UNIVERSITY OF CALIFORNIA,
IRVINE

Comparative and Population Genomics

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Ecology and Evolutionary Biology

by

Edwin Solares

Dissertation Committee:
Distinguished Professor Brandon Gaut, Chair
Assistant Professor JJ Emerson
Professor Anthony Long

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my parents, family, son, Christopher and all of my colleagues and friends. I would also like to give a special thanks to Dr's JJ Emerson and Richard Lathrop for helping me throughout the entirety of my PhD education. I would also like to give many thanks and appreciation to my committee members Dr's JJ Emerson and Tony Long. As well as many, many thanks to my awesome advisor, Dr. Brandon S. Gaut.

Chapter 1: We would like to thank Christina Lin for assistance with the Bionano scaffolding, Melanie Oakes for help generating the Bionano data, and Angela Miller for assistance with editing and figure preparation. We thank Casey Bergman for providing examples of novel TEs in substrains of ISO1. DEM is part of the Oxford Nanopore Educational Initiative and has received free flow cells as part of that program. This material is based upon work supported by the National Science Foundation Bridge to the Doctorate Fellowship under Grant No. 1500284 and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1321846. MC is supported by US National Institutes of Health (NIH) grant R01 RR024862-01A6. JJE is supported by NIH grant R01GM123303-1 and University of California, Irvine setup funds. This work was made possible, in part, through access to the Genomics High-Throughput Facility Shared Resource of the Cancer Center Support Grant CA-62203 at the University of California, Irvine, and NIH shared-instrumentation grants 1S10RR025496-01, 1S10OD010794-01, and 1S10OD021718-01. RSH is an American Cancer Society Research Professor.

Chapter 2: We thank A. Minio from the Cantù Lab at UC Davis for informing us of the haplotig issue in genome assemblies, R. Lathrop, S. Koren and J.J. Emerson at UC Irvine for comments, and R. Kuo from the Roslin Institute, University of Edinburgh for devising the name, HapSolo. EAS is supported by a National Science Foundation (NSF) Graduate Research Program Fellowship under Grant No. DGE-1321846. This grant supported his time to formulate and excute the study. The work was also supported by NSF Grant NSF Grant No. 1741627 to BSG; this grant supported the empirical work that was the motivation for this study, it contributed data, and it funded publication charges. YT and ADL were supported by NIH Grants R01OD010974 and R01GM115562; these grants contributed data and supported personnel time for Hi-C analyses. This work also used the Extreme Science and Engineering Discovery Environment, which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center under award no. TG-MCB180035. These grants provided computational resources that were used in this study.

I would also like to acknowledge Genetics Society of America and BioMed Central Ltd for allowing for the publication of my previous work as well as their permission to incorporate that work into my dissertation.

VITA

Edwin Solares

EDUCATION

Doctor of Philosophy in Biological Sciences **2021**
University of California, Irvine *Irvine, CA*

Masters of Science in Biological Sciences **2019**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Computational Sciences **2014**
University of California, Irvine *Irvine, CA*

FELLOWSHIPS, AWARDS AND HONORS

University of California President's Postdoctoral Fellowship **2021–2022**
University of California, Davis

University of California President's Pre-Professoriate Fellowship **2020–2021**
University of California, Irvine

National Science Foundation Graduate Research Fellowship **2017–2020**
University of California, Irvine

National Science Foundation Bridge to the Doctorate Fellowship **2015–2017**
University of California, Irvine

XSEDE Research Allocation: A genetic and population analysis of *Persea americana* local cultivars: Understanding the Genetic and Structural Landscape of Avocados **2019–2021**

XSEDE Start-up Allocation: Rapid Low-Cost Assembly of Food Crops and *Drosophila* for Structural Variant Detection **2019–2022**

XSEDE Start-up Allocation: Rapid Low Cost Assembly of Food Crops and *Drosophila* for Structural Variant Detection **2018–2019**

Plant and Animal Genomics Conference: Asia Travel Award **2017**

Department of Ecology and Evolutionary Biology Travel Award **2017**

IMSD-MBRS Summer Scholar **2015**

Pacific Biosciences Travel Award PAG & ASHG **2014**

Broad Institute RNASeq Workshop Travel Award **2014**

Undergraduate Research Opportunities Program Travel Award **2014**

Information and Computer Science Excellence in Research	**2014**
Information and Computer Science Honors	**2014**
FASEB MARC Travel Award	**2012**
Southern California Edison Scholarship	**2011-2013**

JOURNAL PUBLICATIONS

(8) **Solares, E. A.**, Comparative and Population Genomics. (In preparation) **2021**
University of California, Irvine

(7) **Solares, E. A.**, Morales, A., Focht, E., Ashworth, V., Zhou, Y., Figueroa-Balderas, R., Minio, A., Cantu, D., Apraia, M. L., Gaut, B. S. **2021**
A new avocado reference genome provides insight into genome evolution and the genetic basis of important traits. (In preparation)

(6) Hemming-Schroeder, E., **Solares, E. A.**, Bradley, L., Zhong, D., Lee, M. C., Zhou, G., Yewhalaw, D., Dent, A., Kazura, J. W., Yan, G. **2021**
Highly multiplexed amplicon sequencing panel for accessing relationship inference by microhaplotypes in *Plasmodium vivax.* (In preparation)

(5) **Solares, E. A.**, Martin, G., Muyle, A., Bousios, A., Gaut, B. S. **2021**
Analysis of the effects of secondary structure on siRNA mappings in maize. (In preparation)

(4) **Solares, E. A.**, Tao, Y., Long, A. D., Gaut, B. S **2021**
HapSolo: An optimization approach for removing secondary haplotigs during diploid genome assembly. *BMC Bioinformatics* 22, 9 (2021) `https://doi.org/10.1186/s12859-020-03939-y`

(3) Zhou, Y., Minio, A., Massonnet, M., **Solares, E. A.**, Lv, Y., Beridze, T., Cantu, D., Gaut, B. S **2019**
The population genetics of structural variants in grapevine domestication. *Nat. Plants* 5, 965–979 (2019) `https://doi.org/10.1038/s41477-019-0507-8`

(2) **Solares, E. A.**, Chakraborty, M., Miller, D. E., Kalsow, S., Hall, K., Perera, A. G., ... Hawley, R. S **2018**
Rapid Low-Cost Assembly of the *Drosophila melanogaster* Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3: Genes, Genomes, Genetics* (2018) 8(10), 10 3143–3154. `https://doi.org/10.1534/g3.118.200162`

(1) Clifton, B. D., Librado, P., Yeh, S.-D., **Solares, E. S.**, Real, D. A., Jayasekera, S. U., ... Ranz, J. M. **2017**
Rapid Functional and Sequence Differentiation of a Tandemly Repeated Species-Specific Multigene Family in *Drosophila. Molecular Biology and Evolution.* `https://doi.org/10.1093/molbev/msw212`

PRESENTATIONS AND POSTERS

(4) "RADC: Reduction in Althaps and Duplicate Contigs for Improved Hi-C Scaffolding" (HapSolo) Plant and Animal Genomics Conference (PAG XXVIII). Poster Presentation	**2020**
(3) "Rapid low-cost assembly of Drosophila melanogaster reference genome using low-coverage, Long-Read Sequencing" Seoul National University Departmental Seminar	**2018**
(2) "New applications of quickmerge allow for improved genome assemblies and use with new sequencing technologies" Plant and Animal Genomics Conference Asia Conference (PAG Asia). Poster Presentation	**2017**
(1) "Web based drug discovery pipeline and automation for in silico screening of small molecules" ISMB Poster Presentations	**2017**

SERVICE AND MEMBERSHIPS

Society for Advancement of Chicanos/Hispanics and Native Americans in Science (SACNAS) - Outreach Chair	**2017-2018**
Society for Advancement of Chicanos/Hispanics and Native Americans in Science (SACNAS) – University of California, Irvine Co-founder	**2017-2018**
Society for Advancement of Chicanos/Hispanics and Native Americans in Science (SACNAS)	**2017-Current**
California Alliance for Minority Participation Orientation, Graduate School and Research Q&A	**2013-2017**
Los Angeles Communities Advocating for Unity, Social Justice, and Action (LA CAUSA) STEM Outreach, Curriculum Advisor and Orientation Speaker	**2014-2015**
Compton Unified School District and 1 Day Paint and Body Corp – Linux PC setup and donation	**2013**
International Society for Computational Biology	**2012-2018**
Society for Molecular Biology and Evolution	**2019-Current**

TEACHING

Computer Science 171 - Introduction to Artificial Intelligence **Summer 2020**
Teaching Assistantship (1 Lecture)

Biological Sciences 94 – Organisms to Ecosystems **Spring 2020**
Teaching Assistantship – 2 Weeks of Lesson Plans

Computer Science 171 – Introduction to Artificial Intelligence **Fall 2019**
Teaching Assistantship – 2 Lectures

Computer Science 171 – Introduction to Artificial Intelligence **Summer 2019**
Teaching Assistantship – 2 Lectures

Computer Science 171 – Introduction to Artificial Intelligence **Spring 2018**
Teaching Assistantship – 3 Lectures

Computer Science 171 – Introduction to Artificial Intelligence **Winter 2018**
Teaching Assistantship – 3 Lectures

Ecology and Evolutionary Biology 282 – Introduction to Bioinformatics **Fall 2018**
Lecturer

Ecology and Evolutionary Biology 283 – Advanced Topics in Bioinformatics **Winter 2019**
Lecturer

Computer Science 184A – Introduction to Bioinformatics **F2017 & F2018**
Lecturer

Computer Science 184B – Advanced Topics in Bioinformatics **W2018 & W2019**
Teaching Assistantship

Computer Science 189– Project in Bioinformatics **S2018 & S2019**
Teaching Assistantship

California Alliance for Minority Participation High Performance Computing and Research Seminar **Winter 2017**
Seminar Organizer & Speaker

MENTORSHIP *Undergraduate Students (Bolded for Completed Honors Thesis)*: Eunbi Yang, Miguel Escobar, Kathleen Leon He, Carolina Rojas, **Khalid el Assad**, Jeanelle Guardado-Mendez, **Agnes Jang**, Jacky Dai, Zexi Sun, Eric Gamarra, Jose Eduardo Corona, Joshua Arias, Kai Chang, **Anastasia Miles**, Sabrina Will, Beoung Lee, Ashlyn Kimura, Jiadong Yang, Ran Duan, Renhao Luo, Chinmay Raut, Vama Jhumkhawala, Joshua Costa, Youjia Yang, Yuting Lu
High School Students: Vineet Disay, Joseph Kim

ABSTRACT OF THE DISSERTATION

Comparative and Population Genomics

By

Edwin Solares

Doctor of Philosophy in Ecology and Evolutionary Biology

University of California, Irvine, 2021

Distinguished Professor Brandon Gaut, Chair

Structural variants (SVs) are large insertions, deletions, duplications, inversions or translocation of sequences that vary among individuals or chromosomes. These SVs have been shown to play a significant role in important phenotypic traits, but they have been difficult to detect on a genome-wide scale until recently. They have long been known, however, to be important to crop evolution. Fitting examples include the lack of branching in maize, sex determination and berry color in grapes, and coloration in crops. Recent advances in long-read sequencing have led to more continuous and accurate genomes, and empowered scientist's ability to identify these SVs, some of which, until recently, were believed to have been due to single nucleotide polymorphisms. In my dissertation, I explored ways to produce more complete and continuous genomes across a broad range of species, which is a necessary precursor for identification of SVs. I also explored SVs at population and individual levels, with the ultimate goal of finding correlations between genetic mutations such as SVs and important phenotypes. To do this I have applied novel uses of methods and sequencing approaches, as well as created tools for reducing the noise in highly heterozygous genomes. In the first chapter of my thesis, I explored the efficacy of reconstructing a genome using low coverage and inexpensive - but inaccurate - sequencing reads, using a new application of genome assembly methods. We were able to achieve a rapid low-cost reference level assembly, as well as identify novel SVs. Chapter two of my thesis aimed to reduce the

presence of alternative contigs from assembly of diploid genomes using novel methods, for improving downstream analysis such as HiC scaffolding. This culminated in the release of a new software package, HapSolo, that improved on current methods and used hill climbing for optimization of parameters to purge and remove alternative contigs. The application of HapSolo improved HiC scaffolding, resulting in a decrease in the total number of contigs, higher scaffold N50's and more continuous genome assemblies. In my final chapter, I have applied the knowledge gained from my previous chapters to decipher the genome of avocado and to examine standing genetic variation within the three major avocado ecotypes, using resequencing data of three outgroups and 31 avocado accessions. Chapter three has the goal of moving toward identifying SV's among accessions that have implications for phenotypes, as well as providing the scientific community with an annotated chromosome level scaffolded assembly.

INTRODUCTION

Comparative and Population Genomics

Approximately 1.3 billion people worldwide experienced food insecurity at moderate levels in 2019 and in 2020, this number grew to 2.37 billion people, of which, 928 million faced severe levels of food insecurity [55]. The predicted effects of climate change will only exacerbate the issue [151, 150, 67, 66, 65, 53, 3] as our ability to grow food will be severely hindered due to climate change [3]. Meanwhile, some of these effects can already be seen today. Currently, crops such as *Zea mays* (maize), *Persea americana* (avocados), and *Elaeis* species (oil palms) have seen an increase in demand, and are projected to continue to increase as our population grows [153]. To date, this growing demand has accelerated the deforestation of the Amazon Rainforest, water rights conflicts, and a fragile dependence on mono-cultures. In the past, many of these issues have disproportionately impacted the poor and people of color [89]. This only increases the urgency for critical advancements needed for crop yield improvements, and an increase in quality and distribution to populations in need. Science will lead the way and help resolve current and future food security issues, especially in a world with rapidly changing climates and large fluctuations in maximum temperatures (Figure 1). A critical foundation to crop improvement is a better understanding of genetic variation within crops and the potential impacts of genetic variation on phenotypes.

To improve our understanding of genetic variation, we first need to understand how plants have, and will adapt to changing environments. An important, yet poorly studied type of genetic variation is structural variants (SVs), which are large insertions, deletions, duplications, inversions or translocations of DNA sequences that vary among and between individuals and populations [52]. SVs have been shown to play a significant role in phenotypic traits important for adaptation, but until recently they have been difficult to detect on a genome-wide scale [143, 99, 129, 63, 76, 36]. They have long been known to be important for crop evolution [174, 175, 161, 137, 33]. Several significant examples include the lack of branching in maize21, sex determination in grapes[175], and coloration in crops such as grapes[175] and citrus[29].

Identifying SV's benefits greatly from a nearly complete genome that is representative of the genetic material from one parent at any given genome position. While reference chromosomes need not be derived from one parent (i.e. complete haplotype), this is often preferred. Frequently, reference assemblies are amalgams constructed from haplotypes derived from both parents. Haploid genomes are crucial for downstream analysis (i.e. gene and repeat annotation, evolutionary inference, scaffolding, etc), as many analyses and programs assume a haploid reference. The use of a non-haploid genome would generate incorrect inferences and results, as deviates significantly from the assumptions of such analyses. Indeed, it is often possible to recover haplotypes from both parents, permitting researchers to identify genetic variants between parents, allowing us to sample population variation [174, 175, 99].

In my dissertation, I aim to: (1) improve on methods for the curation of a more continuous and complete genome that is also affordable, increasing it's accessibility; (2) develop a tool for purging duplicated or alternate haplotypes in order to isolate isolate a single haploid reference genome; and, (3) apply these methods to a poorly understood perennial crop, avocados, with potential for crop improvement, and identify regions of genetic variation differing among populations of avocado that contain genes implicated in domestication and adaptation.

In the first chapter of my thesis, I explore the efficacy in reconstructing a genome using low coverage and inexpensive - but inaccurate - sequencing reads, using a new application of genome assembly methods. Previously, most genome sequencing consisted of low cost high-throughput short-read sequencing. This revolutionized our ability to assay genome-wide sequence variation. These short reads allowed for identifying single nucleotide polymorphisms (SNPs) and short insertion-deletion (indel) polymorphisms in unique genomic regions, but failed to identify large repetitive portions of the genome [117, 155, 25, 62], leading to highly fragmented and incomplete genomes and gene models [117, 155, 25, 62]), due to the amount of repetitive content present. It is important to note the majority of sequence differences between individuals is due to the presence of SV's [52]. SVs are often longer than short sequencing reads (generally 50–150 bp), meaning genotyping SVs is indirect and relies on features of alignments to a reference genome such as divergent read mappings, split reads, or elevated read coverage [103, 5]. However, comparisons of extremely contiguous *de novo* genome assemblies from humans [71, 34] and *Drosophila melanogaster* [37] revealed that short reads miss 40–80% of SVs. Consequently, methods not susceptible to the shortcomings of short-read sequencing are essential to obtaining a more complete view of genome variation [5]. Creating and comparing highly continuous and accurate *de novo* genome assemblies overcome these limitations and allow us to dramatically improve our understanding of genetic variation [5]. Resolving these difficult genomic regions would require longer reads with sufficient read depth [109, 86, 26, 140]. Single Molecule Real-Time (SMRT) sequencing from Pacific Biosciences (PacBio) and Nanopore sequencing from Oxford Nanopore Technologies (ONT) provide such capabilities and are capable of reconstructing many of these repetitive regions, leading to more accurate and continuous genomes [80, 21, 104, 76, 82]. However, due to their high error rates (~10-15%), creating such genomes requires non-trivial coverage (generally 30x or greater) [82, 36].

Sequencing using ONT reduces costs of sequencing and may produce reads that are hundreds of kilobases in length [76]. To understand how effective assembly using ONT is when applied to *de novo* genome assembly, we sequenced using ONT MiniION, and assembled the genome of a well studied species, *Drosophila melanogaster*. This species already contains a highly curated reference level assembly (ISO1) [70], and arguably the best metazoan genome available. We followed and generalized an assembly merging approach [36] to combine modest long-read coverage from a single ONT MinION flow cell (30x depth of coverage with an average read length of 7,122 bp) and Illumina short-read data. This assembly resulted in a highly contiguous and accurate genome assembly. We examine the structural differences between our assembly and the published reference assembly of the same strain and observe several candidate SVs, the majority of which are transposable element (TE) insertions and copy-number variants (CNVs). These are mutations that must either be: 1) recent mutations that occurred in the genome strain since it was sequenced; 2) segregating in the genome strain due to incomplete isogeny; or 3) errors in one of the assemblies. In this study we were able to achieve a rapid low-cost highly continuous and accurate reference level assembly, as well as identify novel SVs.

Chapter two of my thesis aimed to reduce the presence of alternative contigs from assembly of diploid genomes using novel methods, for improving downstream analysis such as Hi-C scaffolding, annotation and inferences about evolution. In chapter one we analyzed a highly homozygous sample allowing us to benchmark our methodology. Homozygous samples allow for easier sequencing and reconstruction of the genome. For example, the first two plant species targeted for reference quality genomes, *Arabidopsis thaliana* [172] and rice (*Oryza sativa*) [60], were chosen due to their self-fertilizing nature and are therefore highly homozygous. Other early genomes, such as those from *Caenorhabditis elegans* and *Drosophila melanogaster* [1, 2], were also based on inbred, highly homozygous materials. Recent sequencing of additional model and non-model species have continued to rely on near-homozygous materials, either through inbreeding [143, 77] or by focusing on haploid tissue [44, 126]. This, however, differs considerably from conditions outside of a lab setting. In the wild, samples consist of unique genetic content from each parent making up two discrete haplotypes.

The reliance on homozygous samples is fading rapidly, however, for at least three reasons. Firstly, it has become clear that inbred materials can misrepresent a genome's natural state. A dramatic illustration of this fact is that some lines of maize purged a significant 8% of their genome in only six generations of self-fertilization [134]; in broader context, inbred genomes tend to be smaller than those based on outbreeding species [120, 54]. Secondly, many species of interest cannot be manipulated into a homozygous state easily. Many animals fall into this category, such as mosquitoes [135], and domesticated cattle [99], as do many perennial crops like grapes, which are highly heterozygous [175] and can be selfed, but only with substantial fitness costs, limiting homozygosity [105]. Finally, some important features and phenotypes—such as sex determination [100] and other important adaptations—can be identified only by analyzing heterozygous samples.

Fortunately, the resolution of highly heterozygous regions, which often contain large structural variants, is now possible due to improvements in sequencing technologies and their affordability. In theory, long-read sequencing technologies like those from PacBio and ONT provide the capability to resolve distinct haplotypes in heterozygous regions, as they can span these regions bridging loci containing their respective haplotypes, leading to the assembly of reference-quality diploid genomes [143, 76, 36]. One limitation however, is when the length of the reads cannot bridge multiple heterozygous loci, leading to ambiguity to their origin of their respective haplotypes [42]. Several genomes based on highly heterozygous materials have been published recently [175, 157, 42, 58, 81, 130, 128], with many additional ongoing efforts.

Nevertheless, the assembly of heterozygous genomes still presents substantial challenges. One of which is resolving distinct haplotypes in regions of high heterozygosity. Programs that assemble long-reads, such as FALCON [42] and Canu [81], can mistakenly fuse distinct haplotypes into the primary assembly. This haplotype-fusion not only produces genomes with pseudo and alternate haplotypes, but also genomes that are much larger than the expected genome size, requiring manual curation [128]. When haplotypes are fused, either into the same contig (resulting in an amalgam of haplotypes) or as different contigs into the primary assembly (alternate haplotypes), the increased size and complexity of the assembly complicates down-stream approaches, such as scaffolding by Hi-C or optical mapping, variant detection and calling, annotation and evolutionary inferences. In theory, FALCON-unzip

[42] solves some problems by identifying alternative (or 'secondary') haplotigs that represent the second allele in a heterozygous region and then provides a primary assembly without secondary contigs, but still containing pseudo haplotigs.

Although we cannot trivially separate out pseudo haplotigs without additional information, it is an easier, albeit still difficult, problem to identify and remove alternative contigs during assembly. Some suggested solutions for removing alternate contigs, like Redundans [121], identify secondary contigs via similarities between contigs [121] and removes the shorter of two contigs that share some pre-defined level of similarity. Another approach, Purge Haplotigs [132] uses sequence coverage as a criterion to identify regions with two haplotypes [132]. The reasoning behind Purge Haplotigs is that alternative alleles in a heterozygous region should have only half the raw sequence coverage of homozygous regions of a single individual as coverage is expected to be uniform across the genome. Accordingly, the algorithm proceeds by first remapping raw reads to contigs, then flagging contigs with lower than expected read depth, and finally re-mapping and removing low-coverage contigs from the primary haplotype-fused assembly. A more recent approach, implemented in the purge_dups tool [63], builds on the coverage-based approach of Purge Haplotigs, and has been shown to be superior based on a few exemplar assemblies [63]. In this study, we sought a slightly different approach, named HapSolo, by identifying and removing potential secondary and alternate haplotigs. Our approach is similar to Redundans, in that it begins with an all-by-all pairwise alignment among contigs and uses features of sequence alignment as a basis to identify potential alternative haplotigs. However, HapSolo is unique in exploring the parameter space of alignment properties to optimize the primary assembly, using features of BUSCO [141] scores as the optimization target.

This culminated in the release of a new software package, HapSolo [144], that improved on current methods and used hill climbing for optimization of parameters to purge and remove contigs representing alternative haplotypes. The application of HapSolo improved

Hi-C scaffolding relative to purge_dups, resulting in a decrease in the total number of contigs, higher scaffold N50's and more continuous genome assemblies.

In my third and final chapter, I apply the knowledge gained from my previous chapters to decipher the genome of avocado and to examine standing genetic variation within the three major avocado ecotypes, using resequencing data of three outgroups and 31 avocado accessions.

Avocado (*Persea americana Mill.*) is a perennial, subtropical crop that is in ever-increasing demand. In the United States, per capita avocado consumption has tripled over the last two decades. Demand in the U.S. is met partly by domestic production, but principally by imports from Mexico and elsewhere. Mexico is currently the largest producer of avocado where the crop is worth an estimated $2.5 billion per year [127], but other major producers include the Dominican Republic, Peru, Chile, Indonesia, Israel and Kenya (`http://www.fao.org/faostat/en/#data/QC/visualize`). Although the popularity of avocados is primarily a 20th century phenomenon [138], it has quickly grown to be a global commodity.

Remarkably, avocado cultivation is dominated by a single variety (Hass) that represents ~90% of cultivation world-wide [127]. All Hass trees are derived clonally from a tree patented by seed in 1935. Despite the shockingly narrow genetic base of agricultural production, avocado *sensu lato* is quite genetically diverse. Some of this diversity stems from the fact that there are three domesticated botanical races: *P. americana var. americana Mill* (which we will call the 'Lowland' race in recognition that the previously accepted name of West Indian is inaccurate as most are found in lowland coastal areas of the Yucatan peninsula and Central America), *var. drymifolia Blake* (the Mexican race), and *var. guatemalensis Williams* (the Guatemalan race) [18]. The strikingly different fruit morphologies among the races suggest that they may have been domesticated separately, a conjecture supported by genetic data [56, 11, 127]. One practical consequence is that each race likely contains separate alleles and/or genes of interest for crop improvement, due to their different domestication histories and ecology. Another consequence is that hybridization between races can produce unique allelic combinations, potentially leading to agronomically useful hybrid offspring. Hass is, in fact, one example; although its precise breeding history is not known, genetic evidence shows that it is a hybrid between Guatemalan and Mexican races [138, 40, 127]. Avocados also require long growing periods prior to producing fruit (5 to 8 years before production); [85], requiring substantial space, water, and financial resources [10]. Additionally, avocado is predominantly out-crossing, due to synchronous dichogamy. There are two flowering types in this system: A and B. Type A trees are female (receptive to pollen) in the morning and shed pollen as males in the afternoon of the following day. In contrast, type B trees are male in the morning of the first day and female in the afternoon of the next day. The system is complicated by the fact that there is some leakiness of flower type that depends on the environmental cues [47]. As a result of these complications, avocado breeding has historically relied on open-pollinated and inter-racial hybridization to the extent that most individual varieties lack accurate breeding records [48, 11, 139].

These complications argue that genomics and molecular breeding are central for the continued improvement of avocado. For example, molecular markers for flowering types may be particularly useful, because type B avocados are crucial for pollination but typically less productive than type A varieties [47]. Recently, Rendon-Ayana et al. (2019) made an important contribution toward molecular breeding by producing draft genomes of Hass and a wild Mexican accession (*P. americana ssp. drymifolia*). Using the Hass reference, Rendon-Ayana et al. (2019) also explored the hybrid history of Hass and a few aspects of the evolutionary genomics of avocado. Nonetheless, several important features of the evolutionary genomics of avocado remain unexplored, including characterizing diploid chromosomes in this highly heterozygous ancestor, using sweep mapping to identify potential regions of agronomic interest, and focusing on genomic diversity in the context of interesting traits, like A vs. B flowering types and change in skin color due to ripening.

In the study described in my third chapter, we produce and assemble the genome of the Gwen variety and use that genome as a reference for evolutionary analyses. Gwen has also been the subject of intensive breeding efforts for three decades. Our Gwen genome vastly improves contiguity relative to the Hass genome, providing a better platform to explore the evolutionary genomics of avocado. More specifically, we intend to use the data outlined in this chapter to focus on four sets of questions. First, what does the Gwen genome tell us about patterns of heterozygosity within an avocado accession? Genomic analysis of highly heterozygous grape (*Vitis vinifera*), another perennial clonally propagated crop, revealed that as many as one in seven genes are hemizygous, perhaps due to structural mutations that have accrued during clonal propagation [175]. Is avocado similar? Second, we use the Gwen genome as a reference to explore genetic diversity within avocado, specifically to recapitulate the three races and to assess the hybrid origin of well-known cultivars. This last question builds on several previous investigations of genetic diversity [11, 40, 39] but extends the work to a genomic scale. Third, we investigate features of avocado domestication. Do the three races demonstrate a cost of domestication - i.e., an accumulation of deleterious alleles - relative to wild accessions, as is common for domesticates [57, 110]? And do the three racial groups share regions of selective sweeps consistent with parallel selection on genic regions associated with specific traits? Finally, we investigate genetic diversity between the A and B flowering types, with the goal of identifying genomic regions that may contribute to synchronous dichogamy. We believe many of these differences are due to the presence and absence of SV's. In particular we find an SV with implications for flowering time in avocados.

Chapter 1

Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing

1.1 Abstract

Accurate and comprehensive characterization of genetic variation is essential for deciphering the genetic basis of diseases and other phenotypes. A vast amount of genetic variation stems from large-scale sequence changes arising from the duplication, deletion, inversion, and translocation of sequences. In the past 10 years, high-throughput short reads have greatly expanded our ability to assay sequence variation due to single nucleotide polymorphisms. However, a recent *de novo* assembly of a second *Drosophila melanogaster* reference genome has revealed that short read genotyping methods miss hundreds of structural variants, including those affecting phenotypes. While genomes assembled using high-coverage long reads can achieve high levels of contiguity and completeness, concerns about cost, errors, and low yield have limited widespread adoption of such sequencing approaches. Here we resequenced the reference strain of *D. melanogaster* (ISO1) on a single Oxford Nanopore MinION flow cell run for 24 hr. Using only reads longer than 1 kb or with at least 30x coverage, we assembled a highly contiguous *de novo* genome. The addition of inexpensive paired reads and subsequent scaffolding using an optical map technology achieved an assembly with completeness and contiguity comparable to the *D. melanogaster* reference assembly. Comparison of our assembly to the reference assembly of ISO1 uncovered a number of structural variants (SVs), including novel LTR transposable element insertions and duplications affecting genes with developmental, behavioral, and metabolic functions. Collectively, these SVs provide a snapshot of the dynamics of genome evolution. Furthermore, our assembly and comparison to the *D. melanogaster* reference genome demonstrates that high-quality *de novo* assembly of reference genomes and comprehensive variant discovery using such assemblies are now possible by a single lab for under $1,000 (USD).

1.2 Introduction

The characterization of comprehensive genetic variation is crucial for the discovery of mutations affecting phenotypes. In the last 10 years, the exponential decline in cost of high-throughput short-read sequencing has revolutionized our ability to assay genome-wide sequence variation. Short reads excel at identifying single nucleotide polymorphisms (SNPs) and short insertion-deletion (indel) polymorphisms in unique genomic regions. However, the majority of the sequence difference between individuals is caused by duplications, deletions, inversions, or translocation of sequences—collectively known as structural variants (SVs) [52]. SVs are often longer than short sequencing reads (generally 50–150 bp), meaning genotyping of SVs is indirect and relies on features of alignments to a reference genome such as divergent read mappings, split reads, or elevated read coverage [103, 5]. However, comparisons of extremely contiguous *de novo* genome assemblies from humans [71, 34] and *Drosophila melanogaster* [37] revealed that short reads miss 40–80% of SVs. Consequently, methods that are not susceptible to the shortcomings of short-read sequencing are essential to obtain a more complete view of genome variation [5]. We propose one approach—comparison of contiguous and accurate *de novo* genome assemblies—that would overcome these limitations and drastically improve our understanding of genetic variation [5].

Although short reads have been used extensively for *de novo* genome assembly, they fail to resolve repetitive regions in genomes, leaving errors and gaps while assembling such regions [117, 155, 25]. Such fragmented draft-quality assemblies are therefore poorly suited for identification of SVs [5] and lead to incomplete and/or missing gene models [62]. Theoretical considerations of the genome assembly problem predict that, with sufficient read depth and length, genome assemblies can resolve even difficult regions [109, 86, 26, 140]. Consistent with this, long reads produced by Single Molecule Real-Time sequencing from Pacific Biosciences (PacBio) and Nanopore sequencing from Oxford Nanopore Technologies (ONT) provide data capable of achieving remarkably contiguous *de novo* genome assemblies [80, 21, 104, 76, 82]. However, due to high error rates (~10–15%), generation of reliable assemblies with these reads requires non-trivial coverage (generally 30x or greater) [82, 36]. Nevertheless, until recently, long-read methods have required prohibitively expensive reagents, and technologies like PacBio also require substantial capital investment related to the housing and maintenance of equipment necessary to perform the sequencing. Combined with concerns about high error rates, widespread adoption of long-molecule sequencing for *de novo* assembly and variant detection has been tentative.

Sequencing using ONT may produce reads that are hundreds of kilobases in length [76], though their application to *de novo* assembly of reference-grade multicellular eukaryotic genomes is not yet routine. To understand how effective assembly using ONT is when applied to *de novo* genome assembly of a metazoan like *Drosophila*, we measured the contiguity, completeness, and accuracy of a *de novo* assembly constructed with ONT reads [143]. To accomplish this, we resequenced the *D. melanogaster* reference genome strain (ISO1) using the ONT MinION and compared the resulting assembly with the latest release of the *D. melanogaster* reference assembly [70], which is arguably the best metazoan reference genome available. We followed an assembly merging approach [36] to combine modest long-read coverage from a single ONT MinION flow cell (30x depth of coverage with an average read length of 7,122 bp) and Illumina short-read data. This assembly resulted in a highly contiguous and accurate genome assembly. Notably, with this approach, the majority of the euchromatin of each chromosome arm is represented by a single contiguous sequence (contig). Collectively, the assembly recovered 97.7% of Benchmarking Universal Single Copy Orthologs (BUSCOs). This is similar to the 98.3% BUSCOs recovered in the most recent release of the *D. melanogaster* genome (version 6.16). Scaffolding of the assembly with Bionano optical maps led to further improvements in contiguity. Finally, we examined the structural differences between our assembly and the published assembly of the same strain and observed several candidate SVs, the majority of which are transposable element (TE) insertions and copy-number variants (CNVs). These are mutations that must be either: 1) recent mutations that occurred in the genome strain since it was sequenced; 2) segregating in the genome strain due to incomplete isogeny; or 3) errors in one of the assemblies.

Overall, we show that high-quality *de novo* genome assembly of *D. melanogaster* genomes is feasible using low-cost ONT technology, enabling an assembly strategy that can be applied broadly to metazoan genomes. This strategy will make high-quality reference assemblies obtainable for species lacking reference genomes. Moreover, *de novo* assemblies for population samples of metazoan species is now feasible, opening the door for studying evolutionary and functional consequences of structural genetic variation in large populations.

1.3 Methods

Stocks

The ISO1 *D. melanogaster* reference stock used for both Nanopore and Illumina sequencing was obtained from the BDGP in 2014 [70]. All flies were kept on standard cornmeal-molasses medium and were maintained at 25°.

DNA isolation and quantification

DNA for Nanopore sequencing was isolated from males and females using the Qiagen Blood & Cell Culture DNA Mini Kit. Briefly, 60–80 flies were placed in two 1.5-mL Eppendorf Lo-Bind tubes and frozen in liquid nitrogen before being homogenized using a pestle in 250 µL of Buffer G2 with RNAse. 750 µL of Buffer G2 and 20 µL of 20 mg/mL proteinase K was then added to each tube and incubated at 50° for 2 hr. After 2 hr, each tube was spun at 5,000 RPM for 5 min, and the supernatant was removed and placed in a new 1.5-mL Lo-Bind tube and vortexed for 10 sec. The supernatant from both tubes was then transferred onto the column and allowed to flow through via gravity. The column was washed 3x with wash buffer and eluted twice with 1 mL of elution buffer into two 1.5-mL Lo-Bind tubes. 700 µl of isopropanol was added and mixed via inversion before being spun at 14,000 RPM for 15 min and 4°. The supernatant was removed and the pellet was washed with 70% ethanol, then spun at 14,000 RPM for 10 min at 4°. The supernatant was removed and 25 µL of ddH$_2$O was added to each tube and allowed to sit at room temperature for 1 hr. Both tubes were then combined into one. DNA was quantified on a Nanodrop and Qubit. DNA for Illumina sequencing was isolated from males and females using the Qiagen DNeasy Blood and Tissue Kit according to the manufacturer's instructions and quantified on a Qubit.

Library preparation, sequencing, and basecalling

For Nanopore sequencing, 1.5 µg of DNA (5.49 µL of 273 ng/uL DNA) was used to prepare a

1D sequencing library (SQK-LSK108) according to the manufacturer's instructions, including the FFPE repair step. The 75-µl library was then immediately loaded onto a R9.5 flow cell prepared according to the manufacturer's instructions and run for approximately 24 hr. Basecalling was completed using ONT Albacore Sequencing Pipeline Software version 2.0.2. Reads collected during the mux phase of sequencing were not included.

For Illumina sequencing, ~600-bp fragments were generated from 500 ng of DNA using a Covaris S220 sonicator. Fragments of 500–700 bp were selected using a Pippin and libraries were prepared using a KAPA High Throughput Library Preparation kit and Bio Scientific NEXTflex DNA Barcodes. The library was pooled with others and run as a 150-bp paired-end run on a single flow cell of an Illumina NextSeq 500 in medium-output mode using RTA version 2.4.11. bcl2fastq2 v2.14 was then run in order to demultiplex reads and generate FASTQ files.

Genome assemblies

Canu [82] release v1.5 was used to assemble the ONT reads. Canu was run with default parameters in grid mode (Sun Grid Engine) using ONT reads >1 kb and a genome size of 130 Mb. To generate the De Bruijn graph contigs for the hybrid assembly, we used Platanus [78] v1.2.4 with default settings to assemble 67x of Illumina paired-end reads obtained from the DPGP (`http://www.dpgp.org/dpgp2/DPGP2.html`) [119]. The hybrid assembly was generated with DBG2OLC [168] using contigs from the Platanus assembly and the longest 30x ONT reads. DBG2OLC settings (options: k 25 AdaptiveTh 0.01 KmerCovTh 2 MinOverlap 35 RemoveChimera 1) were similar to those used for PacBio hybrid assembly of ISO1 [36], except that the k-mer size was increased to 25 and the MinOverlap to 35 to minimize the number of misassemblies. The consensus stage of DBG2OLC was run with PBDAG-Con [41] and BLASR [35]. Separately, minimap v0.2-r123 (using a minimizer size window of 5, FLOAT fraction of minimizers of 0, and min matching length of 100) and miniasm v0.2-r123 (using default settings) were also used to assemble only the ONT reads [90].

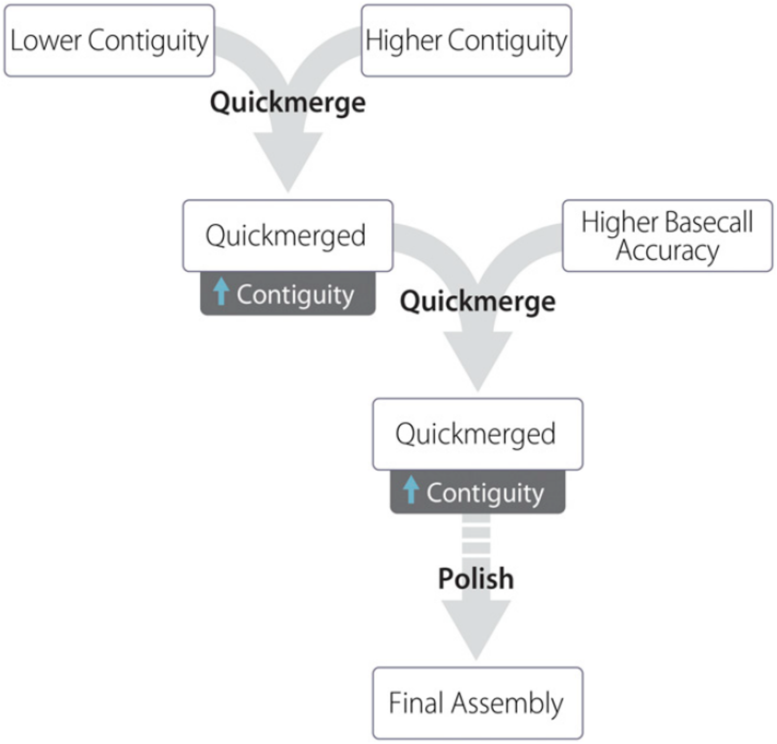

Figure 1.1: Assembly strategy used in this manuscript. A lower-contiguity assembly (Canu) is merged with a higher-contiguity assembly (DBG2OLC). The resulting assembly is again merged with the Canu assembly. The genome is then polished one or more times, here with nanopolish followed by Pilon.

The Canu and DBG2OLC assemblies were merged using quickmerge [36]. First, the two assemblies were merged using the DBG2OLC assembly as the query and the Canu assembly as the reference. Thus, the first quickmerge run (options: `hco 5.0 c 1.5 l 2900000 ml 20000`) filled gaps in the DBG2OLC assembly using sequences from the Canu assembly, giving preference to the Canu assembly sequences at the homologous sequence junctions. The contigs that are unique to the Canu assembly were incorporated in the final assembly by a second round of quickmerge. In the second quickmerge run (options: `hco 5.0 c 1.5 l 2900000 ml 20000`), the merged assembly from the previous step was used as the reference assembly, and the Canu assembly was used as the query assembly (Figure 1).

Assembly polishing

Assembly polishing was performed two ways. First, nanopolish version 0.7.1 [97] was run using the recommended settings with reads longer than 1 kb. Prior to running nanopolish, the merged genome assembly was indexed using bwa, and ONT reads were aligned to the genome using bwa mem. The resulting bam file was then sorted, indexed, and filtered for alignments of size larger than 1 kb using samtools 1.3 [94]. Nanopolish was then run with a minimum candidate frequency of 0.1. Following nanopolish, we polished the assembly twice with Pilon [158] v1.16. For Pilon, we aligned 44x 150-bp and 67x 100-bp Illumina paired-end reads to the assembly using bowtie2 [136] and then ran Pilon on the sorted bam files using default settings.

Bionano scaffolding

Bionano optical map data were collected following Chakraborty et al. [37]. ISO1 embryos less than 12 h old were collected on apple juice/agar Petri dishes, dechorionated using 50% bleach, rinsed with water, then stored at -80°. DNA was extracted from frozen embryos using the Animal Tissue DNA Isolation kit (Bionano Genomics, San Diego, CA). Bionano Irys optical data were generated and assembled with IrysSolve 2.1 at Bionano Genomics. We then merged the Bionano assembly with the final merged assembly contigs using IrysSolve, retaining Bionano assembly features when the two assemblies disagreed.

BUSCO analysis

We used BUSCO v1.22 to evaluate completeness and accuracy of all ISO1 assemblies [141]. We used the Diptera database, which contains 2,799 highly conserved genes, to estimate assembly completeness.

Assembly comparision and quality metrics

Assemblies were compared using alignment dot plots. For dot plots, assembled genomes were aligned to the reference *D. melanogaster genome* (r6.16) using nucmer (using options: `minmatch 100, mincluster 1000, diagfactor 10, banded, diagdiff 5`) [83]. The resulting delta alignment files were used to create the dot plots either with mummerplot (using options: `{fat{filter{ps`) or ggplot. QUAST v4.5.0 was used to compare each generated assembly to the contig reference genome assembly (*D. melanogaster* r6.16) for completeness and errors. QUAST was run in GAGE mode on contigs larger than 1 kb and with the reference assembly as fragmented, as it was originally the scaffolded assembly [136, 64]. GAGE was run on two types of reference assembly: the full reference assembly, and only the part of the reference assembly comprising ordered and oriented contigs on chromosome arms (i.e., Muller elements) and the mitochondrial sequence. Quality score analysis was performed by aligning our Illumina short reads against each genome assembly using Bowtie. The number of variants were summed and divided by the total number of bases (following Berlin et al. 2015) and also by the total number of bases aligned: $P_{error} = Variants/(TotalBases)$, where $TotalBases$ represents the total number of bases in the assembly or the total aligned bases (following Koren et al. 2017). These values are proxies for the probability of error and are used to calculate the QV score for each assembly according to the relation: $QV = -10 \cdot log_{10} P_{error}$.

A modified pipeline originally published in McCoy et al. (2014) and modified following Berlin et al. (2015) and Koren et al. (2017) was used (this pipeline relies on release 5.57 naming conventions) for calculating total genome reconstruction. Nucmer was used to align each assembly to the FlyBase 5.57 reference, and then separated and merged by euchromatic and heterochromatic regions for each chromosome arm using BEDtools [124]. Gene and TE reconstruction was completed similar to Berlin et al. (2015). For measuring the accuracy of gene models and TE reconstruction, both releases 5.57 and 6.16 of the reference were used. Fasta files containing all genes and separately, TEs for both releases were downloaded from FlyBase. In total, 17,730 gene models and 5,392 transposons from release 6.16, as well as 17,294 gene models and 5,409 transposons from release 5.57 were aligned to each genome assembly independently using nucmer. Nucmer was used to align each gene and transposon fasta file to each assembly independently. Alignments greater than 0%, 99% and equal to

Total reads	593,354
Average read length	7,122 bp
Total bases seuenced	4,184,159,334
Genome Coverage	30.2x
Reads > 1 kb	530,466
Genome coverage in reads > 1 kb	29.9x
Reads > 10 kb	145,634
Genome coverage in reads > 10 kb	17.5x

Table 1.1: Statistics of reads used for genome assembly: Only reads with quality scores ≥ 7 were used. A genome size of 140 Mb was used for all calculations.

100% similarity were reported.

Structural variant detection

Large (>100 bp) SVs were detected by aligning the Bionano scaffolds to the FlyBase [50] reference assembly using MUMmer v3.23 (`nucmer -maxmatch`) [83] and then annotating the disagreements between the assemblies as indels and duplications using SVMU v0.2beta (commit 4e65e95) [37]. Insertions overlapping with Repeatmasker 4.0.7 [142] annotated TEs were annotated as TE insertions. SVs were validated with at least two ONT reads spanning the entire genomic feature containing the SVs plus 200 bp on both sides of the SV. TEs were inferred to be segregating when corrected long reads supporting the TE insertions were contradicted by other reads showing absence of the TE. For validation with spanning long reads, we aligned Canu-corrected ONT reads to the FlyBase and Bionano assemblies using BLASR (v 1.3.1) [35] and the sorted alignment bam files were visually examined with IGV [152].

Mitochondrial genome identification

The mitochondrial genome was identified by using BLAST [6] to compare our final assembly (the Bionano assembly) against the mitochondrial genome from r6.16 of the *D. melanogaster* genome. A single contig was identified (tig00000438_pilon_pilon_obj) with 99% identity to the reference mitochondrial genome. This contig contained two copies of the mitochondrial genome in tandem (assembly duplications of circular genomes are not uncommon when doing assembly with long-read data), therefore the first 16,806 and last 2,104 nucleotides were removed from the contig. We used MITOS [22] with default settings, metazoan reference, and invertebrate genetic code to annotate both the reference mitochondrial genome and our assembled mitochondrial genome.

Data availability

The Illumina and basecalled ONT data generated in this study has been uploaded to the National Center for Biotechnology Information (`https://www.ncbi.nlm.nih.gov/`) under bioproject PRJNA433573. Genomes assembled in this study are available at `https://github.com/danrdanny/Nanopore_ISO1`. Releases 6.16 and 5.57 of the *D. melanogaster* genome used in this study are available on FlyBase (`http://www.flybase.org`). Bioinformatic scripts used in this pipeline are available at `https://github.com/esolares/DMI1`. Original data underlying this manuscript can be accessed from the Stowers Original Data Repository at `http://www.stowers.org/research/publications/libpb-1268`. Supplemental material available at Figshare: `https://doi.org/10.25387/g3.6813398`.

1.4 Results

Sequencing results

A 1D sequencing library was loaded onto a release 9.5 flow cell and run for approximately 24 hr (see Methods), generating a total of 663,784 reads. Basecalling was performed with Albacore 2.0.2, with 593,354 (89%) of all reads marked as “pass” (reads having a quality score $\geq$ 7) and an average fragment length of 7,122 bp (Table 1, Figure 2). The read N50 for those that passed filter was 11,840, with 41 reads longer than 50 kb and a maximum read length of 379,978 bp (Figure S1).

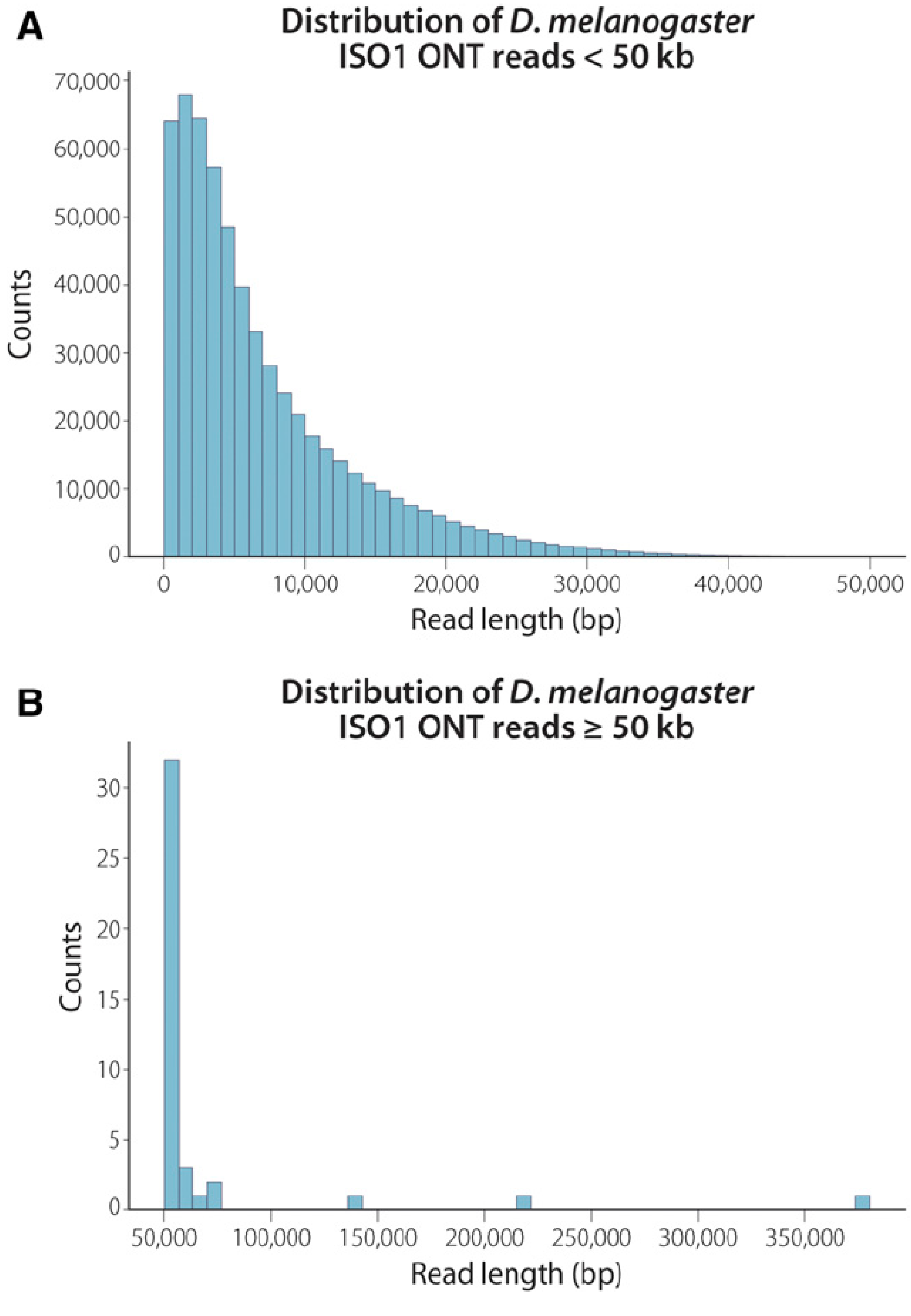

Figure 1.2: Read length distribution for reads with quality scores greater than or equal to 7. Length is the sequence length after basecalling by Albacore, not the length that aligned to the genome. (A) Distribution of read lengths less than 50 kb. (B) Distribution of reads 50 kb or greater. The longest read that passed quality filtering was 380 kb.

Genome assembly

ONT-only assembly using minimap/miniasm: To evaluate ONT data for *de novo* genome assembly, we performed an ONT-only assembly using minimap and miniasm [90]. Together, these programs allow for rapid assembly and analysis of long, error-prone reads. We generated an assembly with a total size of 132 Mb, with 208 contigs and a contig N50 of 3.8 Mb (N50 is the length of the contig such that 50% of the genome is contained within contigs that are equal to or longer than this contig) (Table 2). Evaluation of an alignment dot plot between this assembly and the *D. melanogaster* reference genome revealed high correspondence between our assembly and the reference genome (Figure S2). However, BUSCO analysis of the minimap assembly found only 0.5% of expected single-copy genes present—much lower than the BUSCO score of 98.3% obtained from the current release of the *D. melanogaster genome* (Table 3). BUSCO analysis evaluates the presence of universal single-copy orthologs as a proxy of completeness. Such a low BUSCO score as found in the minimap assembly is unlikely to be measuring low completeness, but rather suggests a high rate of errors that disrupt the genes, making them difficult to properly assay.

ONT-only assemblies using Canu: We also generated an ONT-only assembly with Canu using only reads longer than 1 kb. Alignment dot plots for the assembled genome and the *D. melanogaster* reference genome also revealed large colinear blocks between this assembly and the reference, indicating only one large misassembly on chromosome 2L (this misassembly was broken prior to merging and polishing) (Figure 3A). The Canu assembly was marginally less contiguous (contig N50 = 3.0 Mb) than the minimap assembly, but resulted in a higher BUSCO score of 67.7% (Table S1). The errors in the Canu assembly and the low BUSCO score are consequences of inherently high error rate of ONT reads. However, due to the higher accuracy and completeness of the Canu assembly compared to the minimap assembly, we used the ONT-only Canu assembly for the remainder of our analysis.

Hybrid assembly using ONT and Illumina reads: Modest coverage assemblies of

PacBio long reads can exhibit high contiguity when a hybrid assembly method involving Illumina reads is used [36]. Therefore, we examined whether such assembly contiguity improvements also occur when ONT long reads are supplemented with Illumina paired-end reads. We performed a hybrid assembly using DBG2OLC with the longest 30x ONT reads and contigs from a De Bruijn graph assembly constructed with 67x of Illumina paired-end data. To optimize the parameters, we performed a gridsearch on four parameters (AdaptiveTh, KmerSize, KCovTh and MinOverlap), yielding 36 genome assemblies. We used a range of values recommended by the authors for low-coverage assemblies and verified our KmerSize by looking at meryl's kmer histogram and found it coincided with a value that represented a 99% fraction of all k-mers (`http://kmer.sourceforge.net/wiki/index.php/Getting_Started_with_Meryl`). We selected the best genome based on colinearity and largest N50. The best hybrid assembly was substantially more contiguous (contig N50 = 9.9 Mb) than the ONT-only Canu assembly (contig N50 = 3.0 Mb), had large blocks of contiguity with the reference (Figure 3B), yet had lower BUSCO scores (47.7% compared to the 67.7% observed in the Canu assembly) (Table 3).

Merging of Canu and DBG2OLC assemblies: We have previously shown that merging assemblies constructed using only PacBio long reads with hybrid assemblies constructed with PacBio long reads and Illumina paired-end short reads results in a considerably more contiguous assembly than either of the two component assemblies alone [36]. To examine the effect of assembly merging on assemblies created with ONT long reads, we merged the ONT-only Canu assembly with the DBG2OLC assembly with two rounds of quickmerge (see Methods). The merged assembly (contig N50 = 18.6 Mb) was more contiguous than both the Canu and the DBG2OLC assemblies alone (Table 2), exhibiting large-scale colinearity with the reference genome as seen in the dot plot (Figure 3C). As expected, the BUSCO score of the merged assembly (58.2%) fell between the two component assemblies (Table 3).

Assembly quality

Name	Genome size (bp)	Contigs	Largest contig (bp)	N50 (bp)	L50
FlyBase r6.16*	142,573,024	2,442	27,905,053	21,485,538	3
MiniMap	131,856,353	208	16,991,501	3,866,686	9
DBG2OLC	131,359,678	339	13,129,070	9,907,730	6
Canu	139,205,737	295	14,326,064	2,971,262	11
QuickMerge 2x	138,130,519	250	25,434,901	18,616,266	4
QM2x Nanopolish	139,303,903	250	25,367,201	18,818,677	4
QM2x NP + Pilon x2	140,153,080	250	25,783,280	18,923,871	4
QuickMerge 2x Bionano	142,817,829	231	28,580,427	21,305,147	3

Table 1.2: Genome assembly statistics. *Values are for scaffolds, not contigs.

Polishing: A high number of SNP and indel polymorphisms in all of our assemblies is consistent with other *de novo* assemblies created with noisy long reads with high error rates [97, 36]. Such errors typically lead to error-ridden genic sequence that can give the appearance that many important genes are missing (Table 3; Table S1). Many of these errors can be fixed via assembly polishing, and a number of assembly polishing tools exist. We chose to employ two: nanopolish, which performs consensus-based error correction using ONT reads [97], and Pilon, which performs error correction using Illumina reads [158]. This approach of applying long-read consensus correction followed by short-read polishing has resulted in BUSCO scores comparable to that of the FlyBase reference genome (e.g., Chakraborty et al. 2016 and Chakraborty et al. 2018).

To evaluate this approach for ONT data, we evaluated three different polishing approaches: one using nanopolish alone, one using only two rounds of Pilon, and one using one round of nanopolish followed by two rounds of Pilon (see Methods). Applying nanopolish alone recovered only 79%, 79.2%, and 78.5% complete BUSCOs for the hybrid, ONT-only, and the merged assemblies, respectively (Table S2), suggesting that polishing with ONT reads alone only partially improved assembly quality. On the other hand, polishing all three assemblies twice with Pilon alone fixed a large number of errors as evidenced by improved BUSCO scores of the resulting assemblies [141] (Table 3). The merged assembly was polished once with nanopolish, then twice with Pilon, resulting in nearly all complete BUSCOs (97.9%) being present in our polished assembly, comparable to that of the reference assembly (98.3%)

(Table S2). Variation in BUSCO scores generally tracked the method of polishing more than the type of assembly. However, the hybrid assembly did recover a slightly different subset of BUSCOs than the ONT-only assembly, leading to a slightly higher number of BUSCOs being recovered in the final merged assembly.

QUAST/GAGE metrics of quality: The QUAST output comparing the Canu, DBG2OLC, merged, and Bionano genome assemblies against the *D. melanogaster* reference shows that the quickmerge assembly resulted in intermediate error rates and discordance as compared to the component assemblies (Figure 4). All four assemblies exhibited approximately the same number of SNP and indel errors less than 5 bp, whereas the DBG2OLC assembly resulted in approximately 20% more indels greater than 5 bp in size than the Canu assembly (Figure 4H-I, Table S3, Table S4). Among ONT assemblies, the Canu assembly exhibited superior accuracy and fewer misassemblies, although our quickmerge assembly was nearly as good. When mismatches are measured on a phred scale, the ONT assemblies range in quality from 32 to 33, which is lower than the score of 37 from a PacBio assembly with threefold more data [82] (Table S3, Table S4). Canu also exhibited the lowest raw contiguity for an ONT assembly as measured by N50/NG50 and L50/LG50. The N50 for our merged assembly (19 Mb) exceeded that of Canu (3 Mb) and was nearly double that of DBG2OLC (10 Mb), while maintaining large-scale colinearity with the reference (Figure 3C). However, Canu outperforms the merging approach when considering the contiguity of error-free alignment blocks as demonstrated by the NA50 and NGA50 for Canu, which are greater than that of quickmerge (Table S3). We also observe that half of the assemblies accrue in fewer error-free alignments for Canu than for quickmerge, resulting in slightly lower values in LA50 and LGA50 for Canu than quickmerge (Table S3). As a consequence, when evaluating which approach to employ, we advise users to carefully weigh the tradeoffs between large-scale contiguity and local misassembly and sequence errors and to tailor their decision to the biological questions being addressed. For example, in creating a reference for QTL mapping, there might be a strong preference for high long-range contiguity even at the cost of local misassemblies.

Name	Single copy	Duplicate	Fragmented	Missing	Complete
FlyBase r6.16	2,749 (98.2%)	14 (0.5%)	22 (0.8%)	14 (0.5%)	2,763
MiniMap	14 (0.5%)	0 (0.0%)	31 (1.1%)	2,754 (98.4%)	14
DBG2OLC	1,332 (47.6%)	3 (0.1%)	557 (19.9%)	907 (32.4%)	1,335
Canu	1,884 (67.3%)	11 (0.2%)	557 (19.9%)	347 (12.4%)	1,895
QuickMerge (QM) 2x	1,623 (58.0%)	6 (0.3%)	560 (20.0%)	610 (21.8%)	1,629
QM 2x Nanopolish (NP)	2,189 (78.2%)	8 (0.3%)	400 (14.3%)	202 (7.2%)	2,197
QM 2x NP + Pilon x2	2,726 (97.4%)	14 (0.5%)	39 (1.6%)	20 (0.7%)	2,740
QM 2x Pilon x2	2,718 (97.1%)	14 (0.5%)	45 (1.4%)	22 (0.8%)	2,732
QM 2x Bionano	2,715 (97.0%)	15 (0.5%)	40 (1.4%)	29 (1.0%)	2,730
QM 2x Bionano All	2,720 (97.2%)	16 (0.6%)	40 (1.4%)	23 (0.8%)	2,736

Table 1.3: Busco scores demonstrating genome quality before or after polishing.

On the other hand, when avoiding misassemblies is paramount (as might be the case when characterizing the structure of individual loci), one could make the argument in favor of less long-range contiguity in exchange for fewer misassemblies.

Finally, the low-coverage ONT dataset presented here performed surprisingly well compared to a PacBio dataset three times larger [80, 82] in terms of contiguity (N50, NA50, L50, LA50), completeness (genome fraction), and accuracy (identity, SNPs, InDels, translocations, etc.) (Table S3). In all measures, the merged assembly performed almost as well as the much higher coverage PacBio assembly from Koren et al. (2017). Importantly, however, the GAGE corrected N50 of Koren et al. (2017) was substantially larger than the ONT assemblies, likely due to its superior coverage (Table S3, Table S4).

Completeness: reconstructing genes, TEs, and chromosome arms: We next evaluated the completeness of the assemblies by assessing the reconstruction of elements in the genome, including genes, TEs, and chromosome arms. For all following metrics, Canu and the final Quickmerge assembly performed similar to one another, with DBG2OLC performing somewhat worse. As a result, remaining comparisons will be made between the final Quickmerge ONT assembly and the PacBio assembly produced by Koren et al. 2017, which we use as a standard of comparison.

Our final assembly contained 96.37% of release 6.16 genes greater than 99% complete as compared to 98.4% for the Koren et al. (2017) assembly (Table S5). However, our final assembly was able to reconstruct only 72.71% of gene models compared to 88.14% for the Koren et al. (2017) assembly when only 100% complete gene models are considered (Table S5). Transposon reconstruction followed the same pattern with a 6.66% and a 15.51% difference in counts between the two assemblies for those with greater than 99% identity and complete reconstruction, respectively, with both favoring the higher-coverage PacBio assembly.

Muller element (chromosome arm) reconstruction was more similar between the ONT and PacBio assemblies for the euchromatic regions, with an average difference of 0.47% and a maximum difference of 0.94% in the X chromosome (Table S5). The largest difference was in the number of alignments. Bionano and quickmerge assemblies covered the Muller elements in fewer segments than the Koren et al. (2017) assembly, though they exhibited less total chromosome coverage. The differences were more apparent in the heterochromatic scaffolds, as our final assembly contained less coverage than the Koren et al. (2017) assembly, with an average deficit of 4.44% coverage and a maximum deficit of 7.89% across all heterochromatic scaffolds linked to Muller elements (Table S6). The difference in the Y chromosome was the largest, as expected, for our relatively low coverage of mixed sex flies compared to the high coverage of all males for Koren et al. (2017).

Base quality: We aligned Illumina short reads to measure the sequencing error rate following Berlin et al. (2015). The proportion of reads successfully mapped to assemblies varied from 93.5 to 94.6%. The aligned reads resulted in QV scores ranging from 40.2 to 41.1. A slightly different approach used by Koren et al. (2017) yielded very similar QV scores in the range of 40.1–40.2 (Table S7), suggesting that our assembly was of relatively high quality in areas where reads align well.

Bionano scaffolding

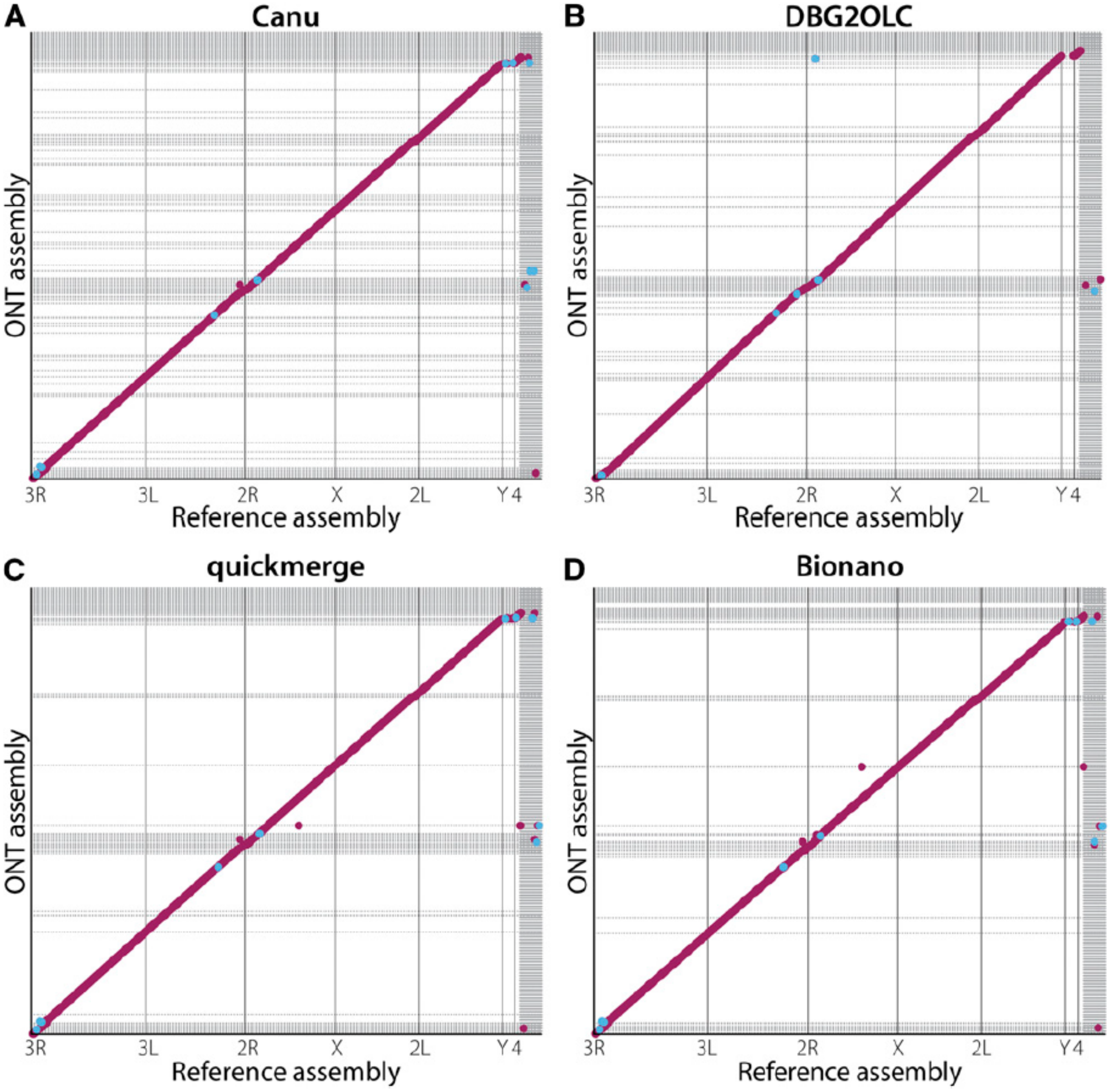

Figure 1.3: Dot plots showing colinearity of our assembled genomes with the current version of the D. melanogaster reference genome. Red dots represent regions where the assembly and the reference aligned in the same orientation; blue dots represent regions where the genomes are inverted with respect to one another. The vertical grid lines represent boundaries between chromosome scaffolds in the reference assembly. Horizontal grid lines represent boundaries between contig (A-C) or scaffolds (D) in the assemblies reported here. (A) Plot of the Canu-only assembly against the reference genome. (B) Plot of the hybrid DBG2OLC Nanopore and Illumina assembly against the reference. (C) Plot of merged DBG2OLC and Canu assemblies showing a more contiguous assembly than either of the component assemblies. (D) Bionano scaffolding of the merged assembly resolves additional gaps in the merged assembly.

Bionano fragments, which are substantially longer than ONT reads, can be used for scaffolding contigs across repetitive regions. We generated 81,046 raw Bionano fragments (20.5 Gb) with an average fragment length of 253 kb. To use these fragments for scaffolding, we first created a Bionano assembly (509 maps, haploid assembly size = 145 Mb) using 78,397 noise-rescaled fragments (19.9 Gb, mean fragment length 253 kb). At 145 Mb, the Bionano assembly is comparable to the reference assembly size (144 Mb). The Bionano maps were used to scaffold the merged assembly. The resulting scaffolded assembly was more contiguous (N50 = 21.3 Mb) than the unscaffolded contigs without substantially changing the number of SNP and indel errors (Figure 3D; Tables S1, S3-S4). Bionano scaffolding of our assembled genome did not change BUSCO scores (Table 3).

Structural variants

One advantage of high quality *de novo* assemblies is that they permit comprehensive detection of large (> 100 bp) SVs. Highly contiguous assemblies, such as the one generated here, allow comparisons between two or more assemblies, revealing novel SVs and facilitating the study of their functional and evolutionary significance. Although we sequenced the reference genome stock, structural differences between our stock and the published assembly are expected due to error, but also to new mutations—especially for TEs, which are the most dynamic structural components of the genome. Indeed, such mutations have been observed in substrains of ISO1 before [171, 108, 125]. Similarly, our assembly revealed the presence of 34 new TE insertions and 12 TE losses compared to the reference genome assembly. Among the 34 TE insertions, 50% (17/34) are LTR TEs comprising chiefly of copia (5/17) and roo elements (6/17). Interestingly, 29% (10/34) of the TE insertions are defective hobo elements that lack an average of 1.7 kb of sequence (base pairs 905–2510 of full-length hobo are absent) from the middle segment of the element encoding the transposase. However, alignment of long reads to the assembly regions harboring the TE insertions revealed that 6/34 insertions are not fixed, but are segregating in the strain (Table S2). The high insertion rate of the LTR

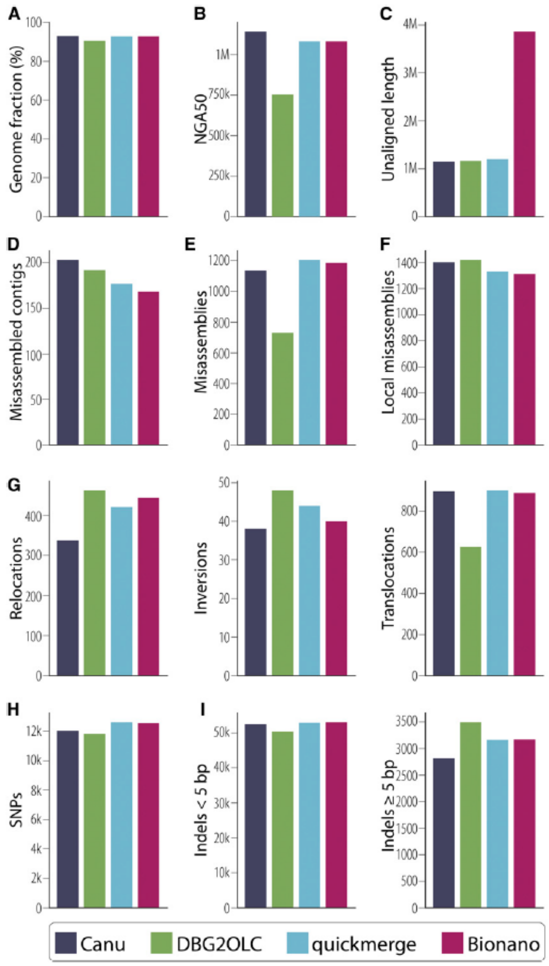

Figure 1.4: QUAST was used to compare each assembly to the D. melanogaster reference genome with selected statistics presented here. (A) Greater than 90% of bases in the reference genome were aligned to each of our four assemblies. (B) The contiguity of assembly blocks aligned to the reference. (C) Total unaligned length includes contigs that did not align to the reference as well as unaligned sequence of partially aligned contigs. (D) The number of contigs that contain misassemblies in which flanking sequences are 1 kb apart, flanking sequences overlap by 1 kb or more, or flanking sequences align to different reference scaffolds. (E) Total count of misassemblies as described in (D). (F) Local misassemblies include those positions in which a gap or overlap between flanking sequence is less than 1 kb [(D) and (E) show those greater than 1 kb] and larger than the maximum indel length of 85 bp on the same reference genome scaffold. (G) Misassemblies can be subdivided into relocations (a single assembled contig aligns to the same reference scaffold but in pieces at least 1 kb apart), inversions (at least one part of a single assembled contig aligns to the reference in an inverted orientation), or translocations (at least one part of a single assembled contig aligns to two different reference scaffolds). Not all misassemblies are captured in these three categories. (H) Total SNPs per assembly are shown and were not significantly different among assemblies. (I) Indels per 100 kb can be divided into small indels (those ,5 bp) and large indels (> 5 bp). Indels .85 bp are considered misassemblies and are shown in panels D, E, or F.

and hobo elements in this single strain mirrors the recent spread of LTR and hobo elements in *D. melanogaster* populations [116, 118, 24]. As expected, the majority (27/34) of the new TE insertions are located within introns, because insertions within exons generally result in gene disruption. Nonetheless, we found five genes (Ance, Pka-C1, CG31826, CG43446, and Ilp6) in which new TEs have inserted within exons (Table S2). We also found 12 TEs present in the reference assembly that are missing from our final scaffolded assembly, among which six are LTR TEs (five 297 elements and one roo element). The high rate of insertion and loss of LTR elements underscores their dynamic evolutionary history in the *D. melanogaster* genome. Because TEs can be locally unique, the presence or absence of such events does not pose a fundamental limitation to assembly when reads are long enough to span the TEs. Consequently, we predict most of these events to be new mutations rather than errors in assembly.

Additionally, we identified several duplications present in our assembly. For example, a tandem duplication of a ~9-kb segment has created partial copies of the genes infertile crescent (ifc) and little imaginal discs (lid). Another ~2-kb tandem duplication has created partial copies of the genes CG10137 and CG33116 (Table S8). Apart from CNVs affecting single copy sequences, our assembly also uncovered copy number increases in tandem arrays with potential functional consequences. For example, we observed copy number increase in a tandem array of a 207-bp segment within the third exon of the chitin-binding protein gene Muc26B. While CNVs such as this have been challenging to identify and validate in the past, at least two Nanopore reads spanning this entire tandem array support the presence of a tandem duplication at this position (Figure 5). Unlike TEs, classifying such tandem events as errors stemming from shorter Sanger reads or an actual mutation is difficult without the access to the original material from which the Sanger data were derived.

Mitochondrial genome identification and identity

After assembly, merging, polishing, and scaffolding, we used BLAST [6] to identify the mi-

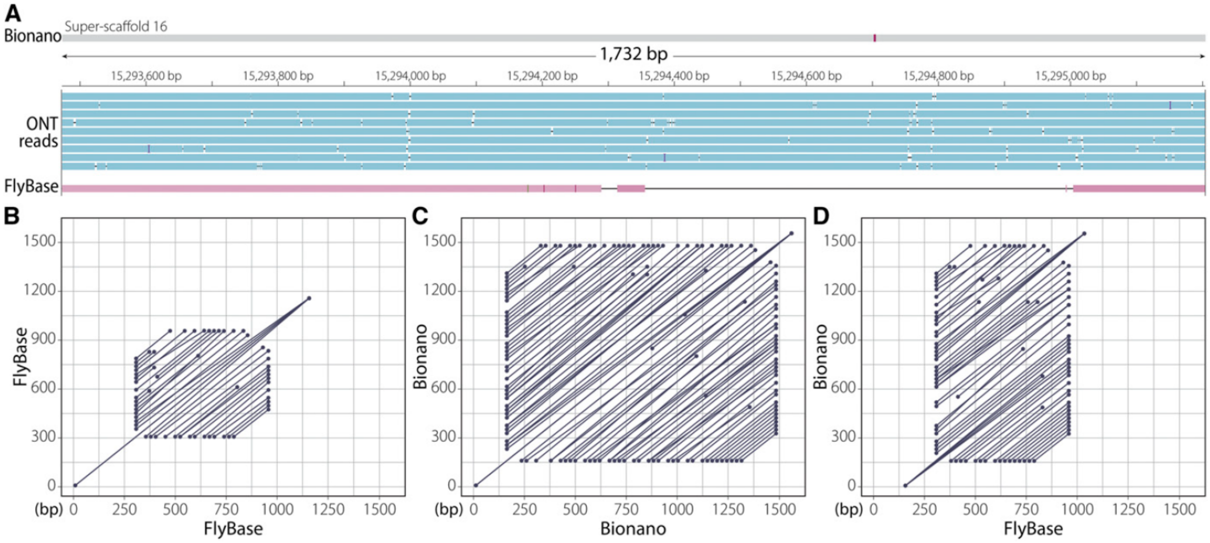

Figure 1.5: Copy number increase in a 207-bp tandem array located inside the third exon of Muc26B. (A) Three tracks showing Bionano assembly (top) with Nanopore long reads (blue) and reference (Flybase) assembly (red) aligned to it. The alignment gap in the reference assembly is due to the extra sequence copies in the Bionano assembly. (B) Alignment dot plot between the reference sequence possessing the tandem array to itself. (C) Alignment dot plot between the genomic region possessing the tandem array in the Bionano assembly to itself. As evidenced by the dot plot, the Bionano assembly has more repeats in this region than the reference assembly in panel B. (D) Alignment dot plot between the reference genomic region (x axis) shown in (B) and the corresponding Bionano genomic region (y axis) shown in (C).

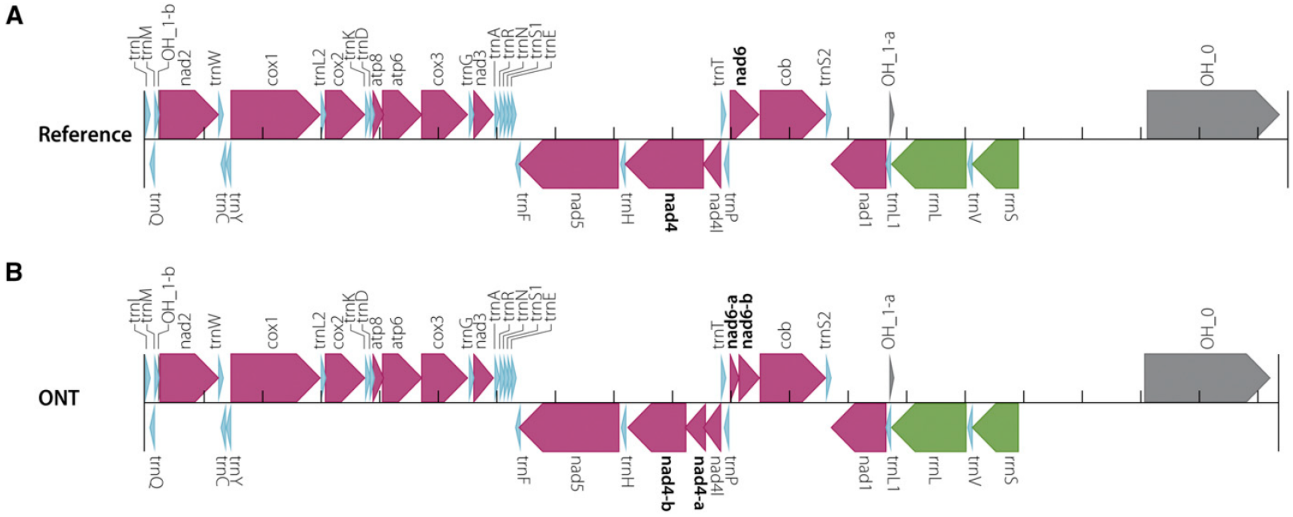

Figure 1.6: Mitochondrial genome annotations generated by MITOS. (A) Annotation of the reference mitochondrial genome. (B) Annotation of the mitochondrial genome assembled in this project is identical to the reference except that nad4 and nad6 in the reference assembly were both annotated as two genes—nad4 as nad4-a and nad4-b, and nad6 as nad6-a and nad6-b.

tochondrial genome. Using the published *D. melanogaster* mitochondrial genome as the subject, we identified one 38,261-bp contig with nearly 100% identity to the reference mitochondrial genome. The first 16,806 and last 2,100 nucleotides of this contig were 99.6% identical to the reference mitochondrial genome, while the middle 19,228 nucleotides were 98% identical to the reference mitochondrial genome. This suggests that our assembled mitochondrial genome had been duplicated during assembly, which commonly occurs when assembling circular genomes using long sequencing reads. For the tandem assembled genomes, nearly all of the SNP and indel polymorphisms occurred in the last 4,000 nucleotides of the reference genome.

To determine if all genes and features that are present in the reference mitochondrial genome were present in our assembled genome, we annotated both genomes using MITOS [22]. Both assemblies were annotated nearly identically, with MITOS reporting that both assemblies were missing the origin of replication for the L region (OL) (Figure 6). The Nanopore assembly also resulted in two split genes not observed in the reference genome assembly:

nad4 and nad6 were each annotated as two continuous genes rather than one single gene.

1.5 Discussion

Drosophila melanogaster was the first genome assembled using a whole-genome shotgun (WGS) strategy [112, 2]. This successful proof-of-principle led to the prevalence of the WGS sequencing approach as a tool in virtually all subsequent metazoan genome assembly projects [87, 160, 60, 170, 8, 59, 68]. While improvements in sequencing technology have led to a precipitous drop in the cost of sequencing [162], stagnation and even regression of read lengths resulted in highly fragmented and incomplete assemblies [5]. Furthermore, complementary approaches (like hierarchical shotgun sequencing and other clone-based approaches) were required to obtain nearly complete and highly contiguous reference genomes, which adds complexity, cost, and time to assembly projects.

Short reads provided by next generation sequencing technologies present limitations to what a pure WGS assembly approach can accomplish [5, 113]. The advent and development of long-read sequencing technologies has led to dramatic increases in read length, permitting assemblies that span previously recalcitrant repetitive regions. Early implementations of these technologies produced reads longer than previous short-read technologies yet still shorter than relatively common repeats, while the cost and error rate remained high compared to the short-read approaches. However, continued improvements in read length overcame many of these difficulties, permitting nearly complete, highly contiguous metazoan genome assemblies with only a WGS strategy [80, 21, 82]. Such approaches led to assemblies comparable in completeness and contiguity to release 6 of the *D. melanogaster* genome assembly [70] for approximately $10,000 USD [36, 37]. However, even with the rapid development of these approaches, substantial capital investment in the form of expensive instrumentation and dedicated genome facility staff was required.

Here, we report an independent resequencing and assembly of the *D. melanogaster* reference strain, ISO1, for less than $1,000 USD in sequencing costs and without the need for extensive capital or personnel investment. The resulting assembly before scaffolding is comparable to release 6 of FlyBase in terms of contiguity and completeness (18.9 Mbp contig N50 and 97.1% complete single copy BUSCOs). We achieved this using 4.2 Gbp of sequence from a single Oxford Nanopore flow cell in conjunction with Illumina short-read data. Such reduction in complexity and cost permits a small team of scientists to shepherd a sequencing project from sample to near-reference-quality assembly in a relatively short amount of time. The addition of optical mapping data permitted ordering and orientation of the contigs, yielding an assembly nearly as contiguous as the published reference (scaffold N50 of 21.3 Mbp).

Comparing this assembly to the FlyBase reference genome shows that it is both accurate (21 mismatches/100 kbp, 36 indels/100 kbp) and colinear (Figure 3D, Figure 4, Table S3). Most of the small-scale differences are expected to be errors introduced by the noisy Oxford Nanopore reads that escaped correction via polishing. It is possible that some of these errors are SVs, which are expected to accumulate because the ISO1 stock has been maintained in the laboratory for approximately 350 generations (assuming 20 gen/year) since initial sequencing in 2000. This allows for the accumulation of new mutations by genetic drift, including ones reducing fitness [12]. Due to the high contiguity in the euchromatic region, our assembly facilitates detection of such euchromatic SVs. Several apparent "assembly errors" in our Bionano assembly are due to TE indels that are supported by spanning long reads. We found 28 homozygous euchromatic TE insertions, which are predominantly LTR and defective hobo elements, suggesting a high rate of euchromatic TE insertions (~0.08 insertion/gen). That we observed a predominance of LTR and hobo elements among the new TE insertions mirrors their recent spread in *D. melanogaster* populations [116, 118, 24, 49]. The abundance of defective hobo elements among the new insertions is particularly interesting given that these hobo elements lack the transposase enzyme necessary for mobilization.

Although most novel TE insertions were found in introns, five were found within exons: the 5 prime UTR of Ance, CG31826, and Ilp6; the 3 prime UTR of Pka-C1; and the coding region of CG43446. Similarly, TE loss primarily involved LTR TEs, including loss of TEs from the 5 prime UTR of the genes Snoo and CG1358. We also observed copy number increases both in unique sequences as well as in tandem arrays (Table S8), with one duplication creating a new copy of the entire coding sequence of the gene lid. Collectively, our assembly provides a snapshot of ongoing genome structure evolution in a metazoan genome, which is often assumed to be approximately invariant for experimental genetics.

A crucial feature of this work is that it is performed in a strain used to generate one of the highest quality reference genomes available, ensuring that our inferences can be judged against a high-quality standard. This approach allowed us to demonstrate that assembly with modest amounts of long-molecule data paired with inexpensive short-read data can yield highly accurate and contiguous reference genomes with minimal expenditure of resources. This demonstration opens myriad opportunities for high-quality genomics in systems with limited resources for genome projects. Moreover, we can now conceive of studying entire populations with high-quality assemblies capable of resolving repetitive structural variants, something previously unattainable with short-read sequencing alone.

Chapter 2

HapSolo: An optimization approach for removing secondary haplotigs during diploid genome assembly and scaffolding

2.1 Abstract

Background

Despite marked recent improvements in long-read sequencing technology, the assembly of diploid genomes remains a difficult task. A major obstacle is distinguishing between alternative contigs that represent highly heterozygous regions. If primary and secondary contigs are not properly identified, the primary assembly will overrepresent both the size and complexity of the genome, which complicates downstream analysis such as scaffolding.

Results

Here we illustrate a new method, which we call HapSolo, that identifies secondary contigs and defines a primary assembly based on multiple pairwise contig alignment metrics. HapSolo evaluates candidate primary assemblies using BUSCO scores and then distinguishes among candidate assemblies using a cost function. The cost function can be defined by the user but by default considers the number of missing, duplicated and single BUSCO genes within the assembly. HapSolo performs hill climbing to minimize cost over thousands of candidate assemblies. We illustrate the performance of HapSolo on genome data from three species: the Chardonnay grape (*Vitis vinifera*), with a genome of 490 Mb, a mosquito (*Anopheles funestus*; 200 Mb) and the Thorny Skate (*Amblyraja radiata*; 2650 Mb).

Conclusions

HapSolo rapidly identified candidate assemblies that yield improvements in assembly metrics, including decreased genome size and improved N50 scores. Contig N50 scores improved by 35%, 9% and 9% for Chardonnay, mosquito and the thorny skate, respectively, relative to unreduced primary assemblies. The benefits of HapSolo were amplified by down-stream analyses, which we illustrated by scaffolding with Hi-C data. We found, for example, that prior to the application of HapSolo, only 52% of the Chardonnay genome was captured in the largest 19 scaffolds, corresponding to the number of chromosomes. After the application of HapSolo, this value increased to ~84%. The improvements for the mosquito's largest three scaffolds, representing the number of chromosomes, were from 61 to 86%, and the improvement was even more pronounced for thorny skate. We compared the scaffolding results to assemblies that were based on PurgeDups for identifying secondary contigs, with generally superior results for HapSolo.

2.2 Introduction

Traditionally, reference genomes have been produced from genetic materials that simplify assembly; for example, the first two plant species targeted for reference quality genomes, *Arabidopsis thaliana* [172] and rice (*Oryza sativa*) [60], were chosen in part because they naturally self-fertilize and are therefore highly homozygous. Other early genomes, such as those from *Caenorhabditis elegans* and *Drosophila melanogaster* [1, 2], were also based on inbred, highly homozygous materials. Recent sequencing of additional model and non-model species have continued to rely on near-homozygous materials, either through inbreeding [143, 77] or by focusing on haploid tissue [44, 126].

The reliance on homozygous materials is fading rapidly, however, for at least three reasons. The first is that it has become clear that inbred materials can misrepresent the natural state of genomes. A dramatic illustration of this fact is that some lines of maize purged 8% of their genome in only six generations of self-fertilization [134]; more generally, inbred genomes tend to be smaller than those based on outbreeding species [120, 54]. The second is that many species of interest cannot be easily manipulated into a homozygous state. Many animals fall into this category, such as mosquitoes [135], as do many perennial crops like grapes, which are highly heterozygous [175] and can be selfed but only with substantial fitness costs that limits homozygosity [105]. Finally, some important features and phenotypes—such as sex determination [100] and other important adaptations—can only be identified by analyzing heterozygous samples.

Fortunately, the resolution of highly heterozygous regions, which often contain large structural variants, is now possible due to improvements in sequencing technologies and their affordability. In theory, long-read sequencing technologies, like those from Pacific Biosciences and Oxford Nanopore, provide the capability to resolve distinct haplotypes in heterozygous regions, leading to the assembly of reference-quality diploid genomes [143, 76, 36]. Several genomes based on highly heterozygous materials have been published recently [175, 157, 42, 58, 81, 130], with many additional efforts ongoing.

Nevertheless, the assembly of heterozygous genomes still presents substantial challenges. One challenge is resolving distinct haplotypes in regions of high heterozygosity. Programs that assemble long-reads, such as FALCON and Canu [81], can fuse distinct haplotypes into the primary assembly. This haplotype-fusion produces genomes that are much larger than the expected genome size. When haplotypes are fused, either into the same contig or as different contigs into the primary assembly, the increased size and complexity of the assembly complicates down-stream approaches, such as scaffolding by Hi-C or optical mapping. In theory, FALCON-unzip [42] solves some problems by identifying alternative (or 'secondary') haplotigs that represent the second allele in a heterozygous region and then providing a primary assembly without secondary contigs.

It remains a difficult problem to identify and remove alternative contigs during assembly, but there are some suggested solutions. For example, Redundans identifies secondary contigs via similarities between contigs [121] and removes the shorter of two contigs that share some predefined level of similarity. Another approach, PurgeHaplotigs uses sequence coverage as a criterion to identify regions with two haplotypes [132]. The reasoning behind PurgeHaplotigs is that alternative alleles in a heterozygous region should have only half the raw sequence coverage of homozygous regions. Accordingly, the algorithm proceeds by first remapping raw reads to contigs, then flagging contigs with lower than expected read depth, and finally re-mapping and removing low-coverage contigs from the primary haplotype-fused assembly. A more recent approach, implemented in the purge_dups tool [63], builds on the coverage-based approach of PurgeHaplotigs. Purge_dups has been compared to PurgeHaplotigs and is superior based on a few exemplar assemblies [63].

Here we report another strategy, which we call HapSolo [144], to identify and remove potential secondary haplotigs. Our approach is similar to Redundans, in that it begins with an all-by-all pairwise alignment among contigs and uses features of sequence alignment as a basis to identify potential alternative haplotigs. However, HapSolo is unique in exploring the parameter space of alignment properties to optimize the primary assembly, using features of BUSCO scores as the optimization target. Here we detail the approach and implementation of HapSolo, demonstrate that it efficiently identifies primary versus secondary haplotigs and show that it improves Hi-C based scaffolding outcomes relative to purge_dups. HapSolo has been implemented in python and is freely available (`https://github.com/esolares/HapSolo`).

2.3 Methods

Pre-processing

Our method begins with the set of contigs from genome assembly. In theory, HapSolo will work for any set of contigs from any assembler and from any sequencing type (i.e., short-read, long-read or merged assemblies). Given the set of contigs, the first steps are to size sort the contigs and then to perform an all-by-all pairwise alignment among all contigs (Fig. 1, steps 1 and 2), using each contig as both a reference and a query. In theory, pre-processing alignments can be performed with any algorithm, with the HapSolo implementation supporting either BLAT [79, 92] or minimap2 [91] input files.

Steps within HapSolo

HapSolo imports alignment results into a PANDAS (`https://pandas.pydata.org/`) dataframe to form a table with rows representing pairs of aligned contigs and columns containing descriptive statistics for each pairwise comparison (Fig. 1, step 3). Columns include the percent nucleotide identity between contigs (ID), a metric similar to those used in previous haplotig reduction programs; the proportion of the query contig length that aligns to the reference contig (Q), which is included to recognize that alignments can be clipped; and the ratio of the proportion of the query aligned to the reference relative to the proportion of the reference aligned to the query (QR). QR is considered because it reflects properties of aligned length and potential structural variant differences between contigs. A downside of QR is that it can reach values > 1.0, as longer variants may exist in either the query or the reference, and it is also non-symmetric. To compensate for this we include a symmetric value, which we define as $QR' = e \cdot -log2(QR)$. The four parameters— ID, Q, QR and QR'—are the basis for filtering query contigs from the table and defining them as putative secondary contigs. For simplicity, however, we will emphasize QR, because QR' is dependent on QR.

In addition to the alignment table, HapSolo generates a table of BUSCO properties [141] for each contig. This BUSCO analysis is performed on each contig of the assembly prior to running HapSolo's reduction algorithm. To perform these analyses, contigs are split into individual FAStA files and then BUSCO v3.0.2 is run on each contig separately so that they can be evaluated in parallel. Ultimately, the BUSCO table generated by HapSolo contains a list of complete (C) and fragmented (F) BUSCO genes for each contig. This table is integral for rapidly evaluating potential candidate assemblies.

Given the alignment table and the BUSCO table, HapSolo begins by assigning threshold values for ID, Q and QR, which we denote as ID_T, Q_T and QR_T. The threshold values can be assigned randomly, with set default values or with values defined by the user. The threshold values are applied to the alignment table to identify query contigs for purging. To be removed, a query contig must be in a pairwise alignment that satisfies three conditions: (1) an ID value $\geq ID_T$; (2) a Q value $\geq Q_T$; and (3) a QR value that falls within the range $\min(QR_T,QR'_T)$ and $\max(QR_T,QR'_T)$. After purging query contigs, HapSolo calculates the number of Fragmented (F), Missing (M), Duplicated (D) and Single-Copy (S) BUSCO genes across all of the primary contigs that remain in the candidate assembly, based on values in the BUSCO table. It then calculates the Cost of the candidate assembly as:

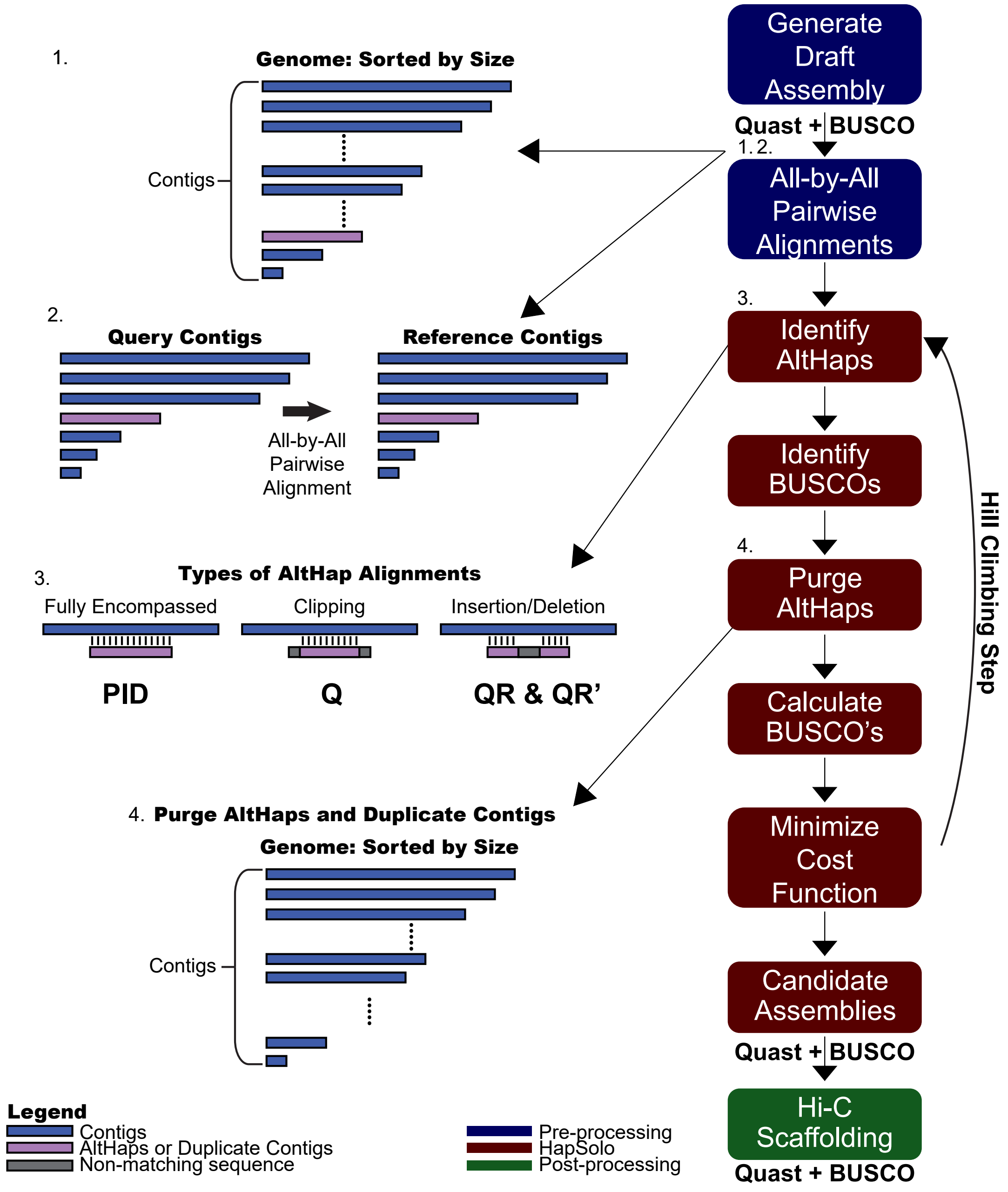

Figure 2.1: A schematic showing the basic workflow and ideas behind HapSolo. The rectangles on the right illustrate the basic steps, including pre-processing (blue rectangles), steps within HapSolo (red rectangles) and post-processing (green rectangle). Some of the HapSolo steps include iterations to perform hill climbing calculations, as described in the text and shown by the arrow. On the left, step 1 shows the contigs from the primary assembly, and step 2 illustrates the all-by-all alignment of contigs. Step 3 provides examples of some properties of potential alignments. The metrics—ID, Q and QR—were defined to help capture some of the variation among these conditions. Step 4 illustrates that new primary assemblies are formed by dropping putative secondary contigs.

$$Cost = (\theta_1 M + \theta_2 D + \theta_3 F)/\theta_4 S$$

where θ_1, θ_2, θ_3, and θ_4 are weights that can vary between 0.0 and 1.0. Weights can be assigned by users; for all of our analyses below, we employ weights of 0.0 for F and 1.0 for M, D and S.

We then employ hill climbing to minimize Cost (Fig. 1). Once Cost is calculated with random starting values, ID_T, Q_T and QR_T are modified at each iteration by a randomized step in the positive direction, which in turn defines a new set of primary contigs for a new cost evaluation. The steps consist of a fixed increment, which can be set by the user but is set to 0.0001 by default, multiplied by a random value sampled from U(0,1). As such, HapSolo utilizes a randomized forward walking agent to traverse the search space. If Cost does not change with new parameter values for a specified number of steps or if parameters increase past their maximum limits of $ID_T = Q_T = 1.00$, then HapSolo assigns new random values of ID_T, Q_T and QR_T. The process is repeated for n total iterations, and the iteration(s) with the smallest Cost are used to define the final set of primary contigs. When there are multiple solutions that minimize Cost, we retain all unique solutions; these additional solutions can be exported by the user for post-processing steps and evaluation. The values that determine the behavior of this minimization—e.g., the threshold for the number of consecutive cost plateaus, the number of x unique best candidate assemblies retained, the increase in step size by a fixed value, and the total number of iterations—can be set by the user.

To retain candidate assemblies with the smallest Cost, we implemented a unique priority queue (UPQ). The UPQ maintains a maximum number of x best assemblies, where x can be set by the user. The UPQ initially takes a list of one set of values, the score, primary contigs and other assembly information. The UPQ then takes the number of primary contigs for each of the candidate assemblies and sorts them by size. It then compares only the candidate assemblies of the same size, because assemblies of unequal size cannot be the same assembly.

Therefore our algorithm, in order to reduce the number of contig set comparisons, only compares contig sets of the same size. Once it is established that the candidate assemblies of the same number of contigs are equal, only the candidate assembly with the lowest score is saved. The list is then sorted by score and returned. This allows retention of the max score of the best x number of assemblies by looking at the score of the last candidate assembly in the list, giving $O(1)$ access to this value. Sorting takes $O(x \cdot log(x))$, where x is the best number of candidate assemblies to return, giving our UPQ a time complexity of $O(x \cdot log(x))$. Since we can instantaneously access the worst of the x candidate assemblies, we then perform an integer comparison of the score of our current candidate assembly with the worst score of our best x number of assemblies, reducing our computational time complexity. Only assemblies with the same or lower scores than the worst candidate assembly are then added to our UPQ. This reduces our total time complexity to $O(i \cdot x \cdot log(x))$ where i is the number of iterations which produce scores lower than our max of x, and x is the number of best candidate assemblies to keep.

Post-processing

Once HapSolo converges on a set (or x sets) of primary contigs that minimize Cost, the contig set is employed for post-processing to evaluate the candidate assembly. Specifically, we run QUAST v4.5 [64] and BUSCO 3.0.2 on the set of primary contigs that represent the best (or set of x best) candidate assemblies. QUAST measures basic genome assembly statistics, such as, N50, total assembly length, L50 and the largest contig size. Although not part of the HapSolo method, we provide scripts that run QUAST and BUSCO to output their results into a single score file.

Implementation and requirements

HapSolo has been implemented and optimized for Python 2.7, but it is also supported under Python 3. However, we recommend using Python 2.7, for faster run times. HapSolo requires the input of a contig assembly (as a FAStA file), the location of a directory for individual contig BUSCO results, and the input of pairwise alignments. It currently supports either BLAT or minimap2 alignment output files (PSL or PAF or compressed PSL.gz or PAF.gz file).

Species and data

The data for the assemblies for *V. vinifera* (cultivar Chardonnay) [175], *A. funestus* (mosquito) [58], and *A. radiata* (thorny skate) [128] were downloaded from public databases (see Data Availability). As mentioned, the contig assemblies were based on PacBio data. The chromosome number for each species was found in various sources [58, 128, 75]. The *P. leucopus* data were published in [98].

Pre-processing

For each genome, pre-processing prior to application of HapSolo consisted of all-by-all pairwise contig alignments, as described above. For this study, we used BLAT v35 [92] and minimap2 [91]. BLAT was run with default options after the reference was compressed into 2 bit format, and it was run using each contig as a separate query to reduce run time. Although not technically a feature of HapSolo, our github release provides a script to run Blat v35 [92] using this parallel approach. After running on individual contigs, the resulting PSL files were concatenated into a single PSL file for input into HapSolo. Minimap2 was used to compare feasibility and results between aligners; it was employed with the options “`-P -k19 -w2 -A1 -B2 -O1,6 -E2,1 -s200 -z200 -N50 -min-occ-floor = 100`”.

Assemblies, Hi-C data and scaffolding HapSolo was applied to with default parameters of 0.70 for ID_T, Q_T and QR_T; hill climbing started with random values of ID_T, Q_T and QR_T and then minimized Cost using hill climbing over 50,000 iterations. In HapSolo, BUSCO is run in geno mode on each contig using the orthoDB9 datasets and the AUGUSTUS species option. BUSCO v.3.0.2 relies on BLAST v.2.2.31+, AUGUSTUS v3.3, and BRAKER v1.9.

We obtained short-read Hi-C data from online public databases for scaffolding [175, 58] (see Data Availability). The Hi-C sequencing data were mapped to their respective assemblies using BWA [92]. The scaffolding of raw assembly and HapSolo processed assemblies were processed with the 3D de novo assembly pipeline v180419 [51], available from `https://github.com/theaidenlab/3d-dna/`. We ran QUAST v4.5 [64] for our post processing example and to assess performance during program development. For Fig. 2 and Additional file 1: Figure S1, the normalized value was calculated by first subtracting the minimum observed Cost min(Cost) from the observed Cost. The numerator [Cost-min(Cost)] was then divided by [max(Cost)-min(Cost)].

Computational resources and processing

For runtime analyses, HapSolo was run on dual CPU Intel E5-2696 V2 Nodes containing 512 GB of RAM. The Blat, minimap2 and BUSCO pre-processing steps were run on these same nodes, but also one the UC Irvine High Performance Computing Cluster, Extreme Science and Engineering Discovery Environment [154], San Diego Supercomputer Center Comet [107] and Pittsburgh Supercomputing Center Bridges [114] clusters.

Availability of data and materials

Vitis vinifera data: NCBI under the BioProject ID PRJNA550461. *Anopheles funestus* data: NCBI under the BioProject ID PRJNA494870. *Amblyraja radiata* data: Genbank ID GCA_010909765.1, `https://vgp.github.io/genomeark/Amblyraja_radiata/`. *Peromyscus leucopus* data: Genbank ID GCA_004664715.2. Software: `https://github.com/esolares/HapSolo`. Publication Version: `https://github.com/esolares/HapSolo/releases/tag/v0.1`

2.4 Results

Primary assemblies

We illustrate the application and results of HapSolo on three diploid genome data sets. The three-including the Chardonnay grape (*Vitis vinifera*), the *Anopheles* mosquito (*A. funestus*) and the Thorny Skate (*Amblyraja radiata*)—represent a range of expected genome sizes, at 490 Mb [131], 200 Mb [58] and 2560 Mb (`https://vgp.github.io/genomeark/Amblyraja_radiata/`), respectively. The three datasets also represent a range of raw sequence coverage (at 58×, 240×, and 128×, respectively), and two different assembly methods—i.e., a hybrid assembly for Chardonnay [175] and Falcon_Unzip for both mosquito [58] and thorny skate [128]. The sequencing data are based on the Pacific Biosciences (PacBio) sequencing platform, but HapSolo should be applicable to any contig assembly drafted from any long-read assembler.

For pre-processing, we utilized pairwise alignments with BLAT and minimap2 for the Chardonnay and mosquito data. To limit run time, we applied BLAT to the Chardonnay and mosquito data without long contigs (>10 Mb) as queries, because we reasoned that >10 Mb contigs are unlikely to represent alternative haplotigs (see "Methods" section). These long contigs were included as references, however, so that they are represented in pairwise alignments. We used only minimap2 for the larger skate genome, due to prohibitively long run times with BLAT.

For each species, we applied HapSolo with and without hill climbing and compared the outcomes to the original unreduced assembly. Table 1 provides assembly statistics, and it illustrates improvements from the unreduced assembly, to the assembly without hill climbing (-HC) based on default values, and finally to the assembly with hill climbing (+HC), which is based on random starting values and 50,000 iterations. Focusing on Chardonnay, for example, the primary contig genome size declined 13% from the unreduced assembly to the -

Species	Chardonnay			Mosquito			Thorny Skate[d]		
Assembly Type	No HapSolo[a]	HapSolo -HC[b]	HapSolo +HC[c]	No HapSolo	HapSolo -HC	HapSolo +HC	No HapSolo	HapSolo -HC	HapSolo +HC
# of Contigs	2,072	1,369	1,155	1,073	674	666	16,218	14,494	12,937
Contig Assembly Size (Mb)	655.2	569.3	539.0	212.0	200.4	200.0	3,229.4	3,147.8	3,031.3
Largest Contig (Mb)	11.6	11.6	11.6	7.6	7.6	7.6	3.4	3.4	3.4
Contig N50 (Mb)	1.1	1.3	1.4	0.6	0.7	0.7	0.4	0.4	0.5
Contig L50	141	106	95	86	77	77	2,022	1,928	1,800

Table 2.1: Contig assembly statistics for three primary assemblies for each of the three species.

[a] Results in this column are based on the primary assembly without application of HapSolo.

[b] Results in this column are based on application of HapSolo without hill climbing (-HC) and with default parameters of ID, Q and QR = 0.70.

[c] Results in this column are based on application of HapSolo with 50,000 cycles of hill climbing (+HC).

[d] Results for HapSolo were generated using minimap2. Chardonnay and mosquito statistics are based on BLAT

HC assembly and another 5% from the -HC assembly to the +HC assembly. Not surprisingly, as genome size decreased, so did the number of contigs included in the assembly, which fell from 2072 to 1369 (-HC) to 1155 (+HC). Moreover, contig N50 increased by 35% from 1.066 Mb to 1.441 Mb. Similar results were achieved after applying HapSolo to contigs from mosquito and thorny skate (Table 1). For both assemblies, the number of contigs, L50 and genome size decreased, while the contig N50 improved by 9% for both mosquito and the thorny skate. We note, however, that hill climbing did not increase N50 for the mosquito assembly much beyond that achieved by applying HapSolo for one iteration with its default values, suggesting that the default values performed well by this measure with this dataset.

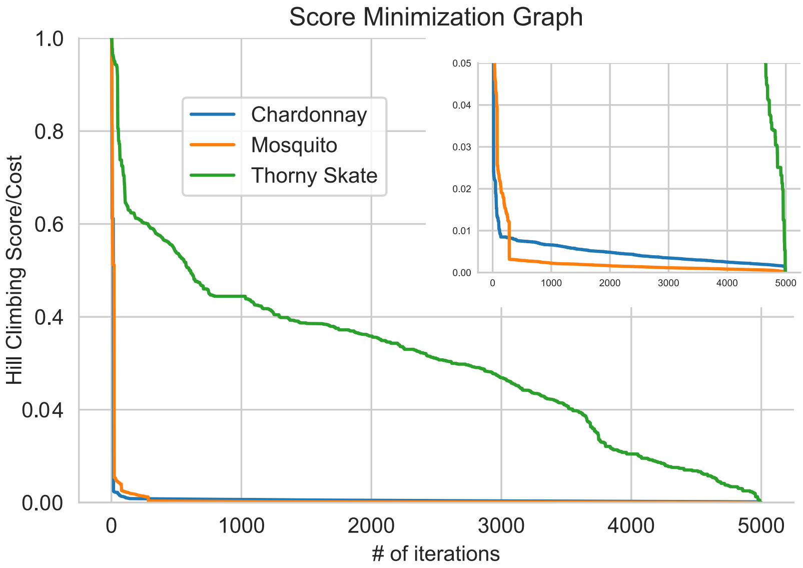

Figure 2.2: A graph of the sorted performance of hill climbing over 5000 iterations, with normalized Cost on the y-axis and the number of iterations on the x-axis. For most of our analyses with HC, we performed 5000 iterations on each of 10 cores; here we are showing results from one core. The top right provides a graph with altered scale for better visualization of Chardonnay and mosquito results.

Although N50 did not decline for the mosquito data, our implementation of hill climbing reduced Cost, as we expected, with the expected effects on BUSCO scores. Figure 2 illustrates a sorted representation of Cost, showing that lower Costs were identified. The behavior of hill climbing is dependent on the assembly, starting values for the three parameters (ID_T, Q_T and QR_T), and the number of local minima in the Cost function. Nonetheless, substantial improvements occurred within the first 1000 iterations for all three datasets (Additional file 1: Figure S1), with only minor improvements thereafter. Overall, the improvement in Cost suggests value in applying hill climbing to new data sets, especially given that the computational requirements are minor (see below).

Table 2 complements information about Cost by reporting BUSCO scores. HapSolo achieved its principal goal, which is to generally increase the representation of single copy (S) BUSCO genes and decrease duplicated (D) genes in reduced compared to unreduced assemblies. Note the differences between the -HC and +HC assemblies, because in some cases the -HC

	No HapSolo		HapSolo (-HC)		HapSolo (+HC)	
Species	GS[a]	BUSCO[b]	GS	BUSCO	GS	BUSCO
Chardonnay	655.2	C:1357 S:1004 D:353	569.3	C:1356 S:1152 D:204	539.0	C:1357 S:1205 D:152
Mosquito	212.0	C:2640 S:2493 D:147	200.4	C:2609 S:2548 D:61	200.0	C:2621 S:2566 D:55
Thorny Skate	3229.4	C:2091 S:1651 D:440	3147.8	C:2087 S:1675 D:412	3031.3	C:2080 S:1715 D:365

Table 2.2: Starting and ending BUSCO values for the three species for primary contig assemblies.

[a] Genome Size (GS) based on the sum of all contigs for the primary assembly

[b] Busco based on all contigs prior to the application of HapSolo. The three values represent the complete (C), the single (S) and duplicated (D) BUSCO genes

assembly had more single copy genes but at the cost of also having more duplicated genes. Thus, the +HC option can produce assemblies with lower Cost but with fewer BUSCO genes.

Figure 3 plots the cumulative contig assembly length for the three assemblies for each of the three species, and it illustrates two important points. First, HapSolo reduced the total assembly length primarily by removing numerous contigs of small size. Second, differences between the -HC and +HC reduced assemblies were more evident for some species (e.g., thorny skate) than for others (e.g., Chardonnay). Nonetheless, when there were differences, hill climbing decreased both assembly size (Table 1) and Cost (Table 2).

Hi-C scaffolding results

HapSolo focuses on the improvement of primary assemblies, but there are potential advantages for removing haplotigs for downstream operations like scaffolding. Failing to remove duplicate haplotigs can cause false joins between duplicate haplotigs or lead to non-

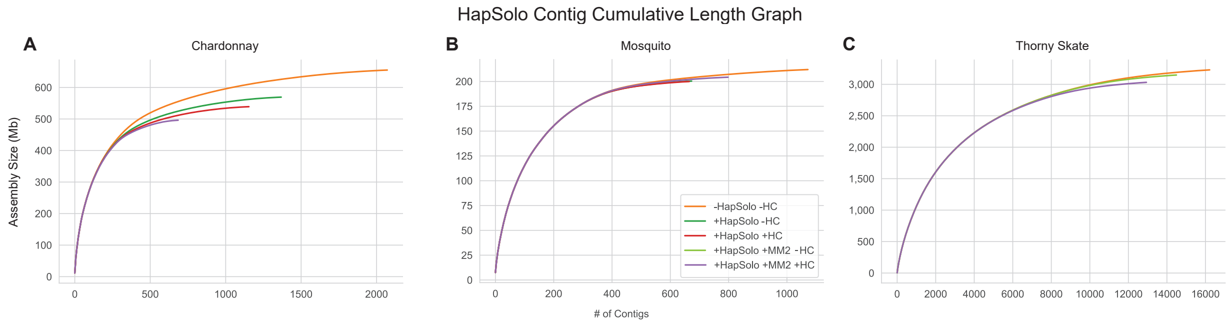

Figure 2.3: The cumulative assembly size (cdf) based on contigs. For Chardonnay (a) and mosquito (b), the five lines depict: an unreduced assembly (-HapSolo -HC), HapSolo applied with default parameter values and no hill climbing (+HapSolo -HC) using BLAT or minimap2 (+Hap Solo +MM2 -HC), HapSolo with random starting values and 50,000 iterations of hill climbing using BLAT (+HapSolo +HC) or minimap2 (+HapSolo +MM2 +HC). For thorny skate (c), three analyses were performed: an unreduced assembly (-HapSolo -HC) and Hapsolo analyses based on minimaps 2 pairwise alignment with and without hillclimbing.

parsimonious joins between duplicate haplotigs and adjacent single copy regions. Here we illustrate the advantage of running HapSolo on primary assemblies prior to Hi-C scaffolding. For these analyses, the unreduced assembly and both reduced assemblies (i.e., -HC and +HC) were scaffolded using the 3D-DNA pipeline [164], resulting in more continuous assemblies overall. We compared the improvements of the two scaffolded HapSolo assemblies against the unreduced scaffolded assembly (Table 3). Gains in improvements to the largest scaffold were clear across all assemblies relative to the unreduced assembly. For example, the largest scaffold increased by 1.71× (-HC) and 1.91× (+HC) for Chardonnay and by 1.22× (-HC) and 2.18× (+HC) for mosquito (Table 3).

Figure 4 illustrates the distribution of scaffolds for each of the three species under various HapSolo implementations. For each scaffold we measured the proportion of the genome that was contained in the k largest scaffolds, where k is the haploid number of chromosomes for each species. For example, Chardonnay has 19 chromosomes, and the 19 largest scaffolds based on the unreduced assembly represented 52% of the genome size. Following HapSolo haplotig reduction, the largest 19 scaffolds encompassed up to 93% of the total expected genome size of 490 Mb. Similar improvements were identified for the two other species, with mosquito improving from 61.9 to 85.7% and thorny skate from 31.5 to 106.3%. The observation that 106.3% of the thorny skate is contained in the largest k scaffolds indicates that the expected genome size is incorrect or that there is a need for additional purging of haplotigs.

HapSolo scaffolded assemblies were always demonstrably superior to the unreduced scaffolded assemblies for all three species, but the additional value of hill climbing varied among datasets. The value of hill climbing was clear for the mosquito, where the first 3 scaffolds (representing k=3 chromosomes) represented ~68% of genome with scaffolded -HC assembly versus 86% for the +HC reduced assembly. In contrast, hill climbing produced a disadvantage for Chardonnay (k=19, 92.6% -HC vs. 88.0% +HC) and only a small improvement for thorny skate (k=49, 104.7% -HC vs. 106.3% +HC). This being said, our metric based on the proportion of the genome in the k largest scaffolds is imperfect. For example, something as simple as a single split chromosome representing two metacentric arms could have a large effect on the metric. We therefore also examined other metrics, like the percentage of the genome encompassed in >10 Mb scaffolds and the longest scaffold. The largest differences were again due to application of HapSolo, with sometimes relatively minor differences associated with hill climbing (Table 3).

Finally, we focused on results based on comparing the two pre-processing alignment algorithms, BLAT and minimap2. We applied both algorithms to Chardonnay and mosquito. For mosquito, the results were similar with either aligner, but the BLAT results were markedly superior for Chardonnay (Figs. 3, 4). We do not know the cause of the discrepancy with Chardonnay, but we note that it is a genome that contains extensive structural variation between haplotypes, such that ~15% of genes are estimated to be in a hemizygous state [175]. We suspect that minimap2 often failed to extend alignments beyond large insertion and deletion events, even though we applied it with low gap and extension penalties substantially (see “Methods” section). Minimap2 is, however, highly preferable for run times, and it can be applied easily to gigabase-scale genomes like thorny skate.

Comparing HapSolo to an alternative method

Other algorithms have been devised to identify and remove alternative haplotigs [121, 63, 132]. In the publication of purge_dups, Guan et al. [63] compared its performance to Purge-Haplotigs and found it to be generally superior. We compared HapSolo to purge_dups, focusing on scaffolding results after Hi-C analysis. Figure 4 indicates that HapSolo generally led to better scaffolded assemblies than purge_dups, but with some caveats. For example, the HapSolo-based Chardonnay assembly was superior to the purge_dups assembly when BLAT was used to perform pre-processing. In this case, the percentage of the genome with >10 Mb scaffolds was 97.7% for HapSolo versus 67.4% purge_dups, with a 32% improvement in largest scaffold (Table 4). However, purge_dups performed similarly to HapSolo for Chardonnay when pairwise alignments were based on minimap2 (Fig. 4). For mosquito, purge_dups performed more poorly than HapSolo with either pre-processing aligner, as long as hill climbing was included in HapSolo analysis. Finally, for the larger thorny skate genome, HapSolo with hill climbing outperformed purge_dups (Fig. 4), resulting in a higher proportion of genomes in k scaffolds, more large (>10 Mb) scaffolds, and a 26% larger 'largest scaffold' (Table 4). Overall, HapSolo performed as well or better than purge_dups, based on the three exemplar datasets.

Applying HapSolo to a genome with low heterozygosity

HapSolo was designed to address a specific problem: the assembly of highly heterozygous genomes with divergent haplotigs. We chose our three exemplars to represent the problem. But how does HapSolo perform on less heterozygous genomes? We applied HapSolo to the mouse, *Peromyscus leucopus*, a mammalian genome from a single diploid individual with low (0.33%) heterozygosity [98]. In samples with low heterozygosity, alternative haplotigs are less likely to exist, and hence we expect fewer benefits with the application of HapSolo.

Species	Chardonnay			Mosquito			Thorny Skate[f]		
Assembly Type	No HapSolo[a]	HapSolo -HC[b]	HapSolo +HC[c]	No HapSolo	HapSolo -HC	HapSolo +HC	No HapSolo	HapSolo -HC	HapSolo +HC
# of Scaffolds	1,332	2,748	2,403	1,211	603	611	14,238	12,269	10,009
% of Genome in *k* largest scaffolds[d,e]	52.0%	89.0%	84.0%	61.9%	68.3%	85.7%	31.5%	104.7%	106.3%
% of Genome in Scaffolds >10Mb[e]	44.0%	94.0%	94.0%	93.5%	94.7%	94.1%	40.4%	103.3%	105.5%
Scaffold Assembly Size (Mb)	656.1	570.4	540.1	212.5	200.8	200.3	3,240	3,158	3,039
Scaffold N50 (MB)	7.2	23.5	20.7	37.9	41.5	41.6	62.1	65.5	69.8
Scaffold L50	28	11	11	3	3	2	16	35	13

Table 2.3: Scaffolded assembly statistics after Hi-C analysis on HapSolo assemblies, for three primary assemblies for each of three species

[a] Results in this column were based on the primary assembly without application of HapSolo

[b] Results in this column were based on application of HapSolo without hill climbing (-HC) and default parameters of ID, Q and QR

[c] Results in this column were based on application of HapSolo with 50,000 cycles of hill climbing (+HC)

[d] Percentage of genome in the largest k scaffolds, where k is equal to the number of chromosomes expected for each species

[e] Percentages normalized using expected genome sizes of 490 Mb, 200 Mb and 2560 Mb for chardonnay, mosquito and thorny skate respectively

[f] Results for HapSolo were generated using minimap2

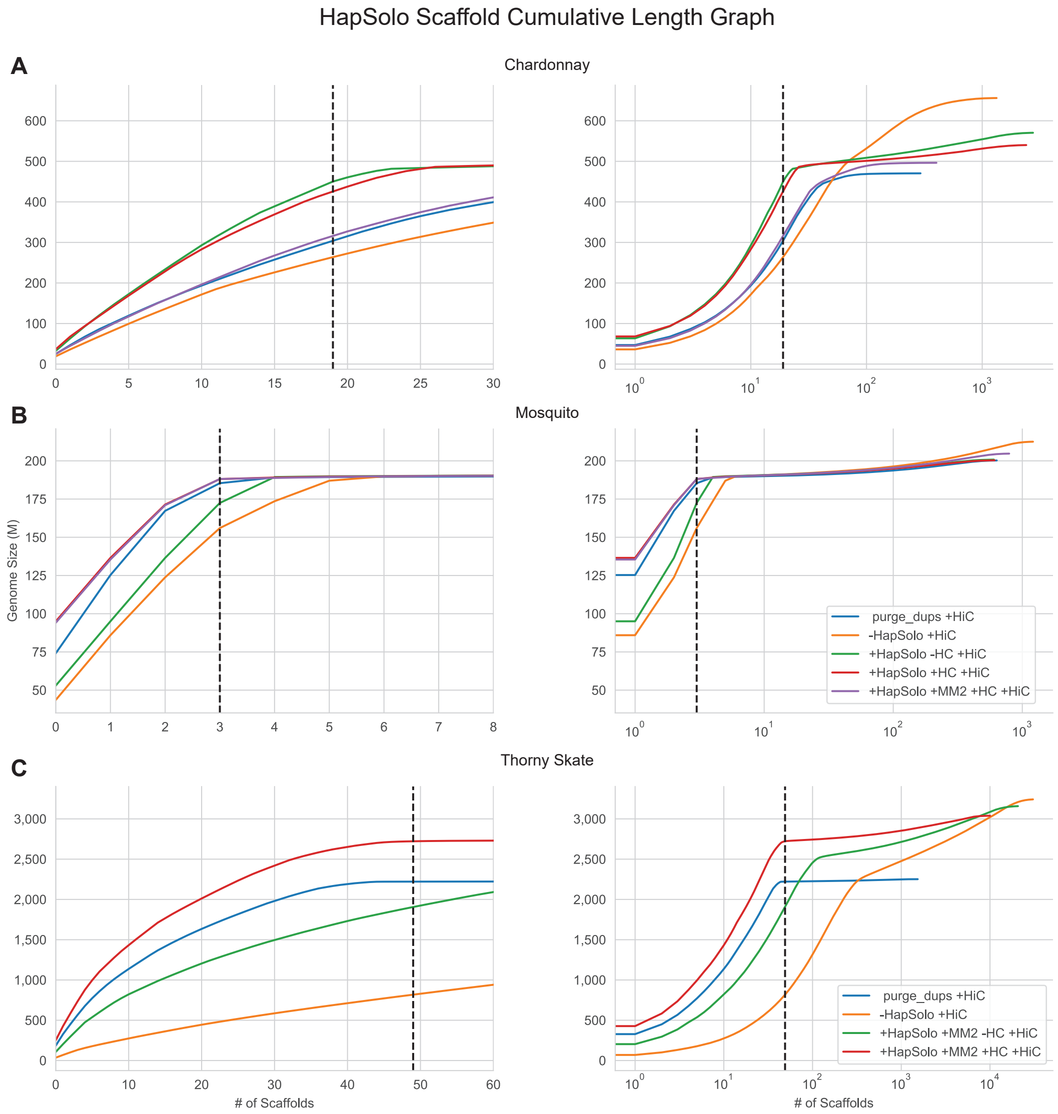

Figure 2.4: The cumulative assembly size (cdf) based on scaffolds for a different species in each row. The three rows represent analyses based on Chardonnay (a), mosquito (b) and thorny skate (c) data. There are two graphs for each species; the one on the left focuses on the chromosome length scaffold portion of the assembly (number of scaffolds), while the one on the right is the complete assembly on a log_{10} (number of scaffolds) scale. Each graph for Chardonnay and mosquito have five lines and follow the key provided in the right-hand graph in panel B for mosquito. The analyses are scaffolded genome based on purge_dups (purge_dups + HiC), the unreduced assembly (-HapSolo -HC +HiC), scaffold based on the HapSolo reduced assembly with BLAT preprocessing and without hill climbing (+HapSolo -HC +HiC), and the scaffold based on the HapSolo reduced assembly with BLAT preprocessing and with hill climbing (+HapSolo +HC +HiC) and, finally, HapSolo reduced assembly with hill climbing using minimap2 (+HapSolo +MM2 +HC +HiC). In all graphs, the dotted line indicates the number of chromosomes for the species. c Reports results for thorny skate, which did not include BLAT processing. The HapSolo analyses are based on minimap2 alignments and presented with (+HC) and without (-HC) hillclimbing.

Species[a]	Chardonnay		Mosquito		Thorny Skate	
Assembly Type	HapSolo +HC[b]	purge_dups	HapSolo +HC	purge_dups	HapSolo +HC[e]	purge_dups
# of Scaffolds	2,403	294	611	635	10,009	1,534
% of Genome in k largest scaffolds[c]	88.0%	58.5%	85.7%	83.7%	106.3%	86.8%
% of Genome in Scaffolds >10Mb[d]	97.7%	67.4%	94.1%	92.7%	105.6%	86.0%
Scaffold Assembly Size (Mb)	540.1	470.4	200.4	200.3	3,039.3	2,251.4
Largest Scaffold (MB)	36.5	24.9	95.0	74.1	250.5	184.1
Scaffold N50 (MB)	20.7	12.7	41.6	51.2	69.8	61.7
Scaffold L50	11	15	2	2	13	11

Table 2.4: Comparison of scaffolded assemblies after Hi-C analysis, based on HapSolo and purge_dups primary assemblies

[a] Data for HapSolo are based on BLAT alignments for Chardonnay and mosquito, and minimap2 alignments for thorny skate

[b] Results in the +HC columns are based on application of HapSolo with 50,000 cycles of hill climbing (+HC)

[c] Percentage of genome in the largest k scaffolds, where k is equal to the number of chromosomes expected for each species. Percentages are normalized using expected genome sizes of 490 Mb, 200 Mb and 2560 Mb for Chardonnay, mosquito and thorny skate respectively

[d] Percentages normalized using expected genome sizes of 490 Mb, 200 Mb and 2560 Mb for chardonnay, mosquito and thorny skate respectively

[e] Results for HapSolo were generated using minimap2

Indeed, we found no benefit. Comparing results between the reference assembly and the HapSolo assembly (with minimap 2 and hill climbing), we found similar proportions of the genome encompassed in 10 Mb scaffolds (96.8% vs. 95.7%) and a substantially smaller proportion of the genome encompassed in k=24 chromosomes (88.5% vs. 71.1%) (Additional file 1: Table S1). The HapSolo assembly was, however, largely contiguous with the reference assembly (Additional file 1: Figure S2). Interestingly, purge_dups did not find any alternative contigs on this assembly and ultimately failed with an error, so we are unable to compare its performance.

In addition to the low heterozygosity, the *P. leucopus* genome has a low percentage of duplicated BUSCOs relative to the complete set of BUSCOs, at 2.1% (Additional file 1: Table S1). In contrast, Chardonnay, mosquito and thorny skate have 26.0%, 5.5% and 21.0%, respectively (Table 2). Perhaps unsurprisingly, given this statistic, mosquito exhibits the least dramatic improvements in assembly statistics after application of HapSolo (Table 3). These observations suggest that there are lower limits at which HapSolo becomes ineffective and perhaps even detrimental. Based on the data we have analyzed, we suggest that ~5% may be a lower limit for the proportion of duplicated BUSCOs. Heterozygosity is likely to define another lower limit. Given that heterozygosity is 0.33% for *P. leucopus*, we expect that HapSolo will not be useful for the assembly of human genomes, because species-wide human heterozygosity is 0.05% [30]. Our results nonetheless suggest that HapSolo is likely to be a helpful tool for assemblies with a high number of duplicate BUSCOs.

Execution time and memory efficiency

To measure runtime, HapSolo was run on dual CPU Intel E5-2696 V2 nodes containing 512 GB of RAM and storage attached via a 40Gbe Infiniband connection. CPU runtime depends on the number of iterations, but it is also dependent on the data and parameter values. We measured runtime across the datasets, measuring different configurations in terms of the number of cores and the number of iterations per core (Additional file 1: Table S2). Under the conditions we used for empirical data (i.e., hill climbing on 10 cores with 5000 iterations per core), the total time was>10 min for both Chardonnay and mosquito but substantially longer at 13 h and 45 min for the much larger thorny skate. Note that memory usage was dependent on the size of the alignment file and independent of the number of iterations, because HapSolo stores alignments in memory for rapid filtering at each step during hill climbing. Nonetheless, the memory and speed requirements are such that HapSolo can be run on a laptop or desktop computer.

2.5 Discussion

We have presented an implementation, HapSolo, that is focused on improving primary assemblies by removing alternative haplotigs. In theory, the HapSolo package can be applied to any set of contigs from any assembly algorithm. The approach implemented in HapSolo is intended to replace laborious manual curation [37], and it follows some of the logic of existing programs, like Redundans [121], PurgeHaplotigs [132] and purge_dups [63]. However, HapSolo differs from competing programs by at least three features. First, it utilizes multiple alignment metrics, so that it is not reliant only on percent identity (ID). The goal of these multiple metrics is to better discriminate among some situations that may yield high identity scores but nevertheless lead to the retention of different contigs in the primary assembly (Fig. 1). Second, when the hill climbing option is utilized, HapSolo relies on a maximization scheme based on BUSCO values. The underlying assumption is that maximizing the number of single-copy BUSCOs establishes more complete and less repetitive genomes. We emphasize that this is an assumption common to the genomics community, because most new genomes are reported with BUSCO scores to reflect their completeness and quality. Third, an important feature of HapSolo is the ability to modify the Cost function, so that the user may choose to weigh duplicated BUSCO genes less heavily or perhaps even ignore them altogether. This flexibility may prove useful for some applications. For example, it may be useful to ignore costs related to duplicated BUSCO genes when assembling polyploid genomes and instead focus only on complete and fragmented genes.

We have illustrated some of the performance features of HapSolo by applying it to data from three species that differ in genome size and complexity: Chardonnay grape, a mosquito, and the thorny skate. The common feature of these species is that diploid assembly is necessary. For all three species, we compared the unreduced primary assembly to two HapSolo assemblies, one that used default values (-HC) and one that used hill climbing minimization (+HC). Both HapSolo assemblies reduced genome size and markedly improved standard statistics like N50 (Table 1 and Fig. 3). The +HC contig assembly was generally better than the -HC assembly, but not always; the most substantial differences occurred between the unreduced assembly and either of the two HapSolo assemblies.

Our reduced assemblies scaffolded faster than unreduced assemblies and also led to more contiguous genomes. For each of our three species, the cumulative genome length associated with first k scaffolds (where k is the chromosome number) was much larger based on reduced vs. unreduced assemblies. The percentage of the genome contained in chromosome length scaffolds increased by at least 25% (Table 3). We conclude that in highly heterozygous samples that potentially have a large number of alternative haplotigs, some reduction step is critical for curating a primary assembly and for downstream scaffolding. This is true even when the primary assemblies are from FALCON_Unzip [42] which has already (in theory but perhaps not always in practice) identified secondary haplotigs. We further advocate for the use of the hill climbing feature in HapSolo, because the computational cost is relatively small but the gains can be large (Fig. 3). Finally, we find that BLAT tends to outperform minimap2 as the pre-processing aligner and advocate for its use. However, it can be time prohibitive on large genomes, and hence HapSolo includes support for minimap2.

2.6 Conclusions

Based on the data in this paper, HapSolo generally led to similar or better outcomes than purge_dups [63], another recently published method to identify and remove haplotigs. That is not to say, however, that HapSolo cannot be improved. We can see three obvious areas for future growth. The first is to consider coverage statistics, which represents a point of departure between our approach and that of both Purge Haplotigs and purge_dups. We predict, but do not yet know, that the inclusion of coverage with our existing alignment statistics could lead to more accurate inferences. A second area of improvement may be to implement alternative maximization algorithms, such as simulated annealing. Finally, it may also be possible to include additional features in the calculations of Cost. Our present reliance on BUSCOs has the advantages of speed and wide acceptance in the genomics community. However, depending on the initial assembly, it is likely that some contigs do not contain a BUSCO gene, are therefore not considered in Cost and do not form the approximation of threshold parameters (ID_T, Q_T and QR_T). It is not yet clear what additional features could be included in the Cost function, but identifying contigs containing an over-representation of shared k-mers is one possibility.

Chapter 3

A new avocado reference genome provides insight into genome evolution and the genetic basis of important traits

3.1 Abstract

Long-read sequencing technologies have allowed researchers to characterize genomes with increasing accuracy, leading to a greater understanding of the genetic basis of phenotypes and adaptation. However, many economically important crops lack suitable genomic tools and annotations. This has been the case for the avocado (*Persea americana*), a perennial crop with high nutritive content. To help address the need for genomic approaches in avocado, we created a reference genome from the Gwen varietal, which is closely related to the agronomically dominant Hass varietal and is a current focus of breeding programs. We produced a

1,032Mb genome assembly with an N50 of 3.37Mb and a BUSCO score of 91%; we also scaffolded the assembly using a genetic map. In addition to producing a scaffolded reference, we resequenced 21 avocados that represented both the three botanical races of P. americana and also three additional closely related outgroup species. Using these data, along with previously published datasets, we: i) characterized the repeat content of the Gwen genome, which constitutes 61% of the genome ii) built a genome annotation with 60,151 genes, which is 2x more than a previous annotation of the fragmented Hass genome assembly, iii) investigated population structure among our sample of 31 avocados, and iv) performed pairwise comparisons of different groups of avocados to identify genomic regions of high differentiation. The pairwise comparisons included two kinds of contrasts. The first was between botanical races; we reasoned that genomic regions of high differentiation may contain genes that affect the phenotypic traits that differ between races. Using this approach, we identified numerous candidate genes of potential agronomic importance with gene ontology functions that included stress responses, disease resistance, and water osmoregulation. The second contrast was between accessions with differing flowering types that define synchronous dichogamy in avocado. Our analyses of flowering types identified candidate genes hypothesized to affect flower development and circadian rhythms.

3.2 Introduction

Avocado (*Persea americana Mill.*) is a perennial, subtropical crop that is in ever-increasing demand. In the United States, for example, per capita avocado consumption has tripled over the last two decades. Demand in the U.S. is met partly by domestic production but principally by imports from Mexico and elsewhere. Mexico is currently the largest producer of avocado, where the crop is worth an estimated $2.5 billion per year (Rendón-Anaya et al. 2019), but other major producers include the Dominican Republic, Peru, Chile, Indonesia, Israel and Kenya (`http://www.fao.org/faostat/en/#data/QC/visualize`). Although the popularity of avocados is primarily a 20th century phenomenon [138], it has quickly grown to be a global commodity.

Remarkably, avocado cultivation is dominated by a single variety (Hass) that represents 90% of cultivation world-wide [127]. All Hass trees are derived clonally from a tree patented in 1935. Despite the shockingly narrow genetic base of agricultural production, avocado *sensu lato* is quite genetically diverse. Some of this diversity stems from the fact that there are three domesticated botanical races: *P. americana var. americana Mill* (which we will call the 'Lowland' race in recognition that the previously accepted name of West Indian is inaccurate), *var. drymifolia Blake* (the Mexican race), and *var. guatemalensis Williams* (the Guatemalan race) [18]. The strikingly different fruit morphologies among the races suggest that they may have been domesticated separately, a conjecture supported by genetic data [56, 11, 127]. One practical consequence is that each race likely contains separate alleles and/or genes of interest for crop improvement, due to their different domestication histories. Another consequence is that hybridization between races can produce unique allelic combinations, potentially leading to agronomically useful hybrid offspring. Hass is, in fact, one example; although its precise breeding history is not known, genetic evidence shows that it is a hybrid between Guatemalan and Mexican races [138, 40, 127].

The high demand for, and economic importance of, avocados motivates breeding efforts, but breeding remains challenging for at least three reasons. First, avocado is a large tree that matures slowly (5 to 8 years before production; [85], requiring substantial space, water, and financial resources [10]. A second major obstacle is the reproductive system. A single tree typically produces more than one million flowers, of which only 0.1% or fewer yield mature fruit [47, 48, 17], making controlled pollinations virtually impossible [38]. Finally, avocado is predominantly out-crossing, due to synchronous dichogamy. There are two flowering types in this system: A and B. Type A trees are female (receptive to pollen) in the morning and shed pollen as males in the afternoon of the following day. In contrast, type B trees are male in the morning of the first day and female in the afternoon of the next day. The system is complicated by the fact that there is some leakiness of flower type that depends on the environmental cues [47]. As a result of these complications, avocado breeding has historically relied on open-pollinated and inter-racial hybridization to the extent that most individual varieties lack accurate breeding records [48, 11, 139].

These complications argue that genomics and molecular breeding are central for the continued improvement of avocado. For example, molecular markers for flowering types may be particularly useful, because type B avocados are crucial for pollination but typically less productive than type A varieties [47]. Recently, Rendon-Ayana et al. (2019) made an important contribution toward molecular breeding by producing draft genomes of Hass and a wild Mexican accession (*P. americana ssp. drymifolia*). They anchored the Hass assembly to a genetic map and ultimately produced a reference genome with 512 scaffolds and a genome size of 419 Mb. Using this reference, Rendon-Ayana et al. (2019) also explored the hybrid history of Hass and a few aspects of the evolutionary genomics of avocado. Nonetheless, several important features of the evolutionary genomics of avocado remain unexplored, including characterizing diploid chromosomes in this highly heterozygous ancestor, using sweep mapping to identify potential regions of agronomic interest, and focusing on genomic diversity in the context of interesting traits, like A vs. B flowering types.

Here we produce and assemble the genome of the Gwen variety and use that genome as a reference for evolutionary analyses. Gwen is a grandchild of Hass with similar flavor characteristics [163] but with higher yields and better fruit storage on the tree [20, 19]. Accordingly, Gwen has been the subject of intensive breeding efforts for three decades. Our Gwen genome vastly improves contiguity relative to the Hass genome, providing a better platform to explore the evolutionary genomics of avocado. More specifically, we intend to use the data outlined in this chapter to focus on four sets of questions. First, what does the

Gwen genome tell us about patterns of heterozygosity within an avocado accession? Genomic analysis of highly heterozygous grape (*Vitis vinifera*), another perennial clonally propagated crop, revealed that as many as one in seven genes are hemizygous, perhaps due to structural mutations that have accrued during clonal propagation [175]. Is avocado similar? Second, we use the Gwen genome as a reference to explore genetic diversity within avocado, specifically to recapitulate the three races and to assess the hybrid origin of well-known cultivars. This last question builds on several previous investigations of genetic diversity [11, 40, 39] but extends the work to a genomic scale. Third, we investigate features of avocado domestication. Do the three races demonstrate a cost of domestication - i.e., an accumulation of deleterious alleles - relative to wild accessions, as is common for domesticates [57, 110]? And do the three racial groups share regions of selective sweeps consistent with parallel selection on genic regions associated with specific traits? Finally, we investigate genetic diversity between the A and B flowering types, with the goal of identifying genomic regions that may contribute to synchronous dichogamy.

3.3 Methods

Sample Preparation and Sequencing protocols

The avocado variety chosen for sequencing was Gwen, a grandchild of Hass which has been central to the University of California Riverside breeding program. High molecular weight genomic DNA (gDNA) was isolated from young leaf materials using the method of [42]. To avoid the co-precipitation of polysaccharides and phenolics with DNA, they were reduced in a pre-washing step with sorbitol. The purity, quantity and integrity of DNA were assessed with a Nanodrop 2000 spectrophotometer (Thermo Scientific, IL, USA), a Qubit 2.0 Fluorometer together with the DNA High Sensitivity kit (Life Technologies, CA, USA), and by pulsed-field gel electrophoresis. SMRTbell libraries were prepared as described in [101] and sequenced on a PacBio RS II (Pacific Biosciences, CA, USA) using P6-C4 chemistry (DNA Technology Core Facility, University of California, Davis).

Genome Assembly, Polishing and Scaffolding

We assembled Gwen Pacific Biosciences (PacBio) SMRT reads with Canu version 2.1, using default settings and including all reads. Once assembled, polishing was performed with PacBio GenomicConsensus v2.33. Two passes of GenomicConsensus were run, followed by two additional passes with Pilon v1.23 [159], using default parameters and 19x coverage of short read Illumina sequencing data. HapSolo v0.1 was then run on the assembly using default parameters and 50,000 iterations [144], producing the "C+H assembly". The genetic map based on a Gwen x Fuerte (G x F) cross [10] was used for orienting assembly contigs into 12 scaffolds. To assemble scaffolds, the G x F map markers were aligned to the Gwen C+H assembly using NCBI BLAST v2.2.31+ [7]. A custom python script was then used to order the alignments based on linkage group ID and cM distance, using only the alignments with the highest percent identity and e-value score to identify contig order. This approach may not lead to proper orientation of contigs, such that contigs with unclear orientation were marked with an asterisk in the scaffolding annotation file. When the orientation of a contig could not be determined, it was placed in the '+' (or positive) direction. All contigs were bridged using 20,000 N's as spacers, so that the spacing would not interfere with analyses that used 20kb windows. MUMmer 3.32 [84] was used to check overlap between contigs, but none could be found.

Gene and Transposable Element Annotation

Repeat annotation was based on RepeatModeler v2.0.1 in conjunction with RepeatMasker v4.1.1. The build can be found as a singularity image at `http://github.com/esolares/singularityimages`. RepeatModeler was run prior to RepeatMasker to generate a repeat database for avocados, since a closely related species was not readily available. This repeat database was built using the option BuildDatabase on the Gwen H+C assembly. RepeatModeler was then run using the database built by the previous execution with the option LTRStruct and run in multi-core mode using the option -pa. The subsequent output files were then used in the RepeatMasker step using the following options

```
-pa ${CORES} -s -lib ${GWENBUILTDATABASE} -a -poly -ace -excln -html
-gff -dir ${OUTDIR} ${GWENGENOME}
```

where the lib option coincides with the repeat library built during the RepeatModeler step, the variables CORES, OUTDIR and GWENGENOME contain the total number of cores requested, the output directory in absolute form, and the Gwen assembly as a fasta file. The Gwen genome was then soft masked using Bedtools v2.29.2 using the maskfasta run option. The resultant soft masked fasta file was used for all subsequent analysis, including gene annotation. We recreated the TE annotation gtf file using the fasta.out file in the RepeatMask output folder, because the original gtf file did not contain the repeat class/family column.

For gene annotation, we mapped the paired end RNASeq reads from previous studies [166, 72, 13] using HiSat2 version 2.2.1 on the repeat masked genome. The resultant BAM files were merged and indexed using Samtools v1.10 [95, 14], which was used to generate hints for the BRAKER v2.1.6 [69] + Augustus [148, 149, 147] v3.4.0 pipeline. The BRAKER pipeline was used in default mode for RNASeq data, with an additional option for softmasked reference assemblies (–softmasking), with avocado as the species option. This option allowed the BRAKER and Augustus to generate the annotation using only the RNASeq mapped data. The BRAKER Augustus image was built using the developers docker script as a reference; dependencies were then built into the image manually. The docker images were created using docker v20.10.0 and imported into singularity v3.7.1 and are available at http://github.com/esolares/singularityimages.

Functional annotations and Gene Ontology analyses were performed with Blast2GO v6.01 [43]. Genes were extracted from the Gwen assembly, based on the gene annotation gff file, excluding (as possible pseudogenes) any genes that have exons overlapping TE's. The remaining gene sets were then mapped to the NCBI_NR, SwissProt and uniref90 databases using NCBI's blastx, blastx-fast option. A Protein family search was also performed using an InterPro scan. These results were then merged into a single annotation for GO mapping, with results based on default options. Functional annotation was also performed using the EggNOG online web submission page (`http://eggnog5.embl.de/#/app/home`). Genes were submitted as nucleotide fasta files using experimental evidence for functional annotation.

Diversity Samples and Sequencing We collected leaf tissue for a total of 20 P. americana accessions and three outgroups (Table S1). For each sample, genomic DNA was extracted from leaf samples with the Qiagen DNeasy plant kit. Paired-end sequencing libraries were constructed with an insert size of 300 bp according to the Nextera Flex (Illumina, Inc) library preparation protocol. Libraries were sequenced using the Illumina NovaSeq platform with cycles to a target coverage of 25X. Raw sequencing data have been deposited in the Short Read Archive at NCBI under BioProject ID: PRJNA758103. We also used Illumina raw reads for 13 accessions that were published previously [127] and downloaded from the Short Read Archive at NCBI.

SNP Calling and Fst Analyses

Short reads were mapped to both the H+C assembly using BWA version 0.7.17 [93] and realigned and recalibrated using GATK v3.7 [102]. Alignment filtering was done using BCFTools v1.10.2 [46] using parameters -s LowQual -e '%QUAL<20 || DP>32'. SNPs were calculated using the GATK Haplotype pipeline using version 3.7. PCA results were generated using VCFTools v0.1.17 git commit 954e607 [45], and PLINK v1.9 [123]. Avocado samples missing over 50% of data were filtered out, and outgroups missing over 75% of data filtered out, resulting in the removal of three samples from the dataset of Rendón-Anaya et al. 2019. The PCA was performed on two datasets: with and without the outgroup sample. The admixture plots and analyses were performed with ANGSD version 0.930 (build Jan 6, 2020 13:30:06). VCFTools was used for the conversion of PLINK results to the BEAGLE format. The online version of CLUMPACK (`http://clumpak.tau.ac.il/bestK.html`) was used for identifying the best number of groups (K) for admixture plots across K=1 to 10, each with 10 replicates.

For calculating Fst between groups, we focused on 20kb non-overlapping windows and used PLINK to calculate Fst values. Groups were created using accessions that contained ¿80% assignment to the Guatemalan, Lowland or Mexican races. For pairwise comparisons between groups, negative Fst values were set to 0.0. We identified Fst windows as peaks when they were within the top 1% tail of the Fst distribution. These windows were then intersected with our gene annotation using Bedtools2 intersect run command with the option -wa, which allows for the preservation of the original gene coordinates. GO pathways for the genes were then generated using Blast2GO, and putative gene functions were retrieved from EggNog results.

3.4 Results

Gwen Genome Assembly and Characterization

Gwen Genome Assembly compared to Hass and *P. americana var. drymifolia*: We isolated high weight DNA from Gwen avocado leaves and sequenced using the long-read Pacific Biosciences (PacBio) platform (see Methods). We generated a total of 81.2G bases, equivalent to roughly 90x coverage, based on the expected 1C genome size of 896Mb [9]. We assembled the PacBio reads using Canu (v 2.1), producing a genome of 1,456Mb with 5,122 contigs. This represents an assembly, however, that includes both of the alternative haplotypes in diploid Gwen. We therefore applied HapSolo [144], a tool designed for identifying primary haplotypes within a diploid assembly, which removes putative secondary contigs (or haplotigs). The Canu+HapSolo (C+H) genome resulted in a primary assembly of 1,032M bases, a longest contig of 17Mb, a BUSCO score of 91% and an N50 of 3.37M bases, which was a 2.4-fold improvement over the unreduced Canu assembly (Table 1). Another useful measure of an assembly is the percentage of the assembly that is encompassed in the largest x contigs, where x represents the number of chromosomes (which is 12 for 1C P. americana). For the C+H assembly, this percentage was 17% - i.e.,the 12 largest contigs represented 17% of the genome.

To improve contiguity, we anchored and scaffolded the C+H assembly using a published genetic map, based on a cross between Gwen and Fuerte [10]. Scaffolding vastly increased contiguity and ultimately resulted in 12 scaffolds that were assigned to 12 linkage groups. Scaffold N50 improved 18-fold (to 61.9Mb) over the 3.37Mb contig N50, and the 12 largest scaffolds represented 78% of the expected genome size. (Table 1). Scaffolding may have incurred a cost, however, because the total size of the scaffolded assembly was 703Mb, which is 22% smaller than the expected haploid size of 896Mb. Although each chromosome could be identified, we interpret this reduction in size to imply that scaffolding missed the incorporation of contigs, perhaps for reasons related to the density of the genetic map. In this context, it is worth noting that the Hass primary assembly decreased substantially in size when it was anchored to a genetic map [127] - e.g., only 47% of the Hass assembly could be anchored, resulting in a scaffolded genome of 421.7Mb. Hence our scaffolded genome is nearly double the size of the previous avocado reference.

We compared our new Gwen genome to the existing Hass and drymifolia assemblies using standard descriptive genome statistics. The Gwen assembly was superior in most respects, including N50, longest contigs (or scaffolds), BUSCO scores, and the proportion of the genome captured in the 12 largest fragments (Table 1). We plotted the cumulative density function of contigs and scaffolds for the three genomes (Figure 1A), providing a virtual comparison among the genomes and illustrating that far more of the Gwen assembly is captured in the largest 12 scaffolds (or contigs) - i.e., 78% for Gwen vs. 4.2% and 2.7% for the Hass and drymifolia assemblies produced previously [127].

Genome Annotation: We annotated the C+H and scaffolded assemblies by first focusing on TEs. After applying RepeatModeler and RepeatMasker (see Methods), we found 61% of the Gwen genome consisted of repetitive elements. Of the 68%, 52% were Long Terminal Repeat (LTR) retroelements, with more than half being Gypsy elements and 40% being unknown repeats (Figure 1B). We then annotated genes by mapping paired-end short-read RNAseq data from previous studies [166, 72, 13] to the C+H and scaffolded assemblies and applied BRAKER2 [69, 27, 147, 149, 148], which automates training of the prediction tools GeneMark-EX [28] and AUGUSTUS and integrates RNASeq information into the gene predictions.

As expected, the gene annotation results differed between the two assemblies, with 57,916 genes on the scaffolded assembly and 87,617 in the C+H assembly. These numbers decrease to 41,237 and 60,151, respectively, when predicted genes were filtered to remove all genes with predicted exons that overlap with annotated TEs. We compared these numbers to the annotation of the Hass genome [127], which reported two values: an initial value of 33,378 genes based on both evidence-based and de-novo prediction and a second value of 24,616 genes based on extensive filtering. Unfortunately, the filtered gene set is not publicly available, so we took the annotation of 33,378 and filtered them to remove duplicates, which yielded 25,211 Hass genes base on our analyses.

Our numbers of annotated genes are nonetheless 1.67 and 2.55-fold more than the 25,211 Hass genes, for the scaffolded and H+C assemblies. The reasons for these discrepancies are not entirely clear, but we can think of four. First, our genome is more contiguous and complete, although it seems unlikely that this alone would lead to 2-fold more gene models. Second, we have employed different RNAseq data than the previous study [127], because their RNAseq data is not publicly available. Our three data sources included a recent dataset that was likely unavailable for the Hass reference analysis [13]. The third is differences in the annotation pipeline, which could lead to different results. Finally, we expect that the Hass gene predictions are too conservative based on our own analyses and a recent transcriptome analysis that infers 63,420 unigenes [32], a number much closer to our gene annotation results. Note, however, that despite the different number of predicted genes between the Hass and Gwen analyses, there is complete overlap, because 100% of the 25,211 Hass genes were present in our Gwen annotation.

Analysis of P. americana Genetic Structure:

Classification of races and hybrid varieties: Given the Gwen reference genome, we performed a preliminary study of genetic variation and evolutionary genomics across avocado. To do so, we amassed a resequencing dataset of 34 accessions with at least 14x coverage (Table 2; Table S_Accessions). Eleven of these accessions came from previous publications [127]; the remaining 24 accessions, including three outgroup species (*P. hintonii, P. donnell-smithi and Ocatea boranthea*), are original to this study. Overall, the sample of 34 accessions were chosen to: i) represent the three botanical races of avocado, ii) sample both flowering types and iii) include historically important cultivars to complement previous studies about potential hybrid origins [39].

Given resequencing data, we mapped reads to the C+H reference and identified 36,250,941 SNPs within avocados and 54,465,536 SNPs when the outgroups were included. SNPs were then subjected to two types of clustering analysis – PCA and Admixture analysis – to define groups of accessions based on whole genome data. We first applied the PCA to the *P. americana* data without outgroups. The results verified many of the expected groupings among accessions but with a few surprises (Figure 1A). The expected groupings included clusters that represent previously identified Mexican, Guatemalan and Lowland races. There were, however, several accessions that were located between groups on the PCA - e.g., Velvick - that hint at their expected hybrid origins (Table 2). A mild surprise came from the location of Hass and Gwen on the PCA. Hass clustered more closely to Guatemalan accessions than was expected of an accession that has been previously defined as ¿50% Mexican by Rendon-Anaya et al [127] and 58% Guatemalan by Chen et al. [39]. Consistent with the placement of Hass, its grandchild Gwen also clustered near Guatemalan accessions, which is again consistent with previous inferences that it is primarily from race guatemalensis. Important varieties like Bacon and Zutano have been inconsistently inferred or assigned as hybrids [39] but appear to have had a hybrid origin based on our whole genome analyses (Figure 1A).

For completeness, we also applied PCA to a dataset that included four outgroup taxa (Table 2; Figure S1). These analyses demonstrated that the genetic placement of three of the outgroups (*P. donnell-smithi, P. hintonii, Ocotea boranthea*) was consistent with their expected outgroup status. However, the fourth potential outgroup (*P. schiedeana*) clustered well within the avocado ingroup, even though it has been used as an outgroup in evolutionary analyses [127]. Our clustering of *P. schiedeana* within the ingroup is consistent with previous evidence suggesting that *P. schiedeana* may hybridize with avocado [11]. Altogether, our PCA results suggest that *P. schiedeana* should not be used as an outgroup for future evolutionary analyses.

In addition to PCA, we investigated relationships among accessions using admixture mapping with K = 2 to 5 clusters. The most highly supported analysis contained four groups: the three previously-recognized races (Lowland, Mexican, Guatemalan) and a series of cultivars related to Gwen and Hass. This last group includes Mendez, a somatic mutation of Hass [73, 138]. We suspect that this last group is recognized in part by sampling several closely related varieties, which may have biased the analysis. Nonetheless, the admixture map identifies other accessions with putative hybrid histories, including Velvick, Fuerte, Bacon, Zutano and others. One especially interesting finding is that an accession (CH-CR-25), which has been thought to represent a new racial ecotype - *var. costaricensis* - [127] is likely to be a hybrid between the Lowland and the Guatemalan groups.

Identifying regions of chromosomal differentiation between races

To investigate one potential feature of domestication among the three races, we assigned individuals to a specific race when their assignment exceeded 80% per group (i.e., $Q_i > 0.8$) [31, 156]. This constructed samples of n=3, n=5 and n=10 for the Mexican, Lowland and Guatemalan races, respectively. For these three sets of samples, which represent potentially distinct domestication events, we investigated divergence between groups based on Fst, applied to 20kb windows across the contigs of the H+C reference. The purpose of this analysis was to identify genomic regions that differ in allelic frequencies between samples, which has the potential to reflect regions of the genome that were under different selective pressures during separate domestication events. These regions may contain genes that contribute to agronomic differences between races.

We began, for example, with a contrast between the Lowland (n=5) and Guatemalan (n=10) races. We calculated Fst based on 20kb non-overlapping windows along each contig, ultimately focusing on regions that include the top 1% of Fst scores (Figure S_manhattan). For this pairwise comparison, we identified 514 peaks containing 756 genes. We expect that this gene set is enriched for genes that differentiate the two races, including genes that contribute to agronomic traits that vary between the two races. Consistent with this expectation, the gene set contained gene with gene ontology (GO) functions that included putative agronomic functions, such as stress response, root meristem activity due to different ion availability, disease resistance, the control of volatile compound (VOCs) release in flowers, water stress response and osmoregulation, and affects on flowering time and promotion. We have highlighted genes that we deem to be particularly interesting in Table 3. Three of the genes in Fst peaks have homologs in *Arabidopsis thaliana* that affect flowering initiation and development and could, presumably, affect differentiation of flowering patterns between races.

Another gene is associated with fruit ripening.

We repeated this analysis for the Lowland (n=5) vs. Mexico (n=3) and for the Mexico (n=3) vs. Guatemalan (n=10) pairs. Focusing again on the top 1% of Fst peaks (Figure S_manhattan), we find that these two comparisons resulted in 773 and 716 genes within regions of high differentiation. These genes again included a variety of GO functions, some of which may contribute to differences in agronomic traits between races. These functions include herbivory resistance, flowering time regulation, resistance to fungal pathogens, pollinator VOC release, stress response pathways, biosynthesis of carotenoids, and flowering promotion, along with others. Altogether, these genes constitute a preliminary list of genes that may deserve additional study for potentially driving divergence – and alternative agronomic traits – among domestication groups.

Investigating potential causative loci that discriminate A and B flowering types

Finally, several of our accessions had known flowering types, with samples of n=13 A types and n=9 B types (Table 2). Importantly, within each flowering type, the samples traversed at least three of the four groups identified by admixture mapping. For example, the A types included samples from each of the three races (Mexican, Guatemalan and Lowland), the Hass/Gwen group, and hybrids. Given the distribution of A and B types across groups, we reasoned that contrasting these two samples are unlikely to be overly confounded by historical groups but instead may provide insights into genomic regions that contribute to this interesting phenotype.

To do so, we performed Fst analyses between the two groups, producing a Manhattan plot with peaks of differentiation between types (Figure 3). Focusing again on the top 1% peaks, we found 786 genes in these differentiated regions. Several genes had functions related to reproduction, including pollen germination, anther development, ovule development and pollen production. However, arguably the most interesting genes include those that affect the flowering time and circadian rhythms. Table 4 highlights four such genes, with homologs of *A. thaliana VOZ1* and *SPA1* among the most interesting. The former (*VOZ1*) regulates *FLC*, a transcription factor that functions as a repressor of floral transition and contributes to temperature compensation of the circadian clock. Given that A- and B-type flowerings respond to some extent to environmental cues, the potential implication of *FLC* regulation in A vs. B differentiation seems reasonable. *SPA1* contributes to the regulation of circadian rhythms in flowering processes, which is again consistent with A- vs. B-type differences.

We further hypothesized that differences between A- and B-types could be detectable as structural variants between types. We performed a preliminary test of this hypothesis by investigating sequencing coverage separately for A and B samples across the *P. americana SPA1* homolog (Figure 4). We initially found there was a drop in PacBio coverage of reads mapping to the 3' of *SPA1* relative to the 5' end. After further investigation, we found that the coverage of Illumina paired end reads mapped to the Gwen C+H genome differed significantly by flowering type (Figure 4). Analysis of the physical reads showed an intriguing pattern: breaks in coverage for avocado accessions between the last three exons with flowering type B at the 3' end (Figure 4), with no such pattern for A type accessions. This result suggests that there are consistent differences between A- and B-type accessions in this region and further suggests (but by no means proves) that *SPA1* could have a role in synchronous dichogamy. Further research will need to investigate the impact of this gene in avocado flowering.

3.5 Discussion

We have completed the assembly and annotation of a complex, diploid plant genome, representing the Gwen accession of *P. americana*. Our Gwen genome is more complete and contiguous than previously published avocado genomes [127], with better BUSCO, N50 and other summary statistics (Table 1). We ultimately produced two assemblies. The C+H assembly is contig-based that we reduced for potential alternative haplotypes using HapSolo [144]. HapSolo is one of a handful of methods designed to aid the assembly of diploid genomes [133, 122, 63] by identifying and removing alternative haplotypes. HapSolo optimizes the assembly with respect to the number of complete, singleton BUSCO genes and generally outperforms competing methods [144]. However, its effectiveness (and that of other methods) likely depends on the evolutionary distance between haplotypes, and it also depends on the accuracy of BUSCO characterization. The size of the C+H assembly, at 1,032Mb, suggests that some alternative haplotigs may still be present in the assembly, given that the expected avocado 1C genome size is ~900 Mbp.

We used the C+H assembly to build the second assembly, which is a scaffold of contigs that was guided by information from a genetic map of Gwen x Fuerte offspring [10]. This scaffolded assembly approaches chromosome level, with the longest 12 scaffolds representing 78% of the expected genome size. However, it suffers from at least two drawbacks that must be kept in mind. First, the use of genetic maps for scaffolding can introduce biases both from the mapping methods and from the identity of cultivars used in the crosses. In this case, we also suspect that the map is not quite dense enough to fully scaffold the genome, given that the assembly size decreased from 1,032Mb in the C+H assembly to 730Mb in the scaffolded assembly. Second, scaffolding by genetic maps provides insights into the order of contigs, but often lacks information about contig orientation. When contig orientation is ambiguous, it may be difficult to study large structural variants, such as inversions between accessions that may affect phenotype [175]. For this reason, we hope to eventually scaffold the Gwen genome using Hi-C or Bionano optical maps. These methods will not only help correct errors in contig orientation but may also help resolve difficult-to-assemble repeat-regions. Despite these two drawbacks, we again have to accentuate that our Gwen genome is far superior to previously available P. americana genomes and that our scaffolded version (at 730Mb) is likely much more complete than the 421Mb Hass genome, which was also scaffolded with a genetic map [127] (Table 1, Figure 1A).

We generated the Gwen genome as a platform for evolutionary genomics and also for downstream breeding applications. Both goals are aided by accurate genome annotation. We first annotated the genome for repeats, identifying 65% of the genome that consists of repetitive elements (Figure 1B). This proportion is not particularly large or notable for plant genomes, especially given that some larger plant genomes (like maize) consist of ¿80% transposable elements [146]. After masking for annotated repetitive elements, we predicted genes using publicly available RNAseq software and the BRAKER pipeline (see Methods). We predicted 57,916 genes on the scaffolded assembly and 87,617 in the C+H assembly, both of which are two-fold more than the 25,211 predicted on the Hass genome. These results initially led us to believe that our approaches had predicted too many genes, despite their reliance on experimental (RNAseq) evidence. If we do have too many genes, some are likely coding regions of transposable elements that were missed by our repeat annotation pipeline. Others are likely to be pseudogenes; we attempted to minimize this category by removing any gene with exons that overlapped with annotated TE's, but of course this has the potential the downside of removing actual genes with function [61, 143]. Nonetheless, while it is certainly possible that we have over-predicted coding regions, our results are more in keeping with a recent transcriptome analysis that predicted 63,000 genes in avocado [32]. We thus suspect that the 25,000 genes annotated on the Hass genome do not represent the complete avocado gene space. That said, the Hass gene annotations appear to be accurate, because all 25,211 genes are present within our annotation set.

Given the Gwen assemblies and annotations, we used the Gwen as a reference for preliminary evolutionary genomic analyses. These analyses were facilitated by a sample of 33 high coverage (>14x) accessions that were chosen to represent the three botanical races of avocado. With the ultimate goal of learning more about the history of the three independent domestication events and the differences among them, our first task was to identify SNPs and then to evaluate genetic relationships among accessions. Both PCA and admixture analyses clearly differentiated among the three botanical races (Figure 2), but they also provided insights into the hybrid origins of some cultivars, representing the first whole genome insights into hybrid origins for most samples. Many of our results confirmed results based on microsatellite and other marker types [11, 40, 39]. For example, many of the accessions in our sample (like Zutano and Bacon) were previously thought to be hybrids, and we have confirmed those inferences here. However, our results also offered some surprises, most notably about the history of Hass, which was traditionally thought to be of Guatemalan origin [11, 39] but has been inferred to be roughly 50% Guatemalan and 50% Mexican from genetic analyses [39, 127]. In our analyses, Hass groups with other Guatemalan accessions (Figure 2A) and is identified as 100% Guatemalan in admixture analyses with K=3 (Figure 2C).

One prosaic explanation for these results could be that we have mislabelled our Hass accession; we think this unlikely because Hass groups with other accessions that are its close relatives (e.g., Gwen and Mendez) (Figure 2). A more nuanced interpretation is that admixture analyses are heavily dependent on the samples used in analyses, and this may affect the assignment of groups with K=4 (Figure 2D). One simple expediency will be to down-sample the number of Hass' close relatives and then repeat admixture analyses to see if K=4 is still the most likely grouping. Nonetheless, both sets of our clustering analyses (i.e., PCA and admixture; Figure 2) suggest that the history of Hass - the most cultivated accession in the world - has not yet been well characterized by genetic analyses. If our results are correct, then it begs the question: why are our results so different from previous inferences? One of the analyses [39] was based on four nuclear loci, which is (in retrospect) a genomic region small enough to provide potentially misleading results. The more recent work on the Hass genome [127] mapped the potential origin of each chromosomal region, so there were few genomic limits to these analyses. However, the Rendon-Anaya et al. (2019) study included few (n=11) samples; this sampling could give misleading results if it did not sufficiently represent the breadth of genetic diversity in *P. americana sensu lato*. As we have shown, their sample also included an inappropriate outgroup. Ultimately, we cannot yet ascribe a definitive cause to differences among studies, but it is possible that more analysis of our existing data (e.g., chromosome painting and analysis of regions that are identical by descent) could prove illuminative.

The admixture analysis permitted the definition of "pure" groups - i.e., accessions that exceeded 80% assignment to one of the three traditional botanical races. This group definition was necessary to remove potential hybrids, but it had the unfortunate effect of greatly reducing the size of the Mexican and Lowland samples (Table 2). Given low sample sizes, we must recognize the high variance associated with analyses like comparisons of Fst between groups. We nonetheless applied Fst to investigate genomic regions of high differentiation between races (Figure S2), with the rationale that regions of high differentiation house genes that contribute to divergent traits among races. Using this approach, we have identified hundreds of plausible candidate genes with potential functions in salt tolerance, drought resistance, climate adaptation, fruit ripening and other physiological responses (Table 3). Our next steps will be to confirm that these are highly divergent regions (using Dxy or similar) and to complement these inferences to search for signals of selective sweeps. It may also be fitting to infer structural variants, which can provide additional signals for localized genomic regions of high divergence. Any remaining candidates will have to be evaluated using experimental or genetic techniques.

Finally, we also applied the same Fst approach to A- vs B- type flower, yielding potentially exciting results. Several flowering time genes were identified among the 800 genes located in peaks of differentiation (Figure 3), including a regulator of *FLC* and a gene known to affect circadian rhythms in flowering (Table 4). This last gene (*PaSPA1*) exhibits an intriguing difference in sequence coverage between A- and B-type accessions (Figure 4), suggesting that a structural variant may differentiate the alleles of the two different flowering types. Nonetheless, these candidate genes - like those identified between botanical races - need to be subjected to additional evolutionary and genetic analyses to better evaluate the strength of their candidacy. It is exciting, however, to speculate that it may be possible to identify genes that contribute to flowering type, because they will be of both fundamental biological interest and also because they may have practical breeding utility.

3.6 Tables and Figures

	Gwen + HapSolo Contigs	Gwen Scaffolds	Hass Contigs	*drymifolia* Contigs
Assembly Size (Mb)	1,032	703	913	823
Number of Fragments	989	12	8,135	42,722
Largest Fragment (kb)	17,080	85,354	2,811	4,611
Percent of Assembly in 12 Largest Fragments	17.00%	78.20%	2.74%	4.32%
Percent of Assembly in Fragments > 10 Mb	17.00%	78.20%	0.00%	0.00%
N50 (Mb)	3.37	61.89	0.30	0.32
L50	176	5	770	502
BUSCO (%)	90.7	87.9	84.9	86.3

Table 3.1: Assembly Metrics for two Gwen assemblies, the Hass assembly and the *drymifolia* assembly.

Historical Classification	Gwen + Accessions^{1,2}
Mexican ($n = 8$)	001-01, 069-02, Bacon$_{\mathrm{B}}$, Ganter$_{\mathrm{B}}$*, Pequeno Charly*, Topa Topa$_{\mathrm{A}}$*, Zutano$_{\mathrm{B}}$
Guatemalan ($n = 14$)	Anaheim$_{\mathrm{A}}$, Carlsbad$_{\mathrm{A}}$, CH-G-7, CH-G-10, CH-G-11, Gwen$_{\mathrm{A}}$*, Linda$_{\mathrm{B}}$, Lyon$_{\mathrm{B}}$, Nabal$_{\mathrm{B}}$*, Nimlioh$_{\mathrm{B}}$, Reed$_{\mathrm{A}}$*, Taft$_{\mathrm{A}}$*, Thille$_{\mathrm{B}}$*, Velvick, Hass$_{\mathrm{A}}$*
Lowland ('West Indies') ($n = 4$)	23-6*, Simmonds$_{\mathrm{A}}$*, VC26$_{\mathrm{A}}$*, Waldin$_{\mathrm{A}}$*
Hybrid ($n = 7$)	Fairchild$_{\mathrm{A}}$, Fuerte$_{\mathrm{B}}$, Hass$_{\mathrm{A}}$, Mendez$_{\mathrm{A}}$, Pinkerton$_{\mathrm{A}}$, Gwen$_{\mathrm{A}}$, CH-CR-25
Outgroups ($n = 4$)	*P. schiedeana, P. donnell-smithi, P. hintonii, Ocotea boranthea*

Table 3.2: A list of accessions used in diversity studies. Additional details are provided in Table S1

[1] Accessions with an asterisk (*) represent samples that were assignable to one of the three historical races based on an assignment that exceeded 80% (i.e., $Qi > 0.8$). These are the accessions used to compare Fst values between groups.

[2] Accessions with a superscripted A or B represent the A and B flowering type samples that we used in analyses.

Gene[1]	Homolog[2]	Homolog Function[3]	Reference[4]
jg81425, jg86221, jg46538	*SGR (Capsicum annuum)*	Triggers chlorophyll degradation during leaf senescence and fruit ripening	[23, 15]
jg16644	*ACCO (Persea americana)*	Fruit ripening via ethylene biosynthesis	N/A
jg36084	*CRY1 (Arabidopsis thaliana)*	Photoreceptor that control floral initiation and regulates other light responses	[96, 4]
jg114403, jg48881, jg74004	*PCFS4 (Arabidopsis thaliana)*	Promotes flowering through the pre-mRNA processing of flowering time genes	[173, 165]
jg97581, jg51837, jg55593, jg46613	*NAC35 (Arabidopsis thaliana)*	Transcription factor that acts as a floral repressor.	[115, 169]

Table 3.3: A partial list of candidate genes found in peaks of Fst differentiation between Guatemalan and Mexican samples.

[1] Gene number in Gwen annotation file.

[2] A homolog to the gene, as identified by functional analyses with SwissProt, with the species in which the homolog was identified.

[3] The inferred function of the homolog.

[4] Citations to studies that describe homolog function.

Gene[1]	Homolog[2]	Homolog Function[3]	Reference[4]
jg92492	*VOZ1 (Arabidopsis thaliana)*	Regulates *Flowering Locus C (FLC)*	[106, 167]
jg63228	*SPA1 (Arabidopsis thaliana)*	Controls normal photoperiodic flowering and regulates circadian rhythms	[176, 88, 74, 16]
jg46090, jg46091	*CCR1 (Petunia hybrida)*	Biosynthesis of volatile compounds in flowers	[111]
jg69989	*EOBI (Petunia hybrida)*	Transcription factor for volatile compounds in flowers	[145]

Table 3.4: Particularly interesting candidate genes found in Fst peaks between samples of A- and B-type flowering accession.
[1] Gene number in Gwen annotation file.
[2] A homolog to the gene, as identified by functional analyses with SwissProt, with the species in which the homolog was identified.
[3] The inferred function of the homolog.
[4] Citations to studies that describe homolog function.

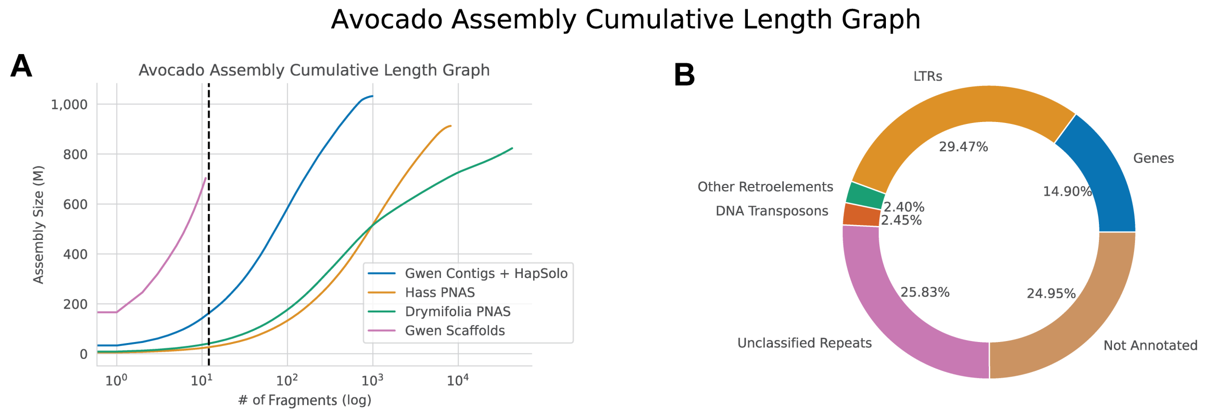

Figure 3.1: (**A**) The cumulative sum assembly graph size (cdf) shows the size of the assemblies with the next largest consecutive contig (or scaffold) being added to the sum along the x-axis. It illustrates cdf for the Hass contig assembly and the drymifolia assembly - both of which were published previously - and the two Gwen assemblies (C+H and scaffolded). (**B**) The annotation results for the Gwen C+H assembly, showing the percentage of the genome attributable to different types of transposable elements, including DNA transposons, unclassified retroelements and Long Terminal Repeat (LTR) retroelements. The percentage of the genome attributable to genes did not include the length of introns.

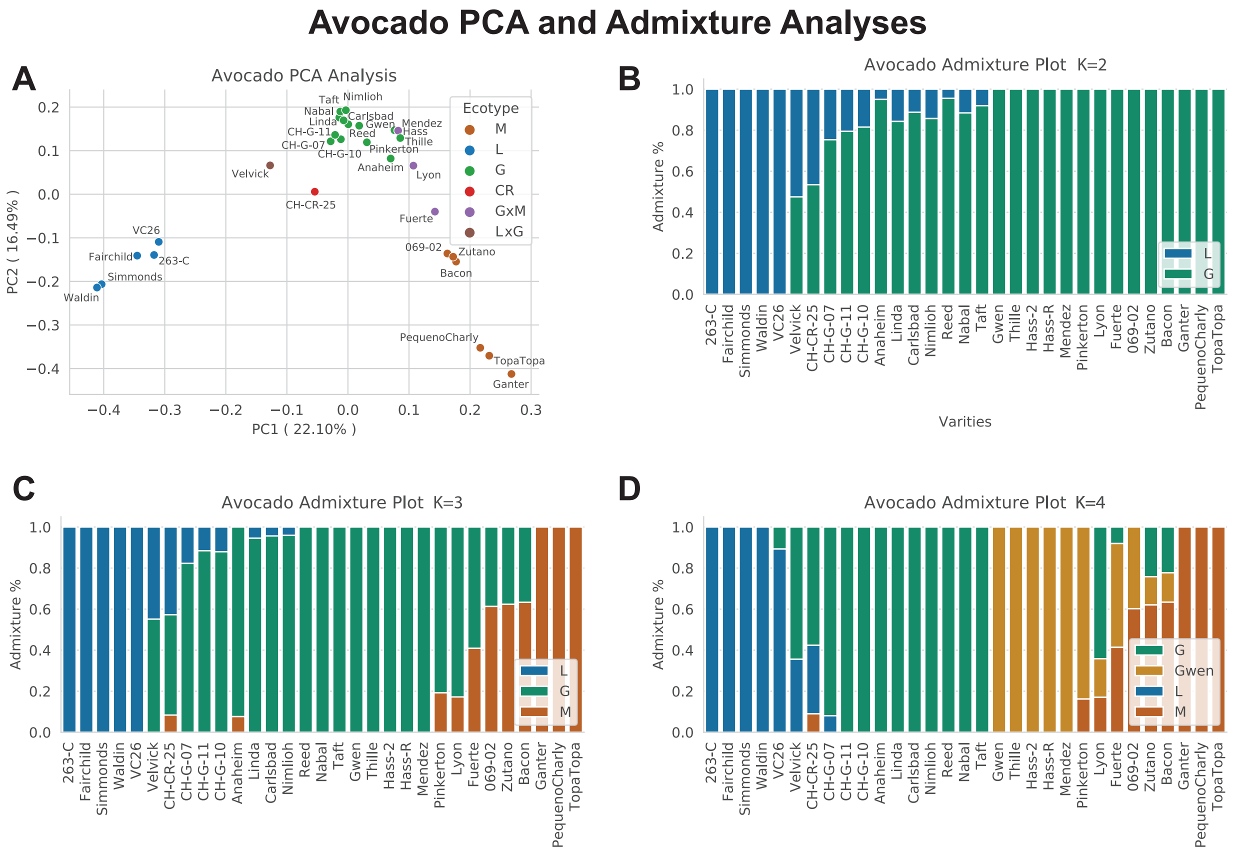

Figure 3.2: (**A**) A PCA analysis of SNP diversity, based on 33 accessions of avocado. Each dot represents an individual, and the color of each dot represents an historical classification of that individual, including Guatemalan x Mexican (GxM) hybrids and Lowland x Guatemalan (LxG) hybrids. (**B,C,D**) Admixture plots were generated using SNPs found in all of the contigs in the Gwen C+H assembly. Plot (**B**) shows the inferred groups with K=2 groups; plot (**C**) represents K=3, which is consistent with the three historical races; and (**D**) represents K=4 groups, which was inferred to be the most likely number of clusters. The K=4 groups include the three historical races and an emergent group showing accessions closely related to Hass.

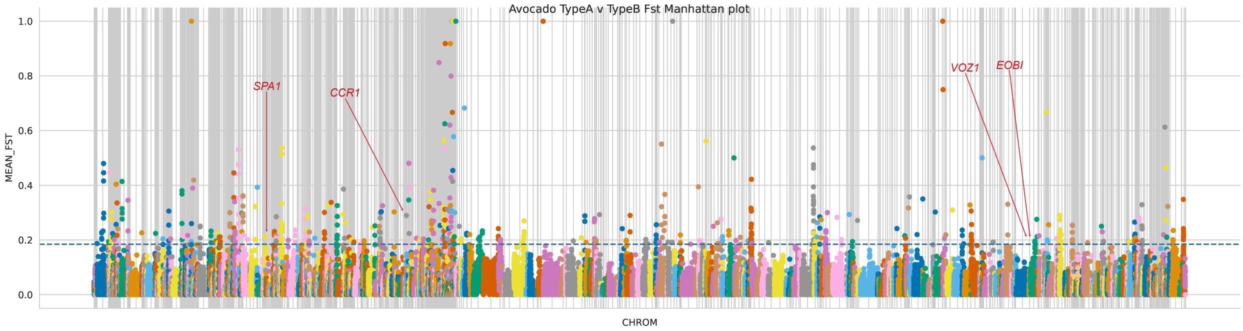

Figure 3.3: A Manhattan plot showing the Fst values across all of the contiguous of the C+H Gwen assembly, focusing on contrasts between the A-type (n=9) and B-type (n=13) flowering samples. The top 1% (horizontal dashed line with a value of 0.184121) of Fst peaks define 786 genes, all of which were annotated functionally. Several of these genes are viable candidates for affecting flowering type; we list four especially interesting genes in Table 3 and indicate their location with an arrow on the Manhattan plot.

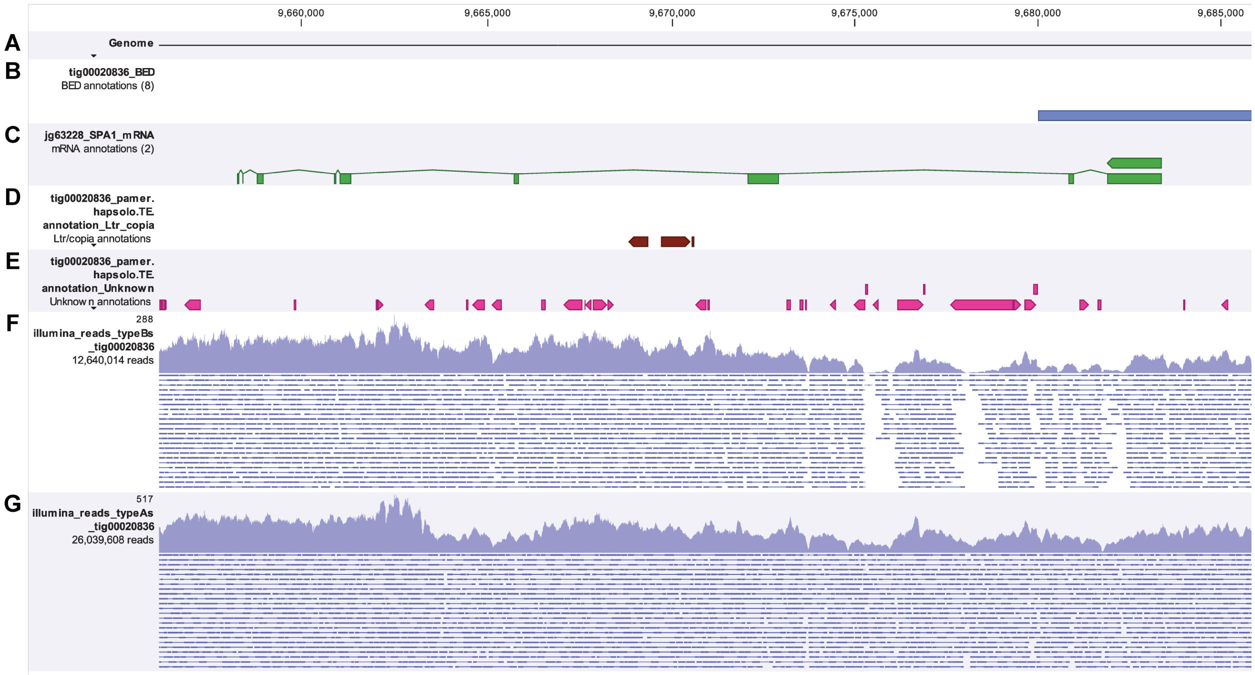

Figure 3.4: Coverage information for the flowering-type candidate gene SPA1 in contig tig00020836. (**A**) The position of the gene region in the contig. (**B**) The BED annotation of the Fst peak containing the *SPA1* gene found in this contig. (**C**) The annotation of the *PaSPA1* gene, which has two predicted transcripts. (**D**) The annotation of LTR copia elements. (**E**) The annotation of unknown/unclassified repeats. (**F**) Illumina read coverage of all type B flowering types of avocados with an average read depth of ~160x at positions <6,665,000, and an average read depth of ~40x coverage at positions <6,675,000, showing a substantial drop in coverage ($\frac{1}{4}$) on the right side of the gene. (**G**) Illumina read coverage of all type A flowering types of avocados with an average read depth of ~300x at positions <6,665,000, and an average read depth of ~150x coverage at positions <6,675,000, also showing a drop in coverage ($\frac{1}{2}$) of the right side of the gene but not as drastic.

CONCLUSION

Comparative and Population Genomics

With the accelerating detrimental effects of climate change, and our ever growing population, a strong solid foundation and understanding of the genetic basis of phenotypic variation that currently exists in food crops is essential for our future and survival. This requires the curation and availability of good genomic tools for researchers seeking to solve these important problems. In short, more contiguous genomes of food crops are required in order to accelerate crop improvement for their adaptation and efficiency in a rapidly changing climate.

In my dissertation I sought to (1) improve on methods for the curation of a more continuous and complete genome that is also affordable making it also more accessible, (2) develop a tool for purging of duplicated or alternate haplotypes in order to isolate these complete/pseudo haplotypes and (3) apply these methods to a poorly understood perennial crop with potential for crop improvement, avocados, and identify regions of genetic variation that differ among populations of avocado that contain genes implicated in domestication and adaptation. This work can be replicated by other researchers and applied to other species important for improving food security.

In the first chapter, I described resequencing of *Drosophila melanogaster*, the first genome assembled using a whole-genome shotgun (WGS) strategy [112, 2]. The original sequencing of this species became a proof-of-principle genome assembly [112, 2] that led to the prevalence of the WGS sequencing approach as a tool in virtually all subsequent metazoan genome assembly projects [87, 160, 60, 170, 8, 59, 68]. Although improvements in sequencing technologies has lowered the cost of sequencing [162], the reduced size and length of reads has resulted in limitations, such as highly fragmented and incomplete genome assemblies

[5, 113]. Even with more advanced approaches (e.g. hierarchical shotgun sequencing and other clone-based approaches) that aided in building nearly complete and highly contiguous reference genomes, they still add complexity, cost and time to assembly projects. This however was changed with the development of new long-read sequencing technologies that drastically improved read lengths, allowing for reads to span entire regions of the genome that were once impossible to resolve. Over time, these long reads improved in accuracy, length and throughput, but with error-rates and costs that were still higher than short read sequencing. Additionally, further developments were made that made it possible to overcome many of the difficulties associated with long-read sequencing, and has led to nearly complete and highly contiguous metazoan genome assemblies using only a whole genome sequencing (WGS) strategy [80, 21, 82]. This however still incurred a significant cost as it required the hiring of highly skilled staff, and a large investment into the platform. For example, the recent resequencing of *D. melanogaster* and other *Drosophila* species ranged in the 10's of thousands of dollars [70, 36, 37]. This cost would only double if one were to sequence a heterozygous diploid individual, as it would require twice as much sequencing for a genome of similar size as Drosophila, let alone sequencing a larger and/or multiple genomes. Such costs make it difficult for research groups to answer important questions that require a reference level assembly. Many of which could be due to the presence of SV's [36, 37, 175]. In this first chapter, we sought to produce a reference level assembly with a small group of researchers, and at a fraction of the cost, ~$1,000, while still retaining completeness, accuracy and contiguity expected from a reference level assembly. By utilizing novel applications of published genome assembly methods [36] and inexpensive long noisy reads, we were able to do just that. With our approach, we were able to identify over 99% of genes but also discovered novel SV's, while preserving over 90% of highly conserved single copy genes (BUSCO's). We achieved this by using 4.2 Gbp of sequencing data from a single Oxford Nanopore flowcell in conjunction with short-read Illumina data. Using this data we utilized the assembly merging approach [36] and polished our assembly using accurate short-reads [159]. Our published as-

sembly exhibited a N50 of 21.3 Mbp which is comparable to the N50 of the FlyBase release 6 of reference assembly (18.9 Mbp) [112, 2], which is considered the gold standard for the *D. melanogaster* research community. Further comparisons of our assembly to the FlyBase reference genome assembly showed that our assembly was both accurate (21 mismatches/100 kbp, 36 indels/100 kbp) and collinear. Many of these errors could be attributed to errors introduced by the noisy ONT reads that escaped polishing, but also due to SV's that have accumulated over time in a laboratory environment for approximately 350 generations (assuming 20 generations/year) since the initial sequencing in 2000. These mutations (including deleterious mutations) can easily be explained by genetic drift [12]. In our assembly we found 28 homozygous euchromatic TE insertions (evidenced by long-reads spanning these regions), which are predominantly LTR and defective hobo elements, suggesting a high rate of euchromatic TE insertions (~0.08 insertion/gen). The observation of a predominance of LTR and hobo elements among the new TE insertions mirrors their recent spread in *D. melanogaster* populations [116, 118, 24, 49] in previous studies. The abundance of defective hobo elements among the new insertions is particularly interesting given that these hobo elements lack the transposase enzyme necessary for mobilization. Collectively, our assembly provides a snapshot of ongoing genome structure evolution in a metazoan genome, which is often assumed to be approximately invariant for experimental genetics.

A crucial feature of this work is that it is performed in a strain used to generate one of the highest quality reference genomes available, ensuring that our inferences can be judged against a high-quality standard. This approach allowed us to demonstrate that assembly with modest amounts of long-molecule data paired with inexpensive short-read data can yield highly accurate and contiguous reference genomes with minimal expenditure of resources (by an order of magnitude lower). This demonstration opens a myriad opportunities for high-quality genomics in systems with limited resources for genome projects. Moreover, we can now conceive of studying entire populations with high-quality assemblies capable of resolving repetitive structural variants, something previously inconceivable and unattainable with short-read sequencing alone.

In the second chapter of my thesis, we expanded upon methods used in chapter 1 for application to more heterozygous samples, specifically, chardonnay wine grapes. As more contiguous genomes are important for obtaining near complete annotations, improved scaffolding and evolutionary inferences, which are all essential for producing genomic resources for crops, understanding population structure and identifying regions of high divergence that contain genes implicated in domestication and adaptation. This led to the construction of a genome assembly with an assembly size nearly double that of the expected size [175]. We were however able to correct this by removing contigs matching alignment thresholds designed to identify alternate haplotypes, as indicated by contigs containing duplicate BUSCO's, and applying them genome wide. This required the development of a novel approach which could be generalized and applied to any metazoan species. The result of this work led to the capture of SV's and homozygous loci implicated in berry color and sex determination [175]. Since the amount of heterozygosity is different and specific to each species, population, and individual, we sought to create an algorithm that could customize a specific set of thresholds that would be tailored to each sample for purging of alternate haplotypes. This optimization was possible using a variant of Hill Climbing, "Random Forward Walking Hill Climbing", along with alignment parameters that could encompass different aspects present in heterozygous loci containing, SNP's, indels and SV's (ID, Q and QR). Our approach is able to capture these properties by approximating these parameters (ID, Q and QR) by classifying contigs with high similarity and containing another BUSCO already present in another contig. We optimized and approximated these alignment parameters using a linear formula that minimizes the number of duplicate, missing and fragmented BUSCO's, while maximizing the number of single copy BUSCO's. Once these parameters are optimized, these new values are set as thresholds for the removal of contigs matching said thresholds, genomewide. This approach, albeit imperfect, was comparable to or better than previously published methods, purge_dups [63], from species across three separate taxa with varying genome sizes. This led to improvements in scaffold N50 (63%, 0%, 13%), largest scaffolds (47%, 28%, 36%), and proportion of assembly in k (where k represents the number of chromosomes per species) largest scaffolds (50%, 2%, 22%) for mosquito, chardonnay grapes and the thorny skate respectively when compared to purge_dups.

Overall, we successfully identified many alternative contigs, thereby reducing the noise that has plagued heterozygous genomes for the past several years, and matched or out performed the best tool currently available across multiple taxa, leading to increased performance of downstream analysis such as Hi-C scaffolding, annotation and evolutionary inferences. That is not to say, however, that HapSolo cannot be improved. We can see three obvious areas for further improvement. The first is to consider coverage statistics, which represents a point of departure between our approach and that of both Purge Haplotigs [132] and purge_dups. We predict, but do not yet know, that the inclusion of coverage with our existing alignment statistics could lead to more accurate inferences, but it is possible that it could reduce the bias dependence of species that perform poorly using BUSCO's. A second area of improvement may be to implement alternative optimization algorithms, such as simulated annealing. Finally, it may also be possible to include additional features in the calculations of Cost. Our present reliance on BUSCO's has the advantages of speed and wide acceptance in the genomics community. However, depending on the initial assembly, it is likely that some contigs do not contain a BUSCO gene, and are therefore not considered in calculating Cost, thereby not contributing to the approximation of threshold parameters (ID_T, Q_T, and QR_T). It is not yet clear what additional features could be included in the Cost function, but identifying contigs containing an over-representation of shared k-mers is one possibility.

In my last and final chapter, we take the knowledge gained from the first two chapters and apply it to a highly complex, heterozygous and interesting diploid perennial food crop, representing the Gwen accession of *P. americana* (avocado), as resequence 21 accessions of avocado using short-read sequencing. We create a highly contiguous decoupled genome assembly, annotate it's genetic features, surpassing contiguity and accuracy of previously published avocado genomes [127], reconstruct the population structure of avocados, as well as identify regions of the genome that differ significantly between groupings of avocado accessions containing unique phenotypes. These regions contain genes implicated in climate adaption, disease resistance, agronomic importance and flowering time. We also identified an SV present in one of the flowering time loci that is present in type A avocados but absent in type B avocados. In this study we produce two genomes, an assembly generated using Canu [82] which was reduced/decoupled (C+H) using methods from chapter 2, HapSolo [144], and a scaffolded assembly containing contigs from the C+H assembly using a genetic map from a previous study [10]. The C+H assembly at approximately 1 Gbp suggests some alternative haplotigs may still be present as the expected 1C genome of avocados is ~900 Mbp. The scaffolded assembly approaches chromosome level, with the longest 12 scaffolds representing 78% of the expected genome size, whereas the previously published avocado scaffolded assembly only represented approximately half of the expected genome size in its entirety. One issue with our scaffolding approach is that it can introduce biases both from the mapping methods and from the identity of the cultivars used in the crosses. In this case the genetic map was generated using crosses between Gwen and Fuerte. We also believe that the decrease of the scaffolded assembly size from 1 Gbp in the C+H assembly to 730 Mbp in the scaffolded assembly could also be due to the lack of density in the genetic map. Another issue lies in orientation of the contigs in the scaffolds, as scaffolding by genetic maps often lacks information about contig orientation. When contig orientation is ambiguous, it may pose a problem to the study of large structural variants, such as inversions between accessions that may affect phenotype [175]. For this reason, we hope to eventually scaffold the Gwen genome using Hi-C or Bionano optical maps. These methods will not only help correct errors in contig orientation but may also help resolve difficult-to-assemble repeat-regions. Despite these two drawbacks, we again have to accentuate that our Gwen genome is far superior to previously available P. americana genomes and that our scaffolded version (at 730Mb) is likely much more complete than the 421Mb Hass genome, which was also scaffolded with a genetic map [127].

One particular interesting aspect of our genome is the presence of repeats. We found that ~65% of the genome consists of repetitive elements. This proportion is not particularly large or notable for plant genomes, especially given that some larger plant genomes (like maize) consist of >80% transposable elements [146]. After masking for annotated repetitive elements, we predicted genes using publicly available RNAseq software and the BRAKER pipeline (see Methods). We predicted 57,916 genes on the scaffolded assembly and 87,617 in the C+H assembly, both of which are two-fold more than the 25,211 predicted on the Hass genome. Although it is certainly possible that we have over-predicted coding regions, our results are more in keeping with a recent transcriptome analysis that predicted ~63,000 genes in avocado [32]. We thus suspect that the ~25,000 genes annotated on the Hass genome do not represent the complete avocado gene space. That said, the Hass gene annotations appear to be accurate, because all 25,211 genes are present within our annotation set.

Given the Gwen assemblies and annotations, we used the Gwen as a reference for preliminary evolutionary genomic analyses. These analyses were facilitated by a sample of 33 high coverage (>14x) accessions that were chosen to represent the three botanical races of avocado. With the ultimate goal of learning more about the history of the three independent domestication events and the differences among them, our first task was to identify SNPs and then to evaluate genetic relationships among accessions. Both PCA and admixture analyses clearly differentiated among the three botanical races, but they also provided insights into the hybrid origins of some cultivars, representing the first whole genome insights into hybrid origins for most samples. Many of our results confirmed results based on microsatellite and other marker types [11, 40, 39]. For example, many of the accessions in our sample (like Zutano and Bacon) were previously thought to be hybrids, and we have confirmed those inferences here. However, our results also offered some surprises, most notably about the history of Hass, which was traditionally thought to be of Guatemalan origin [11, 39] but has been inferred to be roughly 50% Guatemalan and 50% Mexican from genetic analyses [39, 127]. In our analyses, Hass groups with other Guatemalan accessions and is identified as 100% Guatemalan in admixture analyses with K=3. One prosaic explanation for these results could be that we have mislabelled our Hass accession; we think this unlikely because Hass groups with other accessions that are its close relatives (e.g., Gwen and Mendez). A more nuanced interpretation is that admixture analyses are heavily dependent on the samples used in analyses, and this may affect the assignment of groups with K=4. One simple expediency will be to down-sample the number of Hass' close relatives and then repeat admixture analyses to see if K=4 is still the most likely grouping. Nonetheless, both sets of our clustering analyses (i.e., PCA and admixture) suggest that the history of Hass - the most cultivated accession in the world - has not yet been well characterized by genetic analyses. If our results are correct, then it begs the question: why are our results so different from previous inferences? One of the analyses (Chen et al. 2009) was based on four nuclear loci, which is (in retrospect) a genomic region small enough to provide potentially misleading results. The more recent work on the Hass genome [127] mapped the potential origin of each chromosomal region, so there were few genomic limits to these analyses. However, the Rendon-Anaya et al. (2019) study included few (n=11) samples; this sampling could give misleading results if it did not sufficiently represent the breadth of genetic diversity in *P. americana sensu lato*. As we have shown, their sample also included an inappropriate outgroup. Ultimately, we cannot yet ascribe a definitive cause to differences among studies, but it is possible that more analysis of our existing data (e.g., chromosome painting and analysis of regions that are identical by descent) could prove illuminative.

The admixture analysis permitted the definition of "pure" groups - i.e., accessions that exceeded 80% assignment to one of the three traditional botanical races. This group definition was necessary to remove potential hybrids, but it had the unfortunate effect of greatly reducing the size of the Mexican and Lowland samples (n=14, 8 and 4 for G, M and L). Given low sample sizes, we must recognize the high variance associated with analyses like comparisons of Fst between groups. We nonetheless applied Fst to investigate genomic regions of high differentiation between races, with the rationale that regions of high differentiation house genes that contribute to divergent traits among races. Using this approach, we have identified hundreds of plausible candidate genes with potential functions in salt tolerance, drought resistance, climate adaptation, fruit ripening and other physiological responses. Our next steps will be to confirm that these are highly divergent regions (using Dxy or similar) and to complement these inferences to search for signals of selective sweeps. It may also be fitting to infer structural variants, which can provide additional signals for localized genomic regions of high divergence. Any remaining candidates will have to be evaluated using experimental or genetic techniques.

Finally, we also applied the same Fst approach to A-type vs B-type flower, yielding potentially exciting results. Several flowering time genes were identified among the 800 genes located in peaks of differentiation, including a regulator of *FLC* and a gene known to affect circadian rhythms in flowering. This last gene (*PaSPA1*) exhibits an intriguing difference in sequence coverage between A- and B-type accessions, suggesting that a structural variant may differentiate the alleles of the two different flowering types. Nonetheless, these candidate genes - like those identified between botanical races - need to be subjected to additional evolutionary and genetic analyses to better evaluate the strength of their candidacy. It is exciting, however, to speculate that it may be possible to identify genes that contribute to flowering type, because they will be of both fundamental biological interest and also because they may have practical breeding utility.

Our climate is rapidly changing, with higher maximum temperatures (Figure 1), extreme weather conditions and longer drought periods. Many of which will have a detrimental effect on areas important for food production across the globe. Some of which we are experiencing at the time of writing this. These reasons highlight the importance and imperative of having better genomic tools, which are vital for understanding the relationship between genetics and phenotypic variation. Thes requires more continuous, complete, accurate and haplotype decoupled genome assemblies for identifying SV's among haplotypes, individuals and populations. With a better understanding of the genetic variation of SVs and their effects on adaptive traits in crops, we can peer into this unknown and provide critical information for downstream application to crop improvement. Not only will this information be important for future breeding programs for further improvement of avocados, but our methods can be not only used and replicated across food crops, but also across other species facing hurdles in adapting to a rapidly changing climate. For example, with this knowledge, the scientific community can not only replicate this work in other species but also accelerate adaptation of species with more informed and focused breeding programs, but also introduce genetic mutations using methods such as CRISPR. These genetic mutations can in theory accelerate the adaptation of species to rapidly changing environments. Science, more specifically, comparative genomics along with population genetics of SVs will help pave the way to a better understanding of the evolution and adaptation of crops, and other species in general, at the genetic level. This will help give insights to understanding the phenotypic plasticity in traits important for adaptation to a broader range of environments.

Bibliography

[1] Genome sequence of the nematode c. elegans: A platform for investigating biology. *Science*, 282(5396):2012–2018, 1998. cited By 3217.

[2] M. Adams, S. Celniker, R. Holt, C. Evans, J. Gocayne, P. Amanatides, S. Scherer, P. Li, R. Hoskins, R. Galle, R. George, S. Lewis, S. Richards, M. Ashburner, S. Henderson, G. Sutton, J. Wortman, M. Yandell, Q. Zhang, L. Chen, R. Brandon, Y.-H. Rogers, R. Blazej, M. Champe, B. Pfeiffer, K. Wan, C. Doyle, E. Baxter, G. Helt, C. Nelson, G. Gabor Miklos, J. Abril, A. Agbayani, H.-J. An, C. Andrews-Pfannkoch, D. Baldwin, R. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. Beasley, K. Beeson, P. Benos, B. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. Burtis, D. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. Michael Cherry, S. Cawley, C. Dahlke, L. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. Deslattes Mays, I. Dew, S. Dietz, K. Dodson, L. Doup, M. Downes, S. Dugan-Rocha, B. Dunkov, P. Dunn, K. Durbin, C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. Gabrielian, N. Garg, W. Gelbart, K. Glasser, A. Glodek, F. Gong, J. Harley Gorrell, Z. Gu, P. Guan, M. Harris, N. Harris, D. Harvey, T. Heiman, J. Hernandez, J. Houck, D. Hostin, K. Houston, T. Howland, M.-H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. Karpen, Z. Ke, J. Kennison, K. Ketchum, B. Kimmel, C. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. McIntosh, M. McLeod, D. McPherson, G. Merkulov, N. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. Mount, M. Moy, B. Murphy, L. Murphy, D. Muzny, D. Nelson, D. Nelson, K. Nelson, K. Nixon, D. Nusskern, J. Pacleb, M. Palazzolo, G. Pittman, S. Pan, J. Pollard, V. Puri, M. Reese, K. Reinert, K. Remington, R. Saunders, F. Scheeler, H. Shen, B. Christopher Shue, I. Siden-Kiamos, M. Simpson, M. Skupski, T. Smith, E. Spier, A. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. Wang, X. Wang, Z.-Y. Wang, D. Wassarman, G. Weinstock, J. Weissenbach, S. Williams, T. Woodage, K. Worley, D. Wu, S. Yang, Q. Alison Yao, J. Ye, R.-F. Yeh, J. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. Zheng, F. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. Smith, R. Gibbs, E. Myers, G. Rubin, and J. Craig Venter. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000. cited By 4508.

[3] J. A. Aguirre-Liguori, S. Ramírez-Barahona, P. Tiffin, and L. E. Eguiarte. Climate

change is predicted to disrupt patterns of local adaptation in wild and cultivated maize. *Proc. Biol. Sci.*, 286(1906):20190486, July 2019.

[4] M. Ahmad, J. A. Jarillo, O. Smirnova, and A. R. Cashmore. The CRY1 blue light photoreceptor of arabidopsis interacts with phytochrome a in vitro. *Mol. Cell*, 1(7):939–948, June 1998.

[5] C. Alkan, B. Coe, and E. Eichler. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376, 2011. cited By 818.

[6] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. cited By 63753.

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct. 1990.

[8] S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J.-M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M. Sollewijn Gelpke, J. Roach, T. Oh, I. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S. Smith, M. Clark, Y. Edwards, N. Doggett, A. Zharkikh, S. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner. Whole-genome shotgun assembly and analysis of the genome of fugu rubripes. *Science*, 297(5585):1301–1310, 2002. cited By 1172.

[9] K. Arumuganathan and E. D. Earle. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.*, 9(3):208–218, Aug. 1991.

[10] V. E. Ashworth, H. Chen, C. L. Calderón-Vázquez, M. L. Arpaia, D. N. Kuhn, M. L. Durbin, L. Tommasini, E. Deyett, Z. Jia, M. T. Clegg, and Others. Quantitative trait locus analysis in avocado: The challenge of a slow-maturing horticultural tree crop. *J. Am. Soc. Hortic. Sci.*, 144(5):352–362, 2019.

[11] V. E. T. M. Ashworth and M. T. Clegg. Microsatellite markers in avocado (persea americana mill.): genealogical relationships among cultivated avocado genotypes. *J. Hered.*, 94(5):407–415, Sept. 2003.

[12] Z. Assaf, S. Tilk, J. Park, M. Siegal, and D. Petrov. Deep sequencing of natural and experimental populations of drosophila melanogaster reveals biases in the spectrum of new mutations. *Genome Research*, 27(12):1988–2000, 2017. cited By 10.

[13] F. F. Barbier, T. G. Chabikwa, M. U. Ahsan, S. E. Cook, R. Powell, M. Tanurdzic, and C. A. Beveridge. A phenol/chloroform-free method to extract nucleic acids from recalcitrant, woody tropical species for gene expression and sequencing. *Plant Methods*, 15:62, June 2019.

[14] D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Strömberg, and G. T. Marth. BamTools: a c++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):1691–1692, June 2011.

[15] C. S. Barry, R. P. McQuinn, M.-Y. Chung, A. Besuden, and J. J. Giovannoni. Amino acid substitutions in homologs of the STAY-GREEN protein are responsible for the green-flesh and chlorophyll retainer mutations of tomato and pepper. *Plant Physiol.*, 147(1):179–187, May 2008.

[16] R.-L. Baumgardt, K. A. Oliverio, J. J. Casal, and U. Hoecker. SPA1, a component of phytochrome a signal transduction, regulates the light signaling current. *Planta*, 215(5):745–753, Sept. 2002.

[17] G. S. Bender. Avocado flowering and pollination. `https://ucanr.edu/sites/alternativefruits/files/166371.pdf`. Accessed: 2021-3-26.

[18] B. Bergh and N. Ellstrand. Taxonomy of the avocado. *California Avocado Society Yearbook*, 70:135–146, 1986.

[19] B. O. Bergh. Avocado tree–'gwen'. *US Patent*, Oct. 1984.

[20] B. O. Bergh and R. H. Whitsell. Three new patented avocados. *California Avocado Society Yearbook*, 66:51–56, 1982.

[21] K. Berlin, S. Koren, C.-S. Chin, J. Drake, J. Landolin, and A. Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6):623–630, 2015. cited By 497.

[22] M. Bernt, A. Donath, F. Jühling, F. Externbrink, C. Florentz, G. Fritzsch, J. Pütz, M. Middendorf, and P. Stadler. Mitos: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, 69(2):313–319, 2013. cited By 1921.

[23] Y. Borovsky and I. Paran. Chlorophyll breakdown during pepper fruit ripening in the chlorophyll retainer mutation is impaired at the homolog of the senescence-inducible stay-green gene. *Theor. Appl. Genet.*, 117(2):235–240, July 2008.

[24] N. Bowen and J. McDonald. Drosophila euchromatic ltr retrotransposons are much younger than the host species in which they reside. *Genome Research*, 11(9):1527–1540, 2001. cited By 118.

[25] K. Bradnam, J. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.-C. Chou, J. Corbeil, C. Fabbro, T. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. Fonseca, G. Ganapathy, R. Gibbs, S. Gnerre, E. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. Hiatt, I. Ho, J. Howard, M. Hunt, S. Jackman, D. Jaffe, E. Jarvis, H. Jiang, S. Kazakov, P. Kersey, J. Kitzman, J. Knight, S. Koren, T.-W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. MacCallum, M. MacManes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. Otto, B. Paten, O. Paulo, A. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. Ribeiro, S. Richards, D. Rokhsar, J. Ruby, S. Scalabrin, M. Schatz, D. Schwartz, A. Sergushichev, T. Sharpe, T. Shaw, J. Shendure, Y. Shi, J. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. Vieira,

J. Wang, K. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. Korf. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1), 2013. cited By 394.

[26] G. Bresler, M. Bresler, and D. Tse. Optimal assembly for high throughput shotgun sequencing. *BMC bioinformatics*, 14 Suppl 5:S18, 2013. cited By 35.

[27] T. Brůna, K. J. Hoff, A. Lomsadze, M. Stanke, and M. Borodovsky. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform*, 3(1), Jan. 2021.

[28] T. Brůna, A. Lomsadze, and M. Borodovsky. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform*, 2(2):lqaa026, June 2020.

[29] E. Butelli, C. Licciardello, Y. Zhang, J. Liu, S. Mackay, P. Bailey, G. Reforgiato-Recupero, and C. Martin. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell*, 24(3):1242–1255, Mar. 2012.

[30] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. Lane, E. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Daley, and E. Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22(3):231–238, 1999. cited By 1511.

[31] A. Castillo, G. Dorado, C. Feuillet, P. Sourdille, and P. Hernandez. Genetic structure and ecogeographical adaptation in wild barley (hordeum chilense roemer et schultes) as revealed by microsatellite markers. *BMC Plant Biol.*, 10:266, Nov. 2010.

[32] T. G. Chabikwa, F. F. Barbier, M. Tanurdzic, and C. A. Beveridge. De novo transcriptome assembly and annotation for gene discovery in avocado, macadamia and mango. *Sci Data*, 7(1):9, Jan. 2020.

[33] F. J. J. Chain and P. G. D. Feulner. Ecological and evolutionary implications of genomic structural variations. *Front. Genet.*, 5:326, Sept. 2014.

[34] M. Chaisson. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*, 2017. cited By 28.

[35] M. Chaisson and G. Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): Application and theory. *BMC Bioinformatics*, 13(1), 2012. cited By 614.

[36] M. Chakraborty, J. Baldwin-Brown, A. Long, and J. Emerson. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, 44(19), 2016. cited By 122.

[37] M. Chakraborty, N. Vankuren, R. Zhao, X. Zhang, S. Kalsow, and J. Emerson. Hidden genetic variation shapes the structure of functional elements in drosophila. *Nature Genetics*, 50(1):20–25, 2018. cited By 56.

[38] H. Chen, V. E. Ashworth, S. Xu, and M. T. Clegg. Quantitative genetic analysis of growth rate in avocado. *J. Am. Soc. Hortic. Sci.*, 132(5):691–696, 2007.

[39] H. Chen, P. L. Morrell, V. E. Ashworth, M. De La Cruz, and M. T. Clegg. Tracing the geographic origins of major avocado cultivars. *J. Hered.*, 100(1):56–65, 2009.

[40] H. Chen, P. L. Morrell, M. de la Cruz, and M. T. Clegg. Nucleotide diversity and linkage disequilibrium in wild avocado (persea americana mill.). *J. Hered.*, 99(4):382–389, July 2008.

[41] C.-S. Chin, D. Alexander, P. Marks, A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. Eichler, S. Turner, and J. Korlach. Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nature Methods*, 10(6):563–569, 2013. cited By 2482.

[42] C.-S. Chin, P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz, G. R. Cramer, M. Delledonne, C. Luo, J. R. Ecker, D. Cantu, D. R. Rank, and M. C. Schatz. Phased diploid genome assembly with single-molecule real-time sequencing, 2016.

[43] A. Conesa and S. Götz. Blast2GO: A comprehensive suite for functional analysis in plant genomics, 2008.

[44] N. Daccord, J.-M. Celton, G. Linsmith, C. Becker, N. Choisne, E. Schijlen, H. Van De Geest, L. Bianco, D. Micheletti, R. Velasco, E. Di Pierro, J. Gouzy, D. Rees, P. Guérif, H. Muranty, C.-E. Durel, F. Laurens, Y. Lespinasse, S. Gaillard, S. Aubourg, H. Quesneville, D. Weigel, E. Van De Weg, M. Troggio, and E. Bucher. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature Genetics*, 49(7):1099–1106, 2017. cited By 318.

[45] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, Aug. 2011.

[46] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2), Feb. 2021.

[47] T. L. Davenport. Avocado flowering. *Hortic Rev.*, 8(257):89, 1986.

[48] J. Davis, D. Henderson, M. Kobayashi, and M. T. Clegg. Genealogical relationships among cultivated avocado as revealed through RFLP analyses. *J. Hered.*, 89(4):319–323, July 1998.

[49] M. De Freitas Ortiz and E. Loreto. The hobo-related elements in the melanogaster species group. *Genetics Research*, 90(3):243–252, 2008. cited By 13.

[50] G. Dos Santos, A. Schroeder, J. Goodman, V. Strelets, M. Crosby, J. Thurmond, D. Emmert, W. Gelbart, N. Brown, T. Kaufman, M. Werner-Washburne, R. Cripps, K. Broll, L. Gramates, K. Falls, B. Matthews, S. Russo, P. Zhou, M. Zytkovicz, B. Adryan, H. Attrill, M. Costa, S. Marygold, P. McQuilton, G. Millburn, L. Ponting, R. Stefancsik, S. Tweedie, and G. Grumbling. Flybase: Introduction of the drosophila melanogaster release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, 43(D1):D690–D697, 2015. cited By 232.

[51] O. Dudchenko, S. Batra, A. Omer, S. Nyquist, M. Hoeger, N. Durand, M. Shamim, I. Machol, E. Lander, A. Aiden, and E. Aiden. De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017. cited By 322.

[52] L. Feuk, A. Carson, and S. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006. cited By 1306.

[53] S. E. Fick and R. J. Hijmans. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.*, 37(12):4302–4315, Oct. 2017.

[54] J. Fierst, J. Willis, C. Thomas, W. Wang, R. Reynolds, T. Ahearne, A. Cutter, and P. Phillips. Reproductive mode and the evolution of genome size and structure in caenorhabditis nematodes. *PLoS Genetics*, 11(6), 2015. cited By 58.

[55] Food and Agriculture Organization of the United Nations, International Fund for Agricultural Development, United Nations International Children's Emergency Fund, World Food Programme, and World Health Organization. *The State of Food Security and Nutrition in the World 2021: Transforming food systems for food security, improved nutrition and affordable healthy diets for all.* Food & Agriculture Org., July 2021.

[56] G. R. Furnier, M. P. Cummings, and M. T. Clegg. Evolution of the avocados as revealed by DNA restriction fragment variation. *J. Hered.*, 81(3):183–188, May 1990.

[57] B. S. Gaut, D. K. Seymour, Q. Liu, and Y. Zhou. Demography and its effects on genomic variation in crop domestication. *Nat Plants*, 4(8):512–520, Aug. 2018.

[58] J. Ghurye, S. Koren, S. Small, S. Redmond, P. Howell, A. Phillippy, and N. Besansky. A chromosome-scale assembly of the major african malaria vector anopheles funestus. *GigaScience*, 8(6), 2019. cited By 17.

[59] R. Gibbs, G. Weinstock, M. Metzker, D. Muzny, E. Sodergren, S. Scherer, G. Scott, D. Steffen, K. Worley, P. Burch, G. Okwuonu, S. Hines, L. Lewis, C. Deramo, O. Delgado, S. Dugan-Rocha, G. Miner, M. Morgan, A. Hawes, R. Gill, R. Holt, M. Adams, P. Amanatides, H. Baden-Tillson, M. Barnstead, S. Chin, C. Evans, S. Ferriera, C. Fosler, A. Glodek, Z. Gu, D. Jennings, C. Kraft, T. Nguyen, C. Pfannkoch, C. Sitter,

G. Sutton, J. Venter, T. Woodage, D. Smith, H.-M. Lee, E. Gustafson, P. Cahill, A. Kana, L. Doucette-Stamm, K. Weinstock, K. Fechtel, R. Weiss, D. Dunn, E. Green, R. Blakesley, G. Bouffard, P. de Jong, K. Osoegawa, B. Zhu, M. Marra, J. Schein, I. Bosdet, C. Fjell, S. Jones, M. Krzywinski, C. Mathewson, A. Siddiqui, N. Wye, J. McPherson, S. Zhao, C. Fraser, J. Shetty, S. Shatsman, K. Geer, Y. Chen, S. Abramzon, W. Nierman, R. Gibbs, G. Weinstock, P. Havlak, R. Chen, K. Durbin, A. Egan, Y. Ren, X.-Z. Song, B. Li, Y. Liu, X. Qin, S. Cawley, G. Weinstock, K. Worley, A. Cooney, R. Gibbs, L. D'Souza, K. Martin, J. Wu, M. Gonzalez-Garay, A. Jackson, K. Kalafus, M. McLeod, A. Milosavljevic, D. Virk, A. Volkov, D. Wheeler, Z. Zhang, J. Bailey, E. Eichler, E. Tuzun, E. Birney, E. Mongin, A. Ureta-Vidal, C. Woodwark, E. Zdobnov, P. Bork, M. Suyama, D. Torrents, M. Alexandersson, B. Trask, J. Young, D. Smith, H. Huang, K. Fechtel, H. Wang, H. Xing, K. Weinstock, S. Daniels, D. Gietzen, J. Schmidt, K. Stevens, U. Vitt, J. Wingrove, F. Camara, M. Albà, J. Abril, R. Guigo, A. Smit, I. Dubchak, E. Rubin, O. Couronne, A. Poliakov, N. Hübner, D. Ganten, C. Goesele, O. Hummel, T. Kreitler, Y.-A. Lee, J. Monti, H. Schulz, H. Zimdahl, H. Himmelbauer, H. Lehrach, H. Jacob, S. Bromberg, J. Gullings-Handley, M. Jensen-Seaman, A. Kwitek, J. Lazar, D. Pasko, P. Tonellato, S. Twigger, C. Ponting, J. Duarte, S. Rice, L. Goodstadt, S. Beatson, R. Emes, E. Winter, C. Webber, P. Brandt, G. Nyakatura, M. Adetobi, F. Chiaromonte, L. Elnitski, P. Eswara, R. Hardison, M. Hou, D. Kolbe, K. Makova, W. Miller, A. Nekrutenko, C. Riemer, S. Schwartz, J. Taylor, S. Yang, Y. Zhang, K. Lindpaintner, T. Andrews, M. Caccamo, M. Clamp, L. Clarke, V. Curwen, R. Durbin, E. Eyras, S. Searle, G. Cooper, S. Batzoglou, M. Brudno, A. Sidow, E. Stone, J. Venter, B. Payseur, G. Bourque, C. López-Otín, X. Puente, K. Chakrabarti, S. Chatterji, C. Dewey, L. Pachter, N. Bray, V. Yap, A. Caspi, S. Tesler, P. Pevzner, S. Haussler, K. Roskin, R. Baertsch, H. Clawson, T. Furey, A. Hinrichs, D. Karolchik, W. Kent, K. Rosenbloom, H. Trumbower, M. Weirauch, D. Cooper, P. Stenson, B. Ma, M. Brent, M. Arumugam, D. Shteynberg, R. Copley, M. Taylor, H. Riethman, U. Mudunuri, J. Peterson, M. Guyer, A. Felsenfeld, S. Old, S. Mockrin, and F. Collins. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–520, 2004. cited By 1527.

[60] S. Goff, D. Ricke, T.-H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, D. Hadley, D. Hutchison, C. Martin, F. Katagiri, B. Lange, T. Moughamer, Y. Xia, P. Budworth, J. Zhong, T. Miguel, U. Paszkowski, S. Zhang, M. Colbert, W.-L. Sun, L. Chen, B. Cooper, S. Park, T. Wood, L. Mao, P. Quail, R. Wing, R. Deans, Y. Yu, A. Zharkikh, R. Shen, S. Sahasrabudhe, A. Thomas, R. Cannings, A. Gutin, D. Pruss, J. Reid, S. Tavtigian, J. Mitchell, G. Eldredge, T. Scholl, R. Miller, S. Bhatnagar, N. Adey, T. Rubano, N. Tusneem, R. Robinson, J. Feldhaus, T. Macalma, A. Oliphant, and S. Briggs. A draft sequence of the rice genome (oryza sativa l. ssp. japonica). *Science*, 296(5565):92–100, 2002. cited By 2463.

[61] L. G. González and M. K. Deyholos. Identification, characterization and distribution of transposable elements in the flax (linum usitatissimum l.) genome. *BMC Genomics*, 13:644, Nov. 2012.

[62] D. Gordon, J. Huddleston, M. Chaisson, C. Hill, Z. Kronenberg, K. Munson, M. Malig, A. Raja, I. Fiddes, L. Hillier, C. Dunn, C. Baker, J. Armstrong, M. Diekhans, B. Paten, J. Shendure, R. Wilson, D. Haussler, C.-S. Chin, and E. Eichler. Long-read sequence assembly of the gorilla genome. *Science*, 352(6281), 2016. cited By 189.

[63] D. Guan, S. A. McCarthy, J. Wood, K. Howe, Y. Wang, and R. Durbin. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, Jan. 2020.

[64] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. Quast: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013. cited By 2485.

[65] R. J. Hijmans, S. Cameron, J. Parra, P. Jones, A. Jarvis, and K. Richardson. WorldClim-global climate data. *Very High Resolution Interpolated Climate Surfaces for Global Land Areas*, 2005.

[66] R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. The WorldClim interpolated global terrestrial climate surfaces. version 1.3, 2004.

[67] R. J. Hijmans, S. E. Cameron, J. L. Parra, and others. Very high resolution interpolated climate surfaces for global land areas. *J. Appl. Meteorol. Climatol.*, 2005.

[68] L. Hillier, W. Miller, E. Birney, W. Warren, R. Hardison, C. Ponting, P. Bork, D. Burt, M. Groenen, M. Delany, J. Dodgson, A. Chinwalla, P. Cliften, S. Clifton, K. Delehaunty, C. Fronick, R. Fulton, T. Graves, C. Kremitzki, D. Layman, V. Magrini, J. McPherson, T. Miner, P. Minx, W. Nash, M. Nhan, J. Nelson, L. Oddy, C. Pohl, J. Randall-Maher, S. Smith, J. Wallis, S.-P. Yang, M. Romanov, C. Rondelli, B. Paton, J. Smith, D. Morrice, L. Daniels, H. Tempest, L. Robertson, J. Masabanda, D. Griffin, A. Vignal, V. Fillon, L. Jacobbson, S. Kerje, L. Andersson, R. Crooijmans, J. Aerts, J. van der Poel, H. Ellegren, R. Caldwell, S. Hubbard, D. Grafham, A. Kierzek, S. McLaren, I. Overton, H. Arakawa, K. Beattie, Y. Bezzubov, P. Boardman, J. Bonfield, M. Croning, R. Davies, M. Francis, S. Humphray, C. Scott, R. Taylor, C. Tickle, W. Brown, J. Rogers, J.-M. Buerstedde, S. Wilson, L. Stubbs, I. Ovcharenko, L. Gordon, S. Lucas, M. Miller, H. Inoko, T. Shiina, J. Kaufman, J. Salomonsen, K. Skjoedt, G.-S. Wong, J. Wang, B. Liu, J. Wang, J. Yu, H. Yang, M. Nefedov, M. Koriabine, P. Dejong, L. Goodstadt, C. Webber, N. Dickens, I. Letunic, M. Suyama, D. Torrents, C. von Mering, E. Zdobnov, K. Makova, A. Nekrutenko, L. Elnitski, P. Eswara, D. King, S. Yang, S. Tyekucheva, A. Radakrishnan, R. Harris, F. Chiaromonte, J. Taylor, J. He, M. Rijnkels, S. Griffiths-Jones, A. Ureta-Vidal, M. Hoffman, J. Severin, S. Searle, A. Law, D. Speed, D. Waddington, Z. Cheng, E. Tuzun, E. Eichler, Z. Bao, P. Flicek, D. Shteynberg, M. Brent, J. Bye, E. Huckle, S. Chatterji, C. Dewey, L. Pachter, A. Kouranov, Z. Mourelatos, A. Hatzigeorgiou, A. Paterson, R. Ivarie, M. Brandstrom, E. Axelsson, N. Backstrom, S. Berlin, M. Webster, O. Pourquie, A. Reymond, C. Ucla, S. Antonarakis, M. Long, J. Emerson, E. Betrán, I. Dupanloup, H. Kaessmann, A. Hinrichs, G. Bejerano, T. Furey, R. Harte, B. Raney, A. Siepel, W. James Kent, D. Haussler, E. Eyras, R. Castelo, J. Abril, S. Castellano, F. Camara, G. Parra, R. Guigo, G. Bourque, G. Tesler, P. Pevzner, A. Smit, L. Fulton,

E. Mardis, R. Wilson, and I. C. G. S. Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, 2004. cited By 1944.

[69] K. J. Hoff, A. Lomsadze, M. Borodovsky, and M. Stanke. Whole-genome annotation with BRAKER. *Methods Mol. Biol.*, 1962:65–95, 2019.

[70] R. Hoskins, J. Carlson, K. Wan, S. Park, I. Mendez, S. Galle, B. Booth, B. Pfeiffer, R. George, R. Svirskas, M. Krzywinski, J. Schein, M. Accardo, E. Damia, G. Messina, M. Méndez-Lago, B. De Pablos, O. Demakova, E. Andreyeva, L. Boldyreva, M. Marra, A. Carvalho, P. Dimitri, A. Villasante, I. Zhimulev, G. Rubin, G. Karpen, and S. Celniker. The release 6 reference sequence of the drosophila melanogaster genome. *Genome Research*, 25(3):445–458, 2015. cited By 163.

[71] J. Huddleston and E. Eichler. An incomplete understanding of human genetic variation. *Genetics*, 202(4):1251–1254, 2016. cited By 44.

[72] E. Ibarra-Laclette, A. Méndez-Bravo, C. A. Pérez-Torres, V. A. Albert, K. Mockaitis, A. Kilaru, R. López-Gómez, J. I. Cervantes-Luevano, and L. Herrera-Estrella. Deep sequencing of the mexican avocado transcriptome, an ancient angiosperm with a high content of fatty acids. *BMC Genomics*, 16:599, Aug. 2015.

[73] C. Illsley-Granich, R. Brokaw, S. Ochoa-Ascencio, and T. Bruwer. Hass carmen®, a precocious flowering avocado tree. In *Proceedings VII World Avocado Congress*, pages 5–9, 2011.

[74] M. Ishikawa, T. Kiba, and N.-H. Chua. The arabidopsis SPA1 gene is required for circadian clock function and photoperiodic flowering. *Plant J.*, 46(5):736–746, June 2006.

[75] O. Jaillon, J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. Pè, G. Valle, M. Morgante, M. Caboche, A.-F. Adam-Blondon, J. Weissenbach, F. Quétier, and P. Wincker. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467, 2007. cited By 2392.

[76] M. Jain, S. Koren, K. Miga, J. Quick, A. Rand, T. Sasani, J. Tyson, A. Beggs, A. Dilthey, I. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. Olsen, B. Pedersen, A. Rhie, H. Richardson, A. Quinlan, T. Snutch, L. Tee, B. Paten, A. Phillippy, J. Simpson, N. Loman, and M. Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, 2018. cited By 583.

[77] Y. Jiao, P. Peluso, J. Shi, T. Liang, M. Stitzer, B. Wang, M. Campbell, J. Stein, X. Wei, C.-S. Chin, K. Guill, M. Regulski, S. Kumari, A. Olson, J. Gent, K. Schneider, T. Wolfgruber, M. May, N. Springer, E. Antoniou, W. McCombie, G. Presting, M. McMullen, J. Ross-Ibarra, R. Dawe, A. Hastie, D. Rank, and D. Ware. Improved maize reference genome with single-molecule technologies. *Nature*, 546(7659):524–527, 2017. cited By 419.

[78] R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, and T. Itoh. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, 24(8):1384–1395, 2014. cited By 562.

[79] W. Kent. Blat - the blast-like alignment tool. *Genome Research*, 12(4):656–664, 2002. cited By 5565.

[80] K. Kim, P. Peluso, P. Babayan, P. Yeadon, C. Yu, W. Fisher, C.-S. Chin, N. Rapicavoli, D. Rank, J. Li, D. Catcheside, S. Celniker, A. Phillippy, C. Bergman, and J. Landolin. Long-read, whole-genome shotgun sequence data for five model organisms. *Scientific Data*, 1, 2014. cited By 77.

[81] S. Koren, A. Rhie, B. Walenz, A. Dilthey, D. Bickhart, S. Kingan, S. Hiendleder, J. Williams, T. Smith, and A. Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology*, 36(12):1174–1182, 2018. cited By 89.

[82] S. Koren, B. Walenz, K. Berlin, J. Miller, N. Bergman, and A. Phillippy. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5):722–736, 2017. cited By 1902.

[83] S. Kurtz, A. Phillippy, A. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12, 2004. cited By 2955.

[84] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biol.*, 5(2), 2004.

[85] E. Lahav and U. Lavi. Avocado genetics and breeding.

[86] K.-K. Lam, A. Khalak, and D. Tse. Near-optimal assembly for shotgun sequencing with noisy reads. *BMC Bioinformatics*, 15(9), 2014. cited By 19.

[87] E. Lander, L. Linton, B. Birren, C. Nusbaum, M. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitzhugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. Levine, P. McEwan, K. McKernan, J. Meldrim, J. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter,

A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. Waterston, R. Wilson, L. Hillier, J. McPherson, M. Marra, E. Mardis, L. Fulton, A. Chinwalla, K. Pepin, W. Gish, S. Chissoe, M. Wendl, K. Delehaunty, T. Miner, A. Delehaunty, J. Kramer, L. Cook, R. Fulton, D. Johnson, P. Minx, S. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. Gibbs, D. Muzny, S. Scherer, J. Bouck, E. Sodergren, K. Worley, C. Rives, J. Gorrell, M. Metzker, S. Naylor, R. Kucherlapati, D. Nelson, G. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, M. Hong, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. Davis, N. Federspiel, A. Abola, M. Proctor, B. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. McCombie, M. De La Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. Brown, C. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. Copley, T. Doerks, S. Eddy, E. Eichler, T. Furey, J. Galagan, J. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. Johnson, T. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. Kent, P. Kitts, E. Koonin, I. Korf, D. Kulp, D. Lancet, T. Lowe, A. McLysaght, T. Mikkelsen, J. Moran, N. Mulder, V. Pollara, C. Ponting, G. Schuler, J. Schultz, G. Slater, A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. Wolf, K. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. Guyer, J. Peterson, A. Felsenfeld, K. Wetterstrand, R. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. Cox, M. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. Evans, M. Athanasiou, R. Schultz, A. Patrinos, and M. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. cited By 16278.

[88] S. Laubinger, V. Marchal, J. Le Gourrierec, S. Wenkel, J. Adrian, S. Jang, C. Kulajta, H. Braun, G. Coupland, and U. Hoecker. Arabidopsis SPA proteins regulate photoperiodic flowering and interact with the floral inducer CONSTANS to regulate its stability. *Development*, 133(16):3213–3222, Aug. 2006.

[89] B. S. Levy and J. A. Patz. Climate change, human rights, and social justice. *Ann Glob Health*, 81(3):310–322, May 2015.

[90] H. Li. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, 2016. cited By 382.

[91] H. Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018. cited By 1228.

[92] H. Li and R. Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5):589–595, 2010. cited By 5442.

[93] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, Mar. 2010.

[94] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009. cited By 23528.

[95] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug. 2009.

[96] E. Liscum and R. P. Hangarter. Arabidopsis mutants lacking blue Light-Dependent inhibition of hypocotyl elongation. *Plant Cell*, 3(7):685–694, July 1991.

[97] N. Loman, J. Quick, and J. Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, 2015. cited By 483.

[98] A. Long, J. Baldwin-Brown, Y. Tao, V. Cook, G. Balderrama-Gutierrez, R. Corbett-Detig, A. Mortazavi, and A. Barbour. The genome of peromyscus leucopus, natural host for lyme disease and other emerging infections. *Science Advances*, 5(7), 2019. cited By 10.

[99] W. Y. Low, R. Tearle, R. Liu, S. Koren, A. Rhie, D. M. Bickhart, B. D. Rosen, Z. N. Kronenberg, S. B. Kingan, E. Tseng, F. Thibaud-Nissen, F. J. Martin, K. Billis, J. Ghurye, A. R. Hastie, J. Lee, A. W. C. Pang, M. P. Heaton, A. M. Phillippy, S. Hiendleder, T. P. L. Smith, and J. L. Williams. Haplotype-resolved genomes provide insights into structural variation and gene content in angus and brahman cattle. *Nat. Commun.*, 11(1):1–14, Apr. 2020.

[100] M. Massonnet, N. Cochetel, A. Minio, A. Vondras, J. Lin, A. Muyle, J. Garcia, Y. Zhou, M. Delledonne, S. Riaz, R. Figueroa-Balderas, B. Gaut, and D. Cantu. The genetic basis of sex determination in grapes. *Nature Communications*, 11(1), 2020. cited By 21.

[101] M. Massonnet, N. Cochetel, A. Minio, A. M. Vondras, J. Lin, A. Muyle, J. F. Garcia, Y. Zhou, M. Delledonne, S. Riaz, R. Figueroa-Balderas, B. S. Gaut, and D. Cantu. The genetic basis of sex determination in grapes, 2020.

[102] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303, Sept. 2010.

[103] P. Medvedev, M. Stanciu, and M. Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11S):S13, 2009. cited By 395.

[104] T. Michael, F. Jupe, F. Bemm, S. Motley, J. Sandoval, C. Lanz, O. Loudet, D. Weigel, and J. Ecker. High contiguity arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications*, 9(1), 2018. cited By 99.

[105] A. Minio, J. Lin, B. Gaut, and D. Cantu. How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Frontiers in Plant Science*, 8, 2017. cited By 23.

[106] N. Mitsuda, T. Hisabori, K. Takeyasu, and M. H. Sato. VOZ; isolation and characterization of novel vascular plant transcription factors with a one-zinc finger from arabidopsis thaliana. *Plant Cell Physiol.*, 45(7):845–854, July 2004.

[107] R. Moore, C. Baru, D. Baxter, G. Fox, A. Majumdar, P. Papadopoulos, W. Pfeiffer, R. Sinkovits, S. Strande, M. Tatineni, R. Wagner, N. Wilkins-Diehr, and M. Norman. Gateways to discovery: Cyberinfrastructure for the long tail of science. 2014. cited By 26.

[108] R. Moschetti, P. Dimitri, R. Caizzi, and N. Junakovic. Genomic instability of i elements of drosophila melanogaster in absence of dysgenic crosses. *PLoS ONE*, 5(10), 2010. cited By 13.

[109] A. Motahari, G. Bresler, and D. Tse. Information theory of dna shotgun sequencing. *IEEE Transactions on Information Theory*, 59(10):6273–6289, 2013. cited By 49.

[110] B. T. Moyers, P. L. Morrell, and J. K. McKay. Genetic costs of domestication and improvement. *J. Hered.*, 109(2):103–116, Feb. 2018.

[111] J. K. Muhlemann, B. D. Woodworth, J. A. Morgan, and N. Dudareva. The monolignol pathway contributes to the biosynthesis of volatile phenylpropenes in flowers. *New Phytol.*, 204(3):661–670, Nov. 2014.

[112] E. Myers, G. Sutton, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, S. Kravitz, C. Mobarry, K. Reinert, K. Remington, E. Anson, R. Bolanos, H.-H. Chou, C. Jordan, A. Halpern, S. Lonardi, E. Beasley, R. Brandon, L. Chen, P. Dunn, Z. Lai, Y. Liang, D. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. Rubin, M. Adams, and J. Venter. A whole-genome assembly of drosophila. *Science*, 287(5461):2196–2204, 2000. cited By 1097.

[113] G. Narzisi and M. Schatz. The challenge of small-scale repeats for indel discovery. *Frontiers in Bioengineering and Biotechnology*, 3(JAN), 2015. cited By 26.

[114] N. Nystrom, M. Levine, R. Roskies, and J. Scott. Bridges: A uniquely flexible hpc resource for new communities and data analytics. volume 2015-July, 2015. cited By 76.

[115] H. Ooka, K. Satoh, K. Doi, T. Nagata, Y. Otomo, K. Murakami, K. Matsubara, N. Osato, J. Kawai, P. Carninci, Y. Hayashizaki, K. Suzuki, K. Kojima, Y. Takahara, K. Yamamoto, and S. Kikuchi. Comprehensive analysis of NAC family genes in oryza sativa and arabidopsis thaliana. *DNA Res.*, 10(6):239–247, Dec. 2003.

[116] L. Pascual and G. Periquet. Distribution of hobo transposable elements in natural populations of drosophila melanogaster. *Molecular Biology and Evolution*, 8(3):282–296, 1991. cited By 43.

[117] K. Paszkiewicz and D. Studholme. De novo assembly of short sequence reads. *Briefings in Bioinformatics*, 11(5):457–472, 2010. cited By 124.

[118] G. Periquet, F. Lemeunier, Y. Bigot, M. Hamelin, C. Bazin, V. Ladevèze, J. Eeken, M. Galindo, L. Pascual, and I. Boussy. The evolutionary genetics of the hobo transposable element in the drosophila melanogaster complex. *Genetica*, 93(1-3):79–90, 1994. cited By 32.

[119] J. Pool, R. Corbett-Detig, R. Sugino, K. Stevens, C. Cardeno, M. Crepeau, P. Duchen, J. Emerson, P. Saelao, D. Begun, and C. Langley. Population genomics of sub-saharan drosophila melanogaster: African diversity and non-african admixture. *PLoS Genetics*, 8(12), 2012. cited By 179.

[120] H. Price. Evolution of dna content in higher plants. *The Botanical Review*, 42(1):27–52, 1976. cited By 145.

[121] L. Pryszcz and T. Gabaldón. Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, 44(12):e113, 2016. cited By 160.

[122] L. P. Pryszcz and T. Gabaldón. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.*, 44(12):e113, July 2016.

[123] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, Sept. 2007.

[124] A. Quinlan and I. Hall. Bedtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. cited By 8574.

[125] R. Rahman, G.-W. Chirn, A. Kanodia, Y. Sytnikova, B. Brembs, C. Bergman, and N. Lau. Unique transposon landscapes are pervasive across drosophila melanogaster genomes. *Nucleic Acids Research*, 43(22):10655–10672, 2015. cited By 48.

[126] O. Raymond, J. Gouzy, J. Just, H. Badouin, M. Verdenaud, A. Lemainque, P. Vergne, S. Moja, N. Choisne, C. Pont, S. Carrère, J.-C. Caissard, A. Couloux, L. Cottret, J.-M. Aury, J. Szécsi, D. Latrasse, M.-A. Madoui, L. François, X. Fu, S.-H. Yang, A. Dubois, F. Piola, A. Larrieu, M. Perez, K. Labadie, L. Perrier, B. Govetto, Y. Labrousse, P. Villand, C. Bardoux, V. Boltz, C. Lopez-Roques, P. Heitzler, T. Vernoux, M. Vandenbussche, H. Quesneville, A. Boualem, A. Bendahmane, C. Liu, M. Le Bris, J. Salse,

S. Baudino, M. Benhamed, P. Wincker, and M. Bendahmane. The rosa genome provides new insights into the domestication of modern roses. *Nature Genetics*, 50(6):772–777, 2018. cited By 120.

[127] M. Rendón-Anaya, E. Ibarra-Laclette, A. Méndez-Bravo, T. Lan, C. Zheng, L. Carretero-Paulet, C. A. Perez-Torres, A. Chacón-López, G. Hernandez-Guzmán, T.-H. Chang, and Others. The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. *Proceedings of the National Academy of Sciences*, 116(34):17081–17089, 2019.

[128] A. Rhie, S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim, C. Lee, B. J. Ko, M. Chaisson, G. L. Gedman, L. J. Cantin, F. Thibaud-Nissen, L. Haggerty, I. Bista, M. Smith, B. Haase, J. Mountcastle, S. Winkler, S. Paez, J. Howard, S. C. Vernes, T. M. Lama, F. Grutzner, W. C. Warren, C. N. Balakrishnan, D. Burt, J. M. George, M. T. Biegler, D. Iorns, A. Digby, D. Eason, B. Robertson, T. Edwards, M. Wilkinson, G. Turner, A. Meyer, A. F. Kautt, P. Franchini, H. W. Detrich, 3rd, H. Svardal, M. Wagner, G. J. P. Naylor, M. Pippel, M. Malinsky, M. Mooney, M. Simbirsky, B. T. Hannigan, T. Pesout, M. Houck, A. Misuraca, S. B. Kingan, R. Hall, Z. Kronenberg, I. Sović, C. Dunn, Z. Ning, A. Hastie, J. Lee, S. Selvaraj, R. E. Green, N. H. Putnam, I. Gut, J. Ghurye, E. Garrison, Y. Sims, J. Collins, S. Pelan, J. Torrance, A. Tracey, J. Wood, R. E. Dagnew, D. Guan, S. E. London, D. F. Clayton, C. V. Mello, S. R. Friedrich, P. V. Lovell, E. Osipova, F. O. Al-Ajli, S. Secomandi, H. Kim, C. Theofanopoulou, M. Hiller, Y. Zhou, R. S. Harris, K. D. Makova, P. Medvedev, J. Hoffman, P. Masterson, K. Clark, F. Martin, K. Howe, P. Flicek, B. P. Walenz, W. Kwak, H. Clawson, M. Diekhans, L. Nassar, B. Paten, R. H. S. Kraus, A. J. Crawford, M. T. P. Gilbert, G. Zhang, B. Venkatesh, R. W. Murphy, K.-P. Koepfli, B. Shapiro, W. E. Johnson, F. Di Palma, T. Marques-Bonet, E. C. Teeling, T. Warnow, J. M. Graves, O. A. Ryder, D. Haussler, S. J. O'Brien, J. Korlach, H. A. Lewin, K. Howe, E. W. Myers, R. Durbin, A. M. Phillippy, and E. D. Jarvis. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856):737–746, Apr. 2021.

[129] A. Rhie, B. P. Walenz, S. Koren, and A. M. Phillippy. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.*, 21(1):245, Sept. 2020.

[130] E. Rice, S. Koren, A. Rhie, M. Heaton, T. Kalbfleisch, T. Hardy, P. Hackett, D. Bickhart, B. Rosen, B. Ley, N. Maurer, R. Green, A. Phillippy, J. Petersen, and T. Smith. Continuous chromosome-scale haplotypes assembled from a single interspecies f1 hybrid of yak and cattle. *GigaScience*, 9(4), 2020. cited By 12.

[131] M. Roach, D. Johnson, J. Bohlmann, H. van Vuuren, S. Jones, I. Pretorius, S. Schmidt, and A. Borneman. Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar chardonnay. *PLoS Genetics*, 14(11), 2018. cited By 32.

[132] M. Roach, S. Schmidt, and A. Borneman. Purge haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19(1), 2018. cited By 123.

[133] M. J. Roach, S. A. Schmidt, and A. R. Borneman. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19(1):460, Nov. 2018.

[134] K. Roessler, A. Muyle, C. Diez, G. Gaut, A. Bousios, M. Stitzer, D. Seymour, J. Doebley, Q. Liu, and B. Gaut. The genome-wide dynamics of purging during selfing in maize. *Nature Plants*, 5(9):980–990, 2019. cited By 11.

[135] P. Ross, N. Endersby-Harshman, and A. Hoffmann. A comprehensive assessment of inbreeding and laboratory adaptation in aedes aegypti mosquitoes. *Evolutionary Applications*, 12(3):572–586, 2019. cited By 27.

[136] S. Salzberg, A. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. Treangen, M. Schatz, A. Delcher, M. Roberts, G. Marcxais, M. Pop, and J. Yorke. Gage: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22(3):557–567, 2012. cited By 427.

[137] R. K. Saxena, D. Edwards, and R. K. Varshney. Structural variations in plant genomes. *Brief. Funct. Genomics*, 13(4):296–307, July 2014.

[138] B. A. Schaffer, B. Nigel Wolstenholme, and A. W. Whiley. *The Avocado: Botany, Production and Uses.* CABI, 2013.

[139] R. W. Scora, B. N. Wolstenholme, U. Lavi, and Others. Taxonomy and botany. *The avocado: botany, production and uses*, pages 15–37, 2002.

[140] I. Shomorony, T. Courtade, and D. Tse. *Do Read Errors Matter for Genome Assembly?*, 2016. cited By 1.

[141] F. Simão, R. Waterhouse, P. Ioannidis, E. Kriventseva, and E. Zdobnov. Busco: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015. cited By 3502.

[142] A. Smit, R. Hubley, and P. Green. Repeatmasker open-3.0. *RepeatMasker Open-3.0*, 1996. cited By 1083.

[143] E. A. Solares, M. Chakraborty, D. E. Miller, S. Kalsow, K. Hall, A. G. Perera, J. J. Emerson, and R. S. Hawley. Rapid Low-Cost assembly of the drosophila melanogaster reference genome using Low-Coverage, Long-Read sequencing. *G3*, 8(10):3143–3154, Oct. 2018.

[144] E. A. Solares, Y. Tao, A. D. Long, and B. S. Gaut. HapSolo: an optimization approach for removing secondary haplotigs during diploid genome assembly and scaffolding. *BMC Bioinformatics*, 22(1):9, Jan. 2021.

[145] B. Spitzer-Rimon, M. Farhi, B. Albo, A. Cna'ani, M. M. Ben Zvi, T. Masci, O. Edelbaum, Y. Yu, E. Shklarman, M. Ovadis, and A. Vainstein. The R2R3-MYB-like regulatory factor EOBI, acting downstream of EOBII, regulates scent production by activating ODO1 and structural scent-related genes in petunia. *Plant Cell*, 24(12):5089–5105, Dec. 2012.

[146] N. M. Springer, K. Ying, Y. Fu, T. Ji, C.-T. Yeh, Y. Jia, W. Wu, T. Richmond, J. Kitzman, H. Rosenbaum, A. L. Iniguez, W. B. Barbazuk, J. A. Jeddeloh, D. Nettleton, and P. S. Schnable. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.*, 5(11):e1000734, Nov. 2009.

[147] M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5):637–644, Mar. 2008.

[148] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, 34(Web Server issue):W435–9, July 2006.

[149] M. Stanke, O. Schöffmann, B. Morgenstern, and S. Waack. Gene prediction in eukaryotes with a generalized hidden markov model that uses hints from external sources. *BMC Bioinformatics*, 7:62, Feb. 2006.

[150] K. E. Taylor, R. J. Stouffer, and G. A. Meehl. A summary of the CMIP5 experiment design. 4, Jan. 2007.

[151] K. E. Taylor, R. J. Stouffer, and G. A. Meehl. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.*, 93(4):485–498, Apr. 2012.

[152] H. Thorvaldsdóttir, J. Robinson, and J. Mesirov. Integrative genomics viewer (igv): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, 2013. cited By 3883.

[153] D. Tilman, C. Balzer, J. Hill, and B. L. Befort. Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci. U. S. A.*, 108(50):20260–20264, Dec. 2011.

[154] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. Peterson, R. Roskies, J. Scott, and N. Wilkens-Diehr. Xsede: Accelerating scientific discovery. *Computing in Science and Engineering*, 16(5):62–74, 2014. cited By 1743.

[155] T. Treangen and S. Salzberg. Repetitive dna and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, 2012. cited By 873.

[156] Y. Vigouroux, J. C. Glaubitz, Y. Matsuoka, M. M. Goodman, J. Sánchez G, and J. Doebley. Population structure and genetic diversity of new world maize races assessed by DNA microsatellites. *Am. J. Bot.*, 95(10):1240–1253, Oct. 2008.

[157] A. Vondras, A. Minio, B. Blanco-Ulate, R. Figueroa-Balderas, M. Penn, Y. Zhou, D. Seymour, Y. Zhou, D. Liang, and L. Espinoza. The genomic diversification of clonally propagated grapevines. *Biorxiv*, 2019. cited By 3.

[158] B. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. Cuomo, Q. Zeng, J. Wortman, S. Young, and A. Earl. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, 9(11), 2014. cited By 2242.

[159] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11):e112963, Nov. 2014.

[160] R. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. Brent, D. Brown, S. Brown, C. Bult, J. Burton, J. Butler, R. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. Chinwalla, D. Church, M. Clamp, C. Clee, F. Collins, L. Cook, R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. Delehaunty, J. Deri, E. Dermitzakis, C. Dewey, N. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. Dunn, S. Eddy, L. Elnitski, R. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. Fewell, P. Flicek, K. Foley, W. Frankel, L. Fulton, R. Fulton, T. Furey, D. Gage, R. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. Graves, E. Green, S. Gregory, R. Guigó, M. Guyer, R. Hardison, D. Haussler, Y. Hayashizaki, D. LaHillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. Jaffe, L. Johnson, M. Jones, T. Jones, A. Joy, M. Kamal, E. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. Kent, A. Kirby, D. Kolbe, I. Korf, R. Kucherlapati, E. Kulbokas III, D. Kulp, T. Landers, J. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. Maglott, E. Mardis, L. Matthews, E. Mauceli, J. Mayer, M. McCarthy, W. McCombie, S. McLaren, K. McLay, J. McPherson, J. Meldrim, B. Meredith, J. Mesirov, W. Miller, T. Miner, E. Mongin, K. Montgomery, M. Morgan, R. Mott, J. Mullikin, D. Muzny, W. Nash, J. Nelson, M. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. Pohl, A. Poliakov, T. Ponce, C. Ponting, S. Potter, M. Quail, A. Reymond, B. Roe, K. Roskin, E. Rubin, A. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. Singer, G. Slater, A. Smit, D. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. Vinson, A. von Niederhausern, C. Wade, M. Wall,

R. Weber, R. Weiss, M. Wendl, A. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. Wilson, E. Winter, K. Worley, D. Wyman, S. Yang, S.-P. Yang, E. Zdobnov, M. Zody, and E. Lander. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002. cited By 4981.

[161] M. Wellenreuther, C. Mérot, E. Berdan, and L. Bernatchez. Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.*, 28(6):1203–1209, Mar. 2019.

[162] K. Wetterstrand. *DNA Sequencing Costs: Data*, 0000. cited By 17.

[163] G. Witney and G. Martin. Taking the california avocado breeding program into the next century. In *Proceedings of The World Avocado Congress III*, volume 114, page 118, 1995.

[164] J. Wolff, V. Bhardwaj, S. Nothjunge, G. Richard, G. Renschler, R. Gilsbach, T. Manke, R. Backofen, F. Ramírez, and B. Grüning. Galaxy hicexplorer: A web server for reproducible hi-c data analysis, quality control and visualization. *Nucleic Acids Research*, 46(W1):W11–W16, 2018. cited By 38.

[165] D. Xing, H. Zhao, R. Xu, and Q. Q. Li. Arabidopsis PCFS4, a homologue of yeast polyadenylation factor pcf11p, regulates FCA alternative processing and promotes flowering time. *Plant J.*, 54(5):899–910, June 2008.

[166] L.-Á. Xoca-Orozco, E. A. Cuellar-Torres, S. González-Morales, P. Gutiérrez-Martínez, U. López-García, L. Herrera-Estrella, J. Vega-Arreguín, and A. Chacón-López. Transcriptomic analysis of avocado hass (persea americana mill) in the interaction system Fruit-Chitosan-Colletotrichum. *Front. Plant Sci.*, 8:956, June 2017.

[167] Y. Yasui, K. Mukougawa, M. Uemoto, A. Yokofuji, R. Suzuri, A. Nishitani, and T. Kohchi. The phytochrome-interacting vascular plant one-zinc finger1 and VOZ2 redundantly regulate flowering in arabidopsis. *Plant Cell*, 24(8):3248–3263, Aug. 2012.

[168] C. Ye, C. Hill, S. Wu, J. Ruan, and Z. Ma. Dbg2olc: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific Reports*, 6, 2016. cited By 132.

[169] S. Y. Yoo, Y. Kim, S. Y. Kim, J. S. Lee, and J. H. Ahn. Control of flowering time and cold response by a NAC-domain protein in arabidopsis. *PLoS One*, 2(7):e642, July 2007.

[170] J. Yu, S. Hu, J. Wang, G.-S. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, J. Li, Z. Liu, Q. Qi, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao,

G. Li, H. Gao, T. Cao, W. Zhao, P. Li, W. Chen, Y. Zhang, J. Hu, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, M. Tao, L. Zhu, L. Yuan, and H. Yang. A draft sequence of the rice genome (oryza sativa l. ssp. indica). *Science*, 296(5565):79–92, 2002. cited By 2386.

[171] L. Zakharenko, L. Kovalenko, and S. Mai. Fluorescence in situ hybridization analysis of hobo, mdg1 and dm412 transposable elements reveals genomic instability following the drosophila melanogaster genome sequencing. *Heredity*, 99(5):525–530, 2007. cited By 15.

[172] X. Zhang, J. Yazaki, A. Sundaresan, S. Cokus, S.-L. Chan, H. Chen, I. Henderson, P. Shinn, M. Pellegrini, S. Jacobsen, and J. Ecker. Genome-wide high-resolution mapping and functional analysis of dna methylation in arabidopsis. *Cell*, 126(6):1189–1201, 2006. cited By 1241.

[173] J. Zheng, D. Xing, X. Wu, Y. Shen, D. M. Kroll, G. Ji, and Q. Q. Li. Ratio-based analysis of differential mRNA processing and expression of a polyadenylation factor mutant pcfs4 using arabidopsis tiling microarray. *PLoS One*, 6(2):e14719, Feb. 2011.

[174] Y. Zhou, M. Massonnet, J. S. Sanjak, D. Cantu, and B. S. Gaut. Evolutionary genomics of grape (vitis vinifera ssp. vinifera) domestication. *Proc. Natl. Acad. Sci. U. S. A.*, 114(44):11715–11720, Oct. 2017.

[175] Y. Zhou, A. Minio, M. Massonnet, E. Solares, Y. Lv, T. Beridze, D. Cantu, and B. S. Gaut. The population genetics of structural variants in grapevine domestication. *Nat Plants*, 5(9):965–979, Sept. 2019.

[176] Z. Zuo, H. Liu, B. Liu, X. Liu, and C. Lin. Blue light-dependent interaction of CRY2 with SPA1 regulates COP1 activity and floral initiation in arabidopsis. *Curr. Biol.*, 21(10):841–847, May 2011.