the harvest date of project samples. That is, when we impose a contrast of maturity by picking fruit at intervals, we can then use gene activity data to look back and predict the harvest order, essentially recapitulating the harvest order. During the course of optimization we improved model performance substantially, with the tests approaching the performance level of the training data set (Fig. 5, panels A & B). Furthermore, we can generally order samples by harvest date using gene activity data from a relatively small number of genes, that is, model performance approaches maximum performance fairly quickly as genes are added (Fig. 6A). Additionally, we can see the strong positive effect of adding more data to the models (Fig. 6B), indicating that additional orchard/years of gene activity (i.e. sequencing) data will enhance model performance. In fact, a key feature of our prototypes is that they are updatable - as new data are added, we can update the gene targets that are the basis of potential future tools for risk assessment tools, like NGMIs. Therefore, our approach which differs in key ways from previous efforts (we use deep comparative and evolutionary genomics frameworks, for example), will benefit from USDA funded data that Honaas' group is adding to the models, plus data that the new AP-22-101A project will add from many orchards and cultivars. Overall, our results suggest that gene expression patterns are likely viable biosignatures for a new maturity index, and have possible utility within and across cultivars. Combined with mature RNA sample stabilization technologies, NGMI service models based on PCR are potentially possible.

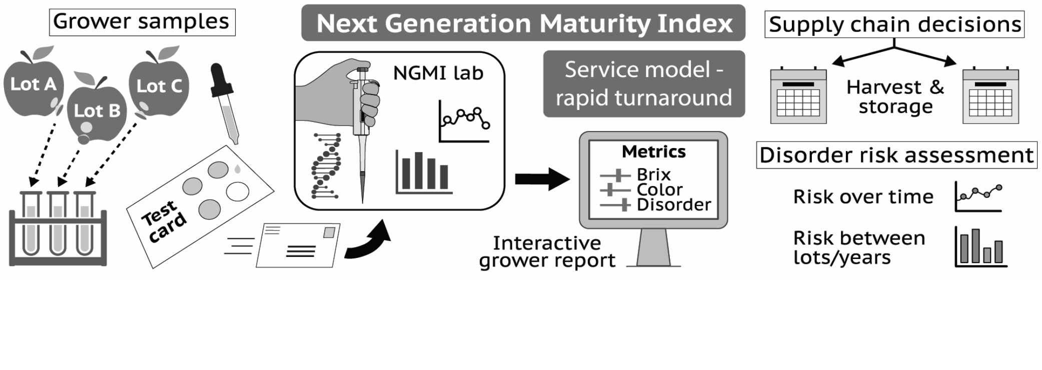
The next steps involve model optimization so we can understand how model performance changes with more data, plus other tweaks that are meant to account for multiple sources of noise. This is because we see clear examples of outliers where, for example, one year in one cultivar shows a divergent pattern - determining how to feed this information back into the model to improve the predictive power is a goal of Honaas' new project, AP-22-101A. The real-world outlook provided by our validation sample set suggests that we can make predictions in cultivars/years/orchards that were not part of model development. The patterns in our validation tests allow us to order fruit samples by pick date 92% of the time on average, but can vary from 70-100% depending on how the data are parsed (Fig. 7A vs. 7B). How to optimize the model to work with new data types and new cultivars remains to be explored - again, this is the goal of Honaas' new project AP-22-101A.

Long term outlook and industry impact

This project has established critical foundational resources and prototype biosignature workflows towards the development of commercially viable risk assessment tools. The compelling preliminary data that this project provides has helped elevate our SCRI proposal and has also helped coalesce a community of stakeholders and scientists around the possibilities of biosignatures for risk assessment in apple fruit. While our models can predict differences that we imposed in our experiments, the next steps involve model optimization to increase model performance. Our eventual aim is to differentiate ostensibly similar fruit before losses in quality occur - indeed our retrospective analysis of fruit quality in the project will show us which lots of fruit had differences in storage potential. Additionally, there are clear outliers in our models and validation tests: there are clear examples of a particular gene, year, orchard, or cultivar that do not always follow the model patterns. What this means is that more data and analyses are needed to understand the structure in the noise. For instance, we need more years of data to determine whether 1) the year or 2) the orchard location has a larger effect on a particular gene in a particular model. What is clear is that commercially viable NGMIs will likely require very sophisticated models based on multiple cultivars (or even species), multiple years, and multiple genes to make reliable predictions.

NGMI concept model

We envision a service based NGMI model based on our prototype biomarkers. The technology for tests in the field is mature and has been deployed commercially. We will use similar methods that include stabilization of fruit extracts on cards and gene measurements based on PCR.



Figures and Tables

Figure 1. The diploid 'Honeycrisp' genome shows high overall structural similarity with the 'Gala' apple genome (Sun et al. 2020). Ribbon plot showing high structural similarity, chromosome by chromosome, of the diploid 'Honeycrisp' genome assembly. A diploid assembly contains two apple genomes, each one called a haplome (abbreviated HAP below). This allows us to study both copies of every apple gene, which substantially increases the number of potential targets for biomarker model development and opens new doors for genetic analysis in apple that will shed light on important fruit traits.

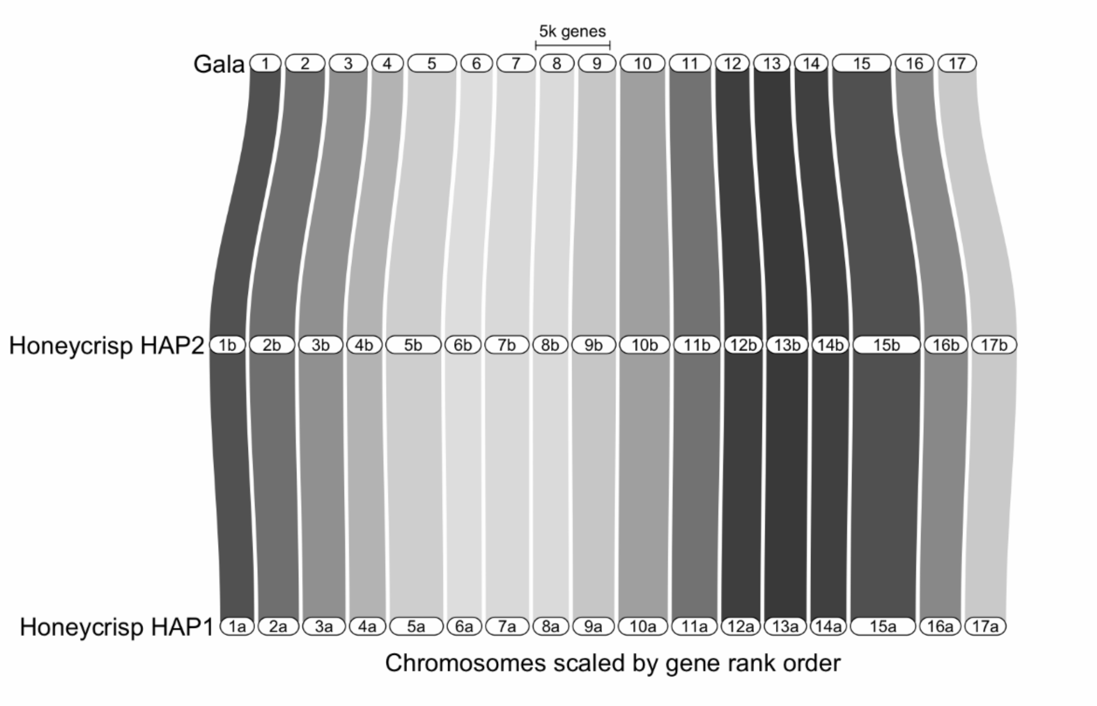


Figure 2. A structural difference in the 'Honeycrisp' genome was inherited by 'WA 38.' Synteny cartoon showing a chromosome inversion that contains ~120 genes (enlarged for detail). This inversion is only in 1 haplome of 'Honeycrisp,' and is therefore only in 1 haplome of 'WA 38.' We do not yet know the impact of this particular change, but it has been well documented that such changes can have large effects on gene activity in or near the inversion, and also on gene structure for genes at the boundaries of the inversion. Changes like this could potentially explain cultivar traits, but also represent potential pitfalls because this inversion is thus far only seen in 'Honeycrisp' and 'WA 38.' Genomes from this project are the first to have fully-phased, perfect, diploid assemblies for apple. This offers new glimpses into haplome structure variation in important apple cultivars.

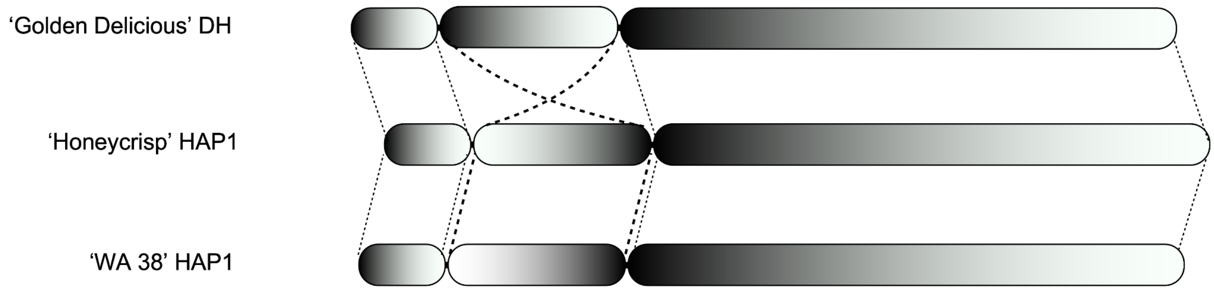


Figure 3. Models for prediction of fruit texture show consistent patterns across orchard/years. When we repeated the ‘Gala’ storage experiment (different year, orchard, and gene activity measurement methodology) the patterns were largely consistent for the example gene below. There were apparent differences in fruit maturity (estimated based on physiological indices; color, starch, texture), which could explain some differences between “Harvest” and “T1” timepoints in Conditions 4 & 5 between each experiment. Validation studies like these suggest our approach may eventually yield robust biomarkers, but at this early stage in development they primarily provide valuable information for model improvement.

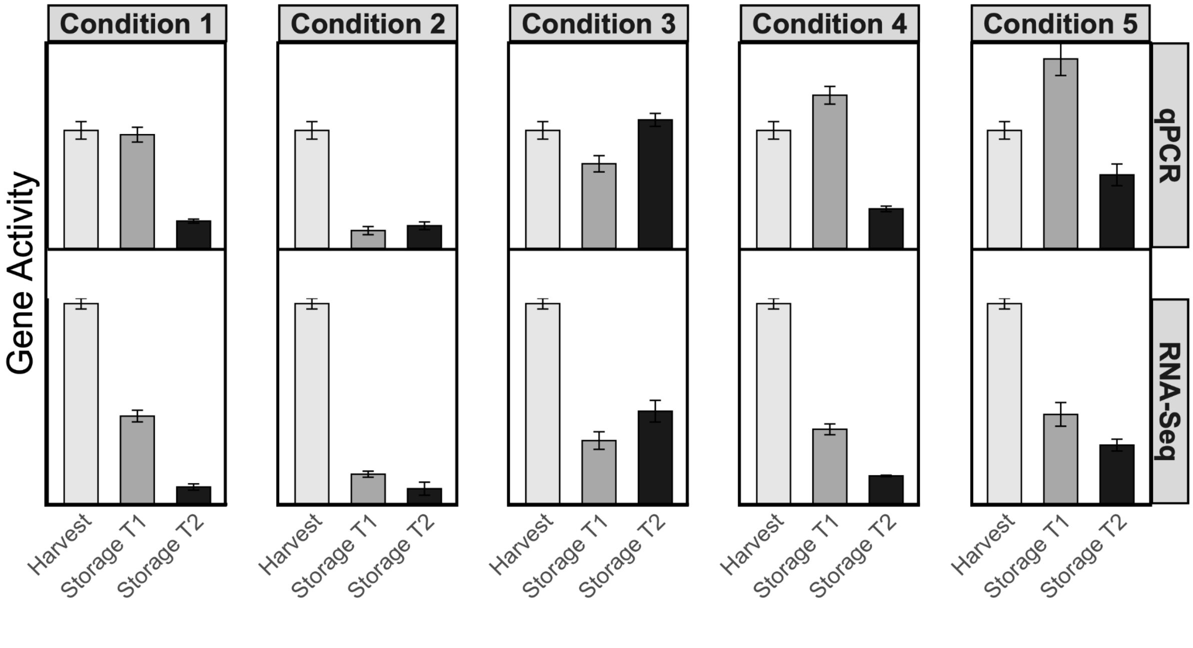


Figure 4. For two "twin" apple genes, a molecular schematic for how plants respond to low O_2 predicts activity for one gene, but not the other. A molecular signaling schematic that was developed in model plants (including Arabidopsis and rice) describes a molecular signaling mechanism that is activated by low oxygen, and may use ethylene as a signal molecule. We used gene family information to find the apple genes that correspond to genes in the published signaling schematic. Because of the complex history of the apple lineage, virtually all genes in apple are present in ratios other than 1:1 to other plant genes, like the ones below that are present in a 2:1 ratio. The experimental treatments that included controlled atmosphere (CA, i.e. low O_2) are shown as ***filled squares***, and all treatments in normal air are shown as ***open circles***. One of the "twin" genes was activated in fruit by low O_2, as expected for CA storage, with the fruit in normal air showing very low activity - **panel A**. However, the other "twin" (**in panel B**) showed an unexpected pattern that included activation in CA, but also in fruit that were stored in normal O_2 levels. This could represent activation of the plant stress pathway for low O_2, or perhaps another role for genes in the model that relates to ripening, rather than just low O_2. Insights like these may represent new opportunities to mitigate negative outcomes that are not well controlled, or are even exacerbated, by long term CA storage.

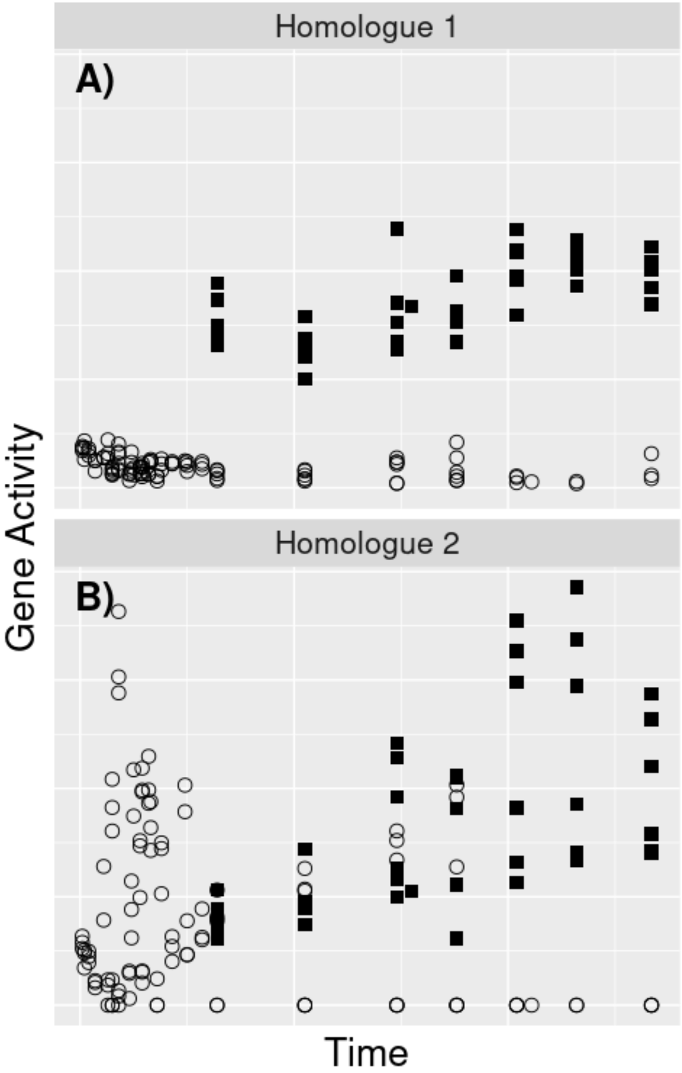


Figure 5. During our project, prototype NGMI model performance has increased. The performance of the NGMI prototypes is generally good enough for us to order picks by date, or even estimate harvest week, in our experiment using only gene activity data. The general scheme is to ***train*** models with a majority of the data, and then ***test*** the models with a portion of data that was set aside - this provides "new" data the model has not seen and allows us to estimate how the model will perform when we carry out real-world tests. Model performance is gauged by linear regression of actual vs. predicted pick date; R^2=1 would indicate a perfect set of predictions. The models performed more-or-less consistently during training (in the left column - **A, C, E**). Our optimizations improved the performance of the test cases starting at R^2=0.786 and increasing to R^2=0.946 (right column - **B, D, F**). The **test** data approached **training** data in model performance (see **B** vs. **A** for our latest model tests). This indicates that our models might be useful as NGMIs with sufficient development.

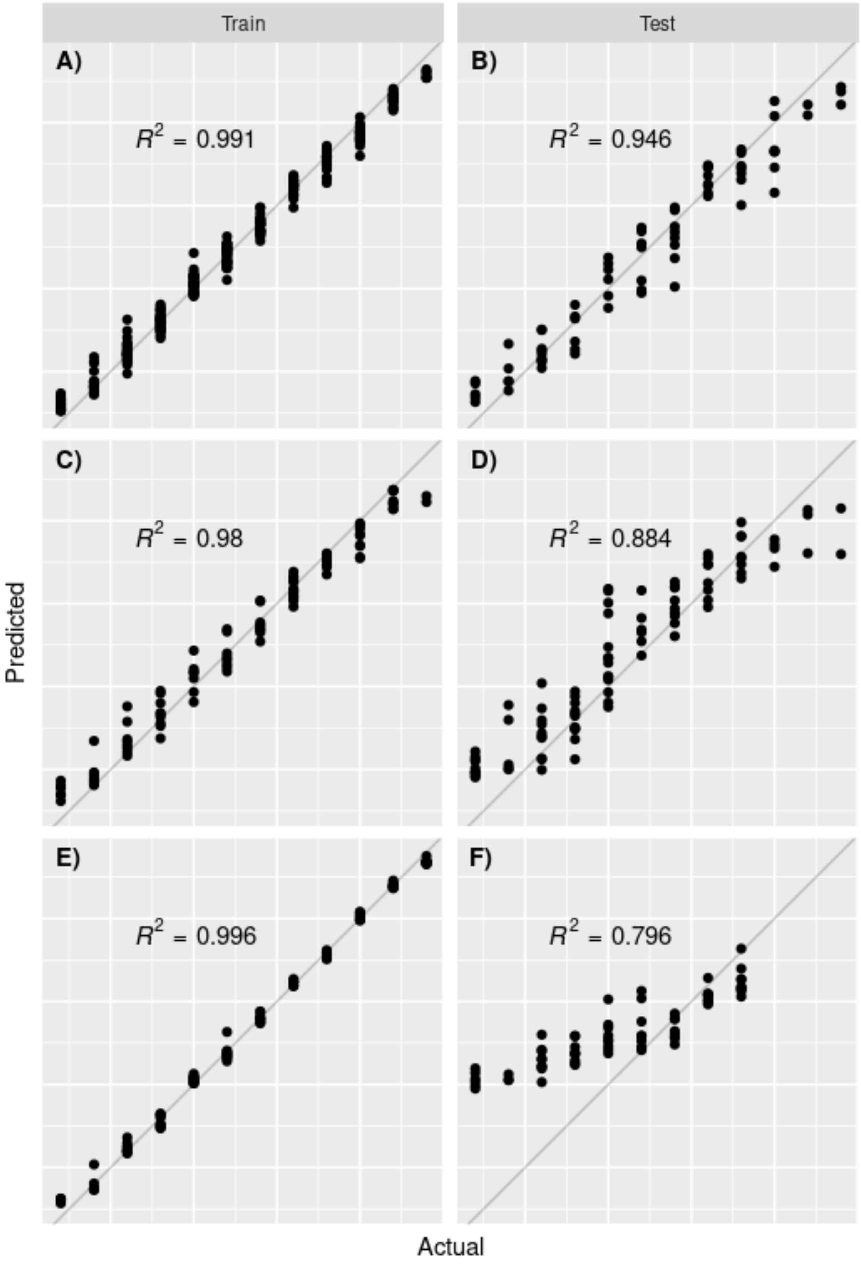


Figure 6. NGMI model performance rapidly increases as more genes and data are included in the model. We tested model performance with varying numbers of genes to explore where the NGMI prototype had changes in performance. The Y-axis shows model performance (i.e. R^2). We used 3 different amounts of training data (dotted, dashed and solid lines in **Panel A - Train** to explore performance as a function of the gene number we use to make predictions. We discovered that for the model, there was a sharp inflection point where model performance stabilized, and that more data increased model performance (indicated by higher stabilization points). Importantly, we saw a similar pattern in the "test" scenarios (**Panel A - Test**), indicating that the model was robust when feeding in new data, and seems to perform well with a tractable number of genes; both important considerations for a future test that is commercially viable. **Panel B** shows a zoomed in view to show detailed differences between **Train** and **Test** in **Panel A**. Note the y-axis scale differences, 0-1.00 in **A**, vs 0.80 - 1.00 in **B**.

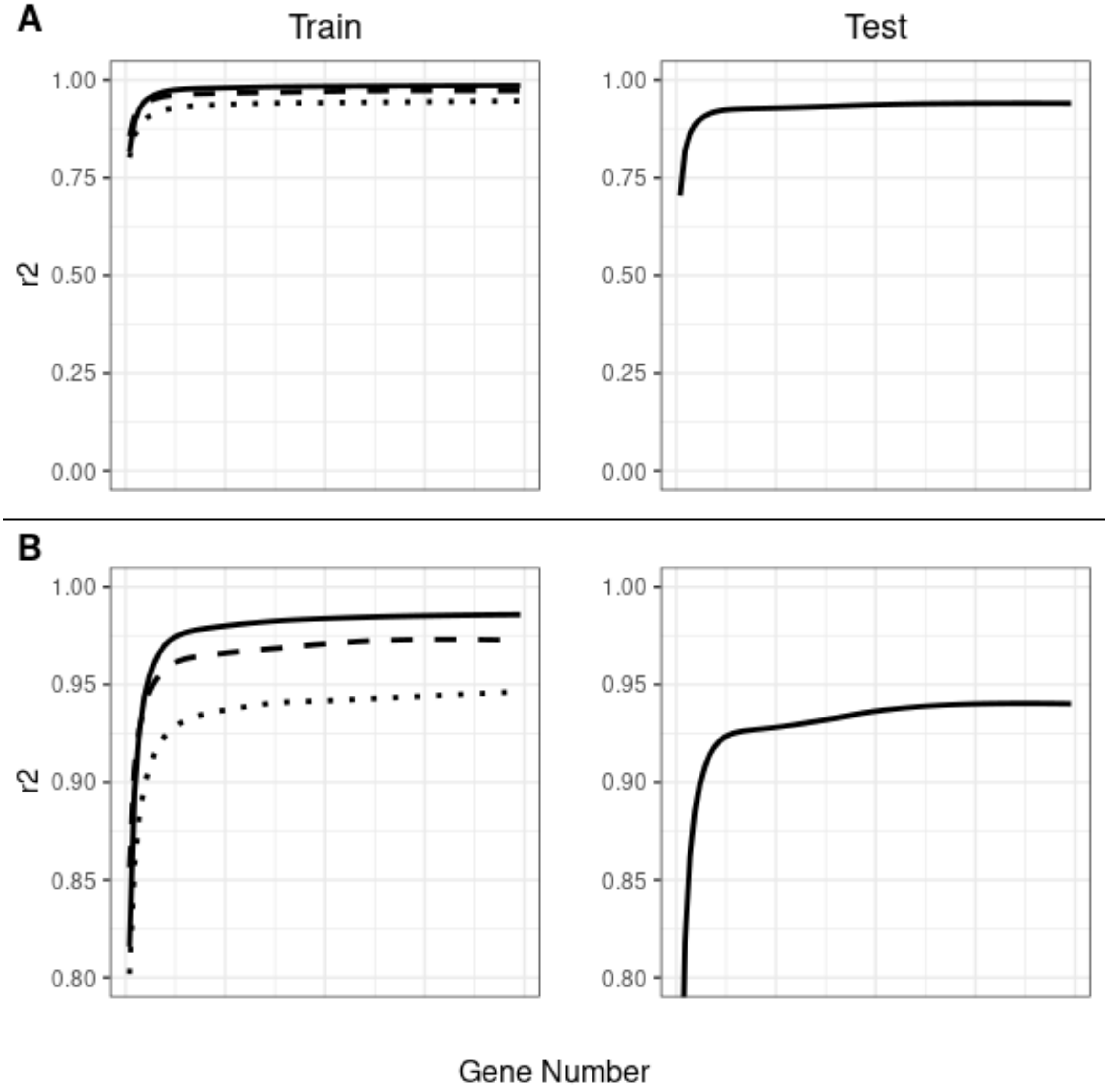


Figure 7. Real world analysis of the NGMI prototypes shows potential for use in many cultivars. When this project was started, we began building a validation sample catalog that now contains hundreds of cryopreserved RNA samples from many cultivar/orchard/years. After we built NGMI models, we tested activity from select fruit samples in our validation catalog to determine if we could recapitulate harvest order with NGMI prototypes. By-and-large, we can predict harvest order >92% of the time on average (panel A), though there are clear examples of outliers (panel B, orchard C). We are exploring how to score validation tests - our initial criteria are shown in Panel C. How these outlier cultivars, orchards, years, and genes can be used to enhance model performance, as well as how different types of data can be integrated into the models, are among the goals of our new WTFRC project AP-22-101A.

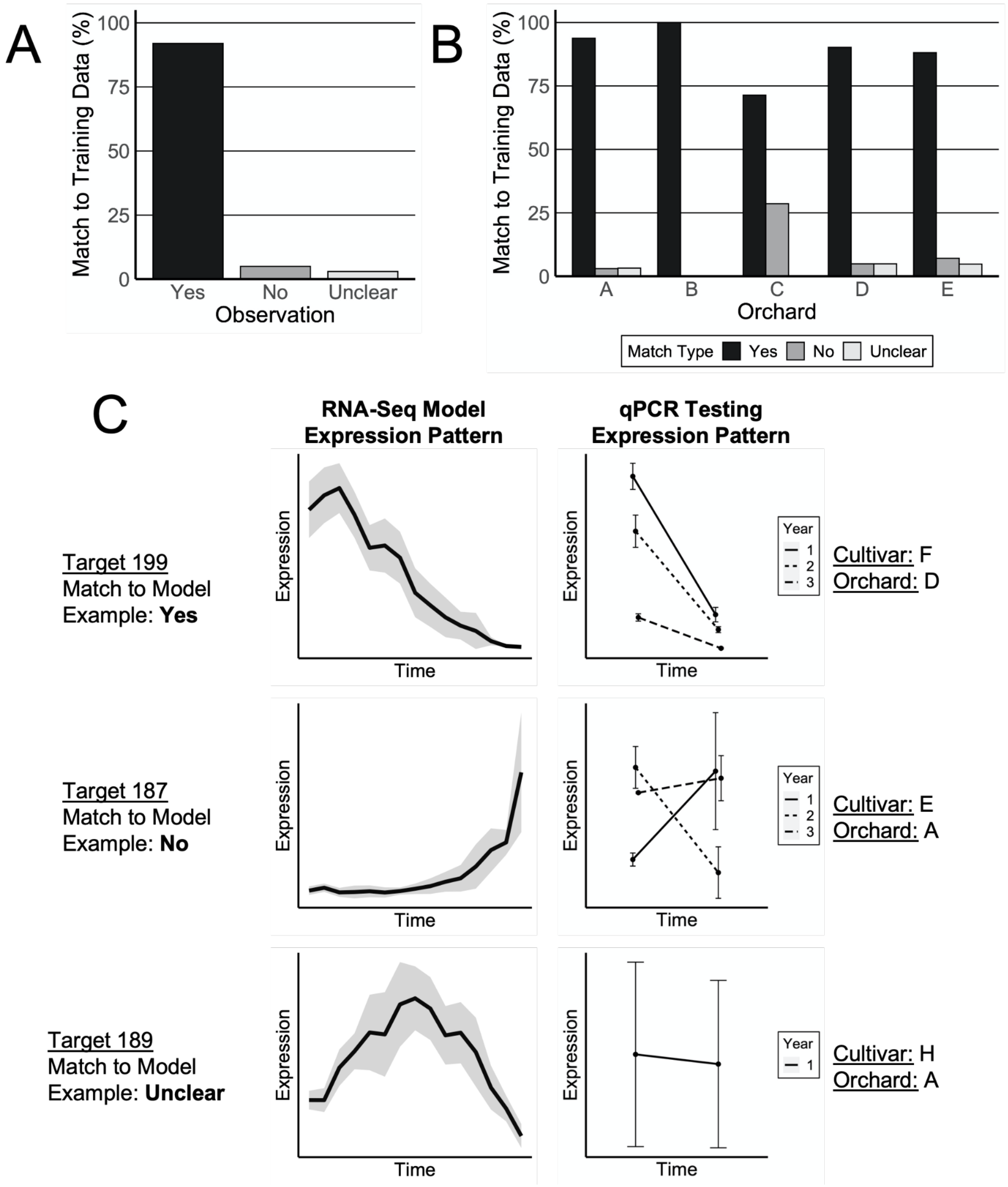


Table 1. Our high-quality apple genomes represent the state-of-the-art in plant genomes, and are likely the best apple genomes to date. Below are genome statistics for project cultivars, and another high-quality apple genome, ‘Gala.’ There are two general categories of genome statistics, ***Assembly*** and ***Annotation.*** Here, a map analogy is appropriate - think of the ***assembly*** as a satellite picture - it’s the raw data for the landscape. Then, for the ***annotation***, imagine adding layers of information that include the boundaries and names of map features, like roads, buildings, rivers, and parks. Reported in millions of base pairs (A, T, G, Cs - abbreviated Mbp) the overall length of our genomes are consistent, yet our assemblies are in larger pieces (larger N50). BUSCOs are a widely used genome benchmark that stands for Benchmarking Universal Single Copy Orthologs; ***translation*** - these are roughly 2,300 plant genes we expect to see in a plant genome, so they are a good watermark for genome quality. Our finished ‘Honeycrisp’ and ‘WA 38’ genomes have the highest BUSCO scores of any apple genome. Our genome ***assembly*** is not much better than ‘Gala,’ but our genome markup strategy that creates the ***annotation*** is philosophically different, and allows us to identify more apple genes than other teams.

Assembly	Length Mbp	N50 Mbp	Assembly BUSCO %	Annotation BUSCO* %
‘Gala’ (Sun et al. 2020) Note: averages shown	673	14	98.4	95.0
‘Honeycrisp’ Haplotype 1	674	33	98.6	96.8
‘Honeycrisp’ Haplotype 2	660	33	98.7	97.4
‘WA 38’ Haplotype 1	678	36	98.7	95.9
‘WA 38’ Haplotype 2	667	37	98.7	97.4
‘Granny Smith’ Haplotype 1	666	38	98.9	90.8
‘Granny Smith’ Haplotype 2	665	37	99.0	90.3

*The annotation process is very labor intensive and requires many iterations that use additional kinds of evidence. The finished annotation results for ‘WA 38’ and ‘Granny Smith’ are expected to meet or exceed those in the ‘Honeycrisp’ genome. In fact, the topic *evaluation of genome quality* is an active area of research - Honaas’ team recently published work that describes novel methods that can be used to evaluate and improve genome resources that have relatively poor assemblies and/or annotations (Zhang et al. 2022, Wafula et al. in press).

References cited

Sun, X., Jiao, C., Schwaninger, H., Chao, C.T., Ma, Y., Duan, N., Khan, A., Ban, S., Xu, K., Cheng, L. and Zhong, G.Y., 2020. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nature genetics*, *52*(12), Pp.1423-1432.

Khan, A., Carey, S.B., Serrano, A, Zhang, H., Hargarten, H., Hale, H., Harkess, A., and Honaas, L.A. A phased, chromosome-scale genome of ‘Honeycrisp’ apple (*Malus domestica*). *Gigabyte*, 2022 https://doi.org/10.46471/gigabyte.69

Zhang, H., Wafula, E.K., Eilers, J., Harkess, A.E., Ralph, P.E., Timilsena, P.R., DePamphilis, C.W., Waite, J.M. and Honaas, L.A., 2022. Building a foundation for gene family analysis in Rosaceae genomes with a novel workflow: A case study in Pyrus architecture genes. *Frontiers in Plant Science*, *13*.

Wafula, E.K., Zhang, H., Von Kuster, G., Leebens-Mack, J.H., Honaas, L.A., and dePamphilis, C.W. PlantTribes2: tools for comparative gene family analysis in plant genomics. *Frontiers in Plant Science*, In Press.

Executive Summary

Project title: Apple genomes for postharvest fruit quality biomarkers

Key words: machine learning, fruit maturity, fruit firmness, RNA-Seq

Abstract: New tools and technologies are needed to help sustain the viability of the tree fruit industry. A key area to innovate is enhancement of supply-chain decision making. By making more informed decisions, losses of fruit quality in the postharvest period could be reduced. Towards this goal, this project developed foundational resources, methods, and datasets that have been used to build prototype biomarker models, or more accurately ***biosignatures*** because multiple targets are required for reliable predictions. We focused on two areas that relate to postharvest fruit quality: at-harvest apple maturity and fruit textural changes during storage. We found that massive datasets (billions of measurements) can be leveraged with state-of-the-art computational methods to build models that are predictive of these two traits. Importantly validation experiments suggest that the models may work beyond the scope of the experiment, and reliable prototype models consist of a tractable number of gene targets.

Objectives:

1. **Exceeded:** Sequence genomes to build variety-specific genomes for 'Honeycrisp,' 'WA 38' (Cosmic Crisp®), and 'Gala'

   NOTE: The 'Gala' genome was published by another group, so we diverted resources from the 'Gala' genome to the 'Granny Smith' genome.
2. **Exceeded:** Refine biomarker discovery pipeline using machine learning algorithms, comparative network analyses, and comparative genomics
3. **Complete:** Begin validation of biomarkers via PCR gene tests in multi-lot, multi-year surveys

Significant findings:

1. Assembled top quality apple genomes, posted to GDR for public access, published 'Honeycrisp'
2. Prototype biomarker models perform well
3. Insights into molecular response of 'Gala' apple fruit to CA - updated molecular model
4. Validation studies generally show expected results in other cultivars/orchards/years
5. Year 4 validation fruit samples obtained, ready for new project AP-22-101
6. New methods to quality check genomes enhance gene studies

Future directions: Some of the next steps are outlined in Honaas' new project AP-22-101A. Briefly, we aim to refine and enhance the biosignature models by exploring new modeling techniques, adding new data, and testing ways to integrate model performance back into model development. The preliminary data that this project developed is used in a proposal to NIFA's Specialty Crops Research Initiative.