LEVERAGING ANGIOSPERM PANGENOMICS TO UNDERSTAND GENOME EVOLUTION

By

Alan E. Yocca

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Biology–Doctor of Philosophy

2022

ABSTRACT

LEVERAGING ANGIOSPERM PANGENOMICS TO UNDERSTAND GENOME EVOLUTION

By

Alan E. Yocca

My dissertation work focused on species-level comparative genomics and pangenomics to describe patterns of genetic variation. I studied multiple systems and unsurprisingly discovered different patterns of variation. Within a species, individuals are genetically diverse. There are some DNA regions present in every individual (core), while others may be specific to a single individual or lineage (variable). The sum of the genetic sequences found across an entire taxonomic group is called the pangenome. This DNA variation greatly contributes to observed phenotypic differences between individuals. Therefore, to understand genome evolution and the link between genotype and phenotype, we must understand the pangenome. In this work, I compare the core and variable genetic regions both coding and noncoding across different flowering plant lineages. I note many consistent features across lineages as well as ways in which each pangenomic pattern is unique. These consistencies and differences can be leveraged in the future to better understand genome evolution as well as how genotype relates to phenotype. Specifically, my dissertation includes four chapters; (1) Evolution of Conserved Noncoding Sequences in *Arabidopsis*

thaliana, (2) Machine learning identifies differences between core and variable genes in *Brachypodium distachyon* and *Oryza sativa*, (3) Current status and future perspectives on the evolution of cis-regulatory elements in plants, and (4) A pangenome for *Vaccinium*.

ACKNOWLEDGEMENTS

My mother Tracy deserves credit for raising five children as a single mother and enabling me to chase any opportunity throughout childhood through higher education. I love you mama. Thank you to my grandmother Dr. Merilyn Davis for supporting me always throughout my education. Thank you to my family for your consistent love. I also want to thank my many friend circles, the boys, the HRs, the council, and fellow MSU graduate students. You make my life fun no matter the circumstance.

I would never have accomplished anything academically without enduring support from mentors. Thank you Claude dePamphilis for allowing me as a teenager with zero experience into your research group. Thank you Huiting Zhang for your trust and patience with me to help with your PhD research. Thank you Robin Buell for being open to me coming to Michigan State and for introducing me to bioinformatics for which the rest of my career will likely be based. Thank you to my committee members Emily Josephs, Robert VanBuren, and Jiming Jiang for your advice and wisdom. Thank you to Patrick Edger. You are the greatest advisor I could have imagined. Your belief and support for all your students is unparalleled. Your support and understanding for me over the past five years, and through two years of a

global pandemic have saved my PhD aspirations and possibly also my marriage. Words cannot fully express my gratitude.

Finally, my wife and partner of more than a decade, Jess (CS). Thank you for following me to dreary Michigan without hesitation. I’m proud of what you were able to accomplish while needing to support me as a graduate student and am excited for the next chapter of our lives.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Current status and future perspectives on the evolution of cis-regulatory elements in plants

The work presented in this chapter is part of the final publication: Yocca AE and Edger PP. 2022. Current status and future perspectives on the evolution of cis-regulatory elements in plants. Curr Opin Plant Biol 65: 102139.

Abstract

Cis-regulatory elements (CREs) are short stretches (~5-15 base pairs) of DNA capable of being bound by a transcription factor and influencing the expression of nearby genes. These regions are of great interest to anyone studying the relationship between phenotype and genotype as these sequences often dictate genes' spatio-temporal expression. Indeed, several associative signals between genotype and phenotype are known to lie outside of protein-coding regions. Therefore, a key to understand evolutionary biology requires their characterization in current and future genome assemblies. In this review, we cover some recent examples of how CRE variation contributes to phenotypic evolution, discuss evidence for the selective pressures experienced by non-coding regions of the genome, and consider several studies on accessible chromatin regions in plants and what they can tell us about CREs. Finally, we discuss how current advances in sequencing technologies will improve our knowledge of CRE variation.

Full text of this work:

https://doi.org/10.1016/j.pbi.2021.102139

CHAPTER 2

Evolution of Conserved Noncoding Sequences in *Arabidopsis thaliana*

The work presented in this chapter is part of the final publication: Yocca AE, Lu Z, Schmitz RJ, Freeling M, and Edger PP. 2021. Evolution of Conserved Noncoding Sequences in *Arabidopsis thaliana*. Molecular Biology and Evolution, 38;7:2692–2703.

Abstract

Recent pangenome studies have revealed a large fraction of the gene content within a species exhibits presence-absence variation (PAV). However, coding regions alone provide an incomplete assessment of functional genomic sequence variation at the species level. Little to no attention has been paid to noncoding regulatory regions in pangenome studies, though these sequences directly modulate gene expression and phenotype. To uncover regulatory genetic variation, we generated chromosome-scale genome assemblies for thirty *Arabidopsis thaliana* accessions from multiple distinct habitats and characterized species level variation in Conserved Noncoding Sequences (CNS). Our analyses uncovered not only PAV and positional variation (PosV) but that diversity in CNS is nonrandom, with variants shared across different accessions. Using evolutionary analyses and chromatin accessibility data, we provide further evidence supporting roles for conserved and variable CNS in gene regulation. Additionally, our data suggests that transposable elements contribute to CNS variation. Characterizing species-level diversity in all functional genomic sequences may later uncover previously unknown mechanistic links between genotype and phenotype.

Full text of this work: https://doi.org/10.1093/molbev/msab042

CHAPTER 3

Machine learning approaches to identify core and dispensable genes in pangenomes

The work presented in this chapter is part of the final publication: Yocca AE and Edger PP. 2022. Machine learning approaches to identify core and dispensable genes in pangenomes, Volume 15, Issue 1.

Abstract

A gene in a given taxonomic group is either present in every individual (core) or absent in at least a single individual (dispensable). Previous pangenomic studies have identified certain functional differences between core and dispensable genes. However, identifying if a gene belongs to the core or dispensable portion of the genome requires the construction of a pangenome, which involves sequencing the genomes of many individuals. Here we aim to leverage the previously characterized core and dispensable gene content for two grass species [*Brachypodium distachyon* (L.) P. Beauv. and *Oryza sativa* L.] to construct a machine learning model capable of accurately classifying genes as core or dispensable using only a single annotated reference genome. Such a model may mitigate the need for pangenome construction, an expensive hurdle especially in orphan crops, which often lack the adequate genomic resources.

Full text of this work: https://doi.org/10.1002/tpg2.20135

CHAPTER 4

A pangenome for *Vaccinium macrocarpon* (cranberry)

Abstract

Vaccinium macrocarpon (Aiton; cranberry) is a native crop to North America and has a relatively short domestication history of less than 200 years. Therefore, characterization of the cranberry breeding gene pool promises to accelerate breeding efforts. Here, we characterize that gene pool in the context of a pangenome. Individuals in a population are genetically diverse. A single reference genome is not sufficient to capture every gene segregating in the population. The sum of all this genetic diversity is termed the pangenome. We find the pangenome of cranberry shows patterns consistent with earlier studies in plants. Furthermore, the pangenome uncovered tens of thousands of novel genes previously undiscovered in the reference genome that may be leveraged in the future for cranberry breeding.

Introduction

Vaccinium macrocarpon (Aiton; cranberry) is a member of the Heath family (Ericaceae) which contains blueberry. It is a diploid native to North America and has been cultivated since at least 1810 (Hancock et al., n.d.). Cranberry is a high value crop, and since its domestication history is much shorter than other crops, there is unexplored breeding potential for this species.

Within a species, individuals are genetically diverse. Sequencing a single reference genotype is insufficient to recover all genetic diversity in a group (Golicz et al. 2020). This was recognized in microbial studies; sets of genes are either found in every member of a population (core) or absent in at least a single individual (dispensable). We choose to refer to dispensable genes as variable genes. Though absent in some individuals, variable genes if lost in combination can be lethal due to either redundancy or epistatic interactions (Marroni, Pinosio, and Morgante 2014). The sum of all core and variable genetic components was termed a pangenome first by Tettelin in 2005 (Tettelin et al. 2005). Since then, several pangenome studies have been conducted in plants such as *Brachypodium distachyon*, *Brassica napus*, soybean, rice, and strawberry (Gordon et al. 2017; Hurgobin et al. 2018; Li et al. 2014; Wang et al. 2018; Qiao et al. 2021). These studies and others produced many consistent and some inconsistent results.

Consistently, core genes are enriched for "housekeeping" functions, or essential metabolic processes such as glycolysis. Variable genes are often enriched for conditional functions. In *Brassica oleracea*, variable genes are strongly enriched for defense response (Golicz et al. 2016). In *Brachypodium distachyon*, variable genes are enriched for development and defense (Gordon et al. 2017). Variable genes often display

signatures of elevated sequence turnover and relaxed selection and are shorter relative to core genes (Yocca and Edger 2022).

Inconsistent findings are most often the proportion of core and variable genes with the focal taxon. This value ranges from 33% to 80%. For example, about 80% of genes in *Oryza sativa* are core, while in Maize, about 60% are variable (Q. Zhao et al. 2018; Hirsch et al. 2014). For a table collecting the proportion of core genes across several plant pangenome studies see (Golicz et al. 2020). The specific factors controlling the proportion of variable genes are unknown, but life history and representative divergence likely play a role (Tao et al. 2019). Lei et al. propose rates of structural variation are key to pangenome size (Lei et al. 2021). For *Vaccinium macrocarpon*, we expected a relatively high proportion of core genes within sub groups due to limited representative divergence time.

Characterization of variable genes is crucial to maximize the value of a breeding program. Genes underlying important traits are often variable. Structural variants can be strongly associated with phenotypes (Tao et al. 2019; Yao et al. 2015; Zhou et al. 2015). GWAS studies uncovered several more candidate loci controlling traits of interest when using a pangenomic framework. Specifically in pigeon pea, GWAS uncovered a gene associated with seed weight that was absent in the reference genome (J. Zhao et al. 2020). Also, Song et al leveraged PAV

information from eight reference quality *Brassica* ecotypes to perform a GWAS and identify transposable element insertions associated with flowering time a silique weight (Song et al. 2020). This illustrates the translational impact of extending analyses beyond reference genome frameworks.

In this chapter, we assemble ten novel genome sequences for ten diverse cranberry genotypes. In conjunction with a previously published reference, we develop a pangenome for this species. We explore the differences between core and variable genes. This resource can be leveraged in the future to improve cranberry breeding programs.

Results

Selection of accessions, sequencing, assembly, and annotation

We selected eleven *Vaccinium macrocarpon* (Aiton; cranberry) genotypes to construct a pangenome for this species. The reference genotype, Stevens, was published previously (Diaz-Garcia et al. 2021). Ten additional accessions were selected based on genetic marker analysis to capture the greatest amount of genetic diversity (data not shown). Accessions were sequenced to an average depth of 110X (Table S1). Genomes for each accession were assembled using a hybrid *de novo* and reference based method. RNA sequencing data was also generated for leaf and berry tissue for the ten non-reference accessions. Our ten non-reference accessions were annotated using MAKER. The ten

non-reference accessions were annotated with an average of 27,856 genes. This is more than the 23,532 in the reference accession. This slight annotation difference should not affect our results as we analyze mostly accession-wide patterns of core and variable genes.

We aligned the eleven cranberry genomes using Progressive Cactus (Armstrong et al. 2020). This alignment was used in conjunction with Orthofinder2 to identify core and variable genes in the *V. macrocarpon* species (Emms and Kelly 2019). Orthofinder2 will often identify members of the same gene family as orthologous. We wanted stricter criteria for core gene identification to allow direct comparisons of orthologs between accessions to study species-level divergence rather than gene family divergence at deeper timescales. Therefore, we filtered out of the Orthofinder2 results and match lacking a proximate alignment from our Progressive Cactus results.

As annotations and genomes were available for all eleven cranberry genotypes, we could label every gene (302,090 total) as core or variable. We find an average of 14,553 core genes and

12,910 variable genes per accession for an average of 53% core genes per accession (Figure 1).

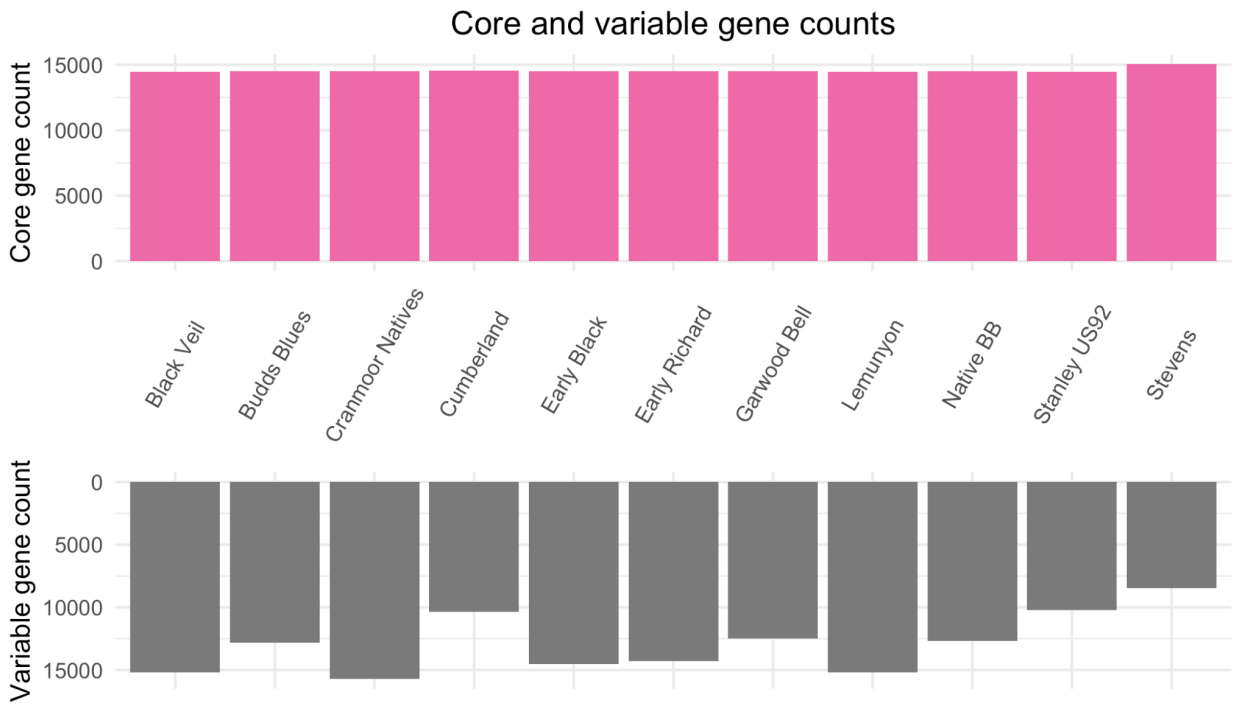

Figure 1 Core and variable gene counts across each cranberry genotype
A mirrored bar plot showing the number of core genes (above; pink) and the number of variable genes (below; grey) for each of eleven cranberry genotypes.

Pangenome modeling

We wanted to model the size of the *V. macrocarpon* core and variable pangenome. The only way to capture all species-level genetic diversity is to sequence every individual. As our collection is only a sample, we need to model the true size of the pangenome. Figure 2 shows our model of the core and variable pangenome. We see overall, there are an average of 14,552 core genes out of an average of 27,462 total genes per accession

(53%). As more genomes are considered, the number of total variable genes increases. We see the amount of variable genes added per accession added decreases. Therefore the cranberry pangenome is considered “closed” where the total number of variable genes will eventually plateau. As our subsampling has not completely plateaued, we believe future sampling of cranberry genomes will uncover greater genetic variation and more novel variable genes.

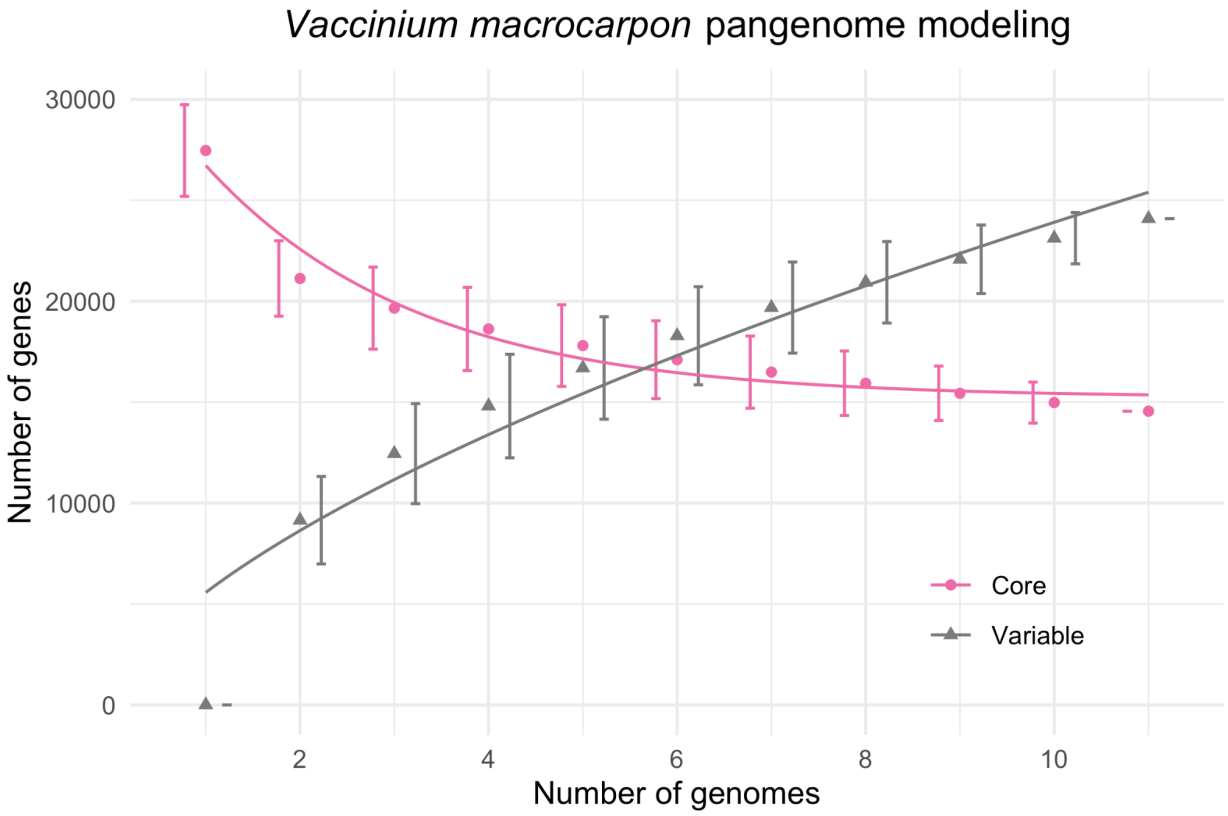

Figure 2 Core and variable genome modeling
A line graph showing a model for the core (pink; circle) and variable (grey; triangle) pangenome. For each point along the x-axis, we take every possible combination of that size from our eleven genome sample and plot the average number of core and variable genes as a point. The error bars represent plus or minus one standard deviation of these counts from the mean. Trend lines were estimated with equations listed in Methods.

What are the differences between variable and core genes

Pangenome studies find consistent trends in the differences between core and variable genes. We uncover similar differences. Figure 3 displays distributional differences between core and variable genes across *V. macrocarpon*. We see variable genes are dramatically shorter, and have both fewer and shorter introns than core genes. We calculated gene expression values from leaf

and berry tissue for our ten non-reference genotypes. Expression of core genes was higher on average across both tissues for all accessions (Table S2). We functionally annotated each gene model using InterPro Scan. This allowed us to compare differences between core and variable genes. There were so many enriched GO terms for both core and variable genes, it is hard to draw specific conclusions. For variable genes in the reference genotype Stevens, we did observe a few expected enriched GO terms such as defense response (GO:0006952) and regulation of response to stress (GO:0080134). However translation (GO:0006412) and photosynthetic electron transport in photosystem II (GO:0009772) was also enriched. For core genes in the reference genotype, core genes were enriched for core biological processes such as cell cycle (GO:0007049) and chromosome organization (GO:0051276). Full GO enrichment results are presented in Table S3 and S4.

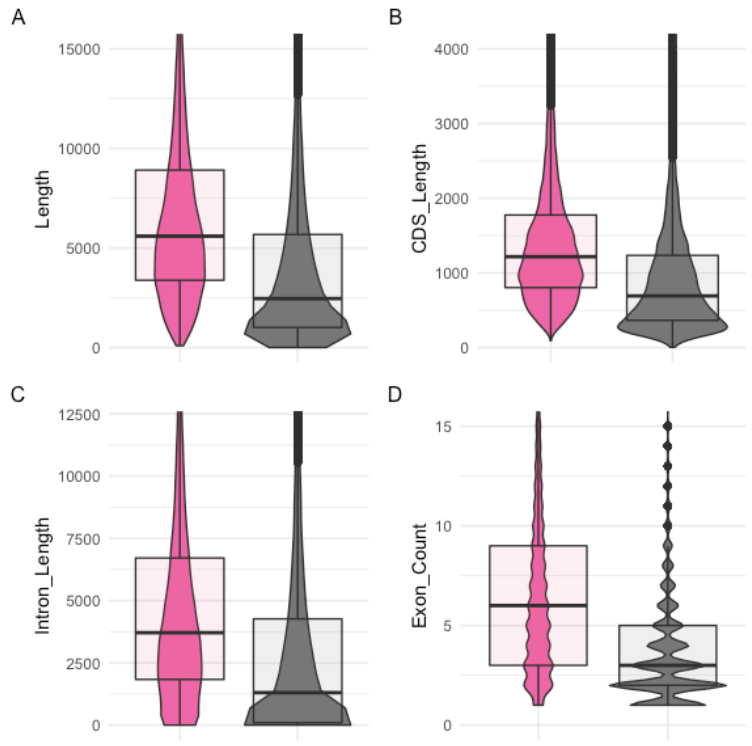

Figure 3 Differences between core and variable genes
Density plots showing the differences between core (pink) and variable (grey) genes. We show (A) gene length (transcription start site to transcription end site), (B) CDS length, (C) intron length, and (D) exon count. This data includes all genes across all accessions.

Discussion

In this study, we define genes in cranberry as core or variable based on their presence or absence across eleven genotypes. We uncovered 53% of all cranberry genes are core, and 47% are variable. Across plant pangenome studies, the proportion of core genes varies (Emms and Kelly 2019; Tao et al. 2019). Two main factors affect the proportion of core genes: divergence time of the genotypes compared, and life history. Cranberry is an outcrossing species which may explain the large proportion of

variable genes despite the short divergence time between our accessions. These estimates still lie within the range of previous observations.

There are characteristic differences between core and variable genes as observed previously. These differences lend insight into core/variable gene function as well as evolution and origin. Both core and variable gene characteristic distributions overlap. Therefore, dichotomies cannot be drawn between these two classes. Rather, as with most biological processes, they exist on a continuum. That being said, our characteristic differences tell us variable genes show patterns of evolutionarily young genes. There are multiple models of gene birth. Shorter sequences for novel genes support two possible models of gene birth: (1) *de novo* emergence, or (2) duplication degeneration.

Determining a specific mechanism of gene birth is beyond the scope of this work. However reports of *de novo* gene origin in yeast and *Drosophila* show evidence of novel genes arising from previously short noncoding DNA sequences (Carvunis et al. 2012; Siepel 2009). Gene duplication presents an abundance of substrate for evolution to shape novel genes (Ohno 1970). One mechanism through which gene duplication leads to novel gene function is through neofunctionalization or subfunctionalization (Lynch and Force 2000; Birchler and Yang 2022). After a gene

duplicates, one copy can explore the evolutionary landscape without detrimental selective impacts as the other copy can compensate for loss or change of function. This can lead to fractionation of either or both copies and possibly reflect the shorter distribution of variable genes. However this hypothesis needs to be followed up on in the future.

We find core and variable gene functions in cranberries follow patterns observed in other studies. Variable genes are enriched for defense response, similar to *Brachypodium distachyon (Gordon et al. 2017)*. Core genes are enriched for basic cellular processes similar to *Brassica oleracea* (Golicz et al. 2016). Variable genes therefore represent an important source of genetic variation to draw from for agronomically relevant traits such as disease resistance.

Importantly, a large proportion of genes in non reference genotypes are variable. This represents a substantial gene pool to leverage for future cranberry breeding efforts. Current applications of pangenome development include pangenome based GWAS (Zhou et al. 2015; Song et al. 2020). This may be a critical next step for cranberry breeding.

Materials and Methods

Genome sequencing, assembly, and annotation

For genomic sequencing, leaf tissue was collected from each of twelve cranberry genotypes selected to best represent both the pedigree of cranberry breeding panels and genetic diversity of the species. DNA was extracted using a Qiagen DNeasy extraction kit. DNA quantity was checked using a Quibit. DNA quantity was insufficient for sequencing of the accessions McFarlin and Wales Henry which was excluded from further analysis. DNA was submitted to the Michigan State University Research Technology Support Facility (MSU RTSF) for library preparation and sequencing. According to their report, "the shotgun genomic libraries were prepared with the Hyper Library construction kit from Kapa Biosystems (Roche)". Libraries were sequenced on an Illumina NovaSeq 6000 using a NovaSeq S4 reagent kit for 151 cycles from each end to generate paired 150 nucleotide long reads.

Genomic reads were quality and adapter trimmed using trimmomatic version 0.38. Reads were then used to generate a hybrid *de novo* and reference based genome assembly for each accession. This assembly method was described in detail previously including tool versions and command line options (Yocca et al. 2021). Briefly, genomic reads were mapped to the reference genome generated previously. Mapped reads were used to

generate a consensus genome sequence iteratively for three rounds. Then, unmapped reads were collected and *de novo* assembled into synthetic long-reads. These long reads were combined with the consensus sequence and incorporated into the genome assembly.

Several tissues were collected for RNA sequencing analysis. Our tissues were young leaf, mature leaf, green berry, and mature berry. Leaf and berry tissues were ground in liquid nitrogen to preserve RNA. Cold ground tissue was transferred to a Qiagen RNA-easy extraction kit. RNA quantity was checked with a Quibit and sent to the MSU RTSF for library construction and sequencing. Libraries were sequenced on an Illumina NovaSeq 6000. Reads were quality and adapter trimmed using trimmomatic version 0.38. They were mapped to their respective genome assemblies using hisat2 version 2.1.0. The resulting SAM files were sorted and converted to BAM files using PicardTools version 2.18.1 SortSam function. From these alignments, transcriptome assemblies were generated using Stringtie version 2.1.3. These transcriptome assemblies were used later for gene annotation.

Each non reference genome was annotated with the same method using the MAKER2 pipeline. We used several lines of evidence for annotation. Proteins from Araport11, and transcriptomes generated above were used as evidence. We also included the *Vaccinium corymbosum* 'Draper' transcriptome as

evidence. We generated two ab initio models trained on 'Draper' gene models, SNAP and Augustus. Augustus models were generated using the script `train_augutsus_draper.sh` on a subset of 4,000 randomly selected gene models. SNAP models were generated using the script `train_snap_draper.sh`.

Identification of core and variable genes

We initially identified orthologs between all eleven cranberry proteomes using Orthofinder2 version 2.4.1 using default parameters (only the working directory specified). We also aligned each genome using progressiveCactus. As Orthofinder2 might identify members of the same gene family as orthologous, we decided to filter out any ortholog calls without an alignment within 5 kilobase-pairs of each other. We used the `filter_orthofinder2.sh` script for ortholog calls.

Gene statistic calculations

We calculated several feature values for each gene model including; gene length, coding sequence length, exon count, intron count, exon length, and intron length. These values were calculated using the `annotate_core_genes_vacc_pan.py` script. Expression values were calculated as described above. However we added the flags `-G` and `-e` so only expression values for annotated genes were reported.

Functional enrichments

Each proteome was functionally annotated using InterPro Scan version 5.28-67.0. We converted the InterPro Scan annotation ID to a gene ontology ID using a manually curated translation table. We performed gene ontology term enrichment difference between core and variable genes in R using the script `vacc_pan_go_enrichment.Rmd`.

REFERENCES

REFERENCES

Armstrong, Joel, Glenn Hickey, Mark Diekhans, Ian T. Fiddes, Adam M. Novak, Alden Deran, Qi Fang, et al. 2020. "Progressive Cactus Is a Multiple-Genome Aligner for the Thousand-Genome Era." *Nature* 587 (7833): 246–51.

Birchler, James A., and Hua Yang. 2022. "The Multiple Fates of Gene Duplications: Deletion, Hypofunctionalization, Subfunctionalization, Neofunctionalization, Dosage Balance Constraints, and Neutral Variation." *The Plant Cell*, March. https://doi.org/10.1093/plcell/koac076.

Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, Michael A. Calderwood, Muhammed A. Yildirim, Nicolas Simonis, Benoit Charloteaux, et al. 2012. "Proto-Genes and de Novo Gene Birth." *Nature* 487 (7407): 370–74.

Diaz-Garcia, Luis, Luis Fernando Garcia-Ortega, Maria González-Rodríguez, Luis Delaye, Massimo Iorizzo, and Juan Zalapa. 2021. "Chromosome-Level Genome Assembly of the American Cranberry (Vaccinium Macrocarpon Ait.) and Its Wild Relative Vaccinium Microcarpum." *Frontiers in Plant Science*. https://doi.org/10.3389/fpls.2021.633310.

Emms, David M., and Steven Kelly. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.

Golicz, Agnieszka A., Philipp E. Bayer, Guy C. Barker, Patrick P. Edger, Hyeran Kim, Paula A. Martinez, Chon Kit Kenneth Chan, et al. 2016. "The Pangenome of an Agronomically Important Crop Plant Brassica Oleracea." *Nature Communications* 7 (November): 13390.

Golicz, Agnieszka A., Philipp E. Bayer, Prem L. Bhalla, Jacqueline Batley, and David Edwards. 2020. "Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications." *Trends in Genetics: TIG* 36 (2): 132–45.

Gordon, Sean P., Bruno Contreras-Moreira, Daniel P. Woods, David L. Des Marais, Diane Burgess, Shengqiang Shu, Christoph Stritt, et al. 2017. "Extensive Gene Content Variation in the Brachypodium Distachyon Pan-Genome Correlates with Population Structure." *Nature Communications* 8 (1): 2184.

Hancock, J. F., P. Lyrene, C. E. Finn, N. Vorsa, and G. A. Lobos. n.d. "Blueberries and Cranberries." *Temperate Fruit Crop Breeding*. https://doi.org/10.1007/978-1-4020-6907-9_4.

Hirsch, Candice N., Jillian M. Foerster, James M. Johnson, Rajandeep S. Sekhon, German Muttoni, Brieanne Vaillancourt, Francisco Peñagaricano, et al. 2014. "Insights into the Maize Pan-Genome and Pan-Transcriptome." *The Plant Cell* 26 (1): 121–35.

Hurgobin, Bhavna, Agnieszka A. Golicz, Philipp E. Bayer, Chon-Kit Kenneth Chan, Soodeh Tirnaz, Aria Dolatabadian, Sarah V. Schiessl, et al. 2018. "Homoeologous Exchange Is a Major Cause of Gene Presence/absence Variation in the Amphidiploid Brassica Napus." *Plant Biotechnology Journal* 16 (7): 1265–74.

Lei, Li, Eugene Goltsman, David Goodstein, Guohong Albert Wu, Daniel S. Rokhsar, and John P. Vogel. 2021. "Plant Pan-Genomics Comes of Age." *Annual Review of Plant Biology* 72 (June): 411–35.

Li, Ying-Hui, Guangyu Zhou, Jianxin Ma, Wenkai Jiang, Long-Guo Jin, Zhouhao Zhang, Yong Guo, et al. 2014. "De Novo Assembly of Soybean Wild Relatives for Pan-Genome Analysis of Diversity and Agronomic Traits." *Nature Biotechnology* 32 (10): 1045–52.

Lynch, M., and A. Force. 2000. "The Probability of Duplicate Gene Preservation by Subfunctionalization." *Genetics* 154 (1): 459–73.

Marroni, Fabio, Sara Pinosio, and Michele Morgante. 2014. "Structural Variation and Genome Complexity: Is Dispensable Really Dispensable?" *Current Opinion in Plant Biology* 18 (April): 31–36.

Ohno, Susumu. 1970. "Evolution by Gene Duplication." https://doi.org/10.1007/978-3-642-86659-3.

Qiao, Qin, Patrick P. Edger, Li Xue, La Qiong, Jie Lu, Yichen Zhang, Qiang Cao, et al. 2021. "Evolutionary History and Pan-Genome Dynamics of Strawberry (Spp.)." *Proceedings of the National Academy of Sciences of the United States of America* 118 (45). https://doi.org/10.1073/pnas.2105431118.

Siepel, Adam. 2009. "Darwinian Alchemy: Human Genes from Noncoding DNA." *Genome Research*.

Song, Jia-Ming, Zhilin Guan, Jianlin Hu, Chaocheng Guo, Zhiquan Yang, Shuo Wang, Dongxu Liu, et al. 2020. "Eight High-Quality Genomes Reveal Pan-Genome Architecture and Ecotype Differentiation of Brassica Napus." *Nature Plants* 6 (1): 34–45.

Tao, Yongfu, Xianrong Zhao, Emma Mace, Robert Henry, and David Jordan. 2019. "Exploring and Exploiting Pan-Genomics for Crop Improvement." *Molecular Plant* 12 (2): 156–69.

Tettelin, Hervé, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, et al. 2005. "Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial 'Pan-Genome.'" *Proceedings of the National Academy of Sciences of the United States of America* 102 (39): 13950–55.

Wang, Wensheng, Ramil Mauleon, Zhiqiang Hu, Dmytro Chebotarov, Shuaishuai Tai, Zhichao Wu, Min Li, et al. 2018. "Genomic Variation in 3,010 Diverse Accessions of Asian Cultivated Rice." *Nature* 557 (7703): 43–49.

Yao, Wen, Guangwei Li, Hu Zhao, Gongwei Wang, Xingming Lian, and Weibo Xie. 2015. "Exploring the Rice Dispensable Genome Using a Metagenome-like Assembly Strategy." *Genome Biology* 16 (September): 187.

Yocca, Alan E., and Patrick P. Edger. 2022. "Machine Learning Approaches to Identify Core and Dispensable Genes in Pangenomes." *The Plant Genome* 15 (1): e20135.

Yocca, Alan E., Zefu Lu, Robert J. Schmitz, Michael Freeling, and Patrick P. Edger. 2021. "Evolution of Conserved Noncoding Sequences in Arabidopsis Thaliana." *Molecular Biology and Evolution* 38 (7): 2692–2703.

Zhao, Junliang, Philipp E. Bayer, Pradeep Ruperao, Rachit K. Saxena, Aamir W. Khan, Agnieszka A. Golicz, Henry T. Nguyen, Jacqueline Batley, David Edwards, and Rajeev K. Varshney. 2020. "Trait Associations in the Pangenome of Pigeon Pea (Cajanus Cajan)." *Plant Biotechnology Journal* 18 (9): 1946–54.

Zhao, Qiang, Qi Feng, Hengyun Lu, Yan Li, Ahong Wang, Qilin Tian, Qilin Zhan, et al. 2018. "Pan-Genome Analysis Highlights the Extent of Genomic Variation in Cultivated and Wild Rice." *Nature Genetics* 50 (2): 278–84.

Zhou, Zhengkui, Yu Jiang, Zheng Wang, Zhiheng Gou, Jun Lyu, Weiyu Li, Yanjun Yu, et al. 2015. "Resequencing 302 Wild and Cultivated Accessions Identifies Genes Related to Domestication and Improvement in Soybean." *Nature Biotechnology* 33 (4): 408–14.