PAN-GENOMICS AND THE STRUCTURAL DIVERSITY OF PLANT GENOMES

by
Michael Alonge

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
October 2021

Thesis advisor: Michael Schatz, Ph.D. Michael C. Alonge

Pan-genomics and the structural diversity of plant genomes

Abstract

A central task of genetics research is to uncover genotypes linked to important phenotypes. However, many genomic loci are incompletely or inaccurately represented in genetics studies, thus obscuring their function and evolution. New technology can accurately and continuously sequence large segments of genomic DNA at affordable cost and unprecedented scale, raising the possibility of complete and accurate representations of genomes across the tree of life. However, new computational methods are required to automatically finish, validate, and curate the forthcoming wave of genome assemblies enabled by these technologies. Researchers must also devise analytical approaches to comparing previously unresolved and usually repetitive genomic loci within and between species. Here, we introduce RaGOO and RagTag, new methods that leverage genome maps to automatically scaffold and improve draft genome assemblies into chromosome-scale representations. By applying these new methods to a bread wheat genome, we show how the established reference falsely collapsed functional paralogs genome-wide. In *Arabidopsis thaliana*, we present a new reference assembly that completely resolves all five centromeres for the first time, revealing centromere architecture, genetics, epigenetics, and evolution. Finally, we present a catalog of natural structural variants (SVs) across 100 diverse tomato accessions revealing exceptional genetic diversity via artificial introgression as well as broad and specific examples of how SVs influence molecular, domestication, and improvement phenotypes. This work underscores the potential to accelerate genetics research with complete and diverse genotype data and apply these findings to plant breeding and engineering.

Committee

Michael Schatz (primary advisor), Zachary Lippman, Steven Salzberg, and Winston Timp

Acknowledgments

Thank you, Sathvika, for moving to Baltimore, adopting Tupper and Billy, marrying me, and being my best friend. Thank you, Dad, for leading our family and loving me completely. Thank you, Greg, for being a lifelong companion and my personal genetics experiment. Thank you, Chris — you know what you did. Thank you, Marlene, Melinda, Ashok, Sandhya, Eric, Jamie, Ananya, Prasad, Charlene, Alan, David, Isabella, and the rest of my loving family who shaped my identity and supported me personally and professionally. Thank you, Kara, Melanie, Rob, Nicole, Shannon, and Rachel for making 2020 bearable. Thank you, Fini Hohnhorst, for welcoming me to the East Coast and helping me start my new home.

Thank you, Judson Ward, for initiating my genomics career, teaching me the fundamentals, and showing me how cool plants are. Thank you, Michael Schatz, for taking a chance on me, guiding me through graduate school, and teaching me so much about genomics. Thank you to the Schatz and Langmead labs for your friendship and for answering all of my questions. Thank you, Zachary Lippman, Xingang Wang, Matthias Benoit, Ian Henderson, and Matthew Naish for fun and successful collaborations and for teaching me so much about plant biology and genetics.

For Michele, who always cheered loudest.

Contents

Listing of tables

Listing of figures

By representing complex genomes through simplified maps, researchers can annotate and compare genomic features to study biological phenomena.

0

Introduction

PLANTS ARE FUNDAMENTAL TO HUMAN EXISTENCE, playing important aesthetic, medicinal, nutritional, and cultural roles in day-to-day life. The domestication of major agricultural staples, a process spanning thousands of years, exemplifies how humans have adopted and manipulated plant systems to sustain and celebrate life. By changing plants to suit their needs, these ancient humans were unknowingly influencing dynamic and complex collections of heritable molecules located within plant cells, known today as "genomes". Modern geneticists seek to uncover the genetic basis for domestication and improvement phenotypes to facilitate future crop improvement and solve contemporary global challenges, such as hunger, nutrition, and climate change. While such research has traditionally excluded large portions of plant genomes, cutting-edge DNA sequencing technology promises to offer complete and accurate genotype information for model and non-model species. Here, I outline some of these technologies and the current state of associated plant genome sequence research. I also present current challenges to representing and comparing genome sequences and my methodological and analytical contributions towards addressing these challenges.

Representing genomes through maps

Genome maps are any genome-wide collection of markers (genomic landmarks) and their relative distances. By representing complex genomes through simplified maps, researchers can annotate and compare genomic features to study biological phenomena. We highlight genome maps here because of their utility in structural variant analysis and genome assembly scaffolding. Researchers have devised many types of maps tailored for specific scientific questions and scenarios and we highlight three relevant examples: (genetic) linkage, spatial proximity, and physical maps (**Figure 0.1**).

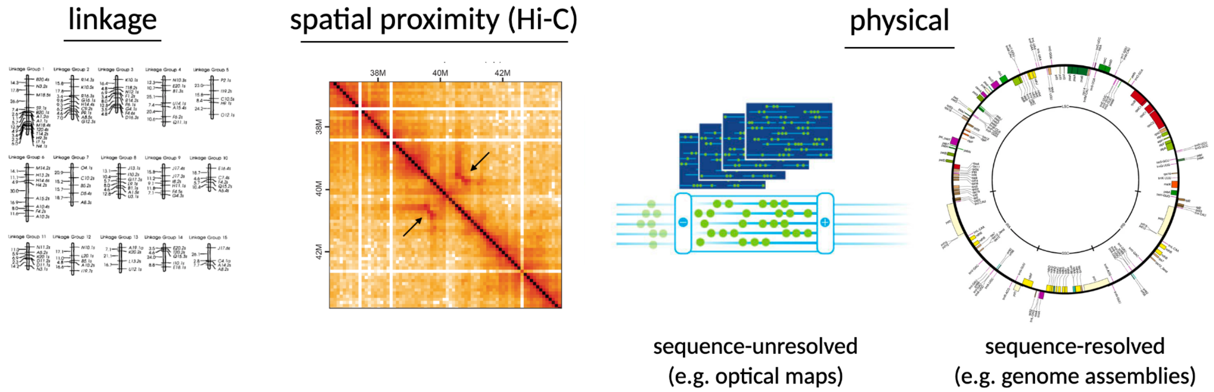


Figure 0.1: Common genome maps. From left to right: A *Saccharum spontaneum* linkage map (original figure from Al-Janabi, S. M. et al, 1993 [1]), a spatial proximity map built from Hi-C data, an optical map (original figure from Bionano: https://bionanogenomics.com/technology/platform-technology/), and a *Potentilla micrantha* chloroplast genome assembly (original figure from Ferrarini, M. et al, 2013 [2]).

Linkage maps are built by observing recombination rates between genetic markers to infer their relative distances [3]. In plants, linkage maps are typically established by observing allele recombination in bi-parental segregating F_2 populations using the "test-cross" method (invented by Gregor Mendel and first applied to pea plants), or variants thereof [4]. While establishing segregating populations can be expensive, laborious, and even impossible in some circumstances, linkage maps are still widely used to identify Quantitative Trait Loci (QTL) via recombination mapping. They are also useful for studying recombination rates between individuals, which is relevant for breeding.

Physical maps refer to a broad class of maps that represent genome sequences. Some physical maps, referred to here as "sequence-unresolved maps" do not represent continuous genomic sequences but only describe how many nucleotides are between markers. These include optical maps, restriction digest maps, linked reads, and mate pairs. Aside from structural variant analysis and genome assembly scaffolding, sequence-unresolved physical maps are useful for inferring the size of genomic loci that are difficult to continuously resolve, such as centromeres [5–7]. Other physical maps, referred to here as "sequence-resolved maps" such as genome assemblies and long DNA sequencing reads, represent continuous genomic sequences. Such maps provide the same utility as sequence-unresolved maps while also directly encoding genotype information and providing references for genomics experiments and oligonucleotide design. Accordingly, sequence-resolved physical maps are the most complete and useful type of physical map and are expected to render sequence-unresolved maps obsolete as sequencing technologies continue to advance. More details on the current state of genome assembly and technical details on genome assembly processes are outlined in "Accurate and complete physical maps via genome assembly and scaffolding".

Optical maps are a specific type of physical map that uses imaging to infer the relative distances of markers. First, high-molecular-weight DNA molecules are fluorescently labeled at specific loci, essentially creating a fingerprint for each molecule. Images are taken of these linearized and labeled molecules and software is used to infer a corresponding physical map. Each physical map corresponding to a molecule can be analyzed individually, or they can be assembled into more contiguous physical maps, similar to genome assembly, but without specific sequence information. While these maps do not contain actual sequences, they are often much longer than DNA sequencing reads and thus provide useful long-range structural information.

Spatial proximity maps convey the relative distance of genetic markers in three-dimensional space. To measure 3D interaction, researchers use Chromatin Conformation Capture followed by sequencing, or Hi-C [8]. Hi-C first employs specialized chemistry techniques to establish a sequencing library with chimeric DNA templates composed of two usually distant loci. By sequencing these chimeric inserts with paired-end sequencing, read pairs can be mapped to a genome assembly and the 3D distance of any two loci can be inferred from read mapping statistics. Spatial proximity maps are primarily used to study the 3D organization of a genome within the nucleus, which plays an important role in gene and genome regulation.

Accurate and complete physical maps via genome assembly and scaffolding

Genome assemblies are, ideally, complete sequence representations of whole genomes. Advances in sequencing technology and assembly algorithms have made genome assembly accessible for researchers studying model and non-model organisms across the tree of life [9,10]. Researchers recently published the first-ever truly complete human genome assembly, indicating that complete genome assemblies for model organisms are on the horizon [11]. Multiple diverse human genomes are increasingly being assembled, and the Vertebrate Genomes Project has begun establishing genome assemblies for every extant vertebrate species [10,12]. In major crops such as soybean, rapeseed, rice, and maize, large collections of genome assemblies have been published representing wild accessions, important cultivars, and mapping population parents [13–16].

Modern genome assemblies are typically derived from long Whole Genome Shotgun (WGS) DNA sequencing reads. Today, these long reads are primarily commercialized by two companies: Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). ONT provides the longest read lengths, ranging from 5kb to well over 100 kbp long. ONT read accuracy has changed dramatically over recent years as basecalling software has improved. Though read accuracy depends on the organism being sequenced, reads with 99% accuracy are now routinely achieved in human samples. PacBio high fidelity long reads (HiFi) are relatively shorter (between 5k and 20k) but are much more accurate (approaching 99.9% accuracy) [17]. PacBio also produces Continuous Long Reads (CLR) that are longer yet noisier than HiFi. While CLR data is still used for legacy projects, it has been mostly rendered obsolete by HiFi and ONT.

The ideal eukaryotic genome assembly completely represents the sequence of every chromosome in a genome. To combine WGS reads into larger, non-redundant contiguous sequences (contigs), genome assemblers approximately or exactly perform all-by-all pairwise comparisons of reads to find sequence overlaps. Assemblers then use this overlap information to build graph data structures that encode sequences and their relationships. Finally, contigs are derived by finding solutions to these graphs that remove redundant or erroneous nodes and edges [18–23]. One major challenge of genome assembly is that DNA is usually sampled from multiple genomes. Samples usually contain DNA from two (one maternal and one paternal) homologous genomes (haplotypes). Homologs can be relatively similar (low rates of heterozygosity) or distinct (high rates of heterozygosity), depending on the species and sample. Additionally, as is common in plant species, samples derived from polyploid genomes contain DNA from multiple homeologous subgenomes. Both autopolyploidy (relatively similar homeologs) and allopolyploidy (relatively distinct homeologs) result from events such as unreduced gametes that increase chromosome copy number. Autopolyploidy occurs when these extra chromosome copies come from the same species, whereas allopolyploidy occurs when they come from distinct species. Aside from mixed genomes, genomic repeats present another major challenge for genome assembly algorithms. When comparing reads to each other, repeats can cause ambiguities in assembly graphs and thus cause assembly errors or omissions.

The problems of mixed genomes and genomic repeats are conceptually similar. Assemblers often confuse haplotypic or subgenome homology with repeats and vice versa [10]. To address both of these challenges, assemblers require input sequencing reads that are capable of resolving the ambiguities caused by these phenomena. If one assumes that reads are perfectly accurate, then assemblers require reads that are long enough to connect ambiguous loci to their surrounding context. For example, in the case of samples with mixed homologous genomes, reads need to be long enough to anchor homozygous sequence to allelic loci (known as "phasing"). For genomic repeats, reads need to be long enough to anchor repetitive sequences to the closest unique sequence. In reality, sequencing reads do not have perfect accuracy, so read length and read accuracy are the two main parameters controlling the ability of reads to resolve ambiguities [24]. ONT reads are exceptionally long, but HiFi reads are exceptionally accurate, and therefore the two technologies can serve complementary roles in genome assembly projects [25].

Aside from using long and accurate reads, researchers have devised alternate techniques to specifically address the challenge of mixed genomes. Firstly, for diploid samples, many assemblies simply choose an allele at each bi-allelic site, either arbitrarily or according to a heuristic, thus creating a "pseudo-haploid" assembly that continuously alternates between haplotypes. Though "pseudo-haploid" assemblies were default for many years, new techniques are now commonly used to produce one assembly for each haplotype ("haplotype resolved assemblies"). For example, trio-binning, gamete-binning, or read mapping and phasing-based approaches are techniques that separate reads into respective haplotypes before assembly [26–28]. For certain model organisms, inbred, double haploid, or haploid samples can be used for sequencing, thus ensuring that samples effectively only represent one genome [11,29].

Though advances in sequencing read length and accuracy combined with improved assembly algorithms have dramatically improved the quality of genome assemblies in recent years, draft assemblies still rarely achieve chromosome-scale (one sequence per chromosome). Instead, each chromosome is usually represented by a collection of unordered and unoriented contigs. “Scaffolding” is the process of ordering and orienting these contigs, placing gaps between adjacent contigs, into chromosome-scale sequences (**Figure 0.2**). Building such chromosome-scale scaffolds is essential for establishing chromosome-sized physical maps, which is crucial for understanding many biological phenomena and creating useful reference genomes. Scaffolding is usually performed by aligning a draft genome assembly to another genome map and then ordering and orienting contigs according to the structure indicated in the map [30–33]. Linkage, spatial proximity, and physical maps (genome assemblies and optical maps) are popular and effective maps for scaffolding [27,34–36]. Because eukaryotic genomes can be structurally diverse, the map and the draft assembly ideally should represent the same genome. However, especially with contiguous draft assemblies, maps representing closely related but distinct genomes can be used while preserving structural variation. This is because contiguous draft assemblies contain most or sometimes all genetic variation within contigs, therefore ordering and orienting contigs does not influence genotype information.

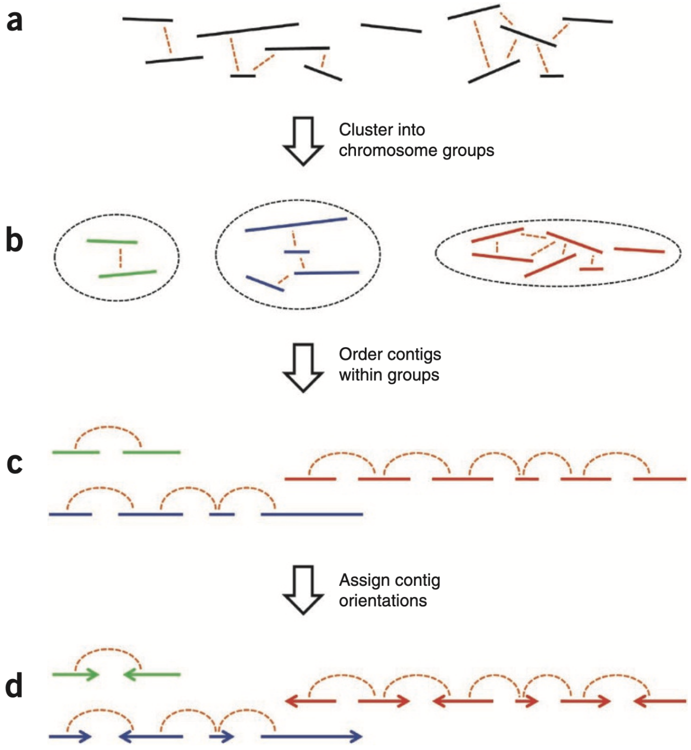

a
Cluster into chromosome groups
b
Order contigs within groups
c
Assign contig orientations
d

Figure 0.2: Genome assembly scaffolding. A diagram describing how sets of genome assembly contigs are ordered and oriented to build chromosome-scale scaffolds. Original figure from Burton, J. N. et al, 2013 [32].

Though genome maps are effective for scaffolding modern genome assemblies, most initial automated scaffolding attempts either fail to achieve complete chromosome-scale or incorrectly order or orient contigs. Chromosome-scale genome assembly physical maps usually enable chromosome-scale scaffolds, but misassemblies or genuine structural variation can lead to errors in the scaffolds. Optical maps may not always achieve chromosome-scale and they often lack markers in extended repetitive sequences such as centromeres. Spatial proximity maps rely on short-read mapping to reference genomes, and therefore repeats with low mappability are not well-represented. Additionally, genuine 3D genome organization can obscure the linear sequence order of a chromosome, often causing misassemblies. Finally, scaffolding algorithms can be computationally inefficient, both in terms of space and time [33,37]. Aside from causing inconvenience and expense, this often necessitates scaffolding algorithms that are not guaranteed to be optimal. Given these scaffolding shortcomings, researchers often manually compare contigs to maps to achieve accurate and chromosome-scale scaffolds [38,39]. While this can be effective, it is laborious and prone to human error.

Genome maps reveal genome structure

Genome maps can be analyzed and compared to study genomic characteristics. One such characteristic is genome "structure", or anything regarding the large-scale (typically >=50 bp) sequence composition of genomes. This is distinct from 3D genome structure, which is a separate field of study and is not addressed here. Here, we outline two specific motivations for studying genome structure, namely "intrinsic" and "extrinsic". Intrinsic genome structure refers to large genomic features that are independently important for genome function and evolution. Researchers study intrinsic genome structure by annotating and analyzing important features of a genome, such as genes, repeats, and regulatory elements. While this encompasses many fields of study, one example of using maps to study intrinsic genome structure is the study of genomic repeats. A common technique used to study genomic repeats is to align a genome assembly to itself. In such a scenario, every locus will align perfectly to itself, but repetitive elements will additionally align to similar sequences across the genome. For example, such analysis can uncover the structure and composition of large satellite repeats, such as centromeres [40,41]. It can also reveal gene paralogs and signatures of ancient polyploidy, a phenomenon that is especially relevant for plant genomes (**Figure 0.3**) [42].

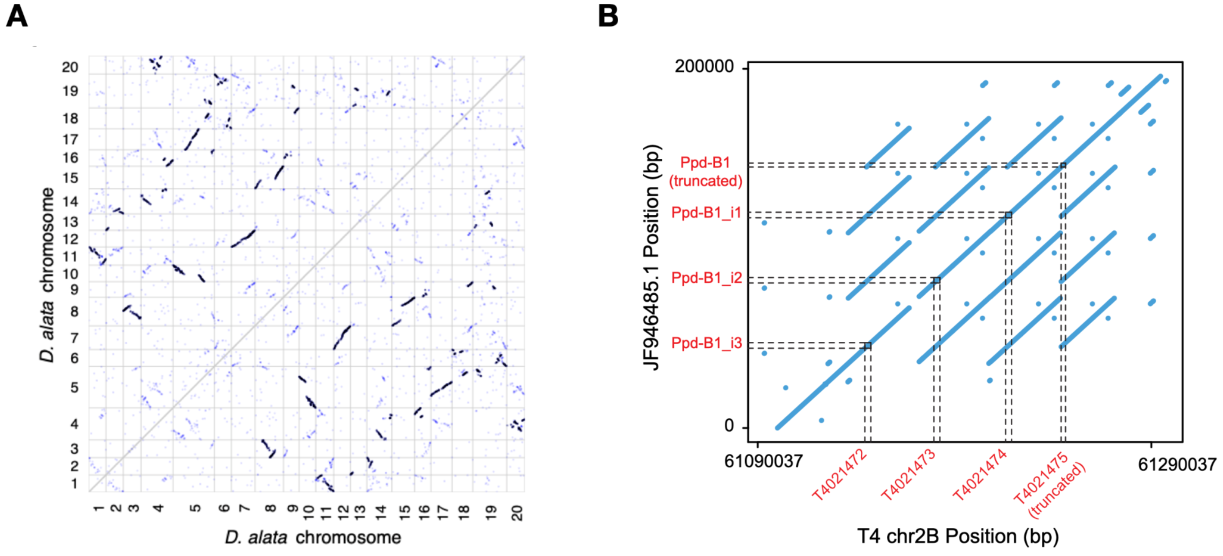


Figure 0.3: Intrinsic repetitive genome structure. (A) Aligning a whole *Dioscorea alata* genome to itself reveals widespread sequence duplication, providing evidence for ancient polyploidy. Original figure from Bredeson, J. V. and Lyons, J. B. et al, 2021 [42]. (B) Aligning a chr2B locus of a *Triticum aestivum* genome assembly to itself reveals tandemly repeated paralogs (dotted lines with red labels) in a segmental duplication. Original figure from Alonge, M. and Shumate, A. et al, 2020 [35].

Extrinsic genome structure refers to the comparison of genome structure across samples and species. This is often referred to as "comparative genomics", which for this dissertation, includes genome comparison both within and between species. Within species, comparative genomics includes the study of structural variants (SVs), or large genetic variants segregating in a population. Structural variants are widespread in eukaryotic genomes, and they underlie large-scale sequence differences between individuals and explain the phenotypic variation of many important traits [43,44]. Genome maps are especially well-suited for this analysis as two genome maps representing distinct genomes can be aligned to each other and large differences in the maps can be interpreted as SVs (**Figure 0.4**) [45]. While all chromosome-scale maps can be effectively used to identify the location of SVs, sequence-resolved physical maps, such as long-reads and genome assemblies are required to characterize specific variable sequences [12]. To study SVs with long reads, reads are typically aligned to a reference genome, and discordant alignment signatures are used to infer SVs [46–48]. To discover SVs from genome assemblies, one can simulate reads from a genome assembly template and then use read mapping-based SV detection strategies. Additionally, one can align the genome assembly to a reference genome with a whole-genome aligner and interpret SVs via discordant alignments [12,49–51]. While such pairwise comparative techniques are currently the most common, efficient, and interpretable forms of SV analysis, multiple whole-genome sequence alignment (comparing >2 maps simultaneously) has matured in recent years and shows promise as a useful comparison technique [52].

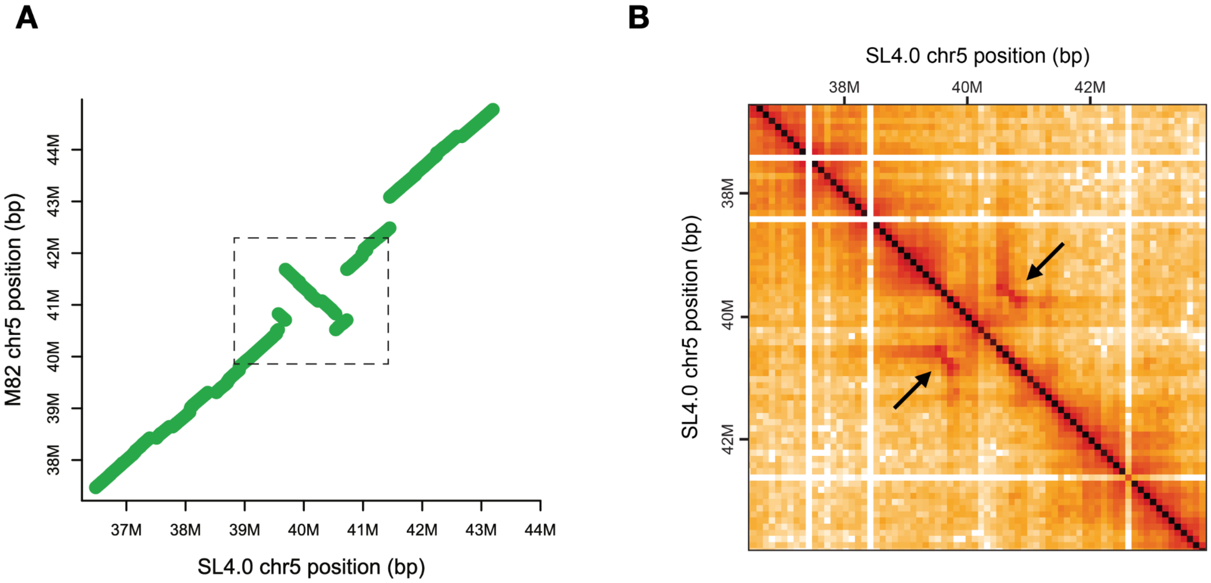


Figure 0.4: Comparing genome maps reveals structural variants. (A) Comparing an M82 tomato genome assembly to the SL4.0 genome assembly reveals a large inversion on chromosome 5, enclosed in the dotted box. (B). Comparing M82 Hi-C data to the SL4.0 genome assembly reveals supporting evidence for the inversion, with black arrows pointing to inversion breakpoints.

A eukaryotic “pan-genome” is a collection of sequence-resolved physical maps for a representative set of individuals within a population or species [9]. Such pan-genomes enable the broad study of intrinsic and extrinsic genome structure at scale. While pan-genome studies have revealed comparatively little genetic diversity in humans, plant and especially crop pan-genomes have revealed dramatic differences between members of the same species [9,53]. Some crops such as tomatoes are especially structurally diverse due to pre-breeding and breeding with wild material [44]. Such intra-species structural genetic diversity has major implications for genomics experiments that typically compare populations to a single reference genome. Namely, sequencing reads emanating from structurally distinct loci will map with lower accuracy, a phenomenon known as “reference bias” [54]. In crops such as tomato with widespread artificial introgressions, reference bias may occur locally at specific introgressed loci. To overcome such reference bias, researchers may employ modified reference genomes or associated indexes such as pan-genome graphs or major allele references [55–63]. The ideal reference genome represents the same genome as the sample of interest, also known as a "personalized" reference genome. This is especially useful for plants with reference genotypes that are maintained via seed stocks and commonly used for genetics experiments. Given the new relative accessibility of establishing high-quality reference genome assemblies, researchers can now plan experiments around efficient and convenient model systems rather than reference genomes.

Here, we first explore new methods to facilitate accurate and automatic genome assembly scaffolding. We first introduce RaGOO, a homology-based scaffolder that uses one genome assembly to scaffold another assembly. We next present RagTag, the successor to RaGOO that introduces a patching and gap-filling feature, as well as a new tool to reconcile many scaffolding proposals for a given genome assembly. While presenting both RaGOO and RagTag, we demonstrate how scaffolded physical maps can be useful for comparative and personalized genomics analysis in tomato and *Arabidopsis thaliana*. We also show how we used RagTag to scaffold a bread wheat genome. Even though both scaffolding and annotation relied on an established reference genome, we uncovered over 1 Gbp of new sequence and over 5,700 new gene copies, thus more comprehensively elucidating the paralog landscape of wheat. Next, we used ONT and HiFi to almost completely assemble and resolve an *Arabidopsis thaliana* genome, including all five centromeres, thus revealing Arabidopsis centromeric architecture and genetics for the first time. Finally, we describe an effort to catalog natural structural variation among 100 diverse tomato genomes, thus revealing how SVs broadly impact gene expression as well as specific instances of SVs influencing important traits. These new methods and applied analyses highlight the unprecedented speed and scale of plant genomics in the pan-genome era and reveal the vast potential for genomics to address the world's most pressing agricultural and nutritional needs.

REFERENCES

1. al-Janabi SM, Honeycutt RJ, McClelland M, Sobral BW. A genetic linkage map of Saccharum spontaneum L. "SES 208." Genetics. academic.oup.com; 1993;134:1249–60.

2. Ferrarini M, Moretto M, Ward JA, Šurbanovski N, Stevanović V, Giongo L, et al. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics. Springer Science and Business Media LLC; 2013;14:670.

3. Sturtevant AH. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. J Exp Zool. Wiley Subscription Services, Inc., A Wiley Company New York; 1913;14:43–59.

4. Grattapaglia D, Sederoff R. Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. Genetics. academic.oup.com; 1994;137:1121–37.

5. Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H. The size and sequence organization of the centromeric region of arabidopsis thaliana chromosome 5. DNA Res. academic.oup.com; 2000;7:315–21.

6. Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H. The size and sequence organization of the centromeric region of Arabidopsis thaliana chromosome 4. DNA Res. academic.oup.com; 2001;8:285–90.

7. Hosouchi T, Kumekawa N, Tsuruoka H, Kotani H. Physical map-based sizes of the centromeric regions of Arabidopsis thaliana chromosomes 1, 2, and 3. DNA Res. academic.oup.com; 2002;9:117–21.

8. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.

9. Lei L, Goltsman E, Goodstein D, Wu GA, Rokhsar DS, Vogel JP. Plant Pan-Genomics Comes of Age. Annu Rev Plant Biol. annualreviews.org; 2021;72:411–35.

10. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46.

11. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV. The complete sequence of a human genome. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1.abstract

12. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science [Internet]. 2021;372. Available from: http://dx.doi.org/10.1126/science.abf7117

13. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-Genome of Wild and Cultivated Soybeans. Cell. Elsevier; 2020;182:162–76.e13.

14. Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants. 2020;6:34–45.

15. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet. nature.com; 2018;50:278–84.

16. Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. biorxiv.org; 2021;373:655–62.

17. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.

18. Myers EW. The fragment assembly string graph. Bioinformatics. academic.oup.com; 2005;21 Suppl 2:ii79–85.

19. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. academic.oup.com; 2016;32:2103–10.

20. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. genome.cshlp.org; 2017;27:722–36.

21. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. nature.com; 2021;18:170–5.

22. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6.

23. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.

24. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36:338–45.

25. van Rengs WMJ, Schmidt MHW, Effgen S, Wang Y. A gap-free tomato genome built from complementary PacBio and Nanopore long DNA sequences reveals extensive linkage drag during breeding. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.08.30.456472.abstract

26. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol [Internet]. nature.com; 2018; Available from: http://dx.doi.org/10.1038/nbt.4277

27. Sun H, Jiao WB, Krause K, Campoy JA, Goel M. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.05.15.444292v1.abstract

28. Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, et al. Chromosome-scale, haplotype-resolved assembly of human genomes. Nat Biotechnol. nature.com; 2021;39:309–12.

29. Scott AD, Zimin AV, Puiu D, Workman R, Britton M, Zaman S, et al. A reference genome sequence for giant sequoia. G3: Genes, Genomes, Genetics. Oxford University Press; 2020;10:3907–19.

30. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20:224.

31. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, et al. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 2015;16:3.

32. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31:1119–25.

33. Pop M, Kosack DS, Salzberg SL. Hierarchical scaffolding with Bambus. Genome Res. 2004;14:149–59.

34. International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators:, Appels R,

Eversole K, Feuillet C, Keller B, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science [Internet]. science.sciencemag.org; 2018;361. Available from: http://dx.doi.org/10.1126/science.aar7191

35. Alonge M, Shumate A, Puiu D, Zimin AV, Salzberg SL. Chromosome-Scale Assembly of the Bread Wheat Genome Reveals Thousands of Additional Gene Copies. Genetics. academic.oup.com; 2020;216:599–608.

36. Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, et al. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. bioRxiv. biorxiv.org; 2019;767764.

37. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;15:e1007273.

38. Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, et al. Significantly improving the quality of genome assemblies through curation. Gigascience [Internet]. academic.oup.com; 2021;10. Available from: http://dx.doi.org/10.1093/gigascience/giaa153

39. Dudchenko O, Shamim MS, Batra S, Durand NC. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv [Internet]. biorxiv.org; 2018; Available from: https://www.biorxiv.org/content/10.1101/254797v1.abstract

40. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, et al. The structure, function, and evolution of a complete human chromosome 8. bioRxiv. 2020;2020.09.08.285395.

41. Naish M, Alonge M, Wlodzimierz P, Tock AJ. The genetic and epigenetic landscape of the Arabidopsis centromeres. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.05.30.446350v2.abstract

42. Bredeson JV, Lyons JB, Oniyinde IO, Okereke NR, Kolade O, Nnabue I, et al. Chromosome evolution and the genetic basis of agronomically important traits in greater yam [Internet]. bioRxiv. 2021 [cited 2021 Sep 21]. p. 2021.04.14.439117. Available from: https://www.biorxiv.org/content/10.1101/2021.04.14.439117v1.abstract

43. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. nature.com; 2013;14:125–38.

44. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. Elsevier; 2020;182:145–61.e23.

45. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. genomebiology.biomedcentral.com; 2019;20:246.

46. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. nature.com; 2018;15:461–8.

47. Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinformatics. academic.oup.com; 2019;35:2907–15.

48. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. genomebiology.biomedcentral.com; 2020;21:189.

49. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016;32:3021–3.

50. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. Springer Science and Business Media LLC; 2019;20:277.

51. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

52. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587:246–51.

53. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. Nat Rev Genet. 2020;21:243–54.

54. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. academic.oup.com; 2009;25:3207–12.

55. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. Genome Biol. Springer; 2020;21:265.

56. Pritt J, Chen N-C, Langmead B. FORGe: prioritizing variants for graph genomes. Genome Biol. genomebiology.biomedcentral.com; 2018;19:220.

57. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. nature.com; 2018;36:875–9.

58. Chen N-C, Solomon B, Mun T, Iyer S, Langmead B. Reference flow: reducing reference bias using multiple population genomes. Genome Biol. Springer Science and Business Media LLC; 2021;22:8.

59. Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O, et al. Simultaneous alignment of short reads against multiple genomes. Genome Biol. genomebiology.biomedcentral.com; 2009;10:R98.

60. Satya RV, Zavaljevski N, Reifman J. A new strategy to reduce allelic bias in RNA-Seq readmapping. Nucleic Acids Res. academic.oup.com; 2012;40:e127.

61. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. nature.com; 2019;37:907–15.

62. Huang L, Popic V, Batzoglou S. Short read alignment with populations of genomes. Bioinformatics. academic.oup.com; 2013;29:i361–70.

63. Maciuca S, del Ojo Elias C, McVean G, Iqbal Z. A Natural Encoding of Genetic Variation in a Burrows-Wheeler Transform to Enable Mapping and Genome Inference. Algorithms in Bioinformatics. Springer International Publishing; 2016. p. 222–33.

RaGOO is a fast and reliable reference-guided scaffolding method ... that orders and orients genome assembly contigs according to Minimap2 alignments to a single reference genome.

1

RaGOO: fast and accurate reference-guided scaffolding of draft genomes

The following sections were previously published in *Genome Biology*:

> Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F.J., Lippman, Z.B., and Schatz, M.C. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol 20, 224 (2019). https://doi.org/10.1186/s13059-019-1829-6

M.C.S. supervised all analysis and manuscript preparation. M.A. and M.C.S. edited all of the text. The results section "Pan-SV analysis of three chromosome-scale tomato genome assemblies" was written by S.S and Z.B.L. The results section "M82 chromosome Hi-C validation, finishing, and annotation" was done by M.A. and S.R. and written by M.A. The results sections "Simulated reference-guided scaffolding", "Pan-SV analysis of 3 tomato accessions", and "Pan-SV analysis of 103 Arabidopsis thaliana genomes" were done by M.A. with help from F.J.S. The methods sections "Plant material and growth conditions", "Genome and transcriptome sequences", "Tissue collection and high molecular weight DNA extraction", and "Hi-C library preparation and sequencing" were done and written by S.S. and X.W. The methods sections "Nanopore library preparation and sequencing" and "10× Genomics library preparation and sequencing" were done and written by S.G. The methods section "Tomato genome annotation" was done by S.R. and written by M.A. and S.R. All other results and methods sections, figures, and tables were done/written by M.A. M.A. wrote the abstract, background, discussion, and conclusions, with editing from M.C.S.

1.1 ABSTRACT

We present RaGOO, a reference-guided contig ordering and orienting tool that leverages the speed and sensitivity of Minimap2 to accurately achieve chromosome-scale assemblies in minutes. After the pseudomolecules are constructed, RaGOO identifies structural variants, including those spanning sequencing gaps. We show that RaGOO accurately orders and orients 3 de novo tomato genome assemblies, including the widely used M82 reference cultivar. We then demonstrate the scalability and utility of RaGOO with a pan-genome analysis of 103 *Arabidopsis thaliana* accessions by examining the structural variants detected in the newly assembled pseudomolecules. RaGOO is available open source at https://github.com/malonge/RaGOO.

1.2 BACKGROUND

Long-read single-molecule sequencing technologies commercialized by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) have facilitated a resurgence of high-quality de novo eukaryotic genome assemblies [1]. Assemblies using these technologies in a variety of plant and animal species have consistently reported contig N50s over 1 Mbp, while also reconstructing higher percentages of target genomes, including repetitive sequences [2,3]. Current long-read sequencers are now able to produce over one terabase of long reads per week, presenting the opportunity for detailed pan-genome analysis of unprecedented scale. Such analyses can include structural variations that are notoriously difficult to detect using short-read sequencing. However, lagging behind the current speed and cost of generating long-read sequencing data are genome assemblers, which are still unable to resolve complex repeats and related structural variants that are widespread in eukaryotic genomes. Thus, there is a need for simplified and faster approaches to scaffold fragmented genome assemblies into chromosome-scale pseudomolecules.

Two common approaches have been used to achieve chromosome-scale assemblies, namely, reference-free (de novo) and reference-guided approaches. One popular reference-free scaffolding approach is to anchor genome assembly contigs to some type of genome map [4], such as a physical or linkage map [5]. This process involves aligning the genomic map to a sequence assembly and scaffolding contigs according to the chromosomal structure indicated in the map. However, contigs not implicated in any alignments will fail to be scaffolded, which can result in incomplete scaffolding. Furthermore, acquiring a genomic map can be expensive, time-consuming, or otherwise intractable depending on the species and the type of map.

Another reference-free method for pseudomolecule construction involves the use of long-range genomic information to scaffold assembled contigs. This includes a large class of technologies such as mate-pair sequencing, Bacterial Artificial Chromosomes (BACs), Linked Reads, and chromatin conformation capture such as Hi-C [6–8]. In particular, Hi-C has recently been shown to be a practical and effective resource for chromosome-scale scaffolding [9–11]. Paired-end Hi-C sequencing reads are aligned to the assembly, and mates which align to different contigs (Hi-C links) are recorded. According to the relative density of such Hi-C links between pairs of contigs, contigs can be ordered and oriented into larger scaffolds, potentially forming chromosome-length pseudomolecules. Also, because misassemblies may be observed by visualizing Hi- C alignments, Hi-C can be used for validation and manual correction of misassemblies [12]. Though Hi-C has been widely adopted, there remain challenges that can impede the ability to form accurate chromosome-scale pseudomolecules with Hi-C alone. Principally, Hi-C data are noisy, and Hi-C-based scaffolders are prone to producing structurally inaccurate scaffolds [13]. Also, because this process relies on the alignment of short Hi-C sequencing reads to the draft assembly, small and repetitive contigs with little or conflicting Hi-C link information often fail to be accurately scaffolded. Finally, the analysis requires deep sequencing coverage and therefore can be expensive and compute-intensive.

Aside from reference-free approaches, there are also a few tools available for reference-guided scaffolding [14]. For example, Chromosomer and MUMmer's "show-tiling" utility leverage pairwise alignments to a reference genome for contig scaffolding and have been used to scaffold eukaryotic genomes [15–17]. RACA is similar, though it also requires paired-end sequencing data to aid scaffolding [18]. Finally, tools such as GOSASM and Ragout2 employ multiple sequence aligners to reconcile multiple, potentially diverse contig sets [19,20]. Though reference-guided scaffolding may introduce erroneous reference bias, it is often substantially faster and less expensive than acquiring the resources for the reference-free methods outlined above. However, current tools for reference-guided scaffolding of eukaryotic genomes have notable shortcomings.

Firstly, these tools depend on slower DNA aligners such as BLAST and Nucmer and accordingly require long compute times of several hours to several days for mammalian-sized genomes [21]. This is especially pronounced in tools like Ragout2 that use multiple sequence aligners, such as Cactus, that can require hundreds of CPU hours for large eukaryotic genomes [22]. These aligners are also not robust to repetitive and/or gapped alignments resulting in a significant portion of contigs being unassigned in pseudomolecules. Finally, many of these methods do not internally offer the ability to correct large-scale misassemblies frequently present in draft assemblies of eukaryotic genomes nor report any metrics on conflicts due to true biological differences in the genomes.

Here, we introduce RaGOO, an open-source method that utilizes Minimap2 [23] alignments to a closely related reference genome to quickly cluster, order, and orient genome assembly contigs into pseudomolecules. RaGOO also provides the option to correct apparent chimeric contigs before pseudomolecule construction. Finally, structural variants (SVs), including those spanning gaps, are identified using an optimized and integrated version of Assemblytics [24], thus enabling rapid pan-genome SV analysis of many genomes at once. This is especially important for detecting large insertions and other complex structural variations that are difficult to detect using read mapping approaches.

We first demonstrate the speed and accuracy of RaGOO scaffolding with simulated data of increasing complexity and show that it outperforms 2 popular alternative methods. We next show the utility of RaGOO by creating high-quality chromosome-scale reference genomes for 3 distinct wild and domesticated genotypes of the model crop tomato using a combination of short and long-read sequencing. Finally, we demonstrate the scalability of RaGOO by ordering and orienting 103 draft *Arabidopsis thaliana* genomes and comparing structural variants across the pan-genome. This uncovers a large number of defense response genes that are highly variable.

1.3 RESULTS

Reference-guided contig ordering and orientation with RaGOO

RaGOO is a fast and reliable reference-guided scaffolding method, implemented as an open-source python command-line utility, that orders and orients genome assembly contigs according to Minimap2 alignments to a single reference genome (**Figure 1.1**) [25]. RaGOO's primary goal is to utilize the large-scale structure of a reference genome to organize assembly contigs, analogous to how a genetic map is used. Therefore, under default settings, RaGOO does not alter or mutate any input assembly sequence but rather arranges them and places gaps for padding between contigs. Additionally, users have the option to break input contigs at points of potential misassembly indicated by discordant alignments to the reference genome. However, these breaks will only fragment the assembly and do not add or remove any sequence content. RaGOO can optionally avoid breaking chimeric intervals at loci within genomic coordinates specified by a gff3 file, to avoid disrupting gene models identified in the de novo assembly.

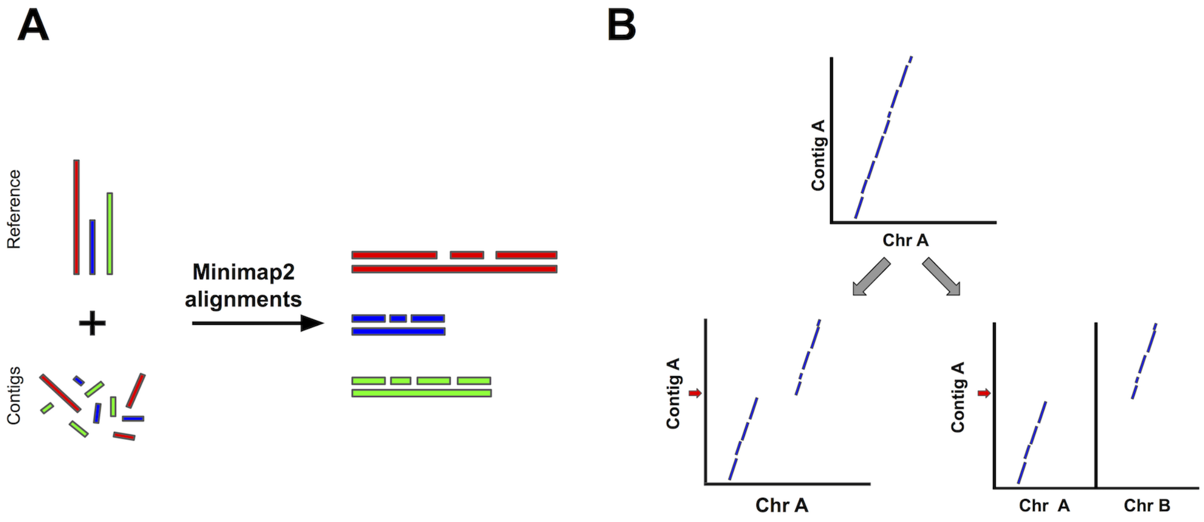


Figure 1.1: The RaGOO pipeline. (A) Contigs are aligned to the reference genome with Minimap2 and are ordered and oriented according to those alignments. (B) Normal alignments between a contig and a reference chromosome (top) and example alignments between a reference chromosome and an intrachromosomal chimera (bottom left) and an interchromosomal chimera (bottom right). Red arrows represent potential contig breakpoints.

Additionally, RaGOO computes confidence scores associated with the clustering, ordering, and orienting of each contig. These scores ultimately strive to measure the fidelity of contig ordering and orienting to the underlying alignments. For example, a contig that aligns with equal coverage to three different chromosomes will have a lower clustering confidence score than a contig that exclusively aligns to a single chromosome. These scores can also be viewed as measuring the level of scaffolding ambiguity present in the alignments. Accordingly, one can compare confidence scores with and without chimeric contig correction to ensure that alignments become less ambiguous after correction (see the "M82 chromosome Hi-C validation, finishing, annotation" section). Furthermore, a poor confidence score distribution can indicate that a draft assembly is too divergent from the reference assembly for optimal scaffolding (see the "Scaffolding a divergent *S. pennellii* genome assembly" section).

After constructing pseudomolecules, RaGOO re-aligns the assembly to the reference and calls structural variants with an integrated version of Assemblytics. We have optimized this approach by replacing the relatively slow single-threaded nucmer alignment phase with the much faster Minimap2 aligner along with the necessary converters between the output formats. Noting that such alignments may traverse gaps in either the reference or the query assembly, we report the percent overlap between each SV and gaps, allowing users to utilize such variants at their discretion. Importantly, the speed of Minimap2 alignments, and therefore RaGOO, facilitates a genome scaffolding and SV analysis at scales previously not feasible with comparable tools. For example, RaGOO scaffolds an *Arabidopsis thaliana* draft assembly in ~ 13 s and a human draft assembly in ~ 12 min and 33 s using eight cores and less than 20 GB of RAM [26].

Simulated reference-guided scaffolding

To assess the efficacy of RaGOO, we used it to scaffold simulated draft eukaryotic genome assemblies of increasing difficulty. To simulate these assemblies, we partitioned the current tomato (*Solanum lycopersicum*) reference genome (Heinz version SL3.0) into variable-length scaffolds [27]. To achieve a realistic distribution of sequence lengths, we sampled the observed contig lengths from a de novo assembly produced with Oxford Nanopore long reads of the *S. lycopersicum* cultivar M82, which is described later in this paper. Given that many of these resulting scaffolds contained a gap sequence ("N" characters) from the reference genome, we also established an assembly comprised of contigs free of sequencing gaps. For this, we split the simulated scaffolds at any stretch of 20 or more "N" characters, excluding the gap sequence. We also excluded any resulting contigs shorter than 10 kbp in length. We refer to these scaffolds and contigs as the "easy" set of simulated data, as they are a partitioning of the reference with no variation. To simulate a "hard" dataset that contained variation, we used SURVIVOR [28] to simulate 10,000 insertion and deletion SVs, ranging in size from 20 bp to 10 kbp, and SNPs at a rate of 1% into the simulated scaffolds. Contigs were then derived from these scaffolds just as with the "easy" contigs.

Utilizing the same SL3.0 reference assembly, we used MUMmer's "show-tiling" utility, as well as Chromosomer and RaGOO to arrange these simulated assemblies into 12 pseudomolecules. To assess scaffolding success, we measured clustering, ordering, and orienting accuracy. Clustering and orienting accuracy is the percentage of localized contigs that were assigned the correct chromosome group and orientation, respectively. To assess the ordering accuracy, the edit distance between the true and predicted contig order was calculated for each pseudomolecule normalized by the true number of contigs in the pseudomolecule. Additionally, for a local measurement of ordering accuracy, the fraction of correct adjacent contig pairs was computed for each pseudomolecule. Finally, to measure the scaffolding completeness, we noted the percentage of contigs and total sequence localized into pseudomolecules.

RaGOO performed best on all datasets, achieving high clustering, ordering, and orienting accuracy on both the "easy" and "hard" datasets, while localizing the vast majority (~ 99.9998% for hard scaffolds) of sequence in only a few minutes (1 min and 15 s for the "hard" scaffolds) (**Figure 1.2**). In all simulations, Chromosomer accurately reconstructed most of the genome, though the presence of gaps in scaffolds and variation in the "hard" assembly degraded the performance to a localization score of 86.65% in the "hard" scaffolds. Show-tiling suffered tremendously from the presence of gaps in scaffolds and accordingly achieved poor localization scores on scaffolds of both the "easy" (8.43%) and "hard" (0.01%) sets. Both Chromosomer and show-tiling took substantially longer to run than RaGOO in all cases and required several hours rather than minutes.

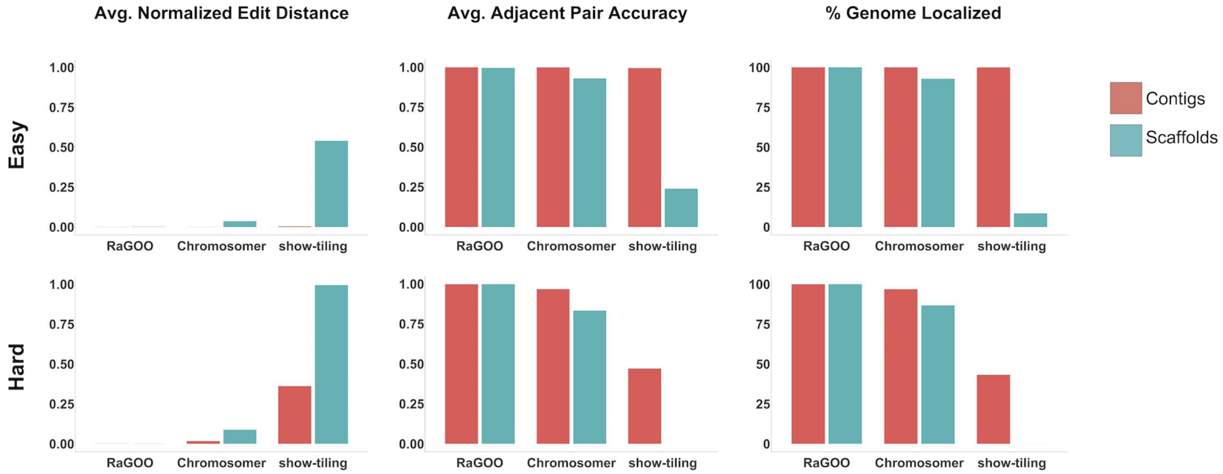


Figure 1.2: Scaffolding simulated assemblies. Ordering and localization results for "easy" and "hard" simulated tomato genome assemblies. Normalized edit distance and adjacent pair accuracy measure the success of contig ordering and are averaged across the 12 simulated chromosomes. The percentage of the genome localized measures how much of the simulated assemblies were clustered, ordered, and oriented into pseudomolecules.

Pan-SV analysis of three chromosome-scale tomato genome assemblies

For more than a decade, the reference genome for tomato (var. "Heinz 1706") has been an invaluable resource in both basic and applied research, but extensive sequence gaps (81.7 Mbp, 9.87%), unlocalized sequence (~ 17.8 Mbp, 2.39%), and limited information on natural genetic variation in the wider germplasm pool impeded its full utilization [27]. To compensate, more than 700 additional accessions have since been sequenced by Illumina short-read technology [29,30]. However, due to the short sequence reads, these studies were limited to evaluating, with reasonable accuracy (depending on variable sequencing quality and coverage), single nucleotide polymorphisms (SNPs), and small insertions and deletions (indels). In contrast, larger structural variations (SVs) that have important and often underestimated functional consequences for genome evolution and phenotypic diversity were largely ignored in this major model crop plant. Critically, without long reads, the complete catalog of structural variations in the species, a pan-SV analysis, is largely incomplete.

To address this knowledge gap and begin constructing a high-quality tomato pan-SV analysis, we used long-read ONT instruments to sequence three distinct genotypes that provide anchor points for wild and domesticated tomato germplasm:

(1) the species *S. pimpinellifolium* is the ancestor of tomato, and the Ecuadorian *S. pimpinellifolium* accession BGV006775 (BGV) represents the group of progenitors that are most closely related to early domesticated types; (2) the *S. lycopersicum* processing cultivar M82 is the most widely used accession in research due to its rich genetic resources; and (3) the *S. lycopersicum* elite breeding line Fla.8924 (FLA) is a large-fruited "fresh market" type that was developed for open-field production in Florida [31,32]. Together, these three accessions provide a foundation for constructing a pan-SV analysis that will enable the identification and classification of thousands of predicted SVs.

Reference-guided and reference-free M82 scaffolding

To evaluate the effectiveness of RaGOO with genuine sequencing data, we first used it along with other reference-guided and reference-free tools to scaffold a highly contiguous assembly of the *S. lycopersicum* cultivar M82. We sequenced the genome with an Oxford Nanopore MinION sequencer to 58.8× fold coverage with an N50 read length of 13.4 kbp (max 1,256,650 bp). The genome was assembled with Canu [33] and was comprised of 1709 contigs with a contig N50 of 1,458,445 bp. To compare RaGOO to other reference-guided tools, the assembly was scaffolded with RaGOO (with chimeric contig correction), MUMmer's "show-tiling" utility, and Chromosomer. Here, a "localized" contig is one that is placed in a pseudomolecule group and is assigned order and orientation. In all cases, the Heinz SL3.0 genome was used as the reference. RaGOO localized the highest portion of sequence, placing 99.01% of sequence into chromosomes compared to 85.6% and 3.17% for Chromosomer and show-tiling, respectively. The resulting RaGOO assembly contained 12 chromosome-length pseudomolecules with only 0.99% of sequence in the ambiguous chromosome 0 (**Figure 1.3**). Additionally, the scaffolding completed in only ~ 3 min for RaGOO, compared to ~ 285 min for show-tiling and ~ 1466 min for Chromosomer.

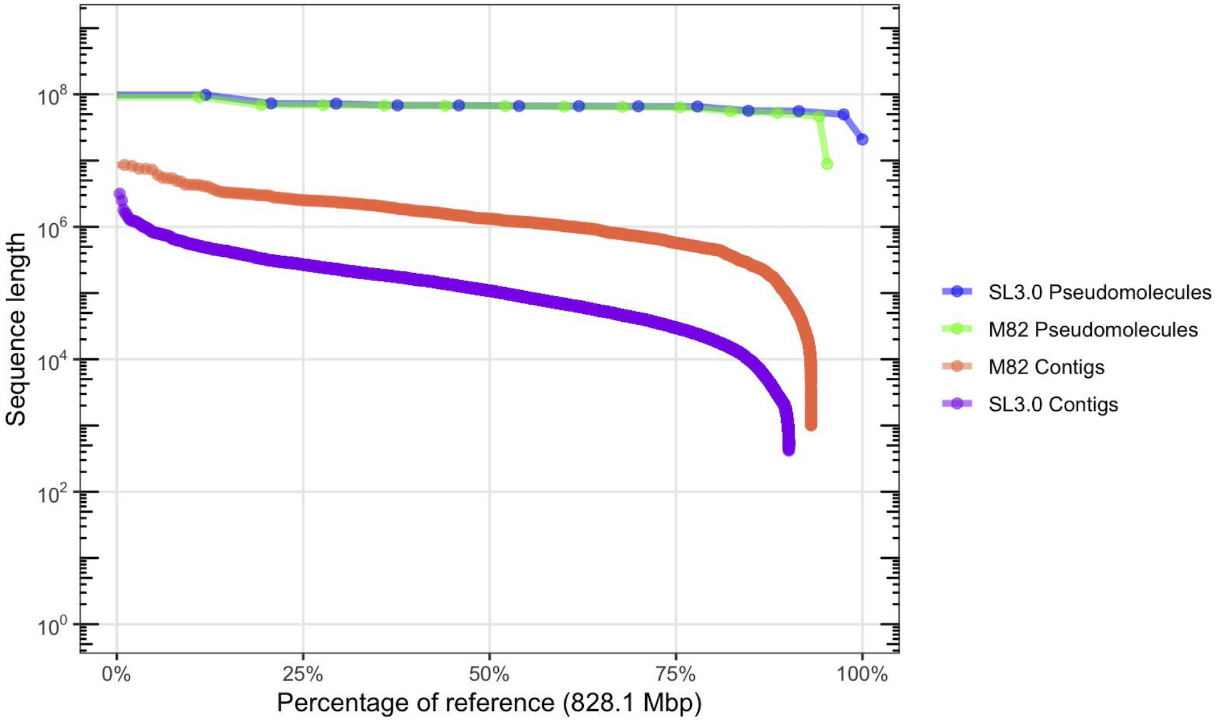


Figure 1.3: M82 assembly contiguity. "Nchart" of the M82 and Heinz contigs and pseudomolecules. M82 pseudomolecules were established by ordering and orienting M82 contigs with RaGOO. Heinz contigs were derived from the SL3.0 pseudomolecules by splitting sequences at stretches of 20 or more contiguous "N" characters.

To compare RaGOO scaffolding to a widely used reference-free approach, we generated Hi-C chromatin conformation data and used SALSA2 [13] to build scaffolds from the M82 contigs. Though SALSA2 does not necessarily build pseudomolecules, it strives to establish chromosome and chromosome-arm length scaffolds as the data allows. SALSA2 utilized Hi-C alignments to the M82 draft assembly along with the M82 Canu assembly graph. Though the scaffolds were highly contiguous compared to the input assembly (scaffold N50 of 18,282,950 bp), they are not chromosome scale.

We further compared the structural accuracy of the RaGOO pseudomolecules to that of the SALSA2 scaffolds by comparing the 12 pseudomolecules of the former and the 12 longest scaffolds of the latter to the Heinz SL3.0 reference (**Figure 1.4**). This shows nearly complete and highly co-linear coverage of the RaGOO pseudomolecules, while highly fragmented and rearranged placements of the SALSA2 scaffolds. Additionally, realigning the same Hi-C data to these pseudomolecules/scaffolds provides a reference-free assessment of the large-scale structural accuracy of these sequences.

Through this analysis, we found that the SALSA2 scaffolds contained many misassemblies, especially false inversions, while the RaGOO pseudomolecules contained very few structural errors (**Figure 1.4**). These Hi-C alignments suggest that most inversions and other large structural differences between the SALSA2 scaffolds and the Heinz reference assembly are likely not biological, but rather are scaffolding errors. They also demonstrate that erroneous reference bias in the RaGOO pseudomolecules, though present, was rare.

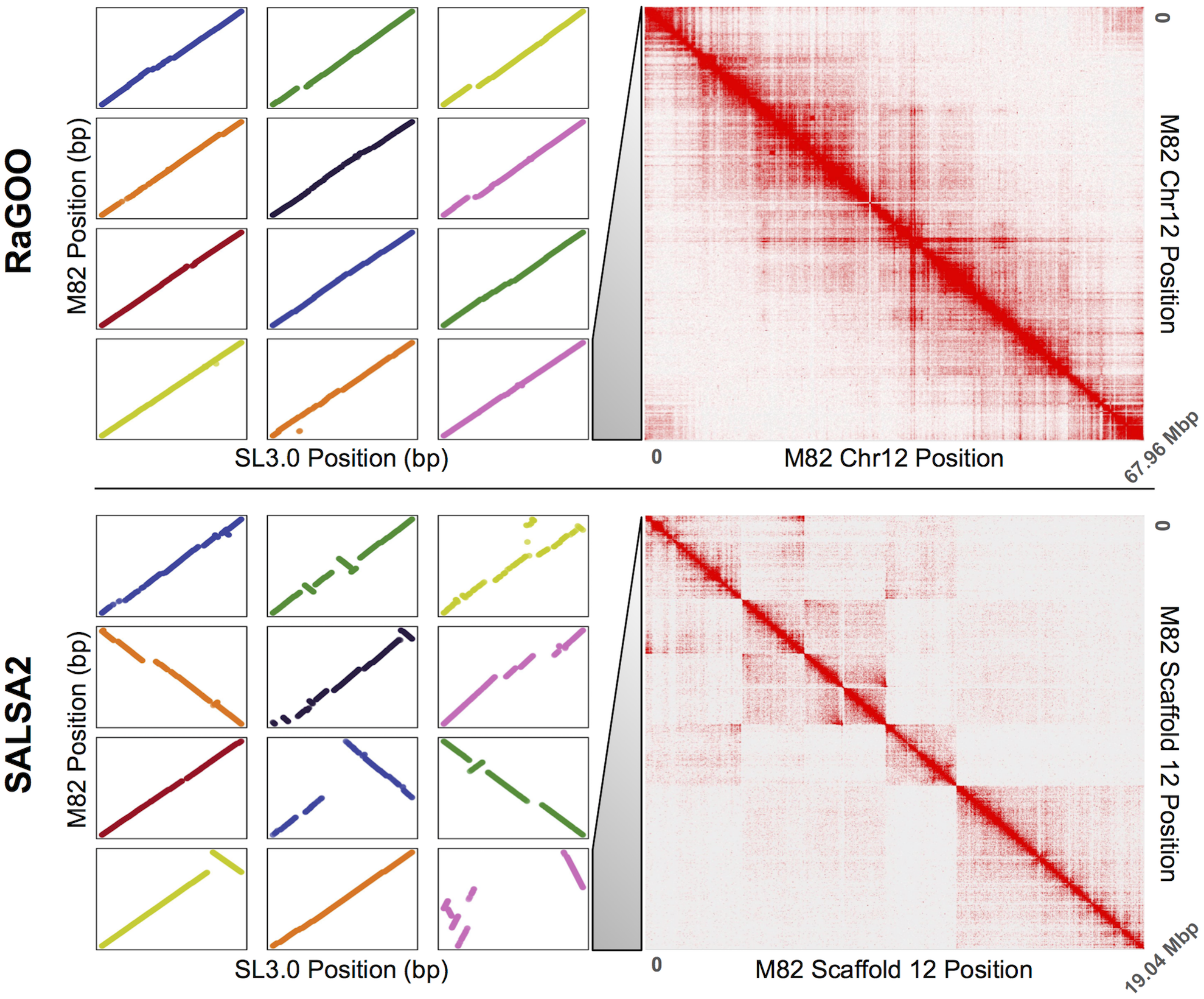


Figure 1.4: Reference-free vs. reference-guided scaffolding of M82. Both the top and bottom panels depict a dotplot (left) and Hi-C heatmap (right). The dotplots are generated from alignments to the Heinz reference assembly. On the top panel is the reference-guided RaGOO assembly dotplot, with chromosomes 1 through 12 depicted from top left to bottom right, and the Hi-C heatmap for chromosome 12. On the bottom is the de novo SALSA scaffolds dotplot, with the 12 largest scaffolds depicted in descending order of length from top left to bottom right and the Hi-C heatmap for the 12th largest scaffold.

M82 chromosome Hi-C validation, finishing, and annotation

To establish a new structurally accurate tomato reference genome, we sought to make further improvements to the RaGOO M82 pseudomolecules, as they provided the best completeness and contiguity with relatively few misassemblies. We first used the abovementioned Hi-C data and Juicebox Assembly Tools to correct apparent lingering misassemblies in the pseudomolecules [12]. A total of three corrections were made: an inversion error correction on chromosome 3 and an ordering error correction on chromosomes 7 and 11. Any "debris" contigs resulting from these alterations were placed in chromosome 0. With these few misassemblies corrected, the pseudomolecules were gap filled with PBJelly and polished with Pilon [34,35]. The final polished assembly had an average identity of 99.56% when compared to the Heinz SL3.0 reference and contained a complete single copy of 94.1% of BUSCO genes [36]. We note that M82 is biologically distinct from Heinz, so we do not expect 100% identity and estimate the overall identity at approximately 99.8 to 99.9%. Additionally, M82 consensus accuracy is reflected in ITAG 3.2 cDNA GMAP alignments, 96.8% of which align with at least 95% coverage and identity [37].

Gene finding and annotation was performed on the finished M82 assembly with the MAKER pipeline [38]. There are 35,957 genes annotated in the M82 assembly, of which 27,624 are protein coding. When comparing M82 and Heinz 1706 ITAG3.2 gene models using gffcompare (https://github.com/gpertea/gffcompare), we found 24,652 gene models with completely matching intron chains. The final M82 assembly contained a total of ~ 46 Mbp novel non-gapped sequence missing from the SL3.0 reference genome. Furthermore, the M82 assembly contained only ~ 8.9 Mbp of unlocalized sequence in chromosome 0 compared to ~ 17.8 Mbp in the Heinz SL3.0 reference.

Pan-SV analysis of 3 tomato accessions

In addition to the M82 cultivar, we also assembled genomes for the BGV and FLA tomato accessions de novo with Oxford Nanopore sequencing reads and the Canu assembler. We sequenced the BGV accession to 33.5× fold coverage with a read N50 length of 27,350 bp (max 192,728 bp) and the FLA accession to 41.6× fold coverage with a read N50 length of 24,225 bp (max 144,350 bp). The FLA assembly contained a total of 750,743,510 bp and had an N50 of 795,751 bp, while the BGV assembly contained a total of 769,694,915 bp and had an N50 of 4,105,177 bp. As with the M82 assembly, RaGOO was then used to establish pseudomolecules and call structural variants for these assemblies. The final FLA and BGV pseudomolecules contained 745,663,382 bp and 765,377,903 bp (99.3% and 99.4%) of the total ungapped sequence localized to chromosomes, respectively. Finally, the assemblies underwent gap filling, polishing, and gene finding using the same methods as M82. A summary of the final assembly statistics for all three accessions is presented in **Table 1.1**. The polished assemblies had 99.4% (FLA) and 98.9% (BGV) average identity compared to the Heinz SL3.0 reference as measured by MUMmer's "dnadiff." These assemblies also demonstrated genome completeness with BGV and FLA containing a single copy of 94.8% and 94.9% of BUSCO genes, respectively.

Accession	Chromosome span (bp)	Chromosome N50 (bp)	Chr0 bases (bp)	Number Contigs	Contig span (bp)	Contig N50 (bp)	Number SVs
Heinz	828,076,956	66,723,567	20,852,292	22,705	746,357,581	133,084	NA
M82	792,934,937	67,021,692	8,891,603	2910	771,143,786	1,458,445	36,191
BGV	794,568,563	67,174,401	4,643,553	638	769,694,915	4,105,177	45,927
FLA	796,004,315	67,650,907	5,490,904	2577	750,743,510	795,751	45,478

Table 1.1: Summary statistics of the reference tomato genome as well as the three novel accessions. Chromosome span indicates the total span of all of the chromosomes, including gaps. Chromosome N50 is the length such that half of the total span is covered in chromosome sequences this length or longer. Chr0 bases report the number of bases assigned to the unresolved chromosome 0. Contig span is the total length of non-gap (N) characters. Contig N50 is the length such that half of the contig span is covered by contigs this length or longer. Number SVs reports the number of SVs reported by RaGOO using the integrated version of Assemblytics.

Together with the M82 genome, we present 3 chromosome-scale assemblies with substantially more sequence content and fewer gaps than the Heinz SL3.0 reference genome. Given the structural variants output by RaGOO, we next used SURVIVOR to determine which variants were shared among these three accessions (**Figure 1.5**). As expected, the most divergent accession, BGV, demonstrated the most structural variant diversity with a total of 45,927 SVs compared to 45,478 and 36,191 SVs in FLA and M82, respectively. The union of these sets of variants yielded 98,988 total structural variants, which overlapped with 19,790 out of 35,768 total ITAG 3.2 genes (with 2 kbp flanking upstream and downstream each gene included). The most variable gene (the gene with the most intersecting SVs), Solyc03g095810.3, is annotated as a member of the GDSL/SGNH-like Acyl-Esterase family, while the second most variable gene, Solyc03g036460.2, is annotated as a member of the E3 ubiquitin-protein ligase. These three chromosome-scale assemblies, along with their associated sets of SVs, establish valuable genomic resources for the Solanaceae scientific community.

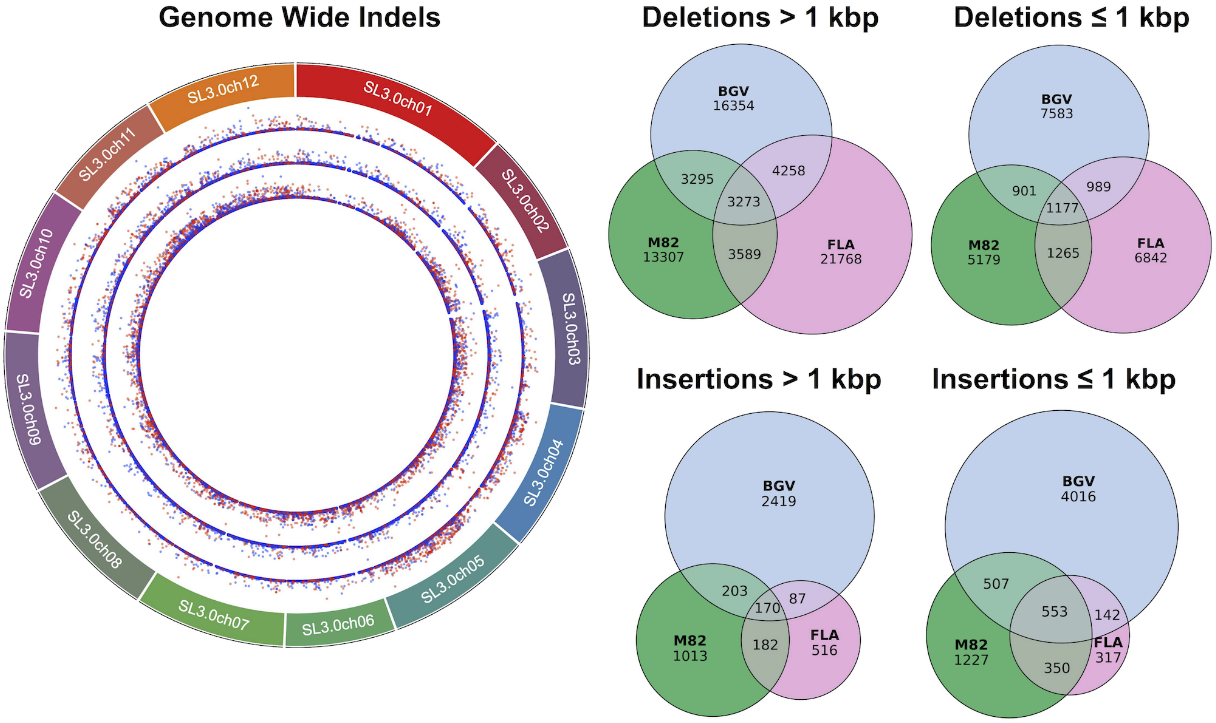


Figure 1.5: The tomato pan-genome. (left) Circos plot (http://omgenomics.com/circa/) depicting the size and type of structural variant. From the outer ring to the inner ring: M82, FLA, and BGV. Point height (y-axis) is scaled by the size of the variant, with red indicating insertions and blue indicating deletions. (right) Euler diagrams (https://github.com/jolars/eulerr) depicting the insertions and deletions shared among the three accessions.

Scaffolding a divergent S. pennellii genome assembly

Reference-guided scaffolding accuracy depends on a shared chromosomal structure between the draft and reference assemblies. This is the case for our three tomato assemblies since they represent either the same species as the reference (*S. lycopersicum*) or a closely related progenitor species (*S. pimpinellifolium*). However, we sought to evaluate the scaffolding success of a more divergent *S. pennellii* draft assembly to assess scenarios where assemblies are not close relatives. To this end, we scaffolded a draft *S. pennellii* genome assembly twice using two distinct reference genomes [39]. First, we scaffolded contigs according to the same *S. lycopersicum* SL3.0 reference genome used thus far in our previous tomato analysis. In addition, we also scaffolded contigs according to an independent, chromosome-scale *S. pennellii* reference genome [40].

If the distantly related *S. lycopersicum* reference is suitable for scaffolding the *S. pennellii* contigs, then the two resulting sets of RaGOO pseudomolecules should be structurally similar. Rather, we found major structural disagreements between the two sets of RaGOO pseudomolecules. Notably, chromosome 0 contained over four times as many bases when using the *S. lycopersicum* reference (26,868,206 bp vs. 6,230,859 bp) indicating that significantly less of the genome had been localized. We further noted the confidence score distributions were appreciably lower when using the *S. lycopersicum* reference. From these results, we conclude that *S. lycopersicum* is too divergent from *S. pennellii* to be used as a guide for scaffolding. Though every case must be examined individually, this analysis shows how confidence scores and localization stats can be used to determine if reference-guided scaffolding is appropriate for divergent assemblies.

Pan-SV analysis of 103 Arabidopsis thaliana genomes

Given the speed of RaGOO, we sought to test its scalability by performing a pan-SV genome analysis on a large population of diverse individuals. To acquire such population-scale data, we examined the sequencing data from the 1001 Genomes Project database, which includes raw short-read sequencing data and small variant calls for 1135 *Arabidopsis thaliana* accessions [41]. We mined the 1001 Genomes Project database for sequencing data amenable to genome assembly with sufficiently deep coverage of paired-end reads. This identified 103 short-read datasets representing a wide range of accessions sampled across 4 continents (**Figure 1.6**). We then established draft de novo assemblies for each accession using SPAdes [42]. Finally, RaGOO utilized the TAIR 10 reference genome to create 103 chromosome-scale assemblies and associated SV calls [43]. Between 85.8 and 98.7% (mean = 96.7%) of sequence was localized into chromosomes per accession, showing that the majority of assembled sequence across the pan-genome was scaffolded into pseudomolecules, even for more divergent accessions. The structural variant calls from this pan-genome provide a database of *A. thaliana* genetic variation previously unreported in the initial 1001 Genomes Project analysis [44].

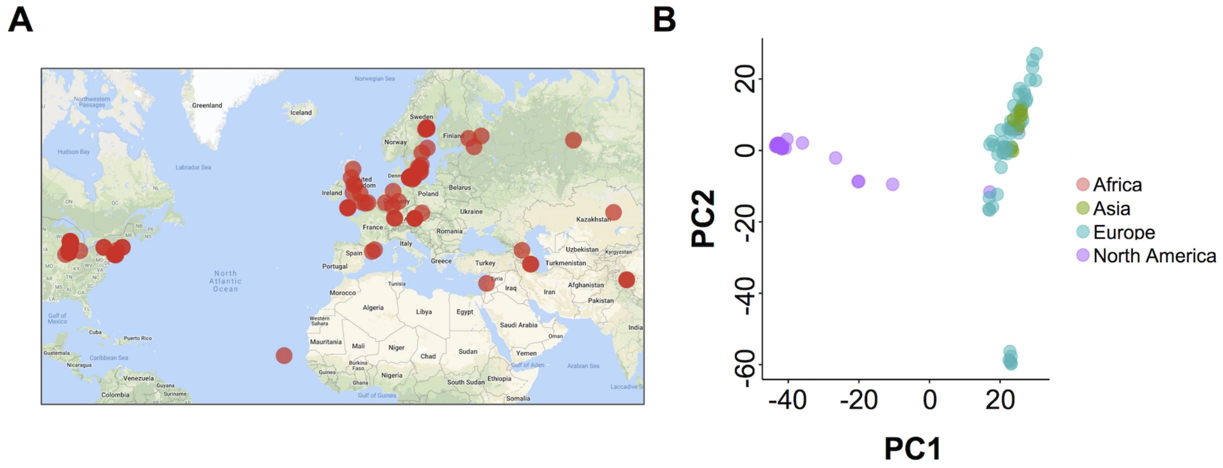


Figure 1.6: The Arabidopsis pan-genome. (A) Map of the 103 Arabidopsis accessions that were assembled in this study. (B) Principal components analysis of the structural variant presence/absence matrix of the 103 Arabidopsis accessions.

SV calls were compared with SURVIVOR, yielding a total of 137,111 merged variants across the pan-genome. From this merged set of variants, we constructed a presence/absence matrix representing which variants were present in which accessions. Principal components analysis of this matrix revealed clustering of accessions according to their geographic location (**Figure 1.6**). Upon further analysis of global trends in the data, we found that SVs were concentrated in pericentromeric regions, consistent with previous findings [45].

We further examined those genes that intersected variants present in small and large numbers of accessions, as these represent rare variants in the population and rare variants in the reference genome, respectively. When including variants present in at least 1, 10, 50, and 100 samples, we found 26,795, 17,593, 7859, and 332 total intersecting protein-coding genes (2 kbp flanking each side), respectively. Since there are a total of 27,416 protein-coding genes in the TAIR 10 database, we conclude that SVs in the pan-genome impact the genomic architecture for the majority of protein-coding genes, though fewer genes are affected by variants present in multiple samples (**Table 1.2**). Interestingly, most of these highly variable genes are defense response genes. Ultimately, our analysis highlights the importance of chromosome-level assembly at a population scale to help understand the broad impact of structural variation.

Gene	Annotation	Number of variants	Normalized number of variants	Number of accessions with variants
AT4G16960	Defense response, chloroplast	62	0.00715605	80
AT1G58602	ADP binding, defense response, ATP binding	57	0.00244101	90
AT3G44400	ADP binding, defense response, cytoplasm, signal transduction	56	0.00621256	89
AT3G44630	Defense response	55	0.00593312	84
AT4G16920	Defense response, chloroplast, cytoplasm	55	0.00522913	79
AT1G62620	*N*,*N*-dimethylaniline monooxygenase activity, flavin adenine dinucleotide binding, NADP binding, monooxygenase activity, nucleus, oxidation-reduction process	54	0.00850796	91
AT4G16950	Defense response to fungus, incompatible interaction, nucleotide binding, defense response, protein binding	54	0.00558486	70
AT1G62630	Defense response, ATP binding, N-terminal protein myristoylation, ADP binding, nucleus	50	0.00748391	93
AT5G41740	Nucleus, defense response, chloroplast	48	0.00565171	91
AT4G16890	Defense response, cytosol, signal transduction, defense response to bacterium, protein binding, ATP binding, defense response to bacterium, incompatible interaction, ADP binding, systemic acquired resistance, salicylic acid-mediated signaling pathway, cytoplasm, intracellular membrane-bounded organelle, nucleus, nucleotide binding, endoplasmic reticulum, response to auxin	48	0.00536373	75

Table 1.2: Summary of the ten most variable genes in the Arabidopsis pan-genome. "Number of variants" is the total number of variants intersecting a given gene, and "Normalized number of variants" is the number of intersecting variants divided by gene length.

1.4 DISCUSSION

We have introduced RaGOO in both a general and focused context for highly accurate genome scaffolding. As a general method, RaGOO may be valuable for chromosome-scale scaffolding in experimental designs where ordering and/or orienting of contigs leveraging an existing reference is available. Ordering and orienting with RaGOO may also facilitate analysis not possible with unlocalized contigs. This is exemplified by the additional sequence found through gap-filling of the M82, BGV, and FLA assemblies or by the identification of structural variants spanning gaps between contigs in the *S.*

lycopersicum and *Arabidopsis thaliana* pan-genomes. Additionally, our pan-genome analysis demonstrates that the speed of RaGOO offers new possibilities as to the scope and size of experiments that require reference-guided scaffolding. Furthermore, the integrated structural variant identification pipeline allows for a rapid survey of gene-related and other variants in the population. This shows that for both tomato and Arabidopsis pan-genomes, the majority of protein-coding genes are associated with the structural variation, highlighting the importance of population-scale assembly and structural variant discovery.

In a more focused analysis, we demonstrate that RaGOO may be a valuable component of a detailed assembly pipeline to establish new high-quality eukaryotic genomic resources. Our use of RaGOO to produce three tomato assemblies highlights a valuable means of organizing contiguous draft assemblies into pseudomolecules. This is especially useful as draft assemblies become more contiguous, and high-quality references become more common, even for non-model species.

For applications that do not have independent data such as Hi-C to validate the accuracy of RaGOO output, it can be challenging to assess the extent to which errors such as reference bias are present in pseudomolecules. However, it is possible to estimate the fidelity of newly created pseudomolecules to the reference. As we show in our *S. pennellii* analysis, the percentage of localized contigs/sequence along with the RaGOO confidence scores can be examined to help determine if scaffolding was successful. In general, if pseudomolecules pass these quality control checks, users can be more confident that RaGOO pseudomolecules are accurate and complete.

1.5 CONCLUSIONS

Our results show that RaGOO is a fast and accurate method for organizing genome assembly contigs into pseudomolecules. They also show that with a closely related reference genome, reference-guided scaffolding may yield substantially better scaffolding results than popular reference-free methods such as scaffolding with Hi-C data. In the process, we produced three tomato genome assemblies that are a valuable resource for the Solanaceae community and were selected to serve as the foundation for many additional tomato accessions we will be sequencing to establish a pan-SV genome for use in biology and agriculture. For this purpose, the M82 assembly has already undergone extensive procedures to provide a complete and accurate assembly with an associated set of gene models and annotations.

1.6 METHODS

Description of RaGOO algorithm and scoring metrics

The RaGOO source code is available on GitHub at https://github.com/malonge/RaGOO and is released under an MIT license. RaGOO is written in Python3 and uses the python packages intervaltree and numpy. It also relies on Minimap2 that is available on GitHub at https://github.com/lh3/minimap2. RaGOO also comes bundled with an integrated implementation of Assemblytics for structural variation analysis.

Scaffolding algorithm overview

RaGOO utilizes alignments to a reference genome to cluster, order, and orient contigs to form pseudomolecules. RaGOO internally invokes Minimap2, with k-mer size and window size both set to 19 bp, to obtain the necessary mappings of contigs to a reference genome. By default, any alignments less than 1 kbp in length are removed. To cluster contigs into chromosome groups, each contig is assigned to the reference chromosome which it covers the most. Coverage here is defined as the total number of reference chromosome base pairs covered in at least one alignment. Next, for each pseudomolecule group, the contigs in that group are ordered and oriented relative to each other. To do this, the longest (primary) alignment for each contig to its assigned reference chromosome is examined. Ordering is achieved by sorting these primary alignments by the start then end alignment position in the reference. Finally, the orientation of that contig is assigned the orientation of its primary alignment. To produce pseudomolecules, ordered and oriented contigs are concatenated, with padding of "N" characters placed between contigs.

Scaffolding confidence scores

Each contig is assigned a confidence score, between 0 and 1, for each of the three stages outlined above. The clustering confidence score is the number of base pairs a contig covered in its assigned reference chromosome divided by the total number of covered base pairs in the entire reference genome. To create a metric associated with contig ordering confidence, we defined a location confidence. First, the smallest and largest alignment positions, with respect to the reference, between a contig and its assigned reference chromosome are found. The location confidence is then calculated as the number of covered base pairs in this range divided by the total number of base pairs in the range. Finally, to calculate the orientation confidence, each base pair in each alignment between a contig and its assigned reference chromosome casts a vote for the orientation of its alignment. The orientation confidence is the number of votes for the assigned orientation of the contig divided by the total number of votes.

Chimeric contig correction

Before clustering, ordering, and orienting, RaGOO provides the option to break contigs that may be chimeric as indicated by discordant alignments to the reference. RaGOO can identify and correct both interchromosomal and intrachromosomal chimeric contigs. Interchromosomal chimeric contigs are contigs that have significant alignments to two distinct reference chromosomes. To identify and break such contigs, all the alignments for a contig are considered. Alignments less than 10 kbp are removed, and the remaining alignments are unique anchor filtered [24]. If there are multiple instances where at least 5% of the total alignment lengths cover at least 100 kbp of a distinct reference chromosome, a contig is deemed chimeric. To break the contig, alignments are sorted with respect to the contig start, then end positions, and the contig is broken where the sorted alignments transition between reference chromosomes.

Intrachromosomal chimeric contigs are contigs that have significant alignments to distant loci on the same reference chromosome. As with interchromosomal chimeric contigs, identification and breaking of intrachromosomal chimeric contigs start with removing short and non-unique alignments. The remaining alignments are sorted with respect to the start then end position in the reference chromosome. Next, the genomic distance between consecutive alignments is calculated, both with respect to the reference and the contig. If any of these distances exceed user-defined thresholds, the contig is broken between the corresponding alignments. Only one intrachromosomal and one interchromosomal break can occur per contig per execution of the software. Importantly, all of the above criteria for breaking contigs are tunable parameters in the RaGOO software. This allows users to specify how large a structural difference between the assembly and the reference must be to consider it an error. Chimeric contig correction should only be used in cases when the user is confident that such large structural differences between the assembly and the reference are more likely to be misassemblies than true, large-scale structural variants. We advise users to validate misassembly correction with independent data to help ensure that true variation is not being masked.

Scaffolding of an Arabidopsis thaliana and human genome

Of our 103 *A. thaliana* assemblies, we highlighted the runtime and scaffolding accuracy of the assembly representing the TFÄ 04 accession (SRR1945711). This assembly was assembled with SPAdes (see below) and had a scaffold N50 of 120,255 bp with a total size of 115,803,138 bp [42]. Additionally, to demonstrate the scaffolding of a mammalian-sized genome, we used RaGOO to order and orient the mixed haplotype human Canu assembly derived from Pacific Biosciences CCS reads. This human assembly had a contig N50 of 22,778,121 bp and a total size of 3,418,171,375 bp. For both the TFÄ 04 and human assemblies, default RaGOO parameters were used and the software was run with 8 threads ("-t 8"). The TAIR 10 and hs37d5 reference genomes were used to scaffold the TFÄ 04 and human assemblies, respectively. RaGOO completed in 12.576 s and 12 min and 33.090 s for TFÄ 04 and human, respectively. The dotplots for both assemblies were made by aligning RaGOO pseudomolecules to the respective reference genomes with nucmer (-l 200 -c 500). Alignments were filtered with delta-filter (-1 -l 20000), and plots were made with Mummerplot (--fat). Only nuclear chromosome and non-alternate sequences are shown in the dotplots.

Simulated reference-guided scaffolding

A simulated *S. lycopersicum* draft genome assembly was created by partitioning the Heinz SL3.0 reference genome, excluding chromosome 0, into scaffolds of variable length. Intervals along each chromosome were successively defined, with each interval length being randomly drawn from the distribution of observed M82 Canu contig lengths. Bedtools [46] was then used to retrieve the sequence associated with these intervals. Finally, simulated scaffolds with more than 50% "N" characters were removed, and half of the remaining contigs were randomly reverse complemented. A second simulated assembly containing contigs, rather than scaffolds, was derived from these simulated scaffolds. Scaffolds were broken at any stretch of "N" characters longer than or equal to 20 bp, excluding the gap sequence. Any resulting contigs less than 10 kbp in length were also excluded. We call this pair of simulated assemblies the "easy" set of simulated data. To simulate a "hard" set of data, we started with the same "easy" scaffolds and added variation. To do this, we used SURVIVOR to simulate 10,000 indels ranging from 20 bp to 10 kbp in size. We also added SNPs at a rate of 1%. Again, we split these scaffolds into contigs resulting in a pair of "hard" simulated assemblies.

Given these "easy" and "hard" simulated scaffolds and contigs, RaGOO, Chromosomer, and MUMmer's "show-tiling" utility were used for reference-guided scaffolding. For RaGOO, chimera breaking was turned off, and default parameters were used except for the padding amount, which was set to zero. Chromosomer utilized Blast alignments with default parameters. Additionally, the "fragmentmap ratio" was set to 1.05, and the padding amount was set to zero. Show-tiling used default parameters. Since RaGOO and Chromosomer rely on aligners that allow for multithreading, both tools were run with eight threads, while show-tiling was run with a single thread.

We recorded various measurements to evaluate the success of these tools in ordering and orienting simulated assemblies. Firstly, we observed the runtime, percentage of localized contigs, and percentage of localized sequence. To assess the clustering and orienting accuracy, we measure the percentage of localized contigs that had been assigned the correct cluster and orientation, respectively. Finally, we used two measurements to assess the ordering accuracy of each pseudomolecule. The first was the edit distance between the true and predicted order of contigs. This edit distance was normalized by dividing by the total number of contigs in the true ordering. The second ordering accuracy measurement was the percentage of correct adjacent contig pairs.

Plant material and growth conditions

Seeds of the *S. lycopersicum* cultivar M82 (LA3475) were from our stocks. Seeds of the *S. pimpinellifolium* accession BGV006775 were provided by E. van der Knaap, University of Georgia. Seeds of the *S. lycopersicum* breeding line Fla.8924 were from the stocks of S. Hutton, University of Florida. Seeds were directly sown and germinated in the soil in 96-cell plastic flats and grown under long-day conditions (16-h light/8-h dark) for 21 days in a greenhouse under natural light supplemented with artificial light from high-pressure sodium bulbs (~ 250 µmol m2 s1). Daytime and nighttime temperatures were 26–28 °C and 18–20 °C, respectively, with a relative humidity of 40–60%.

Genome and transcriptome sequences

Genomic Illumina read data for BGV006775 were downloaded from the NCBI Sequence Read Archive (SRA) database (accession SRS3394566). Genomic Illumina read data for Fla.8924 [32] was provided by S. Hutton, University of Florida. Illumina read data for all transcriptomes were downloaded from ftp://ftp.solgenomics.net/user_requests/LippmanZ/public_releases/by_experiment/Park_etal/ [SeSo1] ftp://ftp.solgenomics.net/transcript_sequences/by_species/Solanum_lycopersicum/libraries/illumina/LippmanZ/; [SeSo2] http://solgenomics.net/[SeSo3]. [SeSo4] [ZBL5].

Tissue collection and high molecular weight DNA extraction

For extraction of high molecular weight DNA, young leaves were collected from 21-day-old light-grown seedlings. Before tissue collection, seedlings were incubated in complete darkness for 48 h. Flash-frozen plant tissue was ground using a mortar and pestle and extracted in five volumes of ice-cold extraction buffer 1 (0.4 M sucrose, 10 mM Tris-HCl pH 8, 10 mM MgCl2, and 5 mM 2-mercaptoethanol). Extracts were briefly vortexed, incubated on ice for 15 min, and filtered twice through a single layer of Miracloth (Millipore Sigma). Filtrates were centrifuged at 4000 rpm for 20 min at 4 °C, and pellets were gently re-suspended in 1 ml of extraction buffer 2 (0.25 M sucrose, 10 mM Tris-HCl pH 8, 10 mM MgCl2, 1% Triton X-100, and 5 mM 2-mercaptoetanol). Crude nuclear pellets were collected by centrifugation at 12,000g for 10 min at 4 °C and washed by re-suspension in 1 ml of extraction buffer 2 followed by centrifugation at 12,000g for 10 min at 4 °C. Nuclear pellets were re-suspended in 500 µl of extraction buffer 3 (1.7 M sucrose, 10 mM Tris-HCl pH 8, 0.15% Triton X-100, 2 mM MgCl2, and 5 mM 2-mercaptoethanol), layered over 500 µl extraction buffer 3, and centrifuged for 30 min at 16,000g at 4 °C. The nuclei were re-suspended in 2.5 ml of nuclei lysis buffer (0.2 M Tris pH 7.5, 2 M NaCl, 50 mM EDTA, and 55 mM CTAB) and 1 ml of 5% Sarkosyl solution and incubated at 60 °C for 30 min. To extract DNA, nuclear extracts were gently mixed with 8.5 ml of chloroform/isoamyl alcohol solution (24:1) and slowly rotated for 15 min. After centrifugation at 4000 rpm for 20 min, ~ 3 ml of aqueous phase was transferred to new tubes and mixed with 300 µl of 3 M NaOAC and 6.6 ml of ice-cold ethanol. Precipitated DNA strands were transferred to new 1.5 ml tubes and washed twice with ice-cold 80% ethanol. Dried DNA strands were dissolved in 100 µl of elution buffer (10 mM Tris-HCl, pH 8.5) overnight at 4 °C. Quality, quantity, and molecular size of DNA samples were assessed using Nanodrop (Thermofisher), Qbit (Thermofisher), and pulsed-field gel electrophoresis (CHEF Mapper XA System, Biorad) according to the manufacturer's instructions.

Nanopore library preparation and sequencing

DNA was sheared to 30 kb using the Megarupter or 20 kb using Covaris g-tubes. DNA repair and end-prep were performed using New England Biosciences kits NEBNext FFPE DNA Repair Kit and Ultra II End-Prep Kit. DNA was purified with a 1× AMPure XP bead cleanup. Next, DNA ligation was performed with NEBNext Quick T4 DNA Ligase, followed by another AMPure XP bead cleanup. DNA was re-suspended in elution buffer and sequenced according to the MinION standard protocol.

10× Genomics library preparation and sequencing

1.12 ng of high molecular weight gDNA was used as input to the 10× Genomics Chromium Genome kit v2 and libraries we prepared according to the manufacturer's instructions. The final libraries, after shearing and adapter ligation, had an average fragment size of 626 bp and were sequenced on an Illumina HiSeq, 2500 2 × 250 bp.

Hi-C library preparation and sequencing

DNA extraction, library construction, and sequencing for Hi-C analyses were performed by Phase Genomics (Seattle, WA) and conducted according to the supplier's protocols. Young leaves from 21-day-old light-grown and 48-h dark-incubated seedlings were wrapped in wet tissue paper and shipped on ice overnight.

Initial de novo assembly of tomato genomes

The Oxford Nanopore sequencing data for M82, BGV, and FLA were assembled with Canu. For all three assemblies, default parameters were used with the expected genome size set to 950 Mbp. Assemblies were submitted to the UGE cluster at Cold Spring Harbor Laboratory for parallel computing. After assembly, it was determined that the M82 assembly contained bacterial contamination. To remove bacterial contigs from the assembly, the Canu contigs were aligned to all RefSeq bacterial genomes (downloaded on June 7, 2018) as well as the Heinz SL3.0 reference genome. If a contig covered more RefSeq bacterial genome base pairs than SL3.0 base pairs, it was deemed a contaminant and removed from the assembly. In this paper, "M82 Canu contigs" refers to the Canu contigs after contaminant contigs had been removed.

Reference-guided and reference-free scaffolding of tomato genomes

The M82 Canu contigs were ordered and oriented into pseudomolecules with RaGOO, Chromosomer, and Nucmer's "show-tiling" utility. The Heinz SL3.0 reference, with chromosome 0 removed, was used for all tools. RaGOO used eight threads with chimeric contig correction turned on and the gap padding size set to 200 bp. We also instructed RaGOO to skip three contigs that had low grouping accuracy scores. Chromosomer used eight threads for BLAST alignments. The Chromosomer fragmentmap ratio was set to 1.05, and the gap padding size was set to 200 bp. Default parameters were used for show-tiling.

For reference-free scaffolding of the M82 assembly, 46,239,525,282 bp (~ 60× coverage of the M82 Canu contigs) of 2 × 101 Hi-C sequencing reads were aligned to the M82 Canu contigs with BWA mem using the "-5" flag [47]. Aligned reads were then filtered with "samtools view" to include alignments where both mates of a pair aligned as primary, non-supplementary alignments (-F 2316) [48]. SALSA2 then utilized these alignments along with the M82 Canu assembly graph to build scaffolds. The SALSA2 "-m" flag was also set to "yes" in order to correct misassemblies in the M82 contigs, and the expected genome size was set to 800 Mbp. Finally, we set "-e GATC" to correspond to the use of Sau3AI in the Hi-C library. The SALSA2 scaffolds were comprised of 2065 scaffolds and had an N50 of 18,282,950 bp and a total size of 827,545,698 bp.

The structural accuracy of the M82 RaGOO pseudomolecules and SALSA2 scaffolds was assessed with dotplots and Hi-C density plots. For dotplots, both sequences were aligned to the Heinz SL3.0 reference (with chromosome 0 removed) with Minimap2 using the "-ax asm5" parameter. Alignments less than 12 kbp in length were excluded. For Hi-C visualization, the same Hi-C data described earlier was aligned to both sequences using the same parameters as were used for SALSA2. These alignments were then visualized with Juicebox [49]. Hi-C mates that mapped to the same restriction fragment were excluded from visualization.

Using the same parameters as M82, RaGOO was also used to order and orient the FLA and BGV Canu assemblies. BGV underwent two rounds of chimeric contig correction. Assemblytics structural variants for each assembly were compared with "SURVIVOR merge," with the "max distance between breakpoints" set to 1 kbp. Variants in chromosome 0 of the SL3.0 reference as well as variants that spanned more than 10% gaps were excluded from the structural variant analysis.

Tomato genome correction and polishing

M82 RaGOO pseudomolecules were manually corrected for misassemblies and/or reference bias. Manual corrections were identified by visualizing Hi-C alignments to the M82 genome described in the previous sections. Firstly, three contigs with spurious alignments were removed from the pseudomolecules. Then, using Juicebox Assembly Tools, an inversion error was corrected on chromosome 3 and two ordering errors were corrected, one on chromosome 7 and one on chromosome 11. Gap filling and polishing were performed on the RaGOO pseudomolecules for the M82, FLA, and BGV tomato accessions. For each assembly, all respective Oxford Nanopore sequencing data used for assembly was used for gap filling with PBJelly.

After gap filling, we sought to find the most effective genome polishing strategy given our data. We used the gap-filled M82 assembly as a starting point for our tests. To polish this genome, we utilized the raw Oxford Nanopore data used for assembly as well as 10× Genomics Illumina Whole Genome Shotgun sequencing reads. We trimmed adapters and primers

(23 bp from the beginning of read 1) and low-quality bases (40 bp from the ends of read 1 and read 2) from these 10× genomics data. With these data, we compared multiple polishing strategies using various alignment and polishing tools. First, we examined assemblies polished with or without Nanopolish [50]. For Nanopolish, the M82 raw Oxford Nanopore read set was aligned to the M82 assembly with Minimap2 using the "map-ont" parameter. Next, we compared assemblies polished with 1 or 2 rounds of Pilon polishing. For each round of polishing, the Illumina data was randomly subsampled to 40× coverage prior to alignment. Finally, we compared bwa mem, Bowtie2, and ngm for short-read alignment prior to Pilon polishing [51,52]. We used bwa mem and ngm with default parameters, while Bowtie2 was run with the "--local" parameter.

We used MUMmer's "dnadiff" utility to compare the efficacy of these polishing pipelines. For dnadiff analysis, polished assemblies and the SL3.0 reference were broken into contigs by breaking sequences at gaps of 20 bp or longer. Then, assemblies were aligned to the reference contigs with nucmer using the "-l 100 -c 500 –maxmatch" parameters. After determining that 2 rounds of Pilon polishing with Bowtie2 yielded the best results, we applied the same pipeline to the BGV and FLA assemblies using ~ 23× coverage and ~ 26× coverage of paired-end Illumina short-read data was used for BGV and FLA, respectively. BUSCO was used to evaluate genome completeness of the polished M82, BGV, and FLA assemblies. The Solanaceae odb10 database was used with the "species" parameter set to "tomato."

Finally, we searched for spurious duplications introduced after gap-filling with PBJelly, since others have reported such phenomena [53]. We first examined the M82, BGV, and FLA assemblies after gap-filling but before polishing. Using these assemblies, we called structural variants with respect to the SL3.0 reference genome using Assemblytics (unique minimum alignment length set to 10 kbp). We then found all "tandem expansions" (duplications) that intersected gaps filled by PBJelly. Finally, for any intersecting tandem expansions, we calculated the average raw ONT read coverage across the variant. For FLA and BGV, all tandem expansions in filled gaps had ample read support (> 15×). For M82, there were two tandem expansions that had less than 1× coverage. Since one variant was only 7 bp long with respect to the M82 assembly, we omitted it from this analysis. The remaining spurious tandem expansion extended 982 bp and was perfectly mapped to the final polished M82 assembly using Minimap2 to M821.3ch09: 21470172-21471154.

Tomato genome annotation

We annotated protein-coding genes in the M82, FLA, and BGV assembly using the Maker v3.0 pipeline on Jetstream by providing repeats, full-length cDNA sequences, and proteins from Heinz 1706 ITAG3.2 assembly [54]. Simple, low-complexity, and unclassified repeats were excluded from masking. We additionally provided Maker with an M82 reference transcriptome derived from 50 M82 RNA-seq libraries. RNA-seq reads were aligned to the M82 genome using STAR, a splice-aware aligner [55]. These alignments were used to assemble transcripts and establish a consensus transcriptome using StringTie and TACO, respectively [56,57]. We ran Maker using parameters est2genome set to 1, protein2genome set to 1 and keep_preds set to 1 to perform the gene annotation. Low consensus gene models with an AED score above 0.5 were filtered from the Maker-predicted gene models. We additionally removed gene models shorter than 62 bp following the cutoffs used for the ITAG3.2 annotation. Putative gene functions were assigned to the MAKER gene models via Interproscan protein signatures and blastp protein homology search [58]. blastp queried the UniProtKB/Swiss-Prot and Heinz 1706 ITAG3.2 protein databases, filtering out alignments with an e value greater than 1e−05 [59]. We further filtered out genes that did not have an associated gene function in either Interproscan, UniprotKB/Swiss-Prot, or ITAG3.2.

S. pennellii genome scaffolding

S. pennellii contigs were scaffolded with both the Heinz 1706 SL3.0 reference and the independent *S. pennellii* reference genome using default RaGOO parameters and excluding chromosome 0 from the reference chromosomes ("-e"). The two resulting sets of pseudomolecules were aligned to each other using Nucmer (-l 200 -c 500). The resulting alignments were filtered with delta-filter (-l 50000 -1) and plotted with mummerplot. The two reference genomes were also aligned to each other using Nucmer (-l 50 -c 100), and the resulting alignments were filtered with delta-filter (-l 10000 -1) and plotted with mummerplot.

Arabidopsis structural variant analysis

The 1001 Genomes Database was mined for accessions for which there was at least 50× coverage of paired-end sequencing data. We also required that the read length be at least 100 bp. For practical reasons, we excluded accessions with excessive coverage. For each of the remaining accessions, the fastq files were randomly subsampled to achieve exactly 50× coverage. Subsampled reads were then assembled with the SPAdes assembler, with k-mer size set to 33, 55, 77, and 99, and otherwise default parameters. These draft assemblies were then ordered and oriented with RaGOO using default parameters (no chimeric contig correction) and the TAIR 10 reference genome (GCA_000001735.1). RaGOO also provided structural variants, with the minimum variant size set to 20 bp. Of the chromosome-scale assemblies, a few assemblies with a genome size greater than 150 Mbp were removed due to putative sample contamination. After this filtering, assemblies and structural variant calls for 103 accessions remained.

Variants that were called in chromosome 0 or the chloroplast/mitochondrial chromosomes were discarded. Also, variants that had more than a 10% overlap with a gap were excluded. To find unique variants across multiple samples, SURVIVOR merge was used such that a variant only had to be present in at least 1 sample for it to be reported. Therefore, given all 103 samples, this yielded the union of all variants present in the pan-genome. To find shared variants across multiple samples, SURVIVOR merge was used such that a variant must have been present in all samples to be reported. This effectively provided the intersection of variants in the pan-genome. In all instances of using SURVIVOR merge, the "max distance between breakpoints" was set to 1 kbp. Also, the strand of the SV was taken into account, while distance based on the size of the variant was not estimated. Finally, the minimum variant size was set to 20 bp to be consistent with the RaGOO parameters. Bedtools was used to find variant/gene intersections.

1.7 ACKNOWLEDGEMENTS

We thank Esther van der Knaap and Sam Hutton for providing seed stocks and W. Richard McCombie for the helpful discussions. We also thank Steven Salzberg and Aleksey Zimin for their help with software features and improvements. This work was supported by the National Science Foundation (DBI-1350041 and IOS-1445025 to MCS; IOS-1732253 to MCS and ZBL) and the US National Institutes of Health (R01-HG006677 to MCS and UM1 HG008898 to FJS). We also thank the NSF XSEDE project for providing compute resources on JetStream for annotation (project MCB180087 to MCS).

1.8 REFERENCES

1. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19:329–46.

2. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36:338–45.

3. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, et al. High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. Nat Commun. 2018;9:541.

4. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, et al. ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 2015;16:3.

5. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546:524–7.

6. Venter JC, Smith HO, Hood L. A new strategy for genome sequencing. Nature. 1996;381:364–6.

7. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27:757–67.

8. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.

9. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31:1119–25.

10. Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. BMC Genomics. 2017;18:527.

11. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5.

12. Dudchenko O, Shamim MS, Batra S, Durand NC. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv [Internet]. biorxiv.org; 2018; Available from: https://www.biorxiv.org/content/10.1101/254797v1.abstract

13. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;15:e1007273.

14. Pop M, Kosack DS, Salzberg SL. Hierarchical scaffolding with Bambus. Genome Res. 2004;14:149–59.

15. Tamazian G, Dobrynin P, Krasheninnikova K, Komissarov A, Koepfli K-P, O'Brien SJ. Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences. Gigascience. 2016;5:38.

16. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

17. Palmieri N, Nolte V, Chen J, Schlötterer C. Genome assembly and annotation of a Drosophila simulans strain from Madagascar. Mol Ecol Resour. 2015;15:372–81.

18. Kim J, Larkin DM, Cai Q, Asan, Zhang Y, Ge R-L, et al. Reference-assisted chromosome assembly. Proc Natl Acad Sci U S A. 2013;110:1785–90.

19. Aganezov S, Alekseyev MA. Multi-genome Scaffold Co-assembly Based on the Analysis of Gene Orders and Genomic Repeats. Bioinformatics Research and Applications. Springer International Publishing; 2016. p. 237–49.

20. Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, et al. Chromosome assembly of large and complex genomes using multiple references. Genome Res. 2018;28:1720–32.

21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

22. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. Cactus: Algorithms for genome multiple sequence alignment. Genome Res. 2011;21:1512–28.

23. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

24. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016;32:3021–3.

25. Alonge M, De Coster W, Sam217pa, Schatz M. malonge/RaGOO: Zenodo integration [Internet]. 2019. Available from: https://zenodo.org/record/3384200

26. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.

27. Consortium TTG, The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485:635–41.

28. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. nature.com; 2017;8:14061.

29. 100 Tomato Genome Sequencing Consortium, Aflitos S, Schijlen E, de Jong H, de Ridder D, Smit S, et al. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant J. Wiley; 2014;80:136–48.

30. Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C, et al. Rewiring of the Fruit Metabolome in Tomato Breeding. Cell. 2018;172:249–61.e12.

31. Blanca J, Montero-Pau J, Sauvage C, Bauchet G, Illa E, Díez MJ, et al. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. BMC Genomics. 2015;16:257.

32. Lee TG, Shekasteband R, Menda N, Mueller LA, Hutton SF. Molecular markers to select for the j-2–mediated jointless pedicel in tomato. HortScience. American Society for Horticultural Science; 2018;53:153–8.

33. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. genome.cshlp.org; 2017;27:722–36.

34. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012;7:e47768.

35. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963.

36. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. academic.oup.com; 2015;31:3210–2.

37. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.

38. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. genome.cshlp.org; 2008;18:188–96.

39. Schmidt MH-W, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, et al. De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. Plant Cell. 2017;29:2336–48.

40. Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, et al. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat Genet. 2014;46:1034–8.

41. Weigel D, Mott R. The 1001 genomes project for Arabidopsis thaliana. Genome Biol. 2009;10:107.

42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.

43. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796–815.

44. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43:956–63.

45. Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell. 2016;166:492–505.

46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

47. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

49. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. 2016;3:99–101.

50. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12:733–5.

51. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

52. Sedlazeck FJ, Rescheneder P, von Haeseler A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 2013;29:2790–1.

53. Dutreux F, Da Silva C, d'Agata L, Couloux A, Gay EJ, Istace B, et al. De novo assembly and annotation of three Leptosphaeria genomes using Oxford Nanopore MinION sequencing. Sci Data. 2018;5:180235.

54. Stewart CA, Cockerill TM, Foster I, Hancock D, Merchant N, Skidmore E, et al. Jetstream: a self-provisioned, scalable

science and engineering cloud environment. Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure. New York, NY, USA: Association for Computing Machinery; 2015. p. 1–8.

55. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

56. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.

57. Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK. TACO produces robust multisample transcriptome assemblies from RNA-seq. Nat Methods. 2017;14:68–70.

58. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.

59. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–69.

… RagTag overcomes the scaffolding bottleneck by leveraging existing genome assemblies to improve new ones, or by collectively drawing from multiple genome maps to build consensus scaffolds.

2

Automating genome assembly scaffolding enables new references for customized plant models

The following sections are not yet published. This work is a collaboration with Zach Lippman's lab at Cold Spring Harbor Laboratory and Sebastian Soyk's lab at University of Lausanne, with contributions from the following authors:

Michael Alonge[1], Ludivine Lebeigle[2], Melanie Kirsche[1], Sergey Aganezov[1], Xingang Wang[3], Zachary B. Lippman[3,4], Michael C. Schatz[1,3,5], and Sebastian Soyk[2]

1. Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
2. Center for Integrative Genomics, University of Lausanne, Lausanne, CH-1015, Switzerland
3. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
4. Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
5. Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA

Z.B.L, M.C.S, and S.S. supervised manuscript preparation. S.S. supervised all Sweet-100 analysis. M.C.S. supervised all RagTag development and genome assembly analysis. M.A., Z.B.L, M.C.S., and S.S. wrote the abstract and edited all of the text. The main sections "Sweet-100 is a new experimental system for functional genomics in tomato" and "Efficient transformation and genome editing in Sweet-100" and the methods sections "Plant material, growth conditions, and phenotyping" and "CRISPR-Cas9 mutagenesis, plant transformation, and identification of mutant alleles" were done by L.L. and S.S and written by S.S. The methods section "Extraction of high-molecular weight DNA and sequencing" was done by the Genomic Technologies Facility at the University of Lausanne and written by S.S. Figures 2.1 and 2.3 were made by L.L. and S.S. The methods section "RagTag 'patch'" was done by M.A. with help from M.K. and written by M.A. The methods section "RagTag merge" was done by M.A. with help from S.A. and written by M.A. The methods section "M82 genome assembly" was done by M.A. with help from X.W. and written by M.A. All other sections and figures were done/made by M.A. and written by M.A, including the discussion.

2.1 ABSTRACT

Advancing crop genomics in the age of pan-genomes requires multiple genetic systems enabled by simplified generation of high-quality genome assemblies. Here, we introduce RagTag, a new method for rapid and accurate automation of genome assembly scaffolding. We use this tool to establish high-quality reference tomato genome assemblies for a widely used genetic system and a new compact, rapid-cycling genotype developed for comparative genomics and genome editing. These references facilitate efficient genome-scale genetic manipulation in tomato and establish strategies for expanding systems in other plant species.

2.2 MAIN

Sweet-100 is a new experimental system for functional genomics in tomato

Recent technological advances in genome sequencing and editing are enabling the deciphering and manipulating of crop genomes and traits with unprecedented accuracy and flexibility. Pan-genomes capture diverse alleles within crop species but studying their phenotypic consequences is limited by efficient functional genetic systems in relevant genotypes. Tomato has been a core crop system to study the dissection of genetic architectures that shape crop domestication and quantitative traits. Sequencing hundreds of tomato genomes has uncovered vast genomic variation [1,2] however, chromosome-scale genomes are only available for few accessions [3–5], and there is a historical discrepancy between the reference genome (Heinz 1706) and the genotypes that are commonly used by the community for genetic and molecular experimentation (e.g. cultivars M82, Moneymaker, Ailsa Craig, etc.). The large-fruited cultivar M82 has been developed into the main reference for genetic and developmental analyses [6,7], however, a high-quality genome assembly has been missing. Furthermore, analyses of quantitative phenotypes in large-fruited cultivars are labor-intensive and require extensive growth facilities to accommodate large plants with long generation times. The ultra-dwarfed variety “Micro-tom” bypasses some of these limitations [8], but its highly mutagenized background, severe hormonal and developmental abnormalities, and low fruit quality undermine its value for studying many quantitative and developmental phenotypes of agronomic importance, such as shoot, inflorescence, and fruit development, for translational research (**Figure 2.1A-E**).

To address these limitations and illustrate how new genomic and genome-editing systems can be rapidly developed, we established the small-fruited tomato cultivar Sweet-100 (S100) into a new system for genome editing and functional genomics studies. Previously, we used CRISPR-Cas9 to engineer mutations in the paralogous flowering repressor genes *SELF PRUNING* (*SP*) and *SELF PRUNING 5G* (*SP5G*) to induce fast flowering and compact growth in S100 (**Figure 2.1A and 2.1B**) [9]. Importantly, null mutations in these genes cause rapid-cycling compact growth without severe developmental abnormalities in shoot, inflorescence, and fruit development (**Figure 2.1C-E**). The first ripe fruits mature 65-70 days after sowing, allowing up to five generations per year compared to three or fewer generations for most other genotypes, which also require more space and resources (**Figure 2.1B**). Indeed, the short generation time and compact growth habit of S100 allows greenhouse and field growth at double the normal density with reduced input. Together, these characteristics make S100 a highly efficient system for genetics in tomato and a valuable addition or alternative to the widely used M82 cultivar for functional genomics and genome editing, but a high-quality reference genome to facilitate wide adoption of S100 has been lacking.

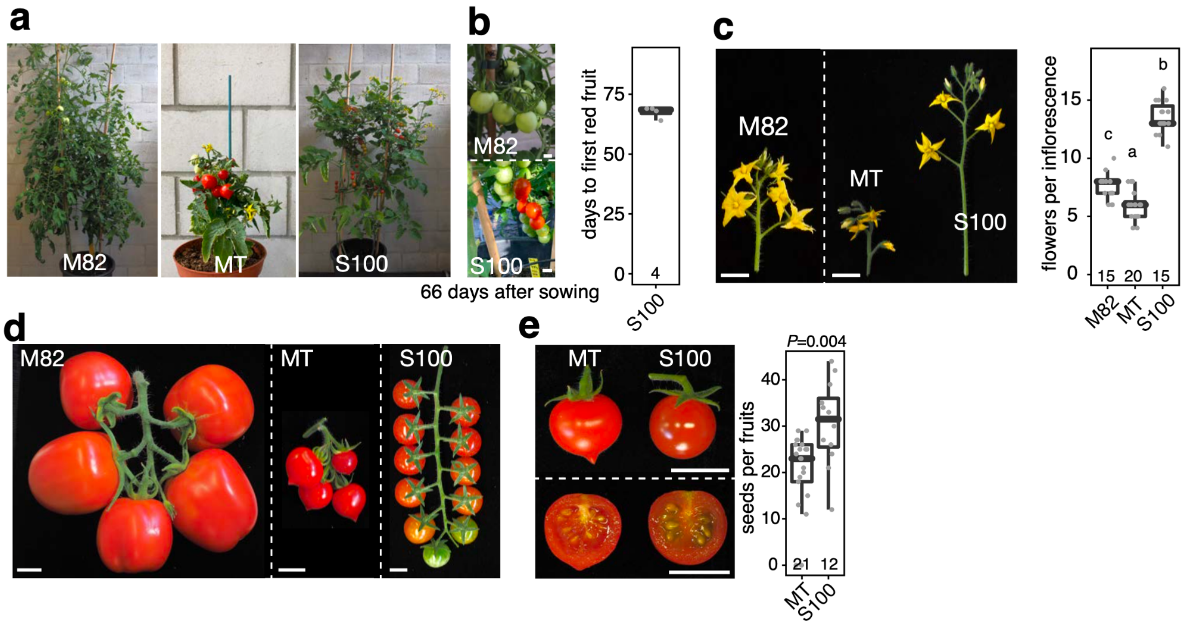


Figure 2.1: The cherry tomato cultivar Sweet-100 shows characteristics of a superior experimental model system. (A) Images of mature M82, Micro-tom (MT), and Sweet-100 (S100) plants ten weeks after sowing. (B) Images of inflorescences from M82 and S100 plants 66 days after sowing, and quantification of fruit ripening in S100. n equals the number of plants. (C) Images of detached inflorescences from M82, MT, and S100 plants and quantification of flower number per inflorescence. n equals the number of inflorescences. (D) Images of detached fruit clusters from M82, MT, and S100. (E) Images of fruits and quantification of seed number per fruit in MT and S100. n equals the number of fruits. Scale bars indicate 1 cm.

RagTag enables improved automated scaffolding of genome assemblies

High-quality genome assemblies are the foundation of genetics and functional genomics analyses. Modern genome assemblies are typically built from PacBio High Fidelity (HiFi) and/or Oxford Nanopore long-reads (ONT) [10]. HiFi reads average 15 kbp in length, are highly accurate (~0.1% error), and can produce highly contiguous draft genome assemblies [11]. However, HiFi-based assemblies often fragment at large and homogenous repeats as well as known sequence-specific coverage dropouts [12]. Built from much longer, though noisier reads with a distinct error-profile, ONT-based assemblies can complement HiFi-based assemblies by resolving some larger repeats or compensating for HiFi coverage dropouts [12]. However, even when using complementary long-read technologies, modern draft genome assemblies rarely achieve complete chromosome scale. Longer and ultimately chromosome-scale sequences are produced by scaffolding, the process of ordering and orienting genome assembly contigs, and placing gaps between adjacent contigs. Scaffolding is usually achieved by comparing a genome assembly to genome maps encoding the relative distances of genomic markers along chromosomes. Linkage, physical, and spatial proximity maps (from Chromatin Conformation Capture, or "Hi-C" data) are popular and effective for scaffolding assemblies. However, because genome maps are noisy and scaffolding methods are fallible, automated scaffolding usually results in incomplete or misassembled scaffolds and researchers often rely on laborious manual curation to correct these shortcomings [13,14].

To overcome these limitations, we developed RagTag, a new method to automate scaffolding and improve modern genome assemblies (**Figure 2.2A)**. RagTag succeeds our previously published RaGOO scaffolder and implements general improvements to the homology-based correction and scaffolding modules [15]. RagTag also provides two new scaffolding tools called "patch" and "merge". RagTag "patch" uses a genome assembly to make scaffolding joins and fill gaps in another genome assembly (**Figure 2.2B**). This is especially useful for genome assembly projects with complementary sequencing technology types, such as HiFi and ONT [12]. RagTag "merge" is an extension of the CAMSA scaffolder that reconciles distinct scaffolding solutions for a given assembly (**Figure 2.2C**) [16]. RagTag "merge" allows users to scaffold an assembly with any map or map-specific technical parameters and synergistically combine results into a single scaffolding solution. RagTag "merge" uses input scaffolding proposals to build a "scaffold graph" and users can optionally use Hi-C data to re-weight the graph and resolve ambiguous paths [17].

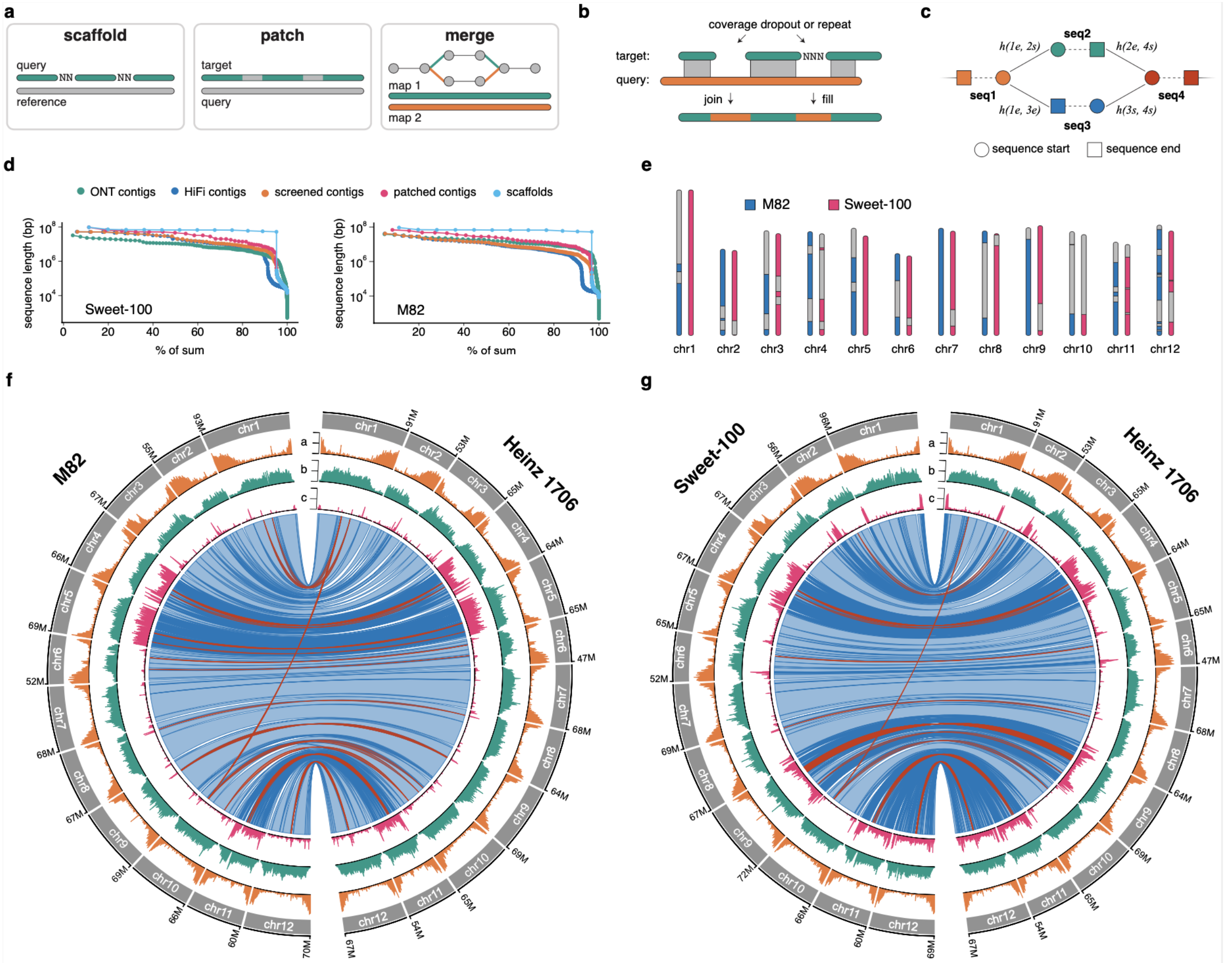


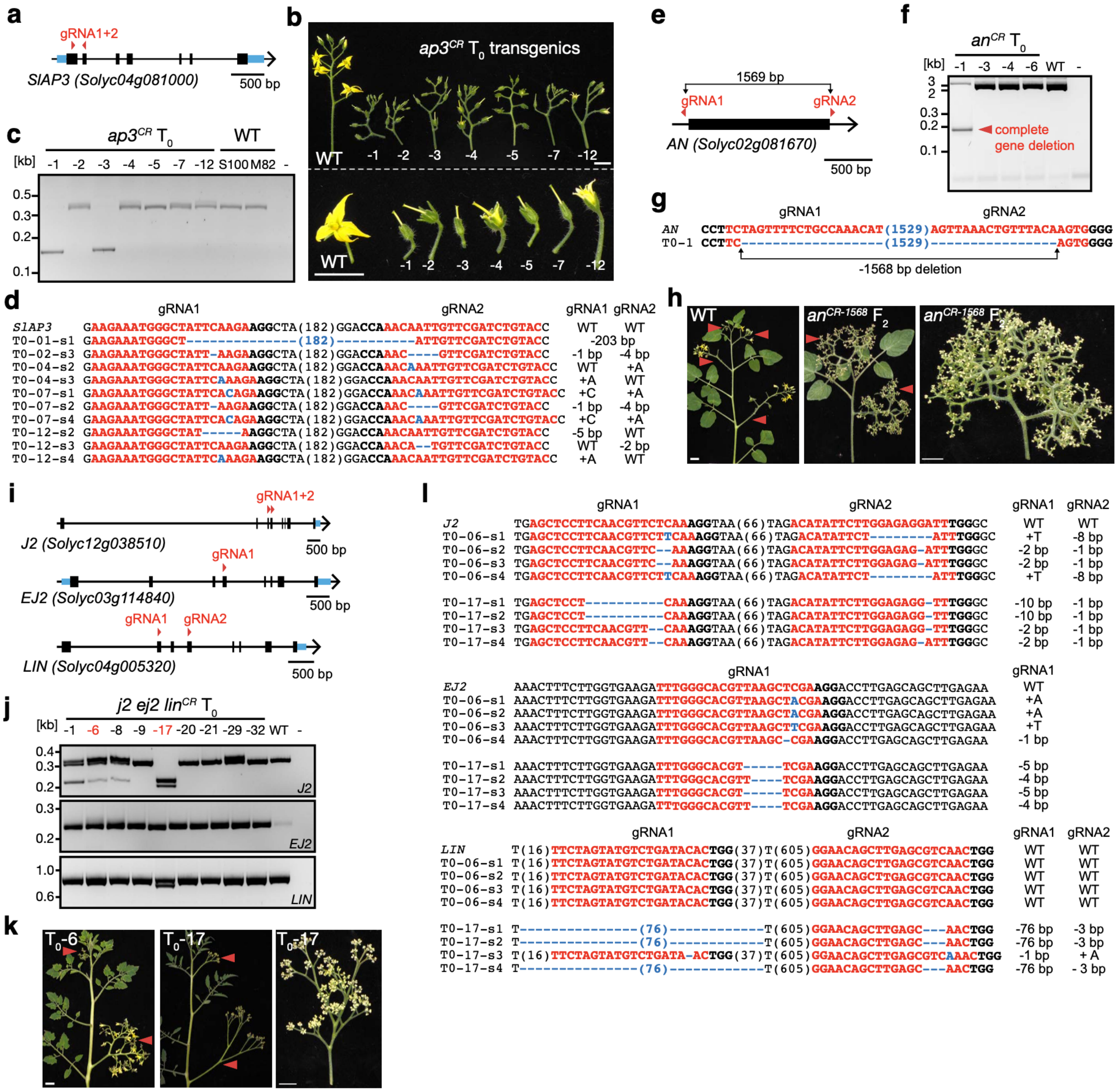
Figure 2.2: RagTag enables new reference genomes for Sweet-100 and M82. (A) An overview diagram describing RagTag "scaffold", "patch", and "merge". (B) A more detailed diagram describing RagTag "patch". Gray bars indicate alignments. (C) A more detailed diagram describing RagTag "merge". The function *h()* maps contig terminus pairs to Hi-C scores (see Methods). (D) nX plots showing the minimum sequence length (*y*-axis) need to constitute a particular percentage of the assembly (*x*-axis). (E) Ideogram showing contig boundaries (alternating color and gray) within the final scaffolds. (F) circos plots comparing M82 to Heinz 1706. Circos quantitative tracks a, b and c are summed in 500 kbp windows and show genes (a, lower tick=0, middle tick=47, upper tick=94), LTR retrotransposons (b, 0, 237, 474), and structural variants (c, 0, 24, 48). The inner ribbon track shows whole-genome alignments, with blue indicating forward-strand alignments and red indicating reverse-strand alignments (inversions) (darker colors indicate alignment boundaries). (G) same as (F) but comparing Sweet-100 to Heinz 1706 and showing genes (a, 0, 48, 96), LTR retrotransposons (b, 0, 269, 538), and structural variants (c, 0, 30, 59) and whole-genome alignment ribbons.

We used RagTag to produce high-quality chromosome-scale reference genomes for M82 and S100. Briefly, for each genotype, we assembled HiFi and ONT data independently. After screening the HiFi primary contigs for bacterial contamination and superfluous organellar sequences, we used RagTag "patch" to patch the HiFi contigs with the ONT contigs, ultimately improving the N50 from 20.1 to 40.8 Mbp and 12.6 to 27.8 Mbp in S100 and M82, respectively, without introducing any gaps (**Figure 2.2D**). After patching, S100 chromosomes 1 and 5 and M82 chromosome 7 were each represented in a single chromosome-scale contig (**Figure 2.2E**). To build 12 chromosome-scale pseudomolecules for each assembly, we used a variety of physical and spatial proximity maps to produce multiple scaffolding hypotheses. We then used RagTag "merge" to reconcile these scaffolding proposals into final scaffolding solutions. Finally, each set of scaffolds were manually corrected with Juicebox Assembly Tools and the assemblies were packaged according to the pan-sol specification (https://github.com/pan-sol/pan-sol-spec) [18].

Using a read mapping approach, we previously reported that S100 and M82 are admixed and are thus structurally distinct from the Heinz 1706 reference genome [2]. When comparing these genomes, we confirmed elevated rates of structural variation across broad chromosomal regions, indicating introgressions from wild relatives during domestication and breeding (**Figures 2.2F and 2.2G**). Nearly whole chromosomes, such as chromosomes 4, 9, 11, and 12 in S100 and chromosomes 4, 5, and 11 in M82 appear to be introgressed from wild relatives. Within introgressions, we found several large inversions in both S100 and M82. The largest inversion, a ~8.6 Mbp inversion observed on chromosome 9 of S100, was previously found in the wild tomato *Solanum pimpinellifolium* LA2093 accession and genotyped in 99% of *S. pimpinellifolium* accessions, reinforcing the contribution of *S. pimpinellifolium* and other wild tomato species to the S100 and M82 genomes [4]. Such widespread structural variation between these three tomato varieties highlights the need for personalized tomato genomes to mitigate reference bias and false signals in genomics experiments.

Efficient transformation and genome editing in Sweet-100

Powerful experimental model systems for genetics and functional genomics allow routine genetic manipulation. Using the new S100 genome assembly as a foundation, we adapted our plant transformation and genome editing protocols to genetically modify S100. We obtained transgenic plants in less than four months, which is comparable to previously published protocols for tomato [8,19]. To test the efficiency of CRISPR-Cas9 genome editing in S100, we targeted the tomato homolog of Arabidopsis *APETALA3* (*SlAP3*, Solyc04g081000) on the chromosome 4 introgression with two guide-RNAs (gRNAs) (**Figure 2.3A**). In Arabidopsis, *AP3* activity is essential for petal and stamen development [20] and we observed abnormal or missing petals and stamen on all seven *ap3*^{CR} first-generation (T0) transgenic plants (**Figure 2.3B**). We isolated multiple *ap3*^{CR} mutant alleles by sequencing and observed germline transmission to the next generation, demonstrating efficient and robust editing in S100 (**Figure 2.3C and 2.3D**). Next, we explored the possibility to delete entire genes by CRISPR-Cas9 to mitigate potential confounding effects from genetic compensation responses, which can be induced by transcribed mutant alleles and lead to upregulation of homologous genes [21]. We targeted the floral identity gene *ANANTHA* (*AN*) [22] with two gRNAs 183 bp upstream and 18 bp downstream of the protein coding sequence (**Figure 2.3E**). From four T0 transgenics we identified the complete 1568 bp gene deletion (*ap*$^{CR\text{-}1568}$), which was transmitted to the next generation (**Figure 2.3F-H**). Second-generation individuals that carried the *ap*$^{CR\text{-}1568}$ allele developed cauliflower-like inflorescence structures that are characteristic of *an* mutants, demonstrating the feasibility of complete gene deletion by CRISPR-Cas9 in S100 (**Figure 2.3H**). Finally, we tested the potential for mutating multigene families and targeted the three floral regulator genes *JOINTLESS2* (*J2*), *ENHANCER OF J2* (*EJ2*), and *LONG INFLORESCENCE* (*LIN*), which belong to the *SEPALLATA* (*SEP*) clade of the MADS-box gene family [9] (**Figure 2.3I**). From eight T0 transgenics, we identified an individual (T0-6) that displayed the *j2*^{CR}*ej2*^{CR} double mutant phenotype with strongly branched inflorescences, and an individual (T0-17) with the *j2*^{CR}*ej2*^{CR}*lin*^{CR} triple mutant phenotype and cauliflower-like inflorescences (**Figure 2.3J**). Sequencing identified mutant *J2* and *EJ2* alleles in addition to wild-type *LIN* alleles in the T0-6 individual, while only mutant alleles for all three genes were detected in individual T0-17 (**Figure 2.3K**). Together, these results illustrate the effectiveness of S100 as an experimental platform for CRISPR-Cas9 genome editing of single genes and multigene families for functional analyses.



```
                              gRNA1                                    gRNA2                     gRNA1  gRNA2
SlAP3      GAAGAAATGGCTATTCAAGAAGGCTA(182)GGACCAAACAATTGTTCGATCTGTACC   WT     WT
T0-01-s1   GAAGAAATGGCT---------------(182)-----------ATTGTTCGATCTGTACC  -203 bp
T0-02-s3   GAAGAAATGGCTATT-AAGAAGGCTA(182)GGACCAAAC----GTTCGATCTGTACC  -1 bp  -4 bp
T0-04-s2   GAAGAAATGGCTATTCAAGAAGGCTA(182)GGACCAAACAAATTGTTCGATCTGTACC  WT     +A
T0-04-s3   GAAGAAATGGCTATTCAAAGAAGGCTA(182)GGACCAAACAATTGTTCGATCTGTACC  +A     WT
T0-07-s1   GAAGAAATGGCTATTCACAGAAGGCTA(182)GGACCAAACAAATTGTTCGATCTGTACC +C     +A
T0-07-s2   GAAGAAATGGCTATT-AAGAAGGCTA(182)GGACCAAAC----GTTCGATCTGTACC  -1 bp  -4 bp
T0-07-s4   GAAGAAATGGCTATTCACAGAAGGCTA(182)GGACCAAACAAATTGTTCGATCTGTACC +C     +A
T0-12-s2   GAAGAAATGGCTAT-----AAGGCTA(182)GGACCAAACAATTGTTCGATCTGTACC  -5 bp  WT
T0-12-s3   GAAGAAATGGCTATTCAAGAAGGCTA(182)GGACCAAACA--TGTTCGATCTGTACC  WT     -2 bp
T0-12-s4   GAAGAAATGGCTATTCAAAGAAGGCTA(182)GGACCAAACAATTGTTCGATCTGTACC  +A     WT
```

```
                 gRNA1                                  gRNA2
AN     CCTTCTAGTTTTCTGCCAAACAT(1529)AGTTAAACTGTTTACAAGTGGGG
T0-1   CCTTC-------------------(1529)-----------------AGTGGGG
                        -1568 bp deletion
```

```
                    gRNA1                              gRNA2                 gRNA1   gRNA2
J2         TGAGCTCCTTCAACGTTCTCAAAGGTAA(66)TAGACATATTCTTGGAGAGGATTTGGGC     WT      WT
T0-06-s1   TGAGCTCCTTCAACGTTCTTCAAAGGTAA(66)TAGACATATTCT--------ATTTGGGC    +T      -8 bp
T0-06-s2   TGAGCTCCTTCAACGTTC--AAAGGTAA(66)TAGACATATTCTTGGAGAG-ATTTGGGC     -2 bp   -1 bp
T0-06-s3   TGAGCTCCTTCAACGTTC--AAAGGTAA(66)TAGACATATTCTTGGAGAG-ATTTGGGC     -2 bp   -1 bp
T0-06-s4   TGAGCTCCTTCAACGTTCTTCAAAGGTAA(66)TAGACATATTCT--------ATTTGGGC    +T      -8 bp

T0-17-s1   TGAGCTCCT----------CAAAGGTAA(66)TAGACATATTCTTGGAGAGG-TTTGGGC     -10 bp  -1 bp
T0-17-s2   TGAGCTCCT----------CAAAGGTAA(66)TAGACATATTCTTGGAGAGG-TTTGGGC     -10 bp  -1 bp
T0-17-s3   TGAGCTCCTTCAACGTT--CAAAGGTAA(66)TAGACATATTCTTGGAGAGG-TTTGGGC     -2 bp   -1 bp
T0-17-s4   TGAGCTCCTTCAACGTT--CAAAGGTAA(66)TAGACATATTCTTGGAGAG-ATTTGGGC     -2 bp   -1 bp

                              gRNA1                                      gRNA1
EJ2        AAACTTTCTTGGTGAAGATTTGGGCACGTTAAGCTCGAAGGACCTTGAGCAGCTTGAGAA     WT
T0-06-s1   AAACTTTCTTGGTGAAGATTTGGGCACGTTAAGCTACGAAGGACCTTGAGCAGCTTGAGAA    +A
T0-06-s2   AAACTTTCTTGGTGAAGATTTGGGCACGTTAAGCTACGAAGGACCTTGAGCAGCTTGAGAA    +A
T0-06-s3   AAACTTTCTTGGTGAAGATTTGGGCACGTTAAGCTTCGAAGGACCTTGAGCAGCTTGAGAA    +T
T0-06-s4   AAACTTTCTTGGTGAAGATTTGGGCACGTTAAGC-CGAAGGACCTTGAGCAGCTTGAGAA     -1 bp

T0-17-s1   AAACTTTCTTGGTGAAGATTTGGGCACGT-----TCGAAGGACCTTGAGCAGCTTGAGAA     -5 bp
T0-17-s2   AAACTTTCTTGGTGAAGATTTGGGCACGTT----TCGAAGGACCTTGAGCAGCTTGAGAA     -4 bp
T0-17-s3   AAACTTTCTTGGTGAAGATTTGGGCACGT-----TCGAAGGACCTTGAGCAGCTTGAGAA     -5 bp
T0-17-s4   AAACTTTCTTGGTGAAGATTTGGGCACGTT----TCGAAGGACCTTGAGCAGCTTGAGAA     -4 bp

                  gRNA1                                gRNA2                 gRNA1   gRNA2
LIN        T(16)TTCTAGTATGTCTGATACACTGG(37)T(605)GGAACAGCTTGAGCGTCAACTGG     WT      WT
T0-06-s1   T(16)TTCTAGTATGTCTGATACACTGG(37)T(605)GGAACAGCTTGAGCGTCAACTGG     WT      WT
T0-06-s2   T(16)TTCTAGTATGTCTGATACACTGG(37)T(605)GGAACAGCTTGAGCGTCAACTGG     WT      WT
T0-06-s3   T(16)TTCTAGTATGTCTGATACACTGG(37)T(605)GGAACAGCTTGAGCGTCAACTGG     WT      WT
T0-06-s4   T(16)TTCTAGTATGTCTGATACACTGG(37)T(605)GGAACAGCTTGAGCGTCAACTGG     WT      WT

T0-17-s1   T---------------(76)------------T(605)GGAACAGCTTGAGC---AACTGG     -76 bp  -3 bp
T0-17-s2   T---------------(76)------------T(605)GGAACAGCTTGAGC---AACTGG     -76 bp  -3 bp
T0-17-s3   T(16)TTCTAGTATGTCTGATA-ACTGG(37)T(605)GGAACAGCTTGAGCGTCAAACTGG    -1 bp   + A
T0-17-s4   T---------------(76)------------T(605)GGAACAGCTTGAGC---AACTGG     -76 bp  - 3 bp
```

Figure 2.3: Assessment of genome editing capabilities in Sweet-100. (A) CRISPR-Cas9 targeting of *SlAP3* using two gRNAs. Black boxes, black lines, and blue boxes represent exonic, intronic, and untranslated regions, respectively. (B) Images of detached inflorescences (top) and flowers (bottom) from wild-type (WT) and independent first-generation (T0) *ap3*^{CR} transgenic plants. (C) and (D) CRISPR-induced mutations in *AP3* identified by agarose gels (C) and Sanger sequencing (D). gRNA and PAM sequences are indicated in red and black bold letters, respectively; deletions are indicated with blue dashes; deletions; sequence gap length is given in parenthesis. (E) Full gene deletion of *AN* by CRISPR-Cas9 using two gRNAs. (F) and (G) Detection of complete deletion of the *AN* gene by agarose gel electrophoresis (F) and Sanger sequencing (G). (H) Images of WT and *an*^{CR} mutant plants in the second (F2) generation. (I) CRISPR-Cas9 targeting of the *SEP4* gene family using five gRNAs. (J) analysis of *j2 ej2 lin*^{CR} T0 plants by agarose gel electrophoresis, (K) images of T0 plants showing *j2 ej2* double (T0-6) and *j2 ej2 lin*^{CR} triple (T0-17) mutant phenotypes. (L) CRISPR-induced mutations in *J2*, *EJ2*, and *LIN* identified by Sanger sequencing. Scale bars indicate 1cm.

2.3 DISCUSSION

Personalized genome assemblies completely represent genotypes and mitigate reference bias in genomics experiments. To meet the high demand for personalized genomes, RagTag overcomes the scaffolding bottleneck by leveraging existing genome assemblies to improve new ones, or by collectively drawing from multiple genome maps to build consensus scaffolds. By using RagTag to produce personalized assemblies for M82 and the new compact, rapid-cycling Sweet-100 genotype, we provide valuable tools for modern functional genomics in tomato while demonstrating a strategy to produce efficient model plant systems and associated genomic resources in other species.

2.4 METHODS

Plant material, growth conditions, and phenotyping

Seeds of *S. lycopersicum* cv. M82 (LA3475), Sweet-100 (S100), and Micro-tom (MT) were from our stocks. Seeds were directly sown and germinated in soil in 96-cell plastic flats. Plants were grown under long-day conditions (16-h light, 8-h dark) in a greenhouse under natural light supplemented with artificial light from high-pressure sodium bulbs (~250 µmol m^{-2} s^{-1}) at 25°C and 50-60% relative humidity. Seedlings were transplanted to soil to 3.5 l (S100 and MT) or 10 l (M82) pots 3-4 weeks after sowing. Analyses of fruit ripening, flower number, seed number, fruit weight, fruit sugar content (Brix), and inflorescence branching were conducted on mature plants grown in pots. Sugar content (Brix) of fruit juice was quantified using a digital refractometer (Hanna Instruments HI96811).

RagTag Overview

RagTag succeeds RaGOO as a homology-based genome assembly correction (RagTag "correct") and scaffolding (RagTag "scaffold") tool [15]. RagTag implements general improvements and conveniences for these features but follows the same algorithmic approach as previously reported. RagTag also provides two new tools called "patch" and "merge" for genome assembly improvement. RagTag "patch" uses one genome assembly to "patch" (continuously join contigs and/or fill gaps) sequences in another assembly. RagTag "merge" reconciles two or more distinct scaffolding solutions for the same assembly. Finally, RagTag offers a variety of command-line utilities for calculating assembly statistics, validating AGP files, and working with genome assembly file formats. RagTag is open source (distributed under the MIT license) and is available on GitHub: https://github.com/malonge/RagTag.

RagTag whole-genome alignment filtering and merging

Most RagTag tools rely on pairwise (a "query" vs. a "reference/target") whole-genome alignments. RagTag supports the use of Minimap2, Unimap, or Nucmer for whole-genome alignment, though any alignments in PAF or delta format can be used [23,24]. RagTag filters and merges whole-genome alignments to extract useful scaffolding information. To remove repetitive alignments, RagTag uses an integrated version of "Unique Anchor Filtering" introduced by Assemblytics [25]. RagTag can also remove alignments based on mapping quality score, when available. Filtered alignments are then merged to identify macro-synteny blocks. For each query sequence, alignments are sorted by reference position. Consecutive alignments within 100 kbp (-d) of each other and on the same strand are merged, taking the minimum coordinate as the new start position and the maximum coordinate as the new end position. Consequently, unmerged alignments are either far apart on the same reference sequence, on different reference sequences, or on different strands. Finally, merged alignments contained within other merged alignments (with respect to the query position) are removed.

RagTag "correct"

RagTag "correct" uses pairwise whole-genome sequence homology to identify and correct putative misassemblies. First, RagTag generates filtered and merged whole-genome alignments between a "query" and a "reference" assembly. The "query" assembly will be corrected and the "reference" assembly will be used to inform correction. Any query sequence with more than one merged alignment is considered for correction. RagTag breaks these query sequences at alignment boundaries provided that the boundaries are not within 5 kbp (-b) from either sequence terminus. Users may optionally choose to only break between alignments to the same or different reference sequences (--intra and --inter). If a GFF file is provided to annotate features in the query assembly, the query assembly will never be broken within a defined feature.

When the query and reference assemblies do not represent the same genotypes, unmerged alignments within a contig can indicate genuine structural variation. To distinguish between structural variation and misassemblies, users can optionally provide Whole Genome Shotgun (WGS) sequencing reads from the same query genotype, such as short accurate reads or long error-corrected reads, to validate putative query breakpoints. RagTag aligns these reads to the query assembly with Minimap2 and computes the read coverage for each position in the query assembly. For each proposed query breakpoint, RagTag will look for exceptionally low (below --min-cov) or high (above --max-cov) coverage within 10 kbp (-v) of the proposed breakpoint. If exceptionally low or high coverage is not observed, the merged alignment boundaries are supposed to be caused by true variation, and the query assembly is not broken at this position.

RagTag "scaffold"

RagTag "scaffold" uses pairwise whole-genome sequence homology to scaffold a genome assembly. First, RagTag generates filtered and merged whole-genome alignments between a "query" and a "reference" assembly. The "query" assembly will be scaffolded and the "reference" assembly will be used to inform scaffolding. The merged alignments are used to compute a clustering, location, and orientation "confidence" score, just as is done in RaGOO, and sequences with confidence scores below certain thresholds are excluded (-i, -a, and -s) [15]. For each query sequence, the longest merged alignment is designated as the "primary" alignment. Primary alignments contained within other primary alignments (with respect to the reference coordinates) are removed. Primary alignments are then used to order and orient query sequences. To order query sequences, sequences are assigned to the reference chromosome to which they primarily align. Then, for each reference sequence, primary alignments are sorted by reference coordinate, establishing an order of query sequences. To orient query sequences, the sequence is assigned the same orientation as its primary alignment. Query sequences with no filtered alignments to the reference assembly ("unplaced" sequences) are output without modification or are optionally concatenated together.

By default, 100 bp gaps are placed between adjacent scaffolded query sequences, indicating an "unknown" gap size according to the AGP specification (https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/). Optionally, RagTag can infer the gap size based on the whole-genome pairwise alignments. Let *seq1* (upstream) and *seq2* (downstream) be adjacent query sequences, and let *aln1* and *aln2* be their respective primary alignments. Let *rs*, *re*, *qs*, and *qe* denote the alignment reference start position, reference end position, query start position, and query end position, respectively. The following function computes the inferred gap length between *seq1* and *seq2*:

$$gapsize() = (aln2_{rs} - aln2_{qs}) - (aln1_{re} + len(seq1) - aln1_{qe})$$

Where $len(seq1)$ is the length of *seq1*. All inferred gap sizes must be at least 1 bp, and if the inferred gap size is too small (-g or less than 1) or too large (-m), it is replaced with an "unknown" gap size of 100 bp.

RagTag "patch"

RagTag "patch" uses pairwise whole-genome sequence homology to make joins, without introducing gaps, and fill gaps in a "target" genome assembly using sequences from a "query" genome assembly. First, RagTag assigns new unique sequence names to all query and target sequences. Next, RagTag breaks all target sequences at gaps and generates filtered and merged whole-genome alignments between the query and target assemblies. Merged alignments that are not close (-i) to a target sequence terminus or are shorter than 50000 bp (-s) are removed. If an alignment is not close to both query sequence termini yet it is not close to either target sequence terminus, meaning the target sequence should be contained within the query sequence, yet large portions of the target sequence do not align to the query sequence, the alignment is discarded.

To ultimately patch the target assembly, RagTag employs a directed version of a "scaffold graph" [16,26]. Nodes in the graph are target sequence termini (two per target sequence), and edges connect termini of distinct target sequences. The graph is initialized with the known target sequence adjacencies originally separated by gaps in the target assembly. Next, merged and filtered alignments are processed to identify new target sequence adjacencies. For each query sequence that aligns to more than one target sequence, alignments are sorted by query position. For each pair of adjacent target sequences, an edge is created in the scaffold graph. The edge stores metadata such as query sequence coordinates in order to continuously join the adjacent target sequences. If an edge already exists due to an existing gap, the gap metadata is replaced with the query sequence metadata so that the gap can be replaced with sequence. If an adjacency is supported by more than one alignment, the corresponding edge is discarded. To find a solution to this graph and output a patched assembly, a maximal weight matching is computed with networkx and any potential cycles are removed [27]. RagTag then iterates through each connected component and iteratively builds a sequence from adjacent target sequences. When target sequences are not overlapping, they are connected with sequence from the supporting query sequence. Unpatched target sequences are output without modification.

RagTag "merge"

RagTag "merge" is a reimplementation and extension of CAMSA, a tool to reconcile two or more distinct scaffolding solutions for a genome assembly [16]. Input scaffolding solutions must be in valid AGP format, and they must order and orient the same set of genome assembly AGP "components". RagTag iteratively builds a scaffold graph to store adjacency evidence provided by each AGP file. First, each AGP file is assigned a weight (1 by default). Then, for each AGP file and for each pair of adjacent components, an edge is added to the scaffold graph, and the edge weight is incremented by the weight of the AGP file, just as is done in CAMSA. After the scaffold graph is created, users can optionally replace native edge weights with Hi-C weights. To do this, Hi-C alignments are used to compute scaffold graph weights according to the SALSA2 algorithm, which uses the same underlying scaffold graph data structure. To find a solution to this graph and to output a merged AGP file, a maximal weight matching is computed with networkx and [27] any potential cycles are removed. RagTag then iterates through each connected component and iteratively builds AGP objects. Unmerged components are output without modification.

Extraction of high-molecular weight DNA and sequencing

Extraction of high-molecular weight genomic DNA, construction of Oxford Nanopore Technology libraries, and sequencing were described previously [2]. Libraries for Pacbio HiFi sequencing were constructed and sequenced at the

Genome Technology Center at UNIL and Genome Center at CSHL. Hi-C experiments were conducted at Arima Genomics (San Diego, CA) from 2 g of flash-frozen leaf tissue.

BLAST databases for screening contigs

We built each BLAST database with makeblastdb (v2.5.0+, -dbtype nucl) [28]. We used all RefSeq bacterial genomes (downloaded on February 11th, 2021) for the bacterial genomes database. We used a collection of Solanum chloroplast sequences for the chloroplast database, and their GenbBank accession IDs are as follows:

MN218076.1, MN218077.1, MN218078.1, MN218079.1, MN218091.1, MN218088.1, MN218089.1, NC_039611.1, NC_035724.1, KX792501.2, NC_041604.1, MH283721.1, NC_039605.1, NC_039600.1, NC_007898.3, MN218081.1, NC_039606.1, NC_030207.1, MT120858.1, MN635796.1, MN218090.1, MT120855.1, MT120856.1, NC_050206.1, MN218087.1, NC_008096.2

We used a collection of Solanum mitochondrial sequences for the mitochondria database, and their GenbBank accession IDs are as follows:

MT122954.1, MT122955.1, MT122966.1, MT122969.1, MT122973.1, MT122974.1, MT122977.1, MT122988.1, NC_050335.1, MT122980.1, MT122981.1, MT122982.1, MT122983.1, MF989960.1, MF989961.1, NC_035963.1, MT122970.1, MT122971.1, NC_050334.1, MW122958.1, MW122959.1, MW122960.1, MT122964.1, MT122965.1, MW122949.1, MW122950.1, MW122951.1, MW122952.1, MW122953.1, MW122954.1, MW122961.1, MW122962.1, MW122963.1, MT122978.1, MT122979.1, MF989953.1, MF989957.1, MN114537.1, MN114538.1, MN114539.1, MT122958.1, MT122959.1

We used a collection of Solanum rDNA sequences for the rDNA database, and their GenbBank accession IDs are as follows:

X55697.1, AY366528.1, AY366529.1, KF156909.1, KF156910.1, KF156911.1, KF156912.1, KF156913.1, KF156914.1, KF156915.1, KF156916.1, KF156917.1, KF156918.1, KF156919.1, KF156920.1, KF156921.1, KF156922.1, KF603895.1, KF603896.1, X65489.1, X82780.1, AF464863.1, AF464865.1, AY366530.1, AY366531.1, AY875827.1

Sweet-100 genome assembly

The following describes the methods used to produce SollycSweet-100_v2.0 assembly. We assembled all HiFi reads with Hifiasm (v0.13-r308, -l0) and we assembled ONT reads at least 30 kbp long (a total of 28,595,007,408 bp) with Flye (v2.8.2-b1689, --genome-size 1g) [29,30]. The Hifiasm primary contigs were screened to remove contaminant or organellar contigs. We first used WindowMasker to mask repeats in the primary contigs (v1.0.0, -mk_counts -sformat obinary -genome_size 882654037) [31]. We then aligned each contig to the bacterial, chloroplast, mitochondria, and rDNA BLAST databases with blastn (v2.5.0+, -task megablast). We only included the WindowMasker file for alignments to the bacterial database (-window_masker_db). For each contig, we counted the percentage of base pairs covered by alignments to each database. If more than 10% of a contig aligned to the rDNA database, we deemed it to be a putative rDNA contig. We then removed any contigs not identified as rDNA contigs that met any of the following criteria: More than 10% of the contig was covered by alignments to the bacterial database; More than 20% of the contig was covered by alignments to the mitochondria database and the contig was less than 1 Mbp long; More than 20% of the contig was covered by alignments to the chloroplast database and the contig was less than 0.5 Mbp long. In total, we removed 1,015 contigs (35,481,360 bp) with an average length of 34,957.005 bp, most of which contained chloroplast sequence.

Even though Sweet-100 is an inbred line, to ensure that the assembly did not contain haplotypic duplication, we aligned all HiFi reads to the screened Hifiasm contigs with Winnowmap2 (v2.0, k=15, --MD -ax map-pb) [23]. We used software from purge_dups to manually inspect the contig coverage distribution, and we determined that haplotypic duplication was not evident in the screened contigs [32].

We used RagTag "patch" to patch the screened Hifiasm contigs with sequences from the ONT flye contigs, and we manually excluded three incorrect patches caused by a misassembly in the Flye contigs. We then scaffolded the patched contigs in three different ways producing three separate AGP files. First, we used RagTag for homology-based scaffolding, once using the SL4.0 reference genome and once using the LA2093 v1.5 reference genome (v2.0.1, --aligner=nucmer --nucmer-params="--maxmatch -l 100 -c 500") [3,4]. In both cases, only contigs at least 100 kbp long were considered for scaffolding, and the reference chromosome 0 sequences were not used for scaffolding. For the third scaffolding proposal, we used Juicebox Assembly Tools to manually scaffold contigs with Hi-C data (using "arima" as the restriction enzyme), and we used a custom script to convert the ".assembly" file to an AGP file. We generated generic Hi-C alignments by aligning the Hi-C reads to the screened contigs with bwa mem (v0.7.17-r1198-dirty) and processing the alignments with the Arima mapping pipeline (https://github.com/ArimaGenomics/mapping_pipeline) (https://broadinstitute.github.io/picard/) [33]. We merged the three AGP files with RagTag "merge" (v2.0.1, -r 'GATC,GA[ATCG]TC,CT[ATCG]AG,TTAA'), using generic Hi-C alignments to weight the Scaffold Graph (-b). Finally, using the merged scaffolds as a template, we made four manual scaffolding corrections in Juicebox Assembly tools. The final assembly contained 12 scaffolds corresponding to 12 chromosomes and 918 unplaced nuclear sequences.

VecScreen did not identify any "strong" or "moderate" hits to the adaptor contamination database (ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/adaptors_for_screening_euks.fa) (https://www.ncbi.nlm.nih.gov/tools/vecscreen/). We packaged the assembly according to the pan-sol v0 specification (https://github.com/pan-sol/pan-sol-spec), and chromosomes were renamed and oriented to match the SL4.0 reference genome. The tomato chloroplast (GenBank accession NC_007898.3) and mitochondria (GenBank accession NC_035963.1) reference genomes were added to the final assembly. To identify potential misassemblies and heterozygous Structural Variants (SVs), we aligned all HiFi reads (v2.0, k=15, --MD -ax map-pb) and ONT reads longer than 30 kbp (v2.0, k=15, --MD -ax map-ont) to the final assembly with Winnowmap2 and we called structural variants with Sniffles (v1.0.12, -d 50 -n -1 -s 5) [34]. We removed any SVs with less than 30% of reads supporting the ALT allele and we merged the filtered SV calls with Jasmine (v1.0.10, max_dist=500 spec_reads=5 --output_genotypes) [35,36].

Sweet-100 gene and repeat annotation

We used Liftoff to annotate the Sweet 100 v2.0 assembly using ITAG4.0 gene models and tomato pan-genome genes as evidence (v1.5.1, -copies) [1,3,37]. Chloroplast and mitochondria annotations were replaced with their original GenBank annotation. Transcript, coding sequence, and protein sequences were extracted using gffread (v0.12.3, -y -w -x) [38]. We annotated transposable elements with EDTA (v1.9.6, --cds --overwrite 1 --sensitive 1 --anno 1 --evaluate 1) [39].

M82 genome assembly

The M82 genome was assembled and annotated following the approach used for the Sweet-100 assembly, with the following distinctions. First, Hifiasm v0.15-r327 was used for assembling HiFi reads. Also, the M82 ONT assembly was polished before patching. M82 Illumina short-reads [15] were aligned to the draft Flye ONT assembly with BWA-MEM (v0.7.17-r1198-dirty) and alignments were sorted and compressed with samtools (v1.10) [33,40]. Small variants were called with freebayes (v1.3.2-dirty, --skip-coverage 480) and polishing edits were incorporated into the assembly with bcftools "consensus" (v1.10.2, -i'QUAL>1 && (GT="AA" || GT="Aa")' -Hla) [41]. In total, two iterative rounds of polishing were used. RagTag "merge" was also used for scaffolding, though the input scaffolding solutions used different methods than the Sweet-100 assembly. First, homology-based scaffolds were generated with RagTag "scaffold", using the SL4.0 reference genome (v2.0.1, --aligner=nucmer --nucmer-params="--maxmatch -l 100 -c 500"). Contigs smaller than 300 kbp were not scaffolded (-j), and the reference chromosome 0 was not used to inform scaffolding (-e). Next, SALSA2 was used to derive Hi-C-based scaffolds. Hi-C reads were aligned to the assembly with the generic pipeline described for Sweet-100. We then produced scaffolds with SALSA2 (-c 300000 -p yes -e GATC -m no) and manually corrected false scaffolding joins in Juicebox Assembly Tools. We reconciled the homology-based and Hi-C-based scaffolds with RagTag "merge" using generic Hi-C alignments to re-weight the scaffold graph (-b). Finally, we made four manual corrections in Juicebox Assembly Tools.

CRISPR-Cas9 mutagenesis, plant transformation, and identification of mutant alleles

CRISPR-Cas9 mutagenesis was performed as described previously [42]. Briefly, guide RNAs (gRNAs) were designed manually or using the CRISPRdirect tool (https://crispr.dbcls.jp/). Binary vectors for plant transformation were assembled using the Golden Gate cloning system as previously described [43]. Final vectors were transformed into the tomato cultivar S100 by A*grobacterium tumefaciens*-mediated transformation according to Gupta and Van Eck (2016) with minor modifications [19]. Briefly, seeds were sterilized for 15 min in 1.3% bleach followed by 10 min in 70% ethanol, and rinsed four times with sterile water before sowing on MS media (4.4 g/l MS salts, 1.5 % sucrose, 0.8 % agar, pH 5.9) in Magenta boxes. Cotyledons were excised 7-8 days after sowing and incubated on 2Z- media [19] at 25°C in the dark for 24 hrs prior to transformation. *A. tumefaciens* were grown in LB media and washed in MS-0.2% media (4.4 g/l MS salts, 2% sucrose, 100 mg/l myo-inositol, 0.4 mg/l thiamine, 2 mg/l acetosyringone, pH5.8). Explants were co-cultivated with *A. tumefaciens* on 2Z- media supplemented with 10 µg/l IAA for 48 hrs at 25°C in the dark and transferred to 2Z selection media (supplemented with 150 mg/l kanamycin). Explants were transferred every two weeks to fresh 2Z selection media until shoot regeneration. Shoots were excised and transferred to selective rooting media [19] (supplemented with 150 mg/l kanamycin) in Magenta boxes. Well-rooted shoots were transplanted to soil and acclimated in a percival growth chamber (~100 µmol m^{-2} s^{-1}, 25°C, 50% humidity) before transfer to the greenhouse. Genomic DNA was extracted from T0 plants using a quick genomic DNA extraction protocol. Briefly, small pieces of leaf tissue were flash-frozen in liquid nitrogen and ground in a bead mill (Qiagen). Tissue powder was incubated in extraction buffer (100 mM Tris-HCl pH9.5, 250 mM KCl, 10 mM EDTA) for 10 min at 95°C followed by 5 min on ice. Extracts were combined with one volume of 3% BSA, vigorously vortexed, and spun at 13.000 rpm for 1 min. One microliter supernatant was used as template for PCR using primers flanking the gRNA target sites. PCR products were separated on agarose gels and purified for Sanger Sequencing (Microsynth) using ExoSAP-IT reagent (NEB). Chimeric PCR products were subcloned before sequencing using StrataClone PCR cloning kits (Agilent).

2.5 ACKNOWLEDGEMENTS

We thank all members of the Lippman, Schatz, and Soyk labs for helpful discussions. We thank B. Tissot-Dit-Sanfin, A. Robadey, and V. Vashanthakumar from UNIL for assistance with plant care. We thank J. Marquis, J. Weber, and M. Dupasquier from the UNIL Genome Technology Facility (GTF) for sequencing support. We thank Steven Salzberg, Aleksey Zimin, Luca Venturini, and Alexander Leonard for their help with software features and improvements. This work was performed with assistance from the US National Institutes of Health Grant S10OD028632-01. M.A. and M.C.S. were supported by the National Science Foundation: DBI-1350041 and IOS-1732253. This work was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC Starting Grant "EPICROP" Grant No. 802008) and a Swiss National Science Foundation (SNSF) Eccellenza Professorial Fellowship (Grant No. PCEFP3_181238) to S.S.

2.6 REFERENCES

1. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet. 2019;51:1044–51.

2. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. Elsevier; 2020;182:145–61.e23.

3. Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, et al. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. bioRxiv. biorxiv.org; 2019;767764.

4. Wang X, Gao L, Jiao C, Stravoravdis S, Hosmani PS, Saha S, et al. Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding. Nat Commun. nature.com; 2020;11:5817.

5. van Rengs WMJ, Schmidt MHW, Effgen S, Wang Y. A gap-free tomato genome built from complementary PacBio and Nanopore long DNA sequences reveals extensive linkage drag during breeding. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.08.30.456472.abstract

6. Eshed Y, Zamir D. An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics. academic.oup.com; 1995;141:1147–62.

7. Menda N, Semel Y, Peled D, Eshed Y, Zamir D. In silico screening of a saturated mutation library of tomato. Plant J. Wiley Online Library; 2004;38:861–72.

8. Meissner R, Jacobson Y, Melamed S, Levyatuv S, Shalev G, Ashri A, et al. A new model system for tomato genetics. Plant J. Wiley; 1997;12:1465–72.

9. Soyk S, Lemmon ZH, Oved M, Fisher J, Liberatore KL, Park SJ, et al. Bypassing Negative Epistasis on Yield in Tomato Imposed by a Domestication Gene. Cell. 2017;169:1142–55.e12.

10. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21:597–614.

11. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.

12. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV. The complete sequence of a human genome. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1.abstract

13. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46.

14. Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, et al. Significantly improving the quality of genome assemblies through curation. Gigascience [Internet]. academic.oup.com; 2021;10. Available from: http://dx.doi.org/10.1093/gigascience/giaa153

15. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20:224.

16. Aganezov SS, Alekseyev MA. CAMSA: a tool for comparative analysis and merging of scaffold assemblies. BMC Bioinformatics. 2017;18:496.

17. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;15:e1007273.

18. Dudchenko O, Shamim MS, Batra S, Durand NC. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv [Internet]. biorxiv.org; 2018; Available from: https://www.biorxiv.org/content/10.1101/254797v1.abstract

19. Gupta S, Van Eck J. Modification of plant regeneration medium decreases the time for recovery of Solanum lycopersicum cultivar M82 stable transgenic lines. Plant Cell Tissue Organ Cult. Springer Nature; 2016;127:417–23.

20. Jack T, Brockman LL, Meyerowitz EM. The homeotic gene APETALA3 of Arabidopsis thaliana encodes a MADS box and is expressed in petals and stamens. Cell. Elsevier; 1992;68:683–97.

21. El-Brolosy MA, Kontarakis Z, Rossi A, Kuenne C, Günther S, Fukuda N, et al. Genetic compensation triggered by mutant mRNA degradation. Nature. nature.com; 2019;568:193–7.

22. Lippman ZB, Cohen O, Alvarez JP, Abu-Abied M, Pekker I, Paran I, et al. The making of a compound inflorescence in tomato and related nightshades. PLoS Biol. journals.plos.org; 2008;6:e288.

23. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

24. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

25. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016;32:3021–3.

26. Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. BMC Genomics. 2017;18:527.

27. Galil Z. Efficient algorithms for finding maximum matching in graphs. ACM Comput Surv. New York, NY, USA: Association for Computing Machinery; 1986;18:23–38.

28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

29. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. nature.com; 2021;18:170–5.

30. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6.

31. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2006;22:134–41.

32. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36:2896–8.

33. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv [q-bio.GN]. 2013. Available from: http://arxiv.org/abs/1303.3997

34. Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, et al. Weighted minimizer sampling improves long read mapping. Bioinformatics. 2020;36:i111–8.

35. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. nature.com; 2018;15:461–8.

36. Kirsche M, Prabhu G, Sherman R, Ni B, Aganezov S, Schatz MC. Jasmine: Population-scale structural variant comparison and analysis [Internet]. bioRxiv. 2021 [cited 2021 Sep 28]. p. 2021.05.27.445886. Available from: https://www.biorxiv.org/content/10.1101/2021.05.27.445886v1.abstract

37. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics [Internet]. 2020; Available from: http://dx.doi.org/10.1093/bioinformatics/btaa1016

38. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020;9:304.

39. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275.

40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

41. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021;10. Available from: http://dx.doi.org/10.1093/gigascience/giab008

42. Brooks C, Nekrasov V, Lippman ZB, Van Eck J. Efficient gene editing in tomato in the first generation using the clustered regularly interspaced short palindromic repeats/CRISPR-associated9 system. Plant Physiol. academic.oup.com; 2014;166:1292–7.

43. Soyk S, Müller NA, Park SJ, Schmalenbach I, Jiang K, Hayama R, et al. Variation in the flowering gene SELF PRUNING 5G promotes day-neutrality and early yield in tomato. Nat Genet. nature.com; 2017;49:162–8.

By scaffolding and annotating our contigs, we created the genomic context needed to quantify and qualify the completeness of the Triticum_aestivum_4.0 assembly, especially relative to its predecessors.

3

Chromosome-Scale Assembly of the Bread Wheat Genome Reveals Thousands of Additional Gene Copies

The following sections were previously published in *Genetics*:

> Alonge, M.*, Shumate, A.*, Puiu, D., Zimin, A.V., and Salzberg, S.L. Chromosome-scale assembly of the bread wheat genome reveals thousands of additional gene copies. Genetics 216.2 (2020): 599-608. https://doi.org/10.1534/genetics.120.303501

A.S. and M.A. contributed equally to this work. S.L.S. supervised all analysis and manuscript preparation. S.L.S and A.V.Z. supervised the genome assembly scaffolding and analysis. A.S., M.A., A.V.Z, and S.L.S. edited all of the text. A.V.Z made Table 3.1. A.S. made Figure 3.4A, 3.4D, and 3.4E. The results section "Annotating the Triticum_aestivum_4.0 genome assembly" was done and written by A.S. The results section "Triticum_aestivum_4.0 accurately represents gene duplications affecting traits" was done by M.A. and A.S. and written by M.A. The methods section "Chloroplast and mitochondria genome assembly" was done by D.P. and written by S.L.S. The methods section "Genome annotation" was done and written by A.S. All other results and methods sections and figures were done/written by M.A. M.A. wrote the abstract, background, and discussion, with editing from A.S., A.V.Z, and S.L.S.

3.1 ABSTRACT

Bread wheat (*Triticum aestivum*) is a major food crop and an important plant system for agricultural genetics research. However, due to the complexity and size of its allohexaploid genome, genomic resources are limited compared to other major crops. The IWGSC recently published a reference genome and associated annotation (IWGSC CS v1.0, Chinese Spring) that has been widely adopted and utilized by the wheat community. Although this reference assembly represents all three wheat subgenomes at chromosome-scale, it was derived from short reads, and thus is missing a substantial portion of the expected 16 Gbp of genomic sequence. We earlier published an independent wheat assembly (Triticum_aestivum_3.1, Chinese Spring) that came much closer in length to the expected genome size, although it was only a contig-level assembly lacking gene annotations. Here, we describe a reference-guided effort to scaffold those contigs into chromosome-length pseudomolecules, add in any missing sequence that was unique to the IWGSC CS v1.0 assembly, and annotate the resulting pseudomolecules with genes. Our updated assembly, Triticum_aestivum_4.0, contains 15.07 Gbp of nongap sequence anchored to chromosomes, which is 1.2 Gbps more than the previous reference assembly. It includes 108,639 genes unambiguously localized to chromosomes, including over 2,000 genes that were previously unplaced. We also discovered >5,700 additional gene copies, facilitating the accurate annotation of functional gene duplications including at the *Ppd-B1* photoperiod response locus.

3.2 BACKGROUND

Bread wheat (*Triticum aestivum*) is a crop of significant worldwide nutritional, cultural, and economic importance. As with most other major crops, there is a strong interest in applying advanced breeding and genomics technologies toward crop improvement. Key to these efforts are high-quality reference genome assemblies and associated gene annotations, which are the foundations of genomics research. However, the bread wheat genome has some notable features that make it especially technically challenging to assemble. One such feature is allohexaploidy (2n = 6× = 42, AABBDD), a result of wheat's dynamic domestication history [1,2]. This polyploidy results from the hybridization of domesticated emmer (*Triticum turgidum*, AABB) with *Aegilops tauschii* (DD). Domesticated emmer—also an ancestor of durum wheat—is itself an allotetraploid resulting from interspecific hybridization between *Triticum urartu* and a relative of *Aegilops speltoides*.

The resulting bread wheat genome is immense, with flow cytometry studies estimating the genome size to be ~16 Gbp [3]. As with most other large plant genomes, repeats, including mostly retrotransposons, make up the majority of the genome, which is estimated to be ~85% repetitive [4]. These repeats make this genome especially difficult to assemble, even given the recent improvements in long-read sequencing and algorithmic advancements in genome assembly technology. Nonetheless, early efforts were made to establish de novo reference genome assemblies for wheat. In 2014, the International Wheat Genome Sequencing Consortium (IWGSC) used flow cytometry-based sorting to sequence and assemble individual chromosome arms, thus removing the repetitiveness introduced by homeologous chromosomes (IWGSC 2014). Despite this approach, this short-read-based assembly was highly fragmented, and only reconstructed ~10.2 Gbp of the genome. Subsequent short-read assemblies using alternate strategies were also developed by the community, though each also struggled to achieve contiguity and completeness [5,6].

In 2017, we released the first-ever long-read-based assembly for bread wheat (Triticum_aestivum_3.1), representing the Chinese Spring variety [7]. With an N50 contig size of 232.7 kbp, Triticum_aestivum_3.1 was far more contiguous than any previous assembly of bread wheat, and with a total assembly size of 15.34 Gbp, it reconstructed the highest percentage of the expected wheat genome size of any assembly. Though this assembly provided a more complete representation of the Chinese Spring genome, its contigs were not mapped onto chromosomes, and, notably, it did not include gene annotation.

In 2018, the IWGSC published a chromosome-scale reference assembly and associated annotations for bread wheat (IWGSC CS v1.0, Chinese Spring), providing the best-annotated reference genome yet [4]. Because that assembly was entirely derived from short reads, it was less complete and more fragmented than Triticum_aestivum_3.1, having a total size of 14.5 Gbp and an N50 contig size of 51.8 kbp. However, a collection of long-range scaffolding data, including physical (BACs, Hi-C), optical (Bionano), and genetic maps, enabled most of the assembled scaffolds to be mapped onto wheat's 21 chromosomes. These pseudomolecules served as a foundation for comprehensive de novo gene and repeat annotation, facilitating investigations into the genomic elements that drove the evolution of genome size, structure, and function in wheat.

Here, we used the IWGSC CS v1.0 assembly (GenBank accession GCA_900519105.1) to inform the scaffolding and annotation of the more complete Triticum_aestivum_3.1 assembly. The new assembly, Triticum_aestivum_4.0, contains 1.1 Gbp of additional nongapped sequence compared to IWGSC CS v1.0 while localizing 97.9% of sequence to chromosomes. Comparative analysis revealed that Triticum_aestivum_4.0 more accurately represents the Chinese Spring repeat landscape, which is heavily collapsed in IWGSC CS v1.0. Our more complete assembly allowed us to anchor ~2000 genes that were previously annotated on unlocalized contigs in IWGSC CS v1.0. We also found 5799 additional gene copies in Triticum_aestivum_4.0, showing extensive collapsing of gene duplicates in the IWGSC CS v1.0 assembly. We highlighted specific examples of these extra gene copies, including at the *Ppd-B1* locus, where Triticum_aestivum_4.0 accurately reflects the expected four copies of pseudo-response regulator (PRR) genes influencing photoperiod sensitivity. We additionally found three extra copies of a MADS-box transcription factor gene in T4, demonstrating the potential to find new gene copy number variants (CNVs) that influence traits. The Triticum_aestivum_4.0 assembly and annotations are available at www.ncbi.nlm.nih.gov/bioproject/PRJNA392179.

3.3 RESULTS

Scaffolding the Triticum_aestivum_3.1 genome assembly

Our goal was to utilize both our previously published Triticum_aestivum_3.1 contigs (T3) and the IWGSC CS v1.0 reference assembly (IW) to establish an improved chromosome-scale genome assembly for the Chinese Spring variety of bread wheat. **Figure 3.1** depicts the pipeline used to derive our final Triticum_aestivum_4.0 (T4) assembly. We started with the T3 contigs because they were highly contiguous (N50 = 232.7 kbp) and contained a total of 1.1 Gbp more nongap sequence compared to the IW assembly. However, we wanted to ensure that our final assembly did not exclude any contigs missing from T3 but present in IW. To incorporate any such “missing” IW contigs, we first derived a set of contigs from the IW assembly by breaking pseudomolecules at gaps. By aligning these IW contigs to the T3 assembly, we identified 4702 IW contigs (89,866,936 bp) with sequence missing from the T3 assembly. These sequences along with the T3 contigs comprised our initial contig set.

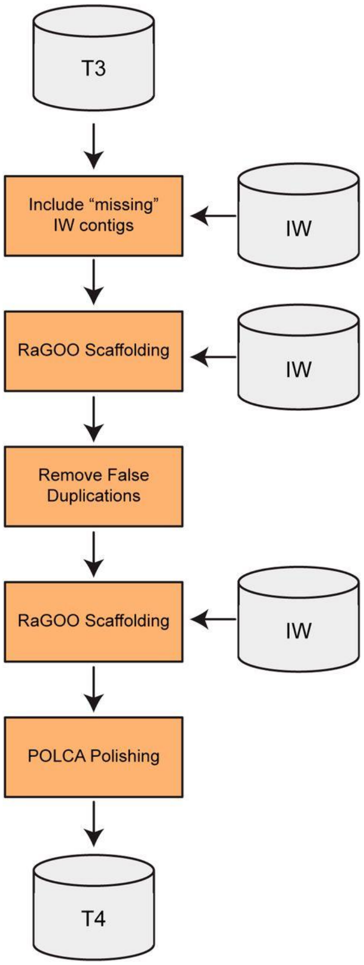


Figure 3.1: The Triticum_aestivum_4.0 assembly scaffolding pipeline. A diagram depicting the Triticum_aestivum_4.0 (T4) assembly scaffolding pipeline, which takes the Triticum_aestivum_3.0 (T3) and IWGSC CS v1.0 (IW) assemblies as input. Gray cylinders represent input or output genome assemblies, while orange boxes show the steps of the scaffolding process.

We used RaGOO [8]—a reference-guided scaffolding tool—to order and orient these contigs into chromosome-length scaffolds. This scenario presents a near-ideal context for reference-guided scaffolding because the contigs and the reference assembly represent the same inbred genotype, and thus we expect no genomic structural differences. Although RaGOO normally utilizes Minimap2 [9] alignments between contigs and a reference assembly, we used NUCmer [10,11] instead, as it offered the necessary flexibility to align these large and repetitive genomes. Specifically, NUCmer provided the specificity needed to unambiguously align contigs to a highly repetitive allohexaploid reference genome. Even with high stringency alignments, RaGOO ordered and oriented most of the assembly (97.67% of bp) into pseudomolecules.

We next sought to remove any false duplications potentially created during the process of incorporating 4702 IW sequences. We aligned these IW contigs to the RaGOO scaffolds and removed 357 IW contigs from the initial set of 4702 that aligned to more than one place in the assembly, and, therefore, were no longer deemed “missing” from T3. This produced our final set of contigs, which included the T3 contigs plus 4345 (84,909,842 bp) contigs from IW that contained sequence missing from T3. The final contigs had an N50 length of 230,687 bp (essentially the same as the T3 assembly) and a sum of 15,429,603,425 bp. We then repeated the RaGOO scaffolding step, and polished the resulting scaffolds with POLCA [12] using the original Illumina reads, yielding the final T4 chromosome-scale assembly. Finally, we removed mitochondria and chloroplast genome sequence from T4 and assembled these genomes separately with Illumina reads.

Despite the highly repetitive nature of the Chinese Spring genome, RaGOO confidence scores indicate that T4 scaffolding was consistent with the reference genome structure. This suggests that our high-specificity NUCmer parameters mitigated erroneous contig ordering and orientation resulting from repetitive alignments. Dotplots further confirm that there are no large-scale structural rearrangements between T4 and IW pseudomolecules. While borrowing its chromosomal structure from IW, T4 demonstrates superior sequence completeness. 97.9% of T4 sequence (15.09 Gbp) was placed onto 21 chromosomes yielding pseudomolecules that had 1.2 Gbp more localized nongapped sequence than the IW reference (**Table 3.1**). This extra sequence was evenly distributed across the genome, with each T4 pseudomolecule containing more sequence (average of 48.8 ± 8.4 Mbp) than its IW counterpart while having substantially fewer gaps (**Figure 3.2**).

Assembly	T4	IW	T3
All sequence (bp)	15,397,713,314	14,271,578,887	15,344,693,583
Anchored sequence (bp)	15,070,919,678	13,840,498,961	N/A

Table 3.1: Wheat genome assembly completeness. Nongapped sequence length of the Triticum_aestivum_4.0 (T4), IWGSC CS v1.0 (IW), and Triticum_aestivum_3.1 (T3) assemblies.

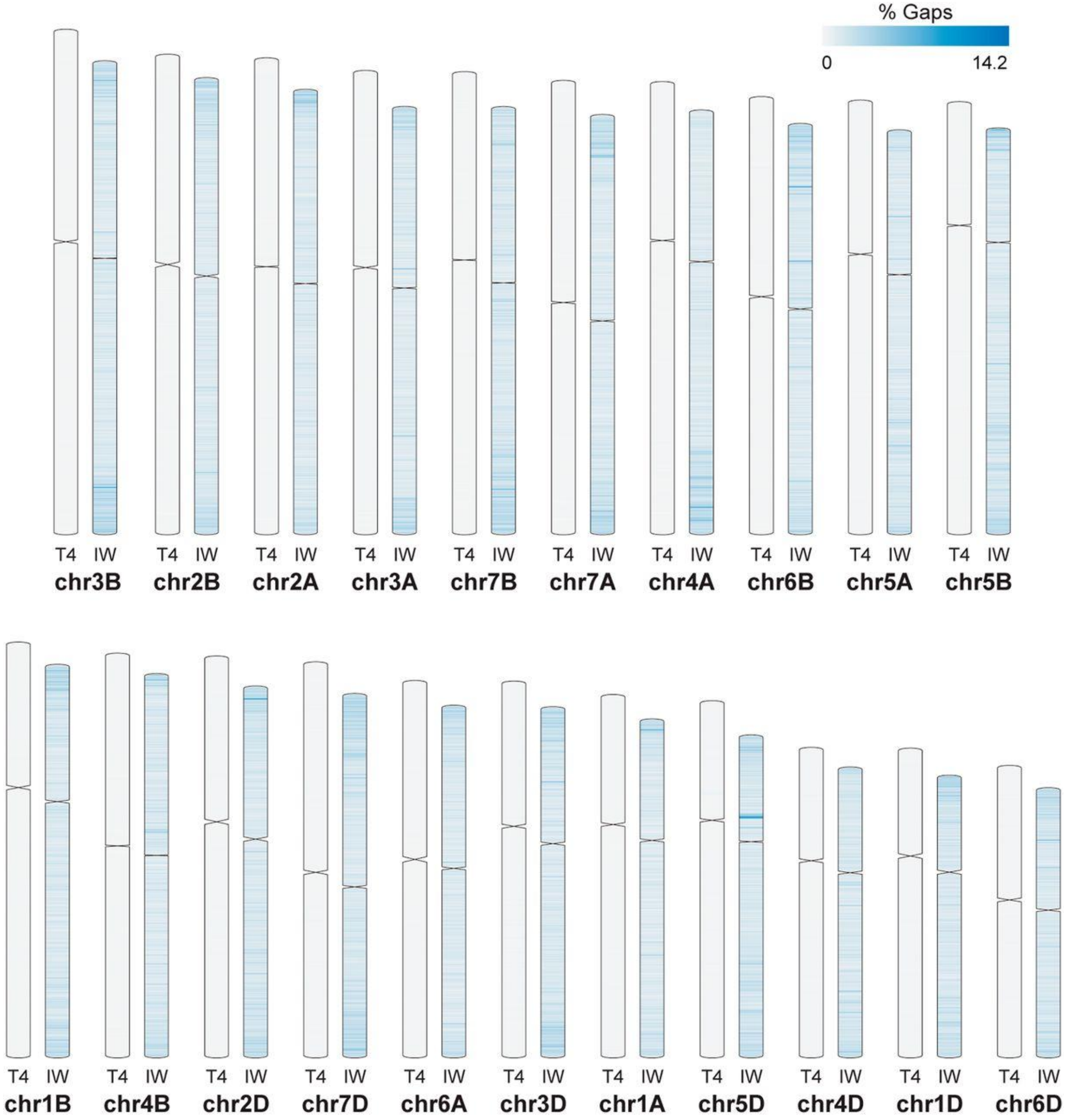


Figure 3.2: A comparison of Triticum_aestivum_4.0 and IWGSC CS v1.0 assembly completeness. An ideogram showing the distribution of gap sequence in the Triticum_aestivum_4.0 (T4) and IWGSC CS v1.0 (IW) assemblies. The heatmap color intensity corresponds to the percentage of gap sequence in nonoverlapping 1 Mbp windows along each chromosome. Chromosomes are sorted by T4 length (left to right, top to bottom), highlighting that each T4 chromosome across all three subgenomes has more sequence and fewer gaps than its IW counterpart.

Because IW was derived from short reads, it is conceivable that some genomic repeats were collapsed during assembly [13]. Therefore, we hypothesized that T4, a long-read-based assembly, more accurately represents the repeat landscape of the Chinese Spring genome. As support for this hypothesis, we observe that 101-mers shared by T4 and IW were present at higher copies in T4 (**Figure 3.3**). This observation holds for a wide range of 101-mer copy numbers, suggesting that T4 more accurately represents both lower-order (duplications) and higher-order (transposable elements) repeats. To investigate a specific instance of repeat collapse in IW, we compared centromere sequence content in the two assemblies. As was done in the original IW publication, we used publicly available CENH3 ChIP-seq data to infer centromere positions in T4 [4,14]. This analysis indicated ChIP-seq peaks corresponding to centromeres for each of the 21 chromosomes. T4 had a total of 39.1 Mbp more centromeric sequence than IW, highlighting that the long-read-based T4 assembly localized more centromeric sequence than IW.

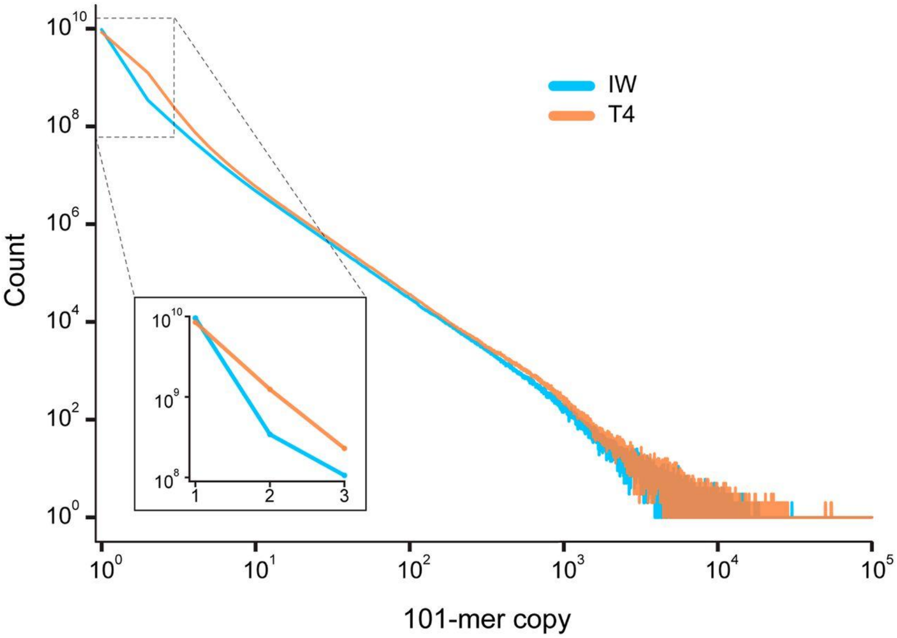


Figure 3.3: Shared assembly k-mer count distribution. Histogram of 101-mer copy number in the Triticum_aestivum_4.0 (T4) and IWGSC CS v1.0 (IW) assemblies. Only 101-mers shared by both assemblies are considered. While IW has more single-copy 101-mers, T4 represents more 101-mers at higher copy numbers.

Annotating the Triticum_aestivum_4.0 genome assembly

We mapped the IW v1.1 high-confidence annotation onto T4 using an annotation lift-over tool we developed called Liftoff [15]. Given a genome annotation, Liftoff aligns all genes, chromosome by chromosome, to a different genome of the same species using BLAST [16]. For all genes that fail to map to the same chromosome, Liftoff attempts to map them across chromosomes. The best mapping for each gene is chosen according to sequence identity and concordance with the exon/intron structure of the original gene model. Out of 130,745 transcripts from 105,200 gene loci annotated on primary chromosomes in IW, we successfully mapped 124,579 transcripts from 100,831 gene loci. We define a transcript as successfully mapped if the mRNA sequence in T4 is at least 50% as long as the mRNA sequence in IW. However, the vast majority of transcripts greatly exceed this threshold, with 92% of transcripts having an alignment coverage of 98% or greater. Sequence identity is similarly high with 92% of transcripts aligning at an identity of 95% or greater. Of the transcripts that failed to map, 4634 had a partial mapping with an alignment coverage <50%, and the remaining transcripts failed to map entirely.

As expected, we observed strong gene synteny between T4 and IW. Of the 100,831 mapped IW genes, 96,148 mapped to the same chromosome in T4. The remaining 4683 mapped to a different chromosome after failing to map to their expected chromosome. There is a clear pattern showing many of these genes mapped to a similar location on the same chromosome of a different subgenome. We also found that the sequence identity of genes mapped to different chromosomes is much lower, with an average identity of 90.7% compared to 99.3% in genes mapped to the same chromosome. We therefore hypothesize that these genes are missing in the T4 assembly, and have instead mapped to paralogs in T4 that are not annotated in IW.

The IW v1.1 annotation also contains 2691 genes annotated on unplaced contigs ("chrUn"). Using Liftoff, we were able to map 2001 of these genes onto a primary chromosome in T4; 1767 genes were confidently placed with a sequence identity of at least 98% while the remaining 234 mapped with a lower identity. To control for differences in annotation pipelines between IW and T4, we used Liftoff to map chrUn genes onto the primary IW chromosomes to look for additional, unannotated, gene copies. Of the 2001 chrUn genes mapped to T4 pseudomolecules, 78 of these were also mapped to primary IW chromosomes. This suggests that ≥1923 genes were placed due to improved assembly completeness rather than differences in annotation methods.

After mapping the IW v1.1 annotation onto T4, we used Liftoff to look for additional gene copies in T4. We required 100% sequence identity in exons and splice sites to map a gene copy. We found 5799 additional gene copies in T4 that are not annotated in IW v1.1. Of these, 4158 genes have one extra copy, and 567 genes have two or more additional copies, with a maximum of 84 additional copies (**Figure 3.4**). IW collapsed most gene copies on the same chromosome rather than across homeologous chromosomes, with 4062 of the 5799 additional gene copies occurring on the same chromosome, and 97 copies occurring on the same chromosome of a different subgenome (**Figure 3.4**); 915 gene copies were placed on different chromosomes. The remaining 725 are extra copies of chrUn genes placed on chromosomes. As was done for unplaced genes, we also looked for additional IW gene copies present elsewhere in IW. Of our 5799 additional gene copies, 159 were also present in IW, suggesting that at least 5640 of T4 copies are strictly the result of improved assembly completeness.

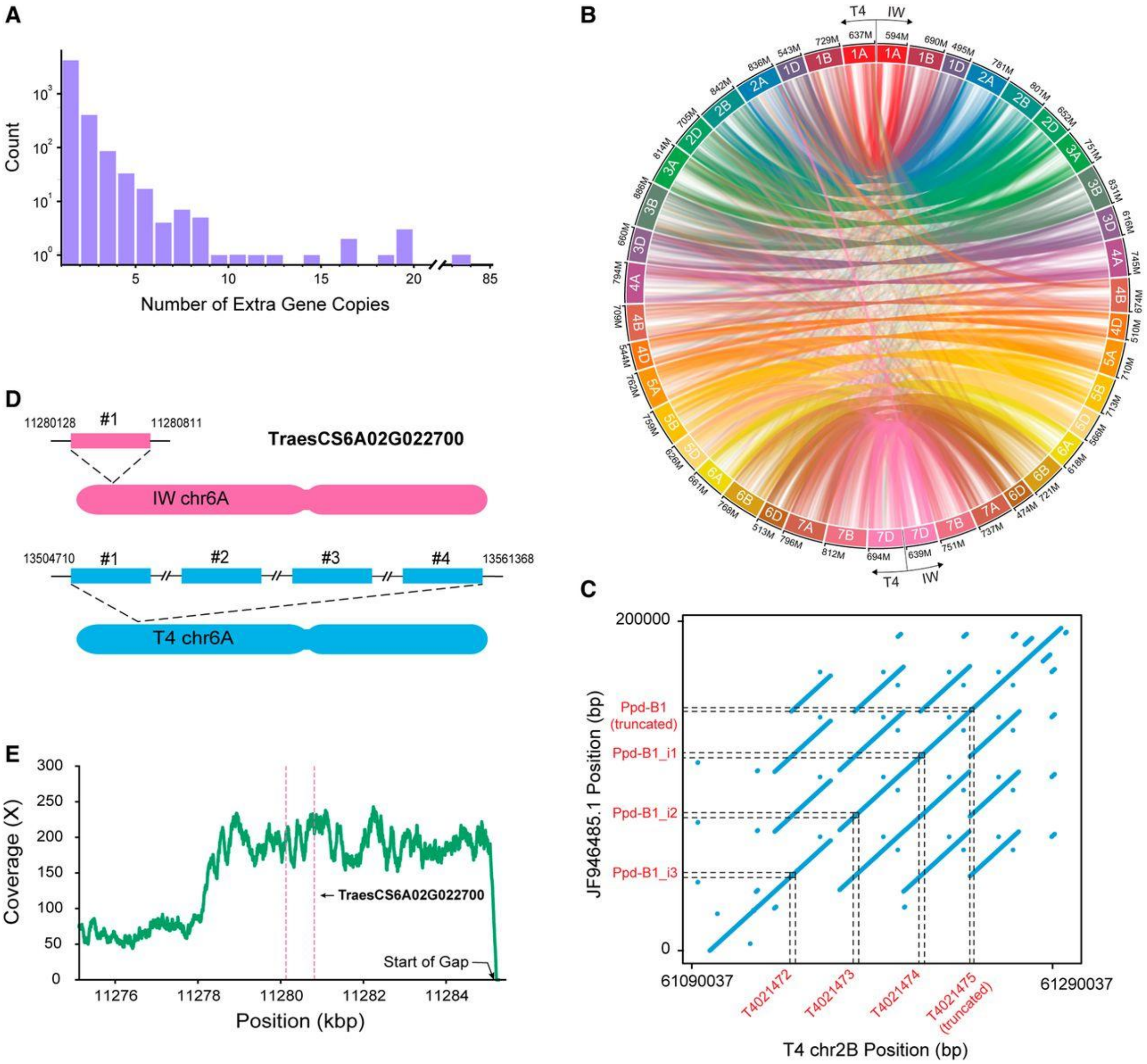


Figure 3.4: Triticum_aestivum_4.0 resolves previously collapsed genic repeats. (A) Histogram depicting the distribution of the number of additional gene copies found in Triticum_aestivum_4.0. (B) Circos plot showing the locations of all additional gene copies (http://omgenomics.com/circa/). Lines are drawn from the location of the gene in IWGSC CS v1.0 (IW) on the right half of the diagram to the location of each copy in Triticum_aestivum_4.0 (T4) on the left half. (C) Dotplot depicting maximal exact matches (MEMs) between T4 *Ppd-B1* (x-axis) and a publicly available Chinese Spring *Ppd-B1* sequence (GenBank accession JF946485.1) (y-axis). Dashed lines indicate the colinear positions of four PRR genes (red labels). (D) Diagram of the MADS-box transcription factor gene, *TraesCS6A02G022700*, present in three additional tandem copies in T4 as relative to IW. Ideograms are not drawn to scale. (E) Plot of the short-read coverage in IW starting 5 kb upstream of *TraesCS6A02G02270* and extending to the first gap downstream of the gene. The pink dashed lines show the location of the gene.

Triticum_aestivum_4.0 accurately represents gene duplications affecting traits

We searched T4 for specific examples of functionally relevant gene duplications previously collapsed or missing in IW. We focused on the *Ppd-B1* locus on chr2B because copy number variation of PRR genes at this locus underlies variation in photoperiod sensitivity among hexaploid wheat varieties [17]. Others have shown that the Chinese Spring variety has four PRR genes at the *Ppd-B1* locus, with one of the copies being truncated [18]. Because the entire ~200 kbp Chinese Spring *Ppd-B1* locus was previously cloned and sequenced, we were able to assess if this region had been accurately assembled in both T4 and IW. IW lacks any PRR genes at the *Ppd-B1* locus, with fragments of three of the four expected paralogs (*TraesCSU02G196100*, *TraesCSU02G221500*, *TraesCSU02G199500*) residing on unplaced chrUn sequence. In contrast, T4 localizes four PRR genes (*T4021472*, *T4021473*, *T4021474*, and *T4021475*) at *Ppd-B1*, matching the expected Chinese Spring copy number state. Alignment of this T4 locus to the known Chinese Spring *Ppd-B1* sequence indicated that the entire locus had been accurately assembled, even correctly representing the three, highly similar, intact PRR genes (**Figure 3.4**). The successful assembly of *Ppd-B1* served as a validation that T4 accurately resolves duplications with high sequence similarity.

The successful resolution of the *Ppd-B1* locus suggested that new functionally relevant CNVs may be discovered among the large number of localized or duplicated genes in T4. One notable example was a MADS-box transcription factor gene, *TraesCS6A02G022700*, which had three additional tandem copies (T4 genes *T4081597*, *T4081598*, *T4081599*, and *T4081600*) on T4 chr6A (**Figure 3.4**). MADS-box transcription factors are known to influence traits such as flowering time and floral organ development [19,20]. Furthermore, MADS-box gene duplications can quantitatively impact gene expression and domestication phenotypes in a dosage-dependent manner [21]. To provide further evidence that this gene is part of a collapsed repeat in IW, we aligned Chinese Spring Illumina reads to IW and calculated the coverage across the gene ±50 kbp of flanking sequence. We observed a spike in coverage indicating a collapsed repeat in IW containing *TraesCS6A02G022700* (**Figure 3.4**). We further note that this region contains 10,205 bp of gap sequence, suggesting that this locus had been misassembled in IW. This duplication of a MADS-box transcription factor gene, as well as our analysis of the *Ppd-B1* locus, highlights how T4, with its superior genome completeness, resolves functionally relevant genic sequence previously misassembled, missing, or unlocalized in IW.

3.4 DISCUSSION

In one critical aspect, the bread wheat genome exemplifies the challenge of eukaryotic genome assembly. Repeats, which remain difficult to assemble, are pervasive in this transposon-rich allohexaploid plant genome. Therefore, the accurate and complete resolution of the bread wheat genome and the subsequent study of genomic structure especially depends on high-quality data and advanced genome assembly techniques. In 2017, we published the first near-complete and highly contiguous representation of the bread wheat genome (Triticum_aestivum_3.1), demonstrating the value of long reads for wheat genome assembly [7]. In our efforts described here, we used Triticum_aestivum_3.1 as our foundation, while leveraging the strengths of the IWGSC CS v1.0 reference genome to establish the most complete chromosome-scale and gene-annotated reference assembly yet created for bread wheat. By scaffolding and annotating our contigs, we created the genomic context needed to quantify and qualify the completeness of the Triticum_aestivum_4.0 assembly, especially relative to its predecessors. Compared to the IWGSC CS v1.0 assembly, Triticum_aestivum_4.0 resolves more repeat sequence, exemplified by the improved centromere localization and by the many additional gene copies. The discovery of these extra gene copies, as well as the localization of 2001 previously unplaced genes, also demonstrates how Triticum_aestivum_4.0 provides an enhanced representation of Chinese Spring genic sequence.

Gene CNVs are pervasive in hexaploid wheat and are associated with traits such as frost tolerance (Fr-A2), vernalization requirement (Vrn-A1), and photoperiod sensitivity (*Ppd-B1*) [18,22–24]. These and other CNVs contributed to the adaptive success of domesticated wheat, which now thrives in diverse conditions and geographies. This is exemplified by the *Ppd-B1* locus, where variation of PRR gene copy number influences photoperiod sensitivity. Our successful assembly of the *Ppd-B1* locus, which was unanchored and incomplete in IWGSC CS v1.0, highlights a specific example where our improved assembly accurately reflected a known CNV genotype in Chinese Spring. This validation suggests that other functional gene duplications may also be directly encoded in the Triticum_aestivum_4.0 assembly and identifiable by our annotation of extra gene copies. We indicated one such potential candidate, the MADS-box transcription factor gene, which appears with three extra copies in Triticum_aestivum_4.0. We expect that further investigation of the extensive gene duplications presented in this work will provide additional insights into the role of CNVs in wheat phenotypes.

Structural variants (SVs), including CNVs, comprise a vast source of natural genetic variation influencing traits. As sequencing technologies continue to advance, plant scientists are increasingly using pan-genome analyses to study genome structure among diverse varieties and ecotypes [25–27]. These studies rely especially on structurally accurate reference genomes to discover SVs. Our work introduces Triticum_aestivum_4.0 as an improved reference genome resource ideal for future structural variant analyses in wheat. Furthermore, our comparative genomics analysis showed that a substantial portion of the Chinese Spring genome was collapsed, missing, or misrepresented when assembled with short reads. This emphasizes the utility of long reads in future wheat pan-genome analyses, where structural accuracy is key. Generally, our work provides a preview of the computational genomics analyses that are possible with an accurate wheat reference genome.

3.5 METHODS

Establishing the initial contig set

We first sought to establish the most complete set of contigs representing the genome of *T. aestivum* Chinese Spring. We started with the Triticum_aestivum_3.1 contigs (T3) [7] because they comprise 1 Gbp of additional nongap sequence compared to the IWGSC CS v1.0 (IW) reference assembly. However, when establishing a set of contigs for downstream scaffolding, we wanted to ensure that we incorporated any contigs unique to the reference assembly, and, therefore, "missing" from the T3 assembly. To do this, we broke the reference assembly into "contigs" by breaking pseudomolecules at gaps (at least 20 "N" characters). We then aligned these reference contigs (query) to the T3 contigs (reference) using NUCmer (-l 250 -c 500) and filtered them using delta-filter (-1 -l 5000) to include only reciprocal best alignments at least 5 kbp long [10]. Of the reference contigs that were at least 10 kbp in length, if under 25% of a contig was covered by alignments, it was deemed a putative "missing" contig.

We then checked to see if these putative missing contigs would indeed be covered by alignments produced with more sensitive parameters. The putative missing contigs (query) were aligned again to the T3 assembly with NUCmer, but with a smaller minimum seed and cluster size (-l 50 -c 200). Alignments were filtered as before, and, if under 25% of a putative missing contig was covered by these more sensitive alignments, they were deemed to be validated as missing from T3. These validated missing IW contigs were combined with the T3 contigs to establish our final set of contigs for downstream scaffolding, which had an N50 length of 230,687 bp and a sum of 15,429,603,425 bp.

RaGOO scaffolding

We performed two rounds of reference-guided scaffolding with RaGOO. We first used RaGOO to look for false sequence duplications, especially those that could have arisen by incorporating "missing" IW contigs. Though RaGOO usually employs Minimap2 [9] to align query contigs to a reference genome, we used NUCmer to produce high-specificity alignments. We aligned our contigs (query) to the IW reference genome (reference) using a very large seed and cluster size (-l 500 -c 1000). Such specificity in alignments was necessary to unambiguously order and orient contigs with respect to the highly repetitive allohexaploid reference genome. The resulting delta file was converted to PAF format using Minimap2's paftools. Next, we ran RaGOO using these alignments rather than the default Minimap2 alignments while also specifying a minimum clustering confidence score of 0.4 (-i). We also excluded any unanchored IW sequence from consideration (-e).

To remove false duplication of missing contig sequence, we observed that such duplications would align to more than one place in these RaGOO pseudomolecules. Conversely, contigs that were truly "missing" should only align once (perfectly) to their ordered and oriented location in the RaGOO scaffolds. We aligned the RaGOO scaffolds (query) to the missing IW contigs (reference) with NUCmer (-l 50 -c 200) and filtered alignments with delta-filter (-q -l 5000) [11]. If a missing contig had more than one alignment with coverage at least 50% and percent identity at least 98%, it was deemed to be a false duplicate and removed from the initial contig set. With false duplicates removed, we proceeded with the second round of RaGOO scaffolding which had all of the same specifications as the first round.

We next sought to remove any unanchored contigs that had duplicated sequences among the anchored contigs. The same previously described process to remove false duplicates was also used here, except that the RaGOO scaffolds along with unanchored contigs (query) were aligned to the unanchored contigs (reference). Also, the minimum coverage was 75% rather than 50%. After removing these unanchored duplications, scaffolds were polished with POLCA [12]. For polishing, we used the Illumina reads from the NCBI SRA accession SRX2994097. POLCA introduced 595,705 bp in substitution corrections and 1,033,593 bp in insertion/deletion corrections. After polishing, the final error rate of the sequence was estimated at <0.008% or <1 error per 10,000 bases. Finally, we removed any redundant mitochondria and chloroplast sequences from unplaced contigs, thus resulting in the final Triticum_aestivum_4.0 (T4) assembly. T4/IW dotplots were made by aligning the polished T4 assembly (query) to the IW reference assembly (reference) with NUCmer (-l 500 -c 1000). Alignments <10 kbp were removed with delta-filter and were plotted with mummerplot (–fat–layout).

Shared k-mer frequency distribution

Groups of 101-mers were counted in T4 and IW using KMC (v3.1.0, -ci1 -cx10000 -cs10000) [28]; 101-mers shared by T4 and IW were then extracted with kmc_tools "simple" using the intersection function. The 101-mer copy frequency distribution of these shared k-mers in both T4 and IW (-ocleft and -ocright) was then plotted in **Figure 3.3**.

Centromere annotation

We annotated centromere sequence in T4 using an approach similar to the original IW publication [4]. First, publicly available Chinese Spring CENH3 ChIP-seq data (SRR1686799) was downloaded from the European Nucleotide Archive [14]. Reads were then trimmed with cutadapt (v1.18, -a AGATCGGAAGAG) and aligned to T4 with bwa mem (v0.7.17-r1198-dirty) [29,30]. Alignments with a mapq score <20 were removed and the remaining alignments were compressed and sorted with samtools view and samtools sort respectively [31]. Alignments were then counted in 100 kbp nonoverlapping windows along the T4 genome using bedtools makewindows and bedtools coverage (v2.29.2) [32]. Any group of two or more consecutive windows with at least three times the genomic average coverage was considered putative centromere sequence, and any such intervals within 500 kbp were merged. These intervals were further merged or removed by comparing them manually with the CENH3 ChIP-seq alignments, resulting in a single inferred centromere annotation for each chromosome. Some IW chromosomes have more than one centromeric position reported in the original IW publication. Accordingly, we picked the longest centromeric interval for each IW chromosome for the comparative analysis presented in this work.

Chloroplast and mitochondria genome assembly

We took the first 20 million Illumina read pairs from the SRR5815659 accession and assembled them with megahit (v1.2.8) [33]. The resulting assembly contained 145,887 contigs (74.41 Mb) with lengths ranging between 200 bp and 56,565 bp. Then we aligned these contigs to the *T. aestivum* reference chloroplast sequence (NC_002762.1) using NUCmer (with -maxmatch switch to align to repeats) and filtered the alignments with delta-filter, keeping the best hits to the reference NC_002762.1. The reference was covered completely by alignments of only five contigs. Then, we aligned these contigs to each other with NUCmer (-maxmatch –nosimplify) and used the alignments to manually order and orient them into a single chloroplast sequence scaffold.

To assemble the mitochondrial genome, we aligned the megahit contigs discussed above to the *T. aestivum* mitochondria reference sequence (MH051716) with NUCmer (-–maxmatch). We then filtered the alignments with delta-filter, keeping the best matches to the MH051716 reference. This revealed 43 nonchloroplast contigs of least 500 bp in length that matched best to the mitochondria reference. We then ordered and oriented these 43 contigs using RaGOO (v1.1), setting the minimum alignment length to 500 bp. The chloroplast and mitochondria sequence are included in our data submission to NCBI.

Genome annotation

We used Liftoff to annotate the T4 genome using the IW v1.1 gene models [15]. Genes were aligned to their same chromosome in T4 using BLASTN v.2.9.0 (-soft_masking False -dust no -word_size 50 -gap_open 3 -gapextend 1

-culling_limit 10). The blast hits were filtered to include only those that contained one or more exons. For each gene, the optimal exon alignments were chosen according to sequence identity and concordance with the exon/intron structure of the gene model in IW. These alignments were used to define the boundaries of each exon, transcript, and gene in T4. We excluded any transcripts that did not map with at least 50% alignment coverage. Any genes without at least one mapped isoform were then aligned against the entire T4 genome using BLASTN with the same parameters and placed given they did not overlap an already placed gene.

To place the chrUn genes, we aligned the genes to the entire T4 genome using the same parameters. We excluded any transcripts that did not meet the 50% alignment coverage threshold or overlapped an already annotated gene.

To find additional gene copies, we aligned all genes (query) to the complete T4 genome (reference) using BLASTN v2.9.0 (-soft_masking False -dust no -word_size 50 -gap_open 3 -gapextend 1 -culling_limit 100, qcov_hsp_perc 100). The notable differences in these parameters are qcov_hsp_perc, which requires 100% query coverage, and culling_limit, which has been increased from 10 to 100 to increase the number of reported alignments for genes with a highly increased copy number. We excluded any alignments that did not have 100% exonic sequence identity or overlapped a previously placed gene. We used gffread to filter out genes with noncanonical splice sites [34].

Finally, using the same methods as described for high confidence genes above, we also used Liftoff to map the IW v1.1 low-confidence annotation onto T4. We successfully mapped 152,900 out of 161,537 low-confidence genes. Another 1581 genes mapped partially below the 50% alignment coverage threshold.

Ppd-B1 haplotype comparison

To find the approximate location of the *Ppd-B1* locus in the T4 and IW assemblies, we aligned a *Ppd-B1* PRR gene sequence (GenBank accession DQ885757.1) to T4 and IW with blastn v2.6.0 (-perc_identity 95) [17]. No matches were found on IW chr2B, though partial matches were found on chrUn. In contrast, four strong matches were found on T4 chr2B, corresponding to genes *T4021472*, *T4021473*, *T4021474*, and *T4021475*. We also aligned the entire Chinese Spring haplotype for this locus, which had been previously cloned and sequenced (GenBank accession JF946485.1), to T4 using blastn v2.6.0 (-perc_identity 95) [18]. We used these alignments to approximately define the genomic coordinates of *Ppd-B1* in T4. To further validate the accuracy of this locus in T4, we aligned the GenBank JF946485.1 sequence to the T4 locus ±10 kbp flanking sequence to find pairwise maximal exact matches (MEMs) at least 50 bp in length. These alignments are depicted in **Figure 3.4** and were generated with mummer v3.23 (-maxmatch -l 50 -b -c). Before alignment, the GenBank JF946485.1 sequence was reverse complemented to refer to the same strand as our T4 chr2B.

Because the PRR gene annotations used to define T4 *Ppd-B1* PRR genes were incomplete in IW, they were also initially incomplete in T4. To correctly annotate these T4 PRR genes, we used Liftoff to lift-over the GenBank JF946485.1 PRR gene annotations to T4. These genes are labeled *T4021472*, *T4021473*, *T4021474*, and *T4021475* in the final annotation.

3.6 ACKNOWLEDGEMENTS

This work was supported in part by the National Institutes of Health (NIH) under grants R01-HG006677 and R35-GM130151, and by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture under grant 2018-67015-28199.

3.7 REFERENCES

1. Petersen G, Seberg O, Yde M, Berthelsen K. Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (Triticum aestivum). Mol Phylogenet Evol. 2006;39:70–82.

2. Dubcovsky J, Dvorak J. Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. Science. 2007;316:1862–6.

3. Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol Biol Rep. Springer Science and Business Media LLC; 1991;9:208–18.

4. International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators:, Appels R, Eversole K, Feuillet C, Keller B, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science [Internet]. science.sciencemag.org; 2018;361. Available from: http://dx.doi.org/10.1126/science.aar7191

5. Chapman JA, Mascher M, Buluç A, Barry K, Georganas E, Session A, et al. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol. 2015;16:26.

6. Clavijo BJ, Venturini L, Schudoma C. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for .... Genome [Internet]. genome.cshlp.org; 2017; Available from: https://genome.cshlp.org/content/27/5/885.short

7. Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. Gigascience. 2017;6:1–7.

8. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20:224.

9. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

10. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

11. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018;14:e1005944.

12. Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol. 2020;16:e1007981.

13. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res. 2010;20:1165–73.

14. Guo X, Su H, Shi Q, Fu S, Wang J, Zhang X, et al. De Novo Centromere Formation and Centromeric Sequence Expansion in Wheat and its Wide Hybrids. PLoS Genet. 2016;12:e1005997.

15. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics [Internet]. 2020; Available from: http://dx.doi.org/10.1093/bioinformatics/btaa1016

16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

17. Beales J, Turner A, Griffiths S, Snape JW, Laurie DA. A pseudo-response regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (Triticum aestivum L.). Theor Appl Genet. Springer; 2007;115:721–33.

18. Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA. Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). PLoS One. journals.plos.org; 2012;7:e33234.

19. Coen ES, Meyerowitz EM. The war of the whorls: genetic interactions controlling flower development. Nature. nature.com; 1991;353:31–7.

20. Ng M, Yanofsky MF. Function and evolution of the plant MADS-box gene family. Nat Rev Genet. nature.com; 2001;2:186–95.

21. Soyk S, Lemmon ZH, Sedlazeck FJ, Jiménez-Gómez JM, Alonge M, Hutton SF, et al. Duplication of a domestication locus neutralized a cryptic variant that caused a breeding barrier in tomato. Nat Plants. nature.com; 2019;5:471–9.

22. Würschum T, Boeven PHG, Langer SM, Longin CFH, Leiser WL. Multiply to conquer: Copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. BMC Genet. Springer; 2015;16:96.

23. Würschum T, Longin CFH, Hahn V, Tucker MR, Leiser WL. Copy number variations ofCBFgenes at theFr-A2locus are essential components of winter hardiness in wheat. Plant J. Wiley; 2017;89:764–73.

24. Würschum T, Langer SM, Longin CFH, Tucker MR, Leiser WL. A three-component system incorporating Ppd-D1, copy number variation at Ppd-B1, and numerous small-effect quantitative trait loci facilitates adaptation of heading time in winter wheat cultivars of worldwide origin. Plant Cell Environ. Wiley Online Library; 2018;41:1407–16.

25. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. Elsevier; 2020;182:145–61.e23.

26. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-Genome of Wild and Cultivated Soybeans. Cell. Elsevier; 2020;182:162–76.e13.

27. Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants. nature.com; 2020;6:34–45.

28. Kokot M, Dlugosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 2017;33:2759–61.

29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

30. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.

31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

33. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.

34. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020;9:304.

[The] long-range sequence information, combined with our optimized assembly and validation pipeline, yielded a nearly closed and highly accurate assembly of the Col-0 genome.

4

The genetic and epigenetic landscape of the Arabidopsis centromeres

The following sections were previously published as a preprint in *bioRxiv* and accepted for publication in *Science*:

> Naish, M.*, Alonge, M.*, Wlodzimierz, P.*, Tock, A.J., Abramson, B.W., Lambing, C., Kuo, P., Yelina, N., Hartwick, N., Colt, K., Kakutani, T., Martienssen, R.A., Bousios, A., Michael, T.P., Schatz, M.C.†, and Henderson, I.R†. The genetic and epigenetic landscape of the Arabidopsis centromeres. Science (2021). In Press. (bioRxiv version available: https://doi.org/10.1101/2021.05.30.446350)

This work is a collaboration with Ian Henderson's lab at the University of Cambridge. M.N., M.A., and P.W. contributed equally to this work. M.C.S. and I.R.H. supervised all analysis and manuscript preparation. M.N., M.A., P.W., M.C.S., and I.R.H. edited all of the text. The background was written by M.N., M.A., P.W., M.C.S., and I.R.H. The results section "Complete assembly of the Arabidopsis centromeres" was done by M.N. and M.A. and written by M.A. Figure 4.1A was made by M.A. with help from A.J.T., A.B., and I.R.H. Figures 4.1B-C were made by M.A. The methods sections "Contig screen", "Contig scaffolding", "Pseudomolecule polishing and gap filling", "Telomere patching", "Assembly curation and preparation", and "Genome annotation" were done and written by M.A. All other results and methods sections, as well as the abstract and discussion, were done and written by authors other than M.A.

4.1 ABSTRACT

Centromeres attach chromosomes to spindle microtubules during cell division and, despite this conserved role, show paradoxically rapid evolution and are typified by complex repeats. We used ultra-long-read sequencing to generate the Col-CEN Arabidopsis thaliana genome assembly that resolves all five centromeres. The centromeres consist of megabase-scale tandemly repeated satellite arrays, which support high CENH3 occupancy and are densely DNA methylated, with satellite variants private to each chromosome. CENH3 preferentially occupies satellites with least divergence and greatest higher-order repetition. The centromeres are invaded by ATHILA retrotransposons, which disrupt genetic and epigenetic organization of the centromeres. Crossover recombination is suppressed within the centromeres, yet low levels of meiotic DSBs occur that are regulated by DNA methylation. We propose that Arabidopsis centromeres are evolving via cycles of satellite homogenization and retrotransposon-driven diversification.

4.2 BACKGROUND

Despite their conserved function during chromosome segregation, centromeres show diverse organization between species, ranging from single nucleosomes to megabase-scale tandem repeat arrays [1]. Centromere 'satellite' repeat monomers are commonly ~100–200 bp, with each repeat capable of hosting a CENPA/CENH3-variant nucleosome [1,2]. CENPA/CENH3 nucleosomes ultimately assemble the kinetochore and position spindle attachment on the chromosome, allowing segregation during cell division [3]. Satellites are highly variable in sequence composition and length when compared between species [2]. The library of centromere repeats present within a genome often shows concerted evolution, yet they have the capacity to change rapidly in structure and sequence within and between species [1,2,4]. However, the genetic and epigenetic features that contribute to centromere evolution are incompletely understood, in large part due to the challenges of centromere sequence assembly and functional genomics of highly repetitive sequences.

De novo assembly of repetitive sequences is challenging. As such, most eukaryotic genome assemblies are in a fragmented state with many repetitive regions completely unresolved, especially the centromeres and other large repeats. Even the most scrutinized genomes, such as the human GRCh38 and Arabidopsis TAIR10 reference genomes, fail to represent centromeres and other large repeats of biological importance. However, recent advances in long-read sequencing, including Oxford Nanopore (ONT) and PacBio single-molecule technologies, have revolutionized the field by enabling substantially more complete and contiguous genome assemblies. Owing to their increased length and accuracy (10 kbp to >100 kbp with 90–99% mean accuracy), the long reads are capable of spanning and assembling repetitive sequences that are too ambiguous to resolve with previous sequencing technologies. Notably, using these technologies, the highly repetitive human centromeres have recently been assembled, leveraging the fact that sequence heterogeneity exists between the satellite repeats, effectively creating regularly spaced unique sequence markers [5–10]. As such, given sufficiently long reads, a genome assembler can effectively bridge from one unique marker to the next, thereby creating a reliable and unambiguous reconstruction. This core concept, combined with more accurate base-calling and consensus generation, is now leading to highly accurate and complete representations of complex genomes for the first time [5,11].

The *Arabidopsis thaliana* genome was first sequenced in 2000, yet the centromeres, telomeres, and ribosomal DNA repeats have remained unassembled, due to their high repetition and similarity [12]. The Arabidopsis centromeres are known to contain millions of base pairs of the *CEN180* satellite repeat, which support CENH3 loading [13–17]. We used ultra-long-read DNA sequencing to establish the Col-CEN reference assembly, which wholly resolves all five Arabidopsis centromeres. The assembly contains a library of 66,129 *CEN180* satellites, with each chromosome possessing largely private satellite variants. Higher-order *CEN180* repetition is prevalent within the centromeres and is also chromosome specific. We identify *ATHILA* LTR retrotransposons that have invaded the satellite arrays and interrupt centromere genetic and epigenetic organization. By analyzing functional data from mutant lines, we demonstrate that DNA methylation epigenetically silences initiation of meiotic DNA double-strand breaks (DSBs) within the centromeres. Together, our data are consistent with satellite homogenization and retrotransposon invasion driving cycles of centromere evolution in Arabidopsis.

4.3 RESULTS

Complete assembly of the Arabidopsis centromeres

The *Arabidopsis thaliana* TAIR10 reference genome is an exceptionally accurate and complete eukaryotic assembly that is an invaluable resource for plant science [12]. However, TAIR10 fails to represent the telomeres, some rDNAs and the centromere satellite arrays. To resolve these remaining sequences, we supplemented existing TAIR10 genomic resources with Oxford Nanopore (ONT) sequencing data from Columbia (Col-0) genomic DNA comprising a total of 73.6 Gbp, and ~55× coverage of ultra-long (>50 kbp) reads. This long-range sequence information, combined with our optimized assembly and validation pipeline, yielded a nearly closed and highly accurate assembly of the Col-0 genome (Col-CEN v1.0). Chromosomes 1 and 3 are wholly resolved from telomere-to-telomere (T2T), chromosomes 2 and 4 are complete apart from the *45S* clusters and adjacent telomeres, and a single gap remains on chromosome 5 (**Figure 4.1**).

After repeat-aware polishing with R9 and R10 ONT reads and selective short-read polishing, the Col-CEN assembly is highly accurate with a QV of 33.95 and 45.58 inside and outside of the centromeres, equivalent to an error rate of 1 in 27,696 and 1 in 63,529 nucleotides, respectively. The assembly is highly concordant with TAIR10, with 95.53% of Columbia BAC contigs aligning with high coverage and identity (>95%), and 99.61% of TAIR10 gene annotations represented in the assembly (97.49% of genes are exactly represented) (**Figure 4.1B**). The Col-CEN assembly includes the *5S* rDNA arrays on chromosomes 3, 4, and 5, as well as a large mitochondrial genome insertion on chromosome 2 (**Figure 4.1A and 4.1C**). Furthermore, the assembly reconstructs all five centromeres spanning 11,787,742 bp of new sequence, 120 and 98 kbp of *45S* rDNA in the chromosome 2 and 4 Nucleolar Organizing Regions (NORs), and the complete telomeres of the 8 chromosome arms without sub-telomeric NORs (**Figure 4.1A–4.1C**). We also identified a thionin gene cluster that was discordant with TAIR10 and after validating the structural accuracy of this locus in Col-CEN, we hypothesize that this may represent a TAIR10 misassembly or a recent structural variant in our Col-0 line.

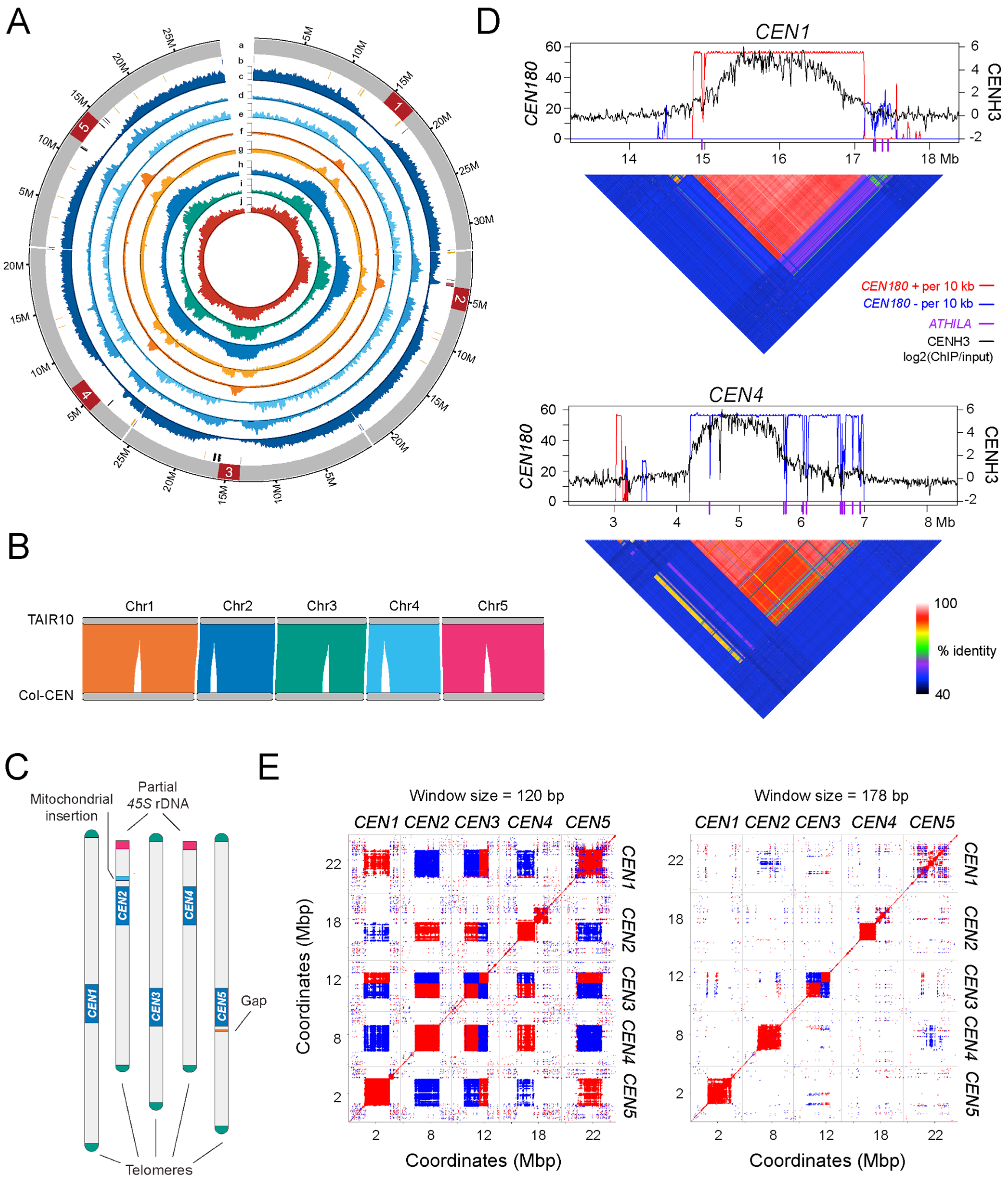


Figure 4.1: Complete assembly of the Arabidopsis centromeres. (A) Genome-wide circos plot of the Col-CEN assembly. Quantitative tracks (c-j) are aggregated in 100 kbp bins and independent y-axis labels are given as (low tick value, mid tick value, high tick value, unit of measurement): (a) chromosome labels with centromeres shown in red; (b) genomic features showing telomeres in blue, *45S* rDNA in yellow, *5S* rDNA in black, and the Chr2 mitochondrial insertion in pink; (c) genes (0, 25, 51, # of genes); (d) transposable elements (0, 85, 171, # of transposable elements); (e) Col×Ler F_2 crossovers (0, 6, 12, # of crossovers); (f) CENH3 (-1, 0, 3, $\log_2$(ChIP/Input)); (g) H3K9me2 (-1, 0, 2, $\log_2$(ChIP/Input)); (h) CpG methylation (0, 40, 80, % methylated); (i) CHG methylation (0, 20, 40, % methylated); (j) CHH methylation (0, 4, 8, % methylated). (B) Plot showing syntenic alignments between the TAIR10 and Col-CEN assemblies. (C) Genome assembly ideogram with annotated chromosome landmarks (not drawn to scale). (D) CENH3 $\log_2$(ChIP/Input (black) plotted over centromeres 1 and 4 [13]. *CEN180* density per 10 kbp is plotted for forward (red) or reverse (blue) strand orientations. *ATHILA* retrotransposons are indicated by purple ticks on the x-axis. Beneath are heatmaps showing pairwise % sequence identity values of adjacent 5 kbp regions. (E) Dotplot analysis comparing the 5 centromere regions, using a search window of 120 or 178 bp. Red and blue shading indicate detection of similarity on the same or opposite strands, respectively.

The assembled centromere sequences are characterized by a repeated 178-bp motif (*CEN180*) that is organized into higher-order repeats (HORs) (**Figure 4.1D, 4.2**). We validated the structural and base-level accuracy of the centromeres using techniques from the Human T2T consortium [5]. Briefly, we aligned our Col-0 ONT reads to the assembly and observed even coverage across the centromeres, with few loci showing plausible alternate base signals. We also observed relatively few 'missing' *k*-mers that are found in the assembly but not in Illumina short reads, which are diagnostic of residual consensus errors from the ONT reads [18]. Notably, the five centromeres are relatively distinct at the sequence level, with each exhibiting chromosome-specific repeats (**Figure 4.1E, 4.2**. This is consistent with our assembly pipeline unambiguously separating the five centromere sequences. We observe that unique 'marker' sequences are relatively frequent, with a maximum distance between consecutive markers in the assembled centromeres of only 28,630 bp, suggesting that our ultra-long reads can confidently span several unique markers and thus reliably assemble centromeric loci.

The Arabidopsis CEN180 satellite repeat library

We performed *de novo* searches for tandem repeats to define the centromere satellite library. We identified 66,129 *CEN180* satellites in total, with between 11,847 and 15,612 copies per chromosome (**Figure 4.2**). The *CEN180* repeats form large stranded arrays, with the exception of centromere 3, which has an inverted structure (**Figure 4.1D**). The length of the repeat monomers is tightly constrained around 178 bp (**Figure 4.2A**). We aligned all unique *CEN180* sequences (*n*=25,192) to derive a genome-wide satellite consensus. Each satellite was then compared to the consensus to calculate a single-nucleotide variant (SNV) score. Substantial sequence variation was observed between satellites, with a mean of 19.6 SNVs per *CEN180* **(Figure 4.2A)**. Each centromere shows essentially private libraries of *CEN180* monomer sequences, with only 0.5% sharing an identical copy on a different chromosome (**Figure 4.1E**). In contrast, there is a high degree of *CEN180* repetition within chromosomes, with 54.2–65.4% showing one or more duplicates. We also observed a minor class of '*CEN160*' tandem repeats found mainly on chromosome 1 (1,289 repeats on Chr1, 43 repeats on Chr4, mean length=158.2 bp) [17].

We aligned CENH3 ChIP-seq data to the assembly and observed on average 10-fold $\log_2$(ChIP/input) enrichment within the *CEN180* arrays, compared to the chromosome arms (**Figure 4.1D**) [13]. CENH3 ChIP-seq enrichment is generally highest in the interior of the main *CEN180* arrays (**Figure 4.1D**). We observed a negative relationship between CENH3 ChIP-seq enrichment and *CEN180* SNV divergence (**Figure 4.2D–4.2E**), consistent with CENH3 nucleosomes preferring to occupy satellites that are closer to the genome-wide consensus. In this respect, centromere 4 is noteworthy, as it consists of two distinct *CEN180* arrays, with the right array showing both higher SNV divergence and lower CENH3 ChIP-seq enrichment (**Figure 4.1D and 4.2D**). Together, this is consistent with satellite divergence leading to loss of CENH3 binding, or vice versa.

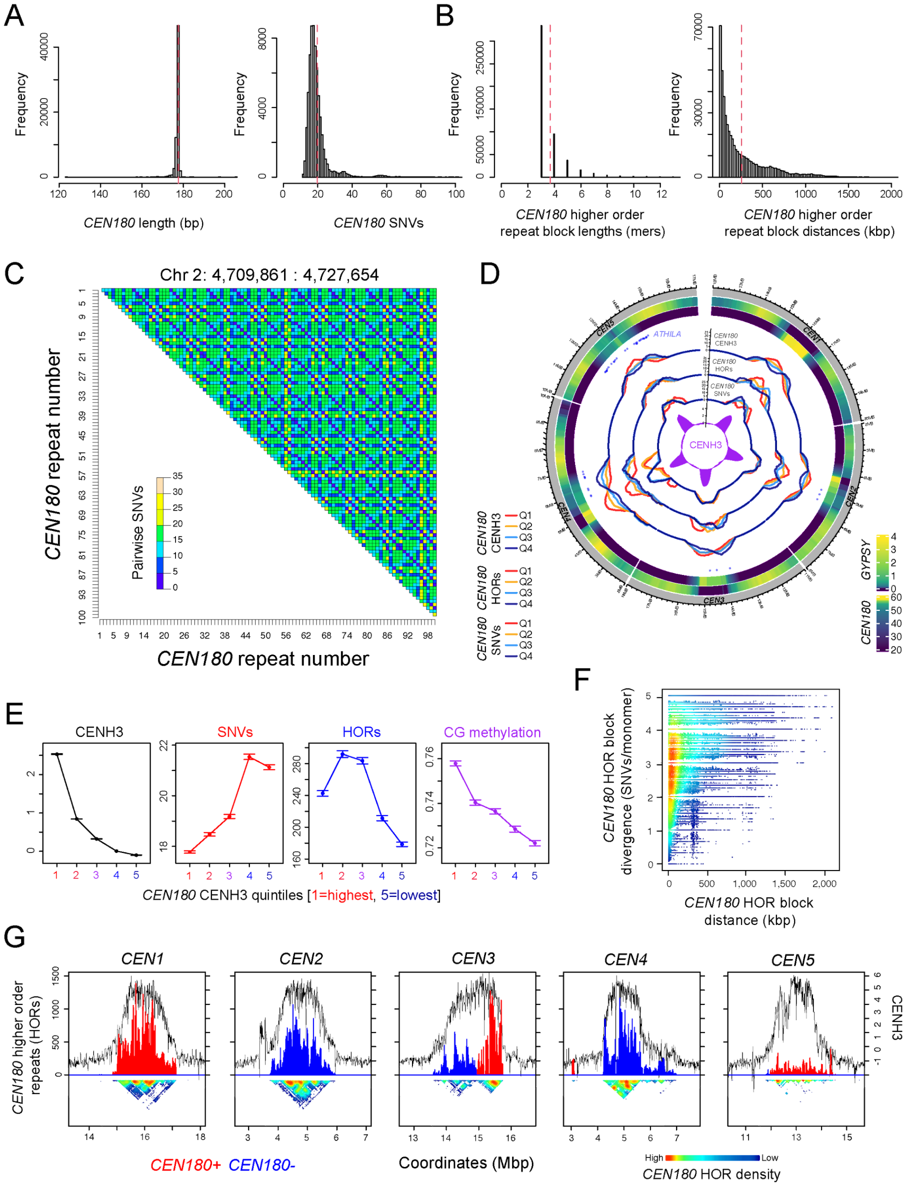


Figure 4.2: The Arabidopsis *CEN180* satellite repeat library. (A) Histograms of *CEN180* monomer lengths (bp), and single nucleotide variants (SNVs) relative to the genome-wide consensus. Mean values are shown by the red dotted line. (B) As for A, but showing widths of *CEN180* higher order repeat (HOR) blocks (monomers, 'mers'), and the distance between HOR blocks (kbp). (C) Heatmap of a representative satellite region within centromere 2, shaded according to pairwise SNVs between *CEN180*. (D) Circos plot showing; (i) *GYPSY* LTR transposon density, (ii) *CEN180* density, (iii) centromeric *ATHILA* rainfall plot, (iv) *CEN180* density grouped by decreasing CENH3 $\log_2$(ChIP/input) (red=high; navy=low), (v) *CEN180* density grouped by decreasing higher order repetition (red=high; navy=low), (vi) *CEN180* grouped by decreasing SNVs (red=high; navy=low) and, (vii) CENH3 $\log_2$(ChIP/input), across the centromere regions. (E) *CEN180* were divided into quintiles according to CENH3 $\log_2$(ChIP/input) and mean values for each group with 95% confidence intervals plotted. The same groups were analyzed for *CEN180* SNVs (red), higher order repetition (blue) and CG context DNA methylation (purple). (F) Plot of the distance between pairs of HOR blocks (kbp) and divergence (SNVs/monomers) between the HOR block sequences. (G) Plots of CENH3 $\log_2$(ChIP/input) (black) across the centromeres, compared to *CEN180* higher order repetition on forward (red) or reverse (blue) strands. A heat map is shown beneath that is shaded according to the density of higher order repeats.

To define *CEN180* higher-order repeats (HORs), monomers were considered the same if they shared 5 or fewer pairwise SNVs. Consecutive repeats of at least 3 monomers below this SNV threshold were identified, yielding 500,833 HORs **(Figure 4.2D)**. Like the *CEN180* monomer sequences, HORs are almost exclusively chromosome specific. The mean number of *CEN180* monomers per HOR was 3.69, equivalent to 656 bp (**Figure 4.2B**), and 91.1% of *CEN180* were part of at least one HOR. HOR block sizes show a negative exponential distribution, with the largest HOR formed of 60 monomers on chromosome 3, equivalent to 10,691 bp (**Figure 4.2B**). Many HORs are in close proximity (42% are <100 kbp apart), although they are distributed along the length of the centromeres. For example, the average distance between HOR blocks was 250.7 kbp and the maximum distance was 2.1 Mbp **(Figure 4.2B)**. We also observed that HOR blocks that were a greater distance apart showed a higher level of SNVs between the blocks (SNVs/monomer) **(Figure 4.2F)**, which is consistent with satellite homogenization being more effective over repeats that are physically closer. The *CEN180* groups with the highest CENH3 occupancy also show the greatest level of higher-order repetition and higher CG DNA methylation frequency **(Figure 4.2D–4.2E and 4.2G)**. However, one notable exception to these trends is centromere 5, which harbors 12–22% of HORs compared to the other centromeres, yet still recruits comparable CENH3 **(Figure 4.2G)**.

Invasion of the Arabidopsis centromeres by ATHILA retrotransposons

We observed that centromere 5 shows both reduced *CEN180* higher-order repetition and was heavily disrupted by breaks in the satellite array (**Figure 4.2G**). Genome-wide, within the main satellite arrays, the vast majority of sequence (>94%) is *CEN180*, with only 69 interspersed sequences larger than 1 kbp. Within these gaps we identified 46 intact and 5 fragmented *ATHILA* LTR retrotransposons of the *GYPSY* superfamily, belonging to the *ATHILA*, *ATHILA2*, *ATHILA5* and *ATHILA6A/6B* subfamilies (**Figure 4.3A)** [19–21]. The intact *ATHILA* elements have a mean length of 10.9 kbp, and the majority have highly similar paired LTRs, target site duplications (TSDs), primer binding sites (PBS), polypurine tracts (PPT) and Gypsy superfamily open reading frames. LTR comparisons indicate that the centromeric *ATHILA* elements are young, with on average 98.39% LTR sequence identity (**Figure 4.3B**), which was higher than *GYPSY* and *COPIA* elements located outside the centromere (**Figure 4.3B**). We also observed 10 *ATHILA* solo LTRs that lacked a downstream PBS or upstream PPT, which is consistent with post-integration intra-element homologous recombination. Interestingly, we also observed 5 instances where gaps containing full-length *ATHILA* or solo LTRs show a duplication on the same chromosome that are between 8.9 and 538.4 kbp apart, consistent with transposon sequences being copied post-integration, potentially via the same mechanism that generates *CEN180* HORs.

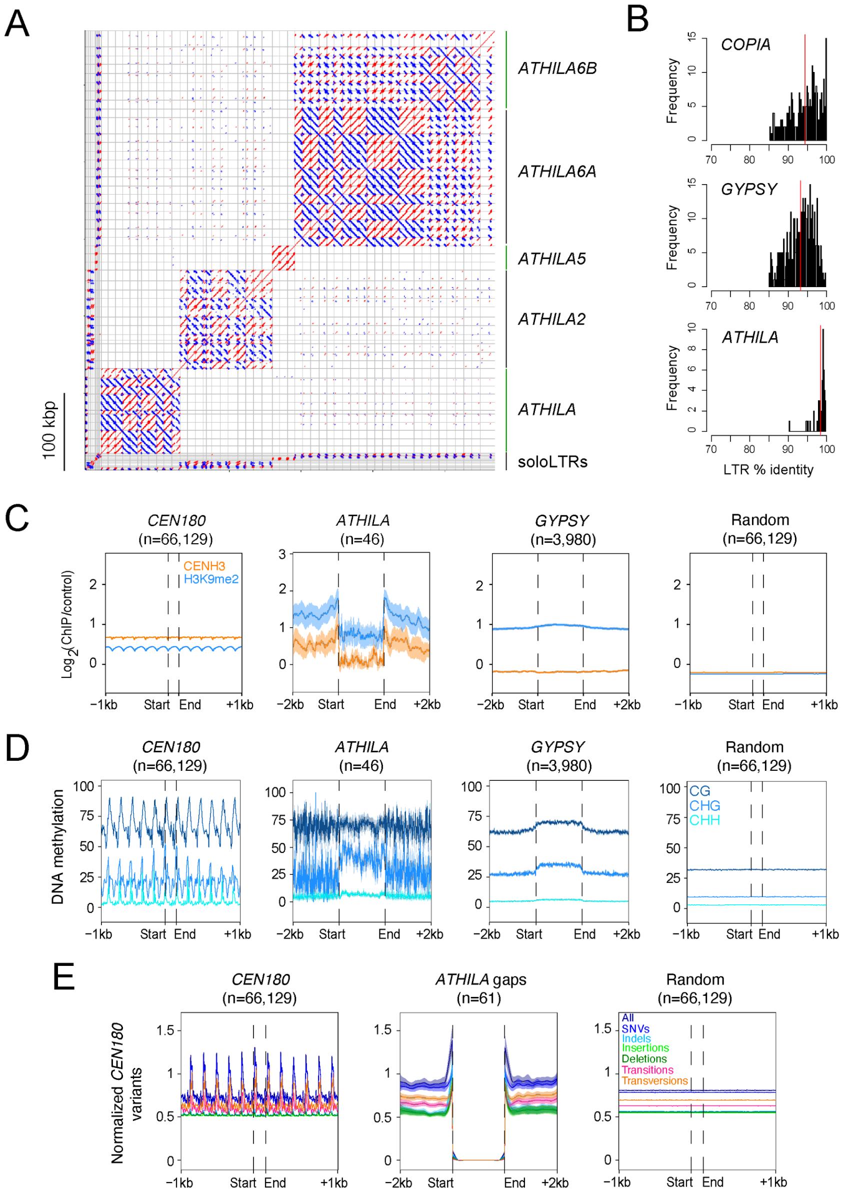


Figure 4.3: Invasion of the Arabidopsis centromeres by *ATHILA* retrotransposons. (A) Dotplot of centromeric *ATHILA* retroelements using a search window of 50 bp. Red and blue indicate forward and reverse strand similarity. The elements assigned to different *ATHILA* families and solo LTRs are indicated. (B) Histograms of LTR percent sequence identity for centromeric *ATHILA* elements, compared to *GYPSY* and *COPIA* elements outside of the centromeres. Mean values are indicated by the red lines. (C) CENH3 (orange) and H3K9me2 (blue) ChIP-seq enrichment (log$_2$(ChIP/input)) over *CEN180 (*n=66,129), centromeric intact *ATHILA (*n=46), *GYPSY* located outside the centromeres (n=3,980) and random positions (n=66,129). Shaded ribbons represent 95% confidence intervals for windowed mean values. (D) As for C, but analyzing ONT-derived percent DNA methylation in CG (dark blue), CHG (blue) and CHH (light blue). (E) The number of *CEN180* sequence edits (insertions, deletions, and mismatches, compared to the *CEN180* consensus) normalized by *CEN180* frequency, in positions surrounding *CEN180* (n=66,129), gaps containing *ATHILA* sequences (n=61), or random positions (n=66,129). All edits (dark blue) are analyzed, in addition to substitutions (SNVs, blue), indels (light blue), insertions (light green), deletions (dark green), transitions (pink) and transversions (orange).

We analyzed the centromeric *ATHILA* elements for CENH3 ChIP-seq enrichment and observed a decrease relative to the surrounding *CEN180*, yet higher levels than observed in *GYPSY* elements located outside the centromere (**Figure 4.3C**). The *ATHILA* elements show greater H3K9me2 enrichment compared to flanking *CEN180* (**Figure 4.3C**). We used our ONT reads to profile DNA methylation over the *ATHILA* and observed dense methylation, at a similar level to the surrounding *CEN180*, although with higher CHG-context methylation (**Figure 4.3D**). Hence, *ATHILA* elements are differentiated from the surrounding satellites at the chromatin level. Interestingly, when we profiled *CEN180* SNVs around gaps containing the *ATHILA* insertions (full length, fragments and solo LTRs), we observed a pronounced elevation in satellite divergence at the insertion boundaries (**Figure 4.3E**). This may indicate that *ATHILA* insertion was mutagenic on the surrounding satellite repeats, or that transposon insertion influenced the subsequent divergence or homogenization of the repeats. Together this indicates that centromeric *ATHILA* insertions interrupt the genetic and epigenetic organization of the Arabidopsis *CEN180* satellite arrays.

Epigenetic organization and meiotic recombination within the centromeres

To assess the genetic and epigenetic features of the centromeres, we analyzed all chromosome arms along their telomere–centromere axes using a proportional scale (**Figure 4.4A**). Centromere midpoints were defined by maximum CENH3 ChIP-seq enrichment. As expected, *CEN180* satellites are highly enriched in proximity to the centromere midpoints (**Figure 4.4A**). Gene density drops precipitously as the centromeres are approached, whereas transposons reciprocally increase until they are replaced by *CEN180* (**Figure 4.4A**). Gene and transposon density are tracked closely by H3K4me3 and H3K9me2 ChIP-seq enrichment, respectively (**Figure 4.4A**). H3K9me2 enrichment is observed in the centromere, although there is a reduction in the center coincident with CENH3 enrichment (**Figure 4.4A**), consistent with reduced H3 occupancy caused by CENH3 replacement. Interestingly, a slight increase in H3K4me3 enrichment is observed within the centromeres, relative to the flanking pericentromeric regions (**Figure 4.4A**). We observed striking biases in base composition over the centromeres, which are relatively GC-rich compared to the AT-rich chromosome arms (**Figure 4.4A**).

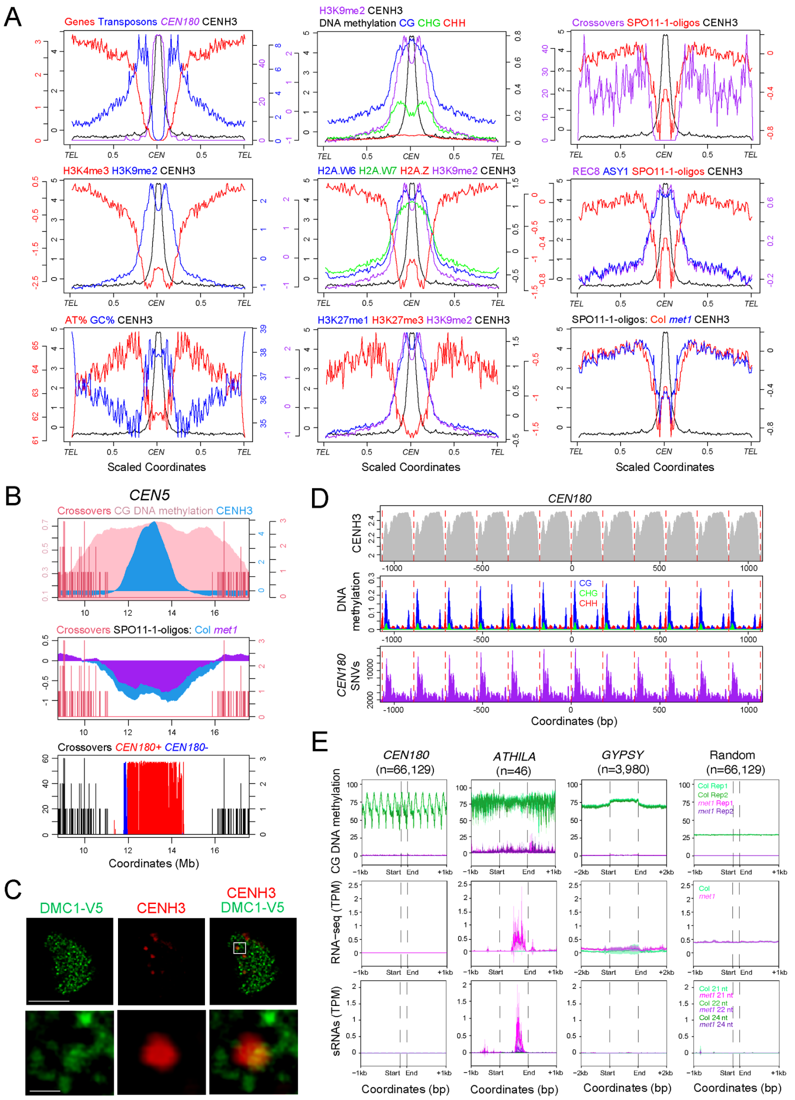


Figure 4.4: Epigenetic organization and meiotic recombination within the centromeres. (A) Data were analyzed along chromosome arms that were proportionally scaled between the telomeres (*TEL*) and centromere midpoint (*CEN*), which was defined by maximum CENH3 ChIP-seq enrichment. Data analyzed were gene, transposon and *CEN180* density, CENH3, H3K4me3, H3K9me2, H2A.W6, H2A.W7, H2A.Z, H3K27me1, H3K27me3, REC8 and ASY1 ChIP-seq (log$_2$(ChIP/input), % AT and GC base composition, DNA methylation, SPO11-1-oligos (in wild type and *met1*) and crossovers. (B) Plot of crossovers (red), CG DNA methylation (pink), CENH3 (blue), SPO11-1-oligos in wild type and *met1* and *CEN180* density along centromere 2 (*CEN2*). (C) Male meiocyte in early prophase I immunostained for CENH3 (red) and V5-DMC1 (green). Scale bars are 10 µM (upper row) and 1 µM (lower row). (D) Plots of CENH3 ChIP enrichment (grey), DNA methylation in CG (blue), CHG (green) and CHH (red) contexts and *CEN180* SNVs (purple), averaged over windows centred on all *CEN180* starts. The red lines show 178 bp increments. (E) CG context DNA methylation in wild type (green) or *met1* (purple) [22], RNA-seq in wild type (green) and *met1* (pink) [23] and siRNA-seq in wild type (green) and *met1* (pink) [22], over *CEN180 (*n=66,129), centromeric intact *ATHILA (*n=46), *GYPSY* located outside the centromeres (n=3,980) and random positions (n=66,129). Shaded ribbons represent 95% confidence intervals for windowed mean values.

Using our ONT sequencing data and DeepSignal-plant, we observed dense DNA methylation across the centromeres in CG, CHG and CHH contexts (**Figure 4.4A**) [24]. However, CHG DNA methylation shows relatively reduced frequency within the centromeres, compared to CG methylation (**Figure 4.4A**). This may reflect depletion of H3K9me2 within the centromeres, which functions to maintain DNA methylation in non-CG contexts [25]. We also observed high ChIP-seq enrichment of the heterochromatic chromatin marks H2A.W6, H2A.W7 and H3K27me1 within the centromeres (**Figure 4.4A**) [26,27]. The Polycomb-group modification H3K27me3 was depleted in the centromeres and found largely in the gene-rich chromosome arms (**Figure 4.4A**). Enrichment of the euchromatic histone variant H2A.Z was low in the centromeres, but similar to H3K4me3, it showed a slight increase in the centromeres, relative to the pericentromeres (**Figure 4.4A**).

To investigate genetic control of DNA methylation in the centromeres, we analyzed bisulfite sequencing (BS-seq) data from wild type and eight mutants defective in the CG and non-CG DNA methylation maintenance pathway [25,28]. Centromeric non-CG methylation is eliminated in a *drm1 drm2 cmt2 cmt3* mutant, and strongly reduced in *kyp suvh5 suvh6*, whereas CG methylation is intact in these lines [25,28]. CG methylation in the centromere is strongly reduced in *ddm1* and *met1*, although non-CG is more greatly reduced in *ddm1* than *met1* [28]. Hence, dense DNA methylation is observed within the centromeres that is maintained by canonical pathways, although CG-context methylation is relatively high compared with non-CG.

Meiotic recombination, including unequal crossover and gene conversion, have been proposed to mediate centromere evolution [4,29]. We mapped 2,042 crossovers from Col×Ler F_2 sequence data that were resolved on average to 1.01 kbp. As expected, crossovers were potently suppressed in proximity to the centromeres (**Figure 4.4A-4.4B**). We observed high centromeric ChIP-seq enrichment of REC8-cohesin and the HORMA domain protein ASY1, which are components of the meiotic chromosome axis (**Figure 4.4A**) [30,31]. To investigate the potential for meiotic DSB formation within the centromeres, we aligned SPO11-1-oligo data from wild type [23]. Overall, SPO11-1-oligos were low within the centromeres, although we observed an increase relative to the flanking pericentromeric heterochromatin, reminiscent of the

H3K4me3 and H2A.Z patterns (**Figure 4.4A**). To investigate the role of DNA methylation, we mapped SPO11-1-oligos sequenced in the CG DNA methylation mutant *met1-3* [23], which showed a gain of DSBs in proximity to the centromere (**Figure 4.4A-4.4B**). To provide cytological evidence of recombination close to the centromeres, we immunostained meiocytes in early prophase I for CENH3 and V5-DMC1, which is a marker of inter-homolog recombination (**Figure 4.4C)**. DMC1-V5 foci were observed along the chromosomes and associated with the surface, but not within, CENH3 foci (**Figure 4.4C**). Hence, despite suppression of crossovers, we observe evidence for low levels of meiotic recombination initiation associated with the centromeres, which are influenced by DNA methylation.

Finally, we analyzed chromatin and transcription around *CEN180* and *ATHILA* retrotransposons at the fine-scale, and compared wild type and the DNA methylation mutant *met1-3*. CENH3 nucleosomes show a strongly phased pattern of enrichment with the *CEN180* satellites, with relative depletion in spacer regions at the start and end of the satellites (**Figure 4.4D**). Interestingly, these CENH3 spacer regions also associate with elevated DNA methylation and *CEN180* SNVs (**Figure 4.4D**), consistent with CENH3-nucleosome occupancy influencing epigenetic modification and genetic divergence of satellites. In *met1*, we observed loss of CG-context DNA methylation in both the *ATHILA* and *CEN180* repeats (**Figure 4.4E**). However, RNA-seq and siRNA-seq counts increased specifically in the *ATHILA* in *met1* (**Figure 4.4E**) [22,23]. Both RNA-seq and siRNA-seq signals increased most strongly in the internal 3′-regions of the *ATHILA* (**Figure 4.4E**), which correspond to 'TSI' transcripts and easiRNA populations previously reported [32–34]. This further indicates that epigenetic regulation of the *CEN180* satellites and *ATHILA* elements are distinct.

4.4 DISCUSSION

The Col-CEN assembly reveals the architecture of the Arabidopsis centromeres, which consist of megabase-scale, stranded *CEN180* arrays, which are invaded by *ATHILA* retrotransposons. Extensive sequence variation is observed between the satellites, and the majority of variant monomer sequences are private to each centromere. This is consistent with satellite homogenization occurring primarily within chromosomes. *CEN180* that are the least divergent and with most higher-order repetition showed the highest CENH3 occupancy. This suggests that CENH3 chromatin may promote recombination pathways that lead to homogenization, including DSB formation and repair via homologous recombination. For example, inter-homolog strand invasion during meiosis has the potential to cause *CEN180* gene conversion. In this respect, we note that *CEN180* higher-order repeats show an average length of 656 bp, which is within the range of observed Arabidopsis meiotic gene conversions [35]. We also see a proximity effect on divergence between higher-order repeats, with repeat blocks further apart showing greater sequence differences. These patterns are reminiscent of human alpha-satellite higher-order repeats, although the alpha-satellite blocks are longer and occur over greater distances [5,36,37]. As meiotic crossover repair is strongly suppressed within the Arabidopsis centromeres, consistent with patterns across eukaryotes [29,38–40], we do not consider unequal crossover to be likely within the centromeres. However, we propose that an ongoing, recombination-based homogenization process maintains the *CEN180* library close to the consensus that is optimal for CENH3 recruitment.

Aside from homogenizing recombination within the *CEN180*, the centromeres have experienced invasion by *ATHILA* retrotransposons. The ability of *ATHILA* elements to insert within Arabidopsis *CEN180* regions is likely determined by their integrase protein [20,41]. Interestingly, the Tal1 *COPIA* element from *Arabidopsis lyrata* shows a strong insertion bias into the *CEN180* when expressed in *Arabidopsis thaliana [42]*, despite satellite sequences varying between these species [43]. The majority of the centromeric *ATHILA* elements appear young, based on LTR identity and possess many features required for transposition, although the centromeres show striking differences in the frequency of *ATHILA* insertions, with centromeres 4 and 5 being the most invaded. *ATHILA* elements show lower CENH3 and higher H3K9me2 and CHG DNA methylation than the surrounding *CEN180*, and associate with increased satellite divergence at their boundaries. Hence, *ATHILA* represent a disruptive influence on the genetic and epigenetic organization of the centromeres. In maize, meiotic gene conversion was observed to act on *CRM2* retrotransposons within the centromeres [29]. Therefore, satellite homogenization pathways may serve as a mechanism to eliminate *ATHILA* insertions. Indeed, a gene conversion mechanism may explain the 5 *ATHILA* intra-chromosome duplications that appear to have occurred post-integration. We also note that the presence of *ATHILA* solo LTRs is consistent with homologous recombination acting on the centromeric retrotransposons following integration. Intriguingly, centromere 5 and the diverged *CEN180* array on chromosome 4, show both high *ATHILA* density and a striking reduction of *CEN180* higher-order repetition. This is consistent with *ATHILA* inhibiting *CEN180* homogenization, or loss of homogenization facilitating *ATHILA* insertion, or both. We propose that each Arabidopsis centromere represents different stages in a cycle of satellite homogenization and disruption by *ATHILA*. These opposing forces provide both a capacity for homeostasis, and a capacity for change, during centromere evolution.

4.5 METHODS

Genomic DNA extraction and ONT sequencing

For genomic DNA extraction, 3 week old Col-0 seedlings were grown on ½ MS media and 1% sucrose and kept in the dark for 48 hours prior to harvesting. Approximately 10 g of tissue was used per 200 ml of MPD-Based Extraction Buffer pH 6 (MEB). Tissue was flash frozen and ground tissue in liquid nitrogen, using a pestle and mortar, and resuspended in 200 ml MEB. Ground tissue was thawed in MEB with frequent stirring. The homogenate was forced through 4 layers of miracloth, and then filtering again through 4 layers of fresh miracloth by gravity. 20% Triton x-100 was added to a final concentration of 0.5% on ice, followed by incubation with agitation on ice for 30 minutes. The suspension was centrifuged at 800*g* for 20 minutes at 4°C. The supernatant was removed and the pellet resuspended using a paintbrush in 10 ml 2-methyl-2,4 pentanediol buffer pH 7.0 (MPDB). The suspension was centrifuged at 650*g* for 20 minutes at 4°C. The supernatant was removed and the pellet was washed with 10 ml of MPDB. Washing and centrifugation was repeated until the pellet appeared white and was finally resuspended in a minimal volume of MPDB. From this point onwards all transfers were performed using wide bore pipette tips. 5 ml CTAB buffer was added to the nuclei pellet and mixed via gentle inversion, followed by incubation at 60°C until full lysis had occurred, taking between 30 minutes and 2 hours. An equal volume of chloroform was added and incubated on a rocking platform, with a speed of 18 cycles per minute, for 30 minutes, followed by centrifugation at 3000*g* for 10 minutes. An equal volume of phenol/chloroform/isoamyl alcohol (PCI, 25:24:1) was added to the lysate, followed by incubation on a rocking platform (18 cycles per minute) for 30 minutes. The lysate was centrifuged at 3000*g* for 10 minutes and the upper aqueous phase was transferred into a fresh tube. The PCI extraction was then repeated. The extraction was then repeated using only chloroform. 1/10th volume of 3M Sodium Acetate was added to the lysate and mixed by gentle inversion. Two volumes of ice cold ethanol were added and mixed by inversion. DNA was precipitated at -20°C for 48 hours. The precipitated DNA was removed using a glass hook and washed three times in fresh 70% ethanol. The DNA was dissolved in 120 µl of 10 mM Tris-Cl (pH 8.5).

Approximately 5 µg of DNA was size selected to be >30 kbp, using the BluePippin™ Size-Selection System (Sage Science) and the 0.75% DF Marker U1 cassette definition, with Range mode and BP start set at 30,000 bp. Library preparation followed the Nanopore SQK-LSK109 protocol and kit**.** Approximately 1.2-1.5 µg of size selected DNA in a volume of 48 µl was used for library preparation. DNA was nic-repaired and end-prepped by the addition of 3.5 µl of NEBNext FFPE Buffer and NEBNext Ultra II End Prep Reaction Buffer, followed by 2 µl of NEBNext DNA Repair Mix and 3 µl NEBNext Ultra II End Prep Enzyme Mix (New England Biolab, E7180S), with incubation for 30 minutes at 20°C, followed by 30 minutes at 65°C. The sample was cleaned using 1×volume AMPure XP beads and eluted in 61 µl of nuclease-free water. Adapters were ligated at room temperature using 25 µl Ligation Buffer, 10 µl NEBNext T4 DNA Ligase and 5 µl Adapter Mix for 2 hours. The library was cleaned with 0.4×volume AMPure XP beads, washed using ONT Long Fragment buffer and eluted in 15 µl elution buffer.

Col-CEN genome assembly

Libraries were sequenced on 6 ONT R9 flow cells and 1 ONT R10 flow cell, and the resulting .fast5 files were basecalled with Guppy (v4.0.15), using the dna_r9.4.1_450bps_hac.cfg and dna_r10.3_450bps_hac.cfg configurations, respectively. This yielded 73.6 Gb of sequence and ~55x coverage of ultra-long reads (>50 kbp). The fastq files of ONT reads used for genome assembly are available for download at ArrayExpress accession E-MTAB-10272 (http://www.ebi.ac.uk/arrayexpress/). We trimmed adapters using Porechop (v0.2.4) and filtered for read lengths greater than 30 kbp and mean read quality scores >90%, using Filtlong (v0.2.0) (https://github.com/rrwick/Filtlong), which yielded 436,146 reads with a mean length of 43.9 kbp (19.15 Gbp), equivalent to 161× coverage of the TAIR10 genome. Flye (version 2.7) was used to assemble the reads, specifying a minimum read overlap of 10 kbp and a *k*-mer size of 17 [44].

Contig screen

We performed a comprehensive contig screen using methods inspired by the Vertebrate Genomes Project (VGP), though adapted for an inbred plant genome [45]. We first aligned Flye contigs to the Columbia reference chloroplast (GenBank accession NC_000932.1) [46], and mitochondria (GenBank accession NC_037304.1) [47] genomes with Minimap2 (v2.17-r941, -x asm5) [48]. Contigs with at least 50% of their bases covered by alignments were considered to be chloroplast or mitochondria genome sequences and were removed from the assembly.

We next used BLAST to screen for contigs representing bacterial contamination. We first masked the Flye assembly with windowmasker (v1.0.0, -mk_counts -genome_size 131405362) [49]. We then aligned the Flye contigs to all RefSeq bacterial genomes (downloaded on 2020/05/21) with megablast (v2.5.0, -outfmt "6 std score"), providing the windowmasker annotations with “-window_masker_db” [50]. We removed BLAST alignments with an E value greater than or equal to 0.0001, a score less than 500, and a Percent Identity less than 98%, and any contigs (four in total) with remaining alignments were manually inspected. Two of the four contigs were already identified as being chloroplast or mitochondria sequence and the other two were clearly nuclear contigs, so we determined that no contigs were derived from bacterial contaminants.

After removing chloroplast and mitochondria contigs, we performed one final screen to remove contigs with low read support. We aligned ONT reads (>=40 kbp) to the contigs with Minimap2 (v2.17-r941, -x map-ont) and removed any contigs (one in total) with more than 50% of its bases covered by fewer than 15 reads. Though we did not use its standard pipeline, we made use of purge_dups scripts for this analysis [51]. After screening, the assembly consisted of 10 contigs with an N50 of 22,078,741 bp.

Contig scaffolding

Though the five Columbia chromosomes were represented by only 10 contigs, we used reference-guide scaffolding to order and orient contigs, assign chromosome labels, and orient pseudomolecules to match the orientation of TAIR10 chromosomes. We ran RagTag (v1.0.1, --debug --aligner=nucmer --nucmer-params='--maxmatch -l 100 -c 500') using TAIR10 as the reference genome, but excluding ChrC and ChrM (-e) [52,53]. Three small contigs (3,200, 90,237 and 8,728 bp) consisting of low complexity sequence were not ordered and oriented and were removed from the assembly. After scaffolding, the 131,388,895 bp assembly was represented in five pseudomolecules corresponding to the five chromosomes of the Columbia genome. Chromosome 1 was gapless, while the other chromosomes contained one to four 100 bp gaps each (9 in total).

Pseudomolecule polishing and gap filling

We corrected misassemblies and filled gaps in the Columbia pseudomolecules with two rounds of Medaka (v1.2.1) ONT polishing (https://github.com/nanoporetech/medaka). For the first round of polishing, we aligned R9 ONT reads (>=50 kbp) to the pseudomolecules with mini_align (minimap2 v2.17-r941, -m). To avoid overcorrection in the centromere satellite sequences, we performed "marker-assisted filtering" to remove alignments not anchored in putatively unique sequences [5] (https://github.com/malonge/T2T-Polish). We defined "marker" *k*-mers as 21-mers that occurred once in the assembly and between 14 and 46 times (inclusive) in the Illumina reads. The first round of polishing was completed using `medaka consensus` (--model r941_min_high_g360 --batch_size 200) and `medaka stitch`. The second round of polishing was performed as for the first round, except we aligned all R10 reads instead of R9 reads and the `medaka consensus` model was set to "r103_min_high_g360". As a result of ONT polishing, the assembly improved from a QV of 32.38 to 33.17 and 34.12 after the first and second rounds, respectively [18]. After medaka polishing, the assembly contained only a single gap on chromosome 2.

Long-read ONT polishing was followed by short-read polishing of non-centromeres with DeepVariant [54]. We first aligned Columbia genomic DNA Illumina reads to the pseudomolecules with bwa mem (v0.7.17-r1198-dirty) and we compressed and sorted alignments with samtools (v1.10) [55,56]. We then created a VCF file of potential polishing edits with DeepVariant (v1.1.0, --model_type=WGS),"bcftools view" (v1.11, -e 'type="ref"' -i 'QUAL>1 && (GT="AA" || GT="Aa")') and "bcftools norm". To avoid error-prone short-read polishing in the centromeres, we used Bedtools to remove polishing edits within the centromeres and we used BCFtools to derive a final consensus FASTA file [57,58]. Though short-read polishing did not alter the centromeres, it improved the overall assembly QV to 41.4616.

Telomere patching

We locally re-assembled and patched telomeric sequences for the 8 Columbia telomeres not adjacent to NORs (all but the beginning of chromosomes 2 and 4). We aligned all R9 reads to the TAIR10 reference with Winnowmap (v1.11, k=15, --MD -ax map-ont) and for each telomere, we collected all reads that aligned once to within 50 bp of the chromosome terminus [9]. Using Bowtie [59] (v1.3.0, -S --all -v 0), we counted the occurrences of the telomeric repeat motif ('CCCTAAA') in each read, and the read with the most occurrences was designated as the "reference" and all other reads were designated as the "query". Local re-assembly was completed by aligning the query reads to the reference read and computing a consensus with `medaka_consensus` (v1.2.1, -m r941_min_high_g360). To patch these telomere consensus sequences into the Columbia pseudomolecules, we identified the terminal BAC sequences for each of the 8 chromosome arms. For each chromosome arm, we aligned the terminal BAC sequence to the Columbia pseudomolecules and the telomere consensus sequence with Nucmer (v3.1, --maxmatch). Using these alignment coordinates, the consensus sequences were manually patched such that everything after the terminal BAC sequence was replaced with telomere consensus sequence. Telomeres were then manually confirmed to be structurally valid.

Assembly curation and preparation

After polishing and telomere patching, we performed final curation steps to correct lingering misassemblies and screen for contamination. First, while it was not straightforward to fill the remaining chromosome 2 gap *de novo*, we were able to replace the gap locus with the corresponding region in TAIR10. We found two BAC sequences flanking the gap locus that aligned concordantly to both the Col-0 pseudomolecules and TAIR10. These BAC contigs were aligned to the pseudomolecules and TAIR10 with Nucmer (v3.1, --maxmatch -l 250 -c 500) and the gap locus between the BAC contigs in the Columbia pseudomolecules was replaced with the corresponding TAIR10 locus between the BAC contigs.

To identify and correct structural misassemblies, we aligned Columbia long-reads to the Columbia pseudomolecules and called structural variants (SVs). First, we used Bedtools `random` (v2.29.2, -l 100000 -n 50000 -seed 23) to simulate 50,000 100 kbp exact reads from TAIR10. These reads, along with R9 (>=50 kbp) and R10 Columbia reads were aligned to the Columbia pseudomolecules with Winnowmap (v1.11, k=15, "--MD -ax map-pb" for TAIR10 reads and "--MD -ax map-ont" for ONT reads). After compressing and sorting alignments with samtools (v1.10), Sniffles (v1.0.12, -d 100 -n -1 -s 3) was used to infer SVs from each of the alignments [60]. SVs with fewer than 30% of reads supporting the ALT allele were removed and the three resulting VCF files were merged with Jasmine (v1.0.10, max_dist=500 spec_reads=3 --output_genotypes) [61]. There were a total of three variants called by all three read sets, including two deletions and one insertion that we corrected. REF and ALT alleles for these SVs were manually refined and validated, and ALT alleles were incorporated into the pseudomolecules using `bcftools consensus`.

Next, we manually inspected all gaps filled by Medaka and found that a 181 bp region containing a 100 bp gap on chromosome 5 was incorrectly replaced with 103 bp of sequence and we manually replaced the filled sequence with the original gap locus. Finally, we used VecScreen to do a final contamination screen. We first aligned the Columbia pseudomolecules to the VecScreen database with blastn (v2.5.0, -task blastn -reward 1 -penalty -5 -gapopen 3 -gapextend 3 -dust yes -soft_masking true -evalue 700 -searchsp 1750000000000 -outfmt "6 std score"). The BLAST alignments did not yield any "moderate" or "strong" matches to the database, so we determined that there was no contamination.

The final assembly contained five pseudomolecules with a single gap on chromosome 5, two missing telomeres, and partially resolved NOR sequence at the beginning of chromosomes 2 and 4. Chromosomes 1 and 3 were gapless and were completely sequence resolved from telomere-to-telomere (T2T). The final Col-CEN assembly FASTA file includes these 5 pseudomolecules and the Columbia chloroplast and mitochondria reference genomes.

Genome annotation

Genes were lifted-over from TAIR10 with Liftoff (v1.5.1, -copies -a 1 -s 1) [62]. Since ChrC and ChrM were directly copied from TAIR10, their lift-over genes were replaced with their original TAIR10 annotations. We then used EDTA (v1.9.6, --sensitive 1 --anno 1 --evaluate 1) to perform *de novo* transposable element (TE) annotation, providing transcripts with "--cds" and the TAIR10 TE library with "--curatedlib" [63,64]. The TE annotation was supplemented with a manual annotation of centromere gaps using dotplot analysis and further manual annotation of the centromeric *ATHILA* elements (see section below). We used LASTZ to identify regions with similarity to *5S*, *45S* rDNA and the mitochondrial genome. To generate similarity heatmaps, the centromere region was divided into adjacent 5 kbp regions which were compared using the pairwiseAlignment (type='global') and pid functions in R, using the Biostrings library. Sequences were compared in forward and reverse directions, and the highest percent sequence identity value kept. These values were then plotted in the heatmap.

CEN180 repeat annotation

To identify repetitive regions, we divided the genome assembly into adjacent 1 kbp windows. In each window, for each position, we defined 12-mers and exactly matched these sequences to the rest of the window. We identified windows where the proportion of non-unique 12-mers was greater than 10%, and merged contiguous windows that were above this threshold. For each region, we generated a histogram of the distances between 12-mers to test for periodic repeats. For example, if a region contains an arrayed tandem repeat of monomer size N, then a histogram of the 12-mer distances will show peaks at values N, N×2, N×3 … . The N value was obtained for each region, using the most frequent 12-mer distance. Next, 5 sequences of length N were randomly chosen from within the region and matched back to the sequence using the R function matchPattern (max.mismatch=N/3 with.indels=T). For each set of matches we identified overlapping repeats. If the overlap was less than 10 nucleotides, the overlap was divided at the midpoint between the repeats. If the overlap was 10 nucleotides or greater, the larger repeat was kept. The set of non-overlapping matches with the highest number was kept for further analysis. These sequence matches were aligned using mafft (--retree 2 --inputorder) [65], and a consensus repeat monomer was derived from the multiple sequence alignment. This consensus sequence was matched back to the region using matchPattern (max.mismatch=N/3 with.indels=T), and overlaps were treated in the same way.

Our approach identified 66,129 *CEN180* repeats with a mean length of 178 bp. The set of unique *CEN180* sequences (n=25,192) were aligned using mafft (--sparsescore 1000 --inputorder) [65]. A consensus sequence was generated from the multiple sequence alignment, which was:
5′-AGTATAAGAACTTAAACCGCAACCCGATCTTAAAAGCCTAAGTAGTGTTTCCTTGTTAGAAGACACAAAGCCAAAGACTCATATGGACTTTGGCTACACCATGAAAGCTTTGAGAAGCAAGAAGAAGGTTGGTTAGTGTTTTGGAGTCGAATATGACTTGATGTCATGTGTATGATTG-3′. In order to analyze *CEN180* diversity, for each position of the multiple sequence alignment (968 positions), we calculated the proportion of A, T, G, C and gaps. The alignment for each monomer at each position was then compared to these proportions and used to calculate a single nucleotide variant (SNV) score for the monomer. For example, if a monomer had an A in the alignment at a given position, and the overall proportion of A at that position was 0.7, the SNV score for that monomer would increase by 1-0.7. This was repeated for each position of the alignment, for each monomer. This ‘weighted’ SNV score was used to assess how similar a given *CEN180* monomer is to the genome-wide consensus. Alternatively, to compare pairwise differences between two specific monomers, the two sequences were compared along the length of the multiple sequence alignment and each instance of disagreement counted to give a ‘pairwise’ SNV score.

To identify higher order repeats (HORs) in a head-to-tail (tandem) orientation, each monomer was taken in turn and compared to all others using a matrix of pairwise SNV scores. If a pair of monomers had an SNV score of 5 or less, and were on the same strand, they were considered a match. For each match, monomers were extended by +1 unit in the same direction on the chromosome, and these were again compared for pairwise SNVs. This process was repeated until the next monomers had a pairwise SNV score higher than threshold, or the repeats were on opposite strands, or the end of the array was reached, with these conditions defining the end of the HOR. We also searched for repeats in head-to-head (inverted) orientation, which was identical apart from that repeats must be on opposite strands, and when monomers are extended to search for HORs, one is extended +1 position along the chromosome, whereas the other decreases -1. HORs were defined for each instance of 3 or more consecutive monomer matches. We define each HOR as consisting of block1 and block2 of *CEN180* monomers. The size of each block was recorded, in terms of monomers and base pairs, in addition to the distance between the block start coordinates. Cumulative pairwise SNVs per *CEN180* monomer were also calculated between each pair of blocks to provide a 'block' SNV score. To measure higher order repetition of each monomer, we summed the HOR block sizes in mers, such that if a monomer was represented in three 5-mer blocks, it would score 15.

ATHILA annotation

To carefully resolve the sequence of the centromeric *ATHILA* elements, we used LTRharvest [66] to complement the EDTA run that was used for the annotation of all Arabidopsis TEs (see above). We ran LTRharvest three times using 'normal', 'strict' and 'very strict' parameters. The parameters were gradually adjusted to allow us to capture the full-length sequence of the *ATHILA* family, based on older studies that reported the total and LTR lengths of intact *ATHILA* elements [20]. These parameters were -maxlenltr 2500 -minltrlen 400 -mindistltr 2000 -maxdistltr 20000 -similar 75 -mintsd 0 -motif TGCA -motifmis 1 for the 'normal' run; -maxlenltr 2000 -minlenltr 1000 -mindistltr 4000 -maxdistltr 16000 -similar 80 -mintsd 3 -motif TGCA -motifmis 1 for the 'strict' run; and -maxlenltr 2100 -minlenltr 1100 -mindistltr 5000 -maxdistltr 14000 -similar 85 -mintsd 4 -motif TGCA -motifmis 1 -vic 20 for the 'very strict' run. Coordinates of predicted full-length elements from EDTA, LTRharvest and the manual dotplot annotation of centromeric TEs were merged and sequences aligned using MAFFT [67]. Through these steps, we were able to pinpoint with base-pair resolution the external junctions of every *ATHILA* element, together with the flanking sequence and the internal junctions of the LTRs with the internal domain (5′-LTR with PBS; PPT with 3′-LTR). Overall, we identified 46 intact elements of which 34 have a detectable target site duplication, 5 fragmented *ATHILA* and 10 solo LTRs. We further identified open reading frames (minimum 150 nt) in the internal domain of the 46 intact elements using getorf in EMBOSS [68], and by running HMMER v3.3.2 (http://hmmer.org/) (-E 0.05 --domE 0.05) using a collection of Hidden Markov Models (HMMs) downloaded from Pfam (http://pfam.xfam.org/) that describe the coding domains of *GYPSY* LTR retrotransposons: PF03732 for gag; PF13650, PF08284, PF13975 and PF09668 for protease; PF00078 for reverse transcriptase; PF17917, PF17919 and PF13456 for RNase-H; PF00665, PF13683, PF17921, PF02022, PF09337 and PF00552 for integrase. *ATHILA* elements may also contain an envelope-like ORF [20]. To identify this domain (and because there is no HMM in Pfam that describes envelope-like genes of LTR retrotransposons), we used a previously published HMM developed by one of the co-authors [69].

ONT DNA methylation analysis

To identify CpG, CHG and CHH methylation contexts we used DeepSignal-plant (v. 0.1) [24], which uses a deep-learning method based on bidirectional recurrent neural network (BRNN) with long short-term memory (LSTM) units to detect DNA 5mC methylation. R9 reads were filtered for length and accuracy using Filtlong (v0.2.0) (--min_mean_q 95, --min_length 30000. https://github.com/rrwick/Filtlong). Basecalled read sequence was annotated onto corresponding .fast5 files, and re-squiggled using Tombo (v 1.5.1). Methylation prediction for the CG, CHG, and CHH contexts were called using Deepsignal-plant using the respective models:

model.dp2.CG.arabnrice2-1_R9.4plus_tem.bn13_sn16.balance.both_bilstm.b13_s16_epoch6.ckpt,

model.dp2.CHG.arabnrice2-1_R9.4plus_tem.bn13_sn16.denoise_sig1nal_bilstm.both_bilstm.b13_s16_epoch4.ckpt

model.dp2.CHH.arabnrice2-1_R9.4plus_tem.bn13_sn16.denoise_signal_bilstm.both_bilstm.b13_s16_epoch7.ckpt.

The script call_modification_frequency.py provided in the Deepsignal-plant package was then used to generate the methylation frequency at each CG, CHG and CHH site.

To identify CpG methylation in Nanopore reads we also used Nanopolish (v 0.13.2), which uses a Hidden Markov model on the nanopore current signal to distinguish 5-methylcytosine from unmethylated cytosine. Reads were first filtered for length and accuracy using Filtlong (v0.2.0) (--min_mean_q 95, --min_length 15000. https://github.com/rrwick/Filtlong). The subset was then indexed to the fast5 files, and aligned to the genome using Winnowmap (v1.11, -ax map-ont). The read fastq, alignment bam, and fast5 files were used as an input to the Nanopolish call-methylation function. The script calculate_methylation_frequency.py provided in the Nanopolish package was then used to generate the methylation frequency at each CG containing *k*-mer.

ChIP-seq and MNase-seq data alignment and processing

Deduplicated paired-end ChIP-seq and MNase-seq reads were processed with Cutadapt v1.18 to remove adapter sequences and low-quality bases (Phred+33-scaled quality <20) [70]. Trimmed reads were aligned to the Col-CEN genome assembly using Bowtie2 v2.3.4.3 with the following settings: --very-sensitive --no-mixed --no-discordant -k 10 [71]. Up to 10 valid alignments were reported for each read pair. Read pairs with Bowtie2-assigned MAPQ <10, including those that aligned equally well to more than one location, were discarded using Samtools v1.9 [56]. For retained read pairs that aligned to multiple locations, with varying alignment scores, the best alignment was selected. Alignments with more than 2 mismatches or consisting of only one read in a pair were discarded. Single-end SPO11-1-oligo reads were processed and aligned to the Col-CEN assembly using an equivalent pipeline without paired-end options, as described [23]. For each data set, bins per million mapped reads (BPM; equivalent to transcripts per million, TPM, for RNA-seq data) coverage values were generated in bigWig and bedGraph formats with the bamCoverage tool from deepTools v3.1.3 [72]. Reads that aligned to chloroplast or mitochondrial DNA were excluded from this coverage normalization procedure.

RNA-seq data alignment and processing

Paired-end RNA-seq reads (2×100 bp) were processed with Trimmomatic v0.38 to remove adapter sequences and low-quality bases (Phred+33-scaled quality <3 at the beginning and end of each read, and average quality <15 in 4-base sliding windows) [23,73]. Trimmed reads were aligned to the Col-CEN genome assembly using STAR v2.7.0d with the following settings: --outFilterMultimapNmax 100 --winAnchorMultimapNmax 100 --outMultimapperOrder Random --outFilterMismatchNmax 2 --outSAMattributes All --twopassMode Basic --twopass1readsN -1 [74]. Read pairs with STAR-assigned MAPQ <3 were discarded using Samtools v1.9 [56]. For retained read pairs that aligned to multiple locations, with varying alignment scores, the best alignment was selected. Alignments with more than 2 mismatches, or consisting of only one read in a pair, were discarded.

Small RNA-seq data alignment and processing

Small RNA-seq reads [22], were processed with BBDuk from BBMap v38.22 [75], to remove ribosomal sequences and Cutadapt v1.18 [70] to remove adapter sequences and low-quality bases (Phred+33-scaled quality <20). Trimmed reads were aligned to the Col-CEN genome assembly using Bowtie v1.2.2, allowing no mismatches [59]. For reads that aligned to multiple locations, with varying alignment scores, the best alignment was selected. For each small RNA size class (18–26 nucleotides), TPM values in adjacent genomic windows were calculated based on the total retained alignments (across all size classes) in the library.

Bisulfite sequencing data alignment and processing

Paired-end bisulfite sequencing reads (2×90 bp) [22,76], were processed with Trim Galore v0.6.4 to remove sequencing adapters, low-quality bases (Phred+33-scaled quality <20) and 3 bases from the 5′ end of each read [77]. Trimmed reads were aligned to the Col-CEN assembly Bismark v0.20.0 [78]. Read pairs that aligned equally well to more than one location and duplicate alignments were discarded. Methylated cytosine calls in CG, CHG and CHH sequence contexts were extracted and context-specific DNA methylation proportions were generated in bedGraph and bigWig formats using the bismark2 bedGraph and UCSC bedGraphToBigWig tools.

Fine-scale profiling around feature sets

Fine-scale profiles around *CEN180* (*n*=66,129), randomly positioned loci of the same number and width distribution (*n*=66,129), centromeric *ATHILA* elements (*n*=50), and non-centromeric *GYPSY* elements (*n*=3,980) were calculated for ChIP-seq, MNase-seq, RNA-seq, small RNA-seq and bisulfite-seq data sets by providing the above-described bigWig files to the computeMatrix tool from deepTools v3.1.3 in 'scale-regions' mode [72]. Each feature was divided into non-overlapping, proportionally scaled windows between start and end coordinates, and flanking regions were divided into 10 bp windows. Mean values for each data set were calculated within each window, generating a matrix of profiles in which each row represents a feature with flanking regions and each column a window. Coverage profiles for a ChIP input sequencing library and a gDNA library were used in conjunction with those for ChIP-seq and SPO11-1-oligo libraries, respectively, to calculate windowed $\log_2$([ChIP+1]/[control+1]) coverage ratios for each feature. Meta-profiles (windowed means and 95% confidence intervals) for each group of features were calculated and plotted using the feature profiles in R version 4.0.0.

Crossover mapping

Total data from 96 Col×Ler genomic DNA F_2 sequencing libraries (2×150 bp) were aligned to the Col-CEN assembly using bowtie2 (default settings), which gave 87.15% overall alignment. Polymorphisms were identified using the alignment files with samtools mpileup (-vu -f) and bcftools call (-mv -Oz). The resulting polymorphisms were filtered for SNPs (n=522,931), which was used as the 'complete' polymorphism set in TIGER. These SNPs were additionally filtered by, (i) removing SNPs with a quality score less than 200, (ii) removing SNPs where total coverage was greater than 300, or less than 50 (mean coverage=170.8), (iii) removing SNPs that had reference allele coverage less than 20 or greater than 150, (iv) removing SNPs that had variant allele coverage greater than 130, (v) masking SNPs that overlapped transposon and repeat annotations and (vi) masking SNPs within the main *CEN180* arrays. This resulted in a 'filtered' set of 171,947 SNPs for use in TIGER. DNA sequencing data from 260 wild type Col×Ler F_2 genomic DNA (192 from ArrayExpress E-MTAB-4657 and 68 from E-MTAB-6577) was aligned to the Col-CEN assembly using bowtie2 (default settings) and the alignment analyzed at the previously defined 'complete' SNPs using samtools mpileup (-vu -f) and bcftools call (-m -T). These sites were used as an input to TIGER, which identifies crossover positions by genotype transitions [79]. A total of 2,042 crossovers were identified with a mean resolution of 1,011 bp.

Epitope tagging of V5-DMC1

The *DMC1* promoter region was PCR amplified from Col-0 genomic DNA using the Dmc1-PstI-fw and Dmc1-SphI-rev oligonucleotides. The remainder of the *DMC1* promoter, gene and terminator were amplified with oligonucleotides Dmc1-SphI-fw and Dmc1-NotI-rev. The resulting PCR fragments were digested with *Pst*I and *Sph*I, or *Sph*I and *Not*I, respectively, and cloned into *Pst*I-*Not*I-digested pGreen0029 vector to yield a pGreen-DMC1 construct. To insert 3 N-terminal V5 epitope tags, first two fragments were amplified with DMC1-Nco-F and 3N-V5-R and 3N-V5-F and

Dmc1-Spe-rev and then used in an overlap PCR reaction using the DMC1-Nco-F and Dmc1-Spe-rev oligonucleotides. The PCR product resulting from the overlap PCR was digested with *Nco*I and *Spe*I and cloned into *Nco*I- and *Spe*I-digested pGreen-DMC1. The resulting binary vector was used to transform *dmc1-3/+* heterozygotes (SAIL_126_F07). We used dmc1-seq11 and Dmc1-Spe-rev oligonucleotides to amplify wild type *DMC1* allele and Dmc1-Spe-rev and LA27 to amplify the *dmc1-3* T-DNA mutant allele. The presence of the *V5-DMC1* transgene was detected with N-screen-F and N-screen-R oligonucleotides. This oligonucleotide pair amplifies a 74 bp product in Col and a 203 bp product in *V5-DMC1*. To identify *dmc1-3* homozygotes in the presence of *V5-DMC1* transgenes, we used DMC1-genot-compl-F and DMC1-genot-compl-R oligonucleotides, which allowed us to distinguish between the wild type *DMC1* gene and *V5-DMC1* transgene.

Immunocytological analysis

Fresh buds at floral stage 8 and 9 were dissected to release the anthers that contain male meiocytes [80]. Chromosome spreads of meiotic and mitotic cells from anthers were performed, followed by immunofluorescent staining of proteins as described [30]. The antibodies used in this study were: α-ZYP1 (rabbit, 1/500 dilution) [81], α-H3K9me2 (mouse, 1/200 dilution) (Abcam, ab1220), α-CENH3 (rabbit, 1/100 dilution) (Abcam, ab72001) and α-V5 (chicken, 1/200 dilution) (Abcam, ab9113). Chromosomes stained with ZYP1, CENH3 and H3K9me2 were visualized with a DeltaVision Personal DV microscope (Applied Precision/GE Healthcare). Chromosomes stained with DMC1-V5 and CENH3 were visualized with a Leica SP8 confocal microscope. Chromosomes stained with H3K9me2 were visualized with a Stimulated emission depletion nanoscopy mounted on an inverted IX71 Olympus microscope.

4.6 ACKNOWLEDGEMENTS

We thank Isabel Thompson for ATHILA analysis. This work was supported by BBSRC grants BB/S006842/1 and BB/V003984/1, European Research Council Consolidator Award ERC-2015-CoG-681987 'SynthHotSpot', Marie Curie International Training Network 'MEICOM' to IRH, US National Science Foundation grants DBI-1350041 and IOS-1732253 to MCS, and Royal Society awards UF160222 and RGF/R1/180006 to AB.

4.7 REFERENCES

1. Malik HS, Henikoff S. Major evolutionary transitions in centromere complexity. Cell. 2009;138:1067–82.

2. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 2013;14:R10.

3. McKinley KL, Cheeseman IM. The molecular basis for centromere identity and function. Nat Rev Mol Cell Biol. 2016;17:16–29.

4. Rudd MK, Wray GA, Willard HF. The evolutionary dynamics of alpha-satellite. Genome Res. 2006;16:88–96.

5. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020;585:79–84.

6. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, et al. The structure, function, and evolution of a complete human chromosome 8 [Internet]. Cold Spring Harbor Laboratory. 2020 [cited 2021 Mar 12]. p. 2020.09.08.285395. Available from: https://www.biorxiv.org/content/10.1101/2020.09.08.285395v1

7. Vollger MR, Dishuck PC, Sorensen M, Welch AE, Dang V, Dougherty ML, et al. Long-read sequence and assembly of segmental duplications. Nat Methods. 2019;16:88–94.

8. Mikheenko A, Bzikadze AV, Gurevich A, Miga KH, Pevzner PA. TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. Bioinformatics. 2020;36:i75–83.

9. Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, et al. Weighted minimizer sampling improves long read mapping. Bioinformatics. 2020;36:i111–8.

10. Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, et al. Linear assembly of a human centromere on the Y chromosome. Nat Biotechnol. 2018;36:321–3.

11. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36:338–45.

12. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796–815.

13. Maheshwari S, Ishii T, Brown CT, Houben A, Comai L. Centromere location in Arabidopsis is unaltered by extreme divergence in CENH3 protein sequence. Genome Res. 2017;27:471–8.

14. Copenhaver GP, Nickel K, Kuromori T, Benito MI, Kaul S, Lin X, et al. Genetic definition and sequence analysis of Arabidopsis centromeres. Science. 1999;286:2468–74.

15. Talbert PB, Masuelli R, Tyagi AP, Comai L, Henikoff S. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. Plant Cell. 2002;14:1053–66.

16. Martinez-Zapater JM, Estelle MA, Somerville CR. A highly repeated DNA sequence in Arabidopsis thaliana. Mol Gen Genet. 1986;204:417–23.

17. Round EK, Flowers SK, Richards EJ. Arabidopsis thaliana centromere regions: genetic map positions and repetitive DNA structure. Genome Res. 1997;7:1045–53.

18. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245.

19. Pélissier T, Tutois S, Deragon JM, Tourmente S, Genestier S, Picard G. Athila, a new retroelement from Arabidopsis thaliana. Plant Mol Biol. 1995;29:441–52.

20. Wright DA, Voytas DF. Athila4 of Arabidopsis and Calypso of soybean define a lineage of endogenous plant retroviruses. Genome Res. 2002;12:122–31.

21. Thompson HL, Schmidt R, Dean C. Identification and distribution of seven classes of middle-repetitive DNA in the Arabidopsis thaliana genome. Nucleic Acids Res. 1996;24:3017–22.

22. Rigal M, Becker C, Pélissier T, Pogorelcnik R, Devos J, Ikeda Y, et al. Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in Arabidopsis thaliana F1 epihybrids. Proc Natl Acad Sci U S A. 2016;113:E2083–92.

23. Choi K, Zhao X, Tock AJ, Lambing C, Underwood CJ, Hardcastle TJ, et al. Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis thaliana transposons and gene regulatory regions. Genome Res. 2018;28:532–46.

24. Ni P, Huang N, Nie F, Zhang J, Zhang Z, Wu B, et al. Genome-wide Detection of Cytosine Methylations in Plant from Nanopore sequencing data using Deep Learning. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.02.07.430077v1.abstract

25. Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, et al. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat Struct Mol Biol. 2014;21:64–72.

26. Jacob Y, Feng S, LeBlanc CA, Bernatavichute YV, Stroud H, Cokus S, et al. ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. Nat Struct Mol Biol. 2009;16:763–8.

27. Yelagandula R, Stroud H, Holec S, Zhou K, Feng S, Zhong X, et al. The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in Arabidopsis. Cell. 2014;158:98–109.

28. Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell. 2013;152:352–64.

29. Shi J, Wolf SE, Burke JM, Presting GG, Ross-Ibarra J, Dawe RK. Widespread gene conversion in centromere cores. PLoS Biol. 2010;8:e1000327.

30. Lambing C, Tock AJ, Topp SD, Choi K, Kuo PC, Zhao X, et al. Interacting Genomic Landscapes of REC8-Cohesin, Chromatin, and Meiotic Recombination in Arabidopsis. Plant Cell. 2020;32:1218–39.

31. Lambing C, Kuo PC, Tock AJ, Topp SD, Henderson IR. ASY1 acts as a dosage-dependent antagonist of telomere-led recombination and mediates crossover interference in Arabidopsis. Proc Natl Acad Sci U S A. 2020;117:13647–58.

32. Slotkin RK. The epigenetic control of the Athila family of retrotransposons in Arabidopsis. Epigenetics. 2010;5:483–90.

33. Steimer A, Amedeo P, Afsar K, Fransz P, Mittelsten Scheid O, Paszkowski J. Endogenous targets of transcriptional gene silencing in Arabidopsis. Plant Cell. 2000;12:1165–78.

34. Lee SC, Ernst E, Berube B, Borges F, Parent J-S, Ledon P, et al. Arabidopsis retrotransposon virus-like particles and their regulation by epigenetically activated small RNA. Genome Res. 2020;30:576–88.

35. Wijnker E, Velikkakam James G, Ding J, Becker F, Klasen JR, Rawat V, et al. The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. Elife. 2013;2:e01426.

36. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. Genomic and genetic definition of a functional human centromere. Science. 2001;294:109–15.

37. Durfy SJ, Willard HF. Patterns of intra- and interarray sequence variation in alpha satellite from the human X chromosome: evidence for short-range homogenization of tandemly repeated DNA sequences. Genomics. 1989;5:810–21.

38. Vincenten N, Kuhl L-M, Lam I, Oke A, Kerr AR, Hochwagen A, et al. The kinetochore prevents centromere-proximal crossover recombination during meiosis. Elife [Internet]. 2015;4. Available from: http://dx.doi.org/10.7554/eLife.10850

39. Hartmann M, Umbanhowar J, Sekelsky J. Centromere-Proximal Meiotic Crossovers in Drosophila melanogaster Are Suppressed by Both Highly Repetitive Heterochromatin and Proximity to the Centromere. Genetics. 2019;213:113–25.

40. Mahtani MM, Willard HF. Physical and genetic mapping of the human X chromosome centromere: repression of recombination. Genome Res. 1998;8:100–10.

41. Malik HS, Eickbush TH. Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. J Virol. 1999;73:5186–90.

42. Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, et al. Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata. Genes Dev. 2012;26:705–13.

43. Kawabe A, Nasuda S. Structure and genomic organization of centromeric repeats in Arabidopsis species. Mol Genet Genomics. 2005;272:593–602.

44. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol [Internet]. 2019; Available from: https://doi.org/10.1038/s41587-019-0072-8

45. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species [Internet]. Cold Spring Harbor Laboratory. 2020 [cited 2021 Mar 12]. p. 2020.05.22.110833. Available from: https://www.biorxiv.org/content/10.1101/2020.05.22.110833v1

46. Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. Complete structure of the chloroplast genome of Arabidopsis thaliana. DNA Res. 1999;6:283–90.

47. Sloan DB, Wu Z, Sharbrough J. Correction of Persistent Errors in Arabidopsis Reference Mitochondrial Genomes. Plant Cell. 2018;30:525–7.

48. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

49. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2006;22:134–41.

50. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24:1757–64.

51. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36:2896–8.

52. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

53. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 2019;20:224.

54. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.

55. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. arXiv:1303.3997 [q-bio.GN]. 2013. Available from: http://arxiv.org/abs/1303.3997

56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

57. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience [Internet]. 2021;10. Available from: http://dx.doi.org/10.1093/gigascience/giab008

58. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

59. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.

60. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461–8.

61. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. 2020;182:145–61.e23.

62. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics [Internet]. 2020; Available from: http://dx.doi.org/10.1093/bioinformatics/btaa1016

63. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275.

64. Buisine N, Quesneville H, Colot V. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics. 2008;91:467–75.

65. Yamada KD, Tomii K, Katoh K. Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees. Bioinformatics. Oxford Academic; 2016;32:3246–51.

66. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18.

67. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.

68. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7.

69. Bousios A, Kourmpetis YAI, Pavlidis P, Minga E, Tsaftaris A, Darzentas N. The turbulent life of Sirevirus retrotransposons and the evolution of the maize genome: more than ten thousand elements tell the story. Plant J. 2012;69:475–88.

70. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. EMBnet

Stichting; 2011;17:10.

71. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

72. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42:W187–91.

73. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

74. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

75. Bushnell B, Egan R, Copeland A, Foster B, Clum A, Sun H, et al. BBMap: a fast, accurate, splice-aware aligner. 2014. Available: sourceforge net/projects/bbmap. 2019;

76. Yang D-L, Zhang G, Tang K, Li J, Yang L, Huang H, et al. Dicer-independent RNA-directed DNA methylation in Arabidopsis. Cell Res. 2016;26:1264.

77. Krueger F. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. 2015;516:517.

78. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.

79. Rowan BA, Patel V, Weigel D, Schneeberger K. Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. G3 . 2015;5:385–98.

80. Armstrong SJ, Jones GH. Meiotic cytology and chromosome behaviour in wild-type Arabidopsis thaliana [Internet]. Journal of Experimental Botany. 2003. p. 1–10. Available from: http://dx.doi.org/10.1093/jxb/erg034

81. Higgins JD, Sanchez-Moran E, Armstrong SJ, Jones GH, Franklin FCH. The Arabidopsis synaptonemal complex protein ZYP1 is required for chromosome synapsis and normal fidelity of crossing over. Genes Dev. 2005;19:2488–500.

Our work uncovers the prevalence and importance of SVs in plant genomes and demonstrates the underexplored roles of SVs in trait variation.

5

Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato

The following sections were previously published in *Cell*:

> Alonge, M.*, Wang, X.*, Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., Levy, Y., Harel, T. H., Shalev-Schlosser, G., Amsellem, Z., Razifard, H., Caicedo, A. L., Tieman, D. M., Klee, H., Kirsche, M., Aganezov, S., Ranallo-Benavidez, T. R., Lemmon, Z. H., Kim, J., Robitaille, G., Kramer, M., Goodwin, S., McCombie, W. R., Hutton, S., Van Eck, J., Gillis, J., Eshed, Y., Sedlazeck, F. J., van der Knaap, E., Schatz, M. C.†, and Lippman, Z. B†. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182.1 (2020): 145-161. https://doi.org/10.1016/j.cell.2020.05.021

This work is a collaboration with Zachary Lippman's lab at the Cold Spring Harbor Laboratory. M.A. and X.W. contributed equally to this work. M.C.S. and Z.B.L. supervised all analysis and manuscript preparation. M.A., X.W., M.C.S., and Z.B.L. edited all of the text and figures, with input from Y.E. The abstract and background were written by M.A., X.W., M.C.S., and Z.B.L. The results section "Long-Read Sequencing of 100 Tomato Accessions Establishes a PanSV Genome" was done by M.A. with help from F.J.S. and M.C.S. and written by M.A. Figures 5.1B-H were made by M.A. The results section "Fourteen New High-Quality Tomato Reference Genomes" was done by M.A. with help from S.R. and written by M.A. The results section "SV Distribution Reveals Extensive Admixture and Introgression" was done by M.A. and written by M.A. and Z.B.L., with input from S.H. Figure 5.2 was made by M.A. The results section "SVs Associated with Genes Have Widespread Impacts on Expression" was done by H.S. with help from M.A. and X.W. and written by M.A., X.W., and H.S. Figures 5.3A-C were made by M.A. The methods section "Long-read Structural Variant Calling, Filtering, and Merging" was done by M.A. with help from M.C.S. and written by M.A. The methods sections "MAS2.0 Genome Assembly", "PI129033 NSGT Local Assembly", "SV Hotspot and Introgression Analysis", and "SV Genomic Feature Annotation" were done and written by M.A. The methods section "MAS 2.0 and SV Repeat Annotation" was done and written by M.A. and F.M. All other results and methods sections as well as the discussion were done and written by authors other than M.A.

5.1 ABSTRACT

Structural variants (SVs) underlie important crop improvement and domestication traits. However, resolving the extent, diversity, and quantitative impact of SVs has been challenging. We used long-read nanopore sequencing to capture 238,490 SVs in 100 diverse tomato lines. This panSV genome, along with 14 new reference assemblies, revealed large-scale intermixing of diverse genotypes, as well as thousands of SVs intersecting genes and *cis*-regulatory regions. Hundreds of SV-gene pairs exhibit subtle and significant expression changes, which could broadly influence quantitative trait variation. By combining quantitative genetics with genome editing, we show how multiple SVs that changed gene dosage and expression levels modified fruit flavor, size, and production. In the last example, higher order epistasis among four SVs affecting three related transcription factors allowed introduction of an important harvesting trait in modern tomato. Our findings highlight the underexplored role of SVs in genotype-to-phenotype relationships and their widespread importance and utility in crop improvement.

5.2 BACKGROUND

Phenotypic variation in crop plants is shaped by genetic variation from their wild ancestors, as well as the selection and maintenance of collections of mutations that impact agricultural adaptations and human preferences [1,2]. The majority of this variation is quantitative, and now more than ever, a major goal of genetics is to identify and understand how specific genes and variants contribute to quantitative trait variation. In particular, this knowledge is necessary for designing and engineering favored alleles in crop improvement, enabled by genome editing [3–5]. Although high-throughput short-read sequencing accelerated the discovery of natural genetic variants among diverse germplasm of major crops, it has also introduced an unavoidable bias: characterized variants are disproportionately skewed toward single-nucleotide polymorphisms (SNPs) and small indels [6]. However, decades of research have shown that structural variations (SVs) (large deletions, insertions, duplications, and chromosomal rearrangements) are important in plant evolution and agriculture, affecting traits such as shoot architecture, flowering time, fruit size, and stress resistance [7]. Compared to SNPs, SVs can cause large-scale perturbations of *cis*-regulatory regions and are therefore more likely to quantitatively change gene expression and phenotypes. SVs can also modify expression levels by directly altering gene copy number. However, despite their importance, identifying SVs with short-read sequencing is notoriously difficult and unreliable, leaving the vast majority of SVs poorly resolved and their molecular and phenotypic impacts largely hidden [8,9].

High-throughput Oxford Nanopore Technologies (ONT) long-read sequencing now enables a broad survey of population-scale SV landscapes. Such resources that capture the diversity of SVs, in combination with expression profiling and genome editing, immediately allow for the direct interrogation of the molecular and phenotypic consequences of SVs. Here, we present the most comprehensive panSV genome for a major crop and study its significance in evolution, domestication, quantitative genetics, and breeding. We used ONT long-read sequencing to identify SVs from a collection of 100 diverse wild and domesticated tomato accessions. Tomato, in addition to its agricultural and economic importance, has extensive genetic resources, well-described phenotypic diversity, and efficient genome editing, making it an ideal system to investigate the significance of SVs in both fundamental plant biology and agriculture. Our data provided continuous long-range information that allowed for the sequence-resolved inference of more than 200,000 SVs, the majority being transposons and related repeat sequences. Patterns of SV distribution revealed extensive admixture and population-scale introgressions. RNA sequencing showed that gene expression is widely impacted by SVs affecting both coding and *cis*-regulatory regions. Establishing high-quality de novo genome assemblies for 14 selected genotypes allowed us to resolve hidden genomic complexity involving SVs. To demonstrate the value of this panSV genome, we directly linked these complex alleles with multiple domestication and improvement traits affecting fruit flavor, size, and productivity. For two of these traits, modest changes in expression originated from gene copy number variation, and we used CRISPR-Cas9 genome editing to demonstrate causal quantitative relationships between gene dosage and phenotype. Our work uncovers the prevalence and importance of SVs in plant genomes and demonstrates the underexplored roles of SVs in trait variation.

5.3 RESULTS

Long-read sequencing of 100 tomato accessions establishes a panSV genome

To deeply survey the landscape of natural structural variation in tomato, we collected long-read sequencing data from a representative population-scale tomato panel (**Figure 5.1A**). To this end, we first used available short-read sequencing data to call SVs from over 800 tomato accessions and then applied the SVCollector algorithm to optimally select 51 diverse modern and early domesticated samples that maximize SV diversity [10]. We then separately selected an additional 49 wild species and modern accessions that are used by tomato research and breeding communities. Our final set of 100 accessions captures phylogenetic diversity spanning the closest wild relatives of domesticated tomato (*S. pimpinellifolium* [SP], *S. cheesmaniae* [CHE], and *S. galapagense* [GAL]), early domesticated forms (*S. lyc. var. cerasiforme* [SLC]), and "vintage" cultivars and modern varieties (*S. lycopersicum*, [SLL]; **Figure 5.1A**).

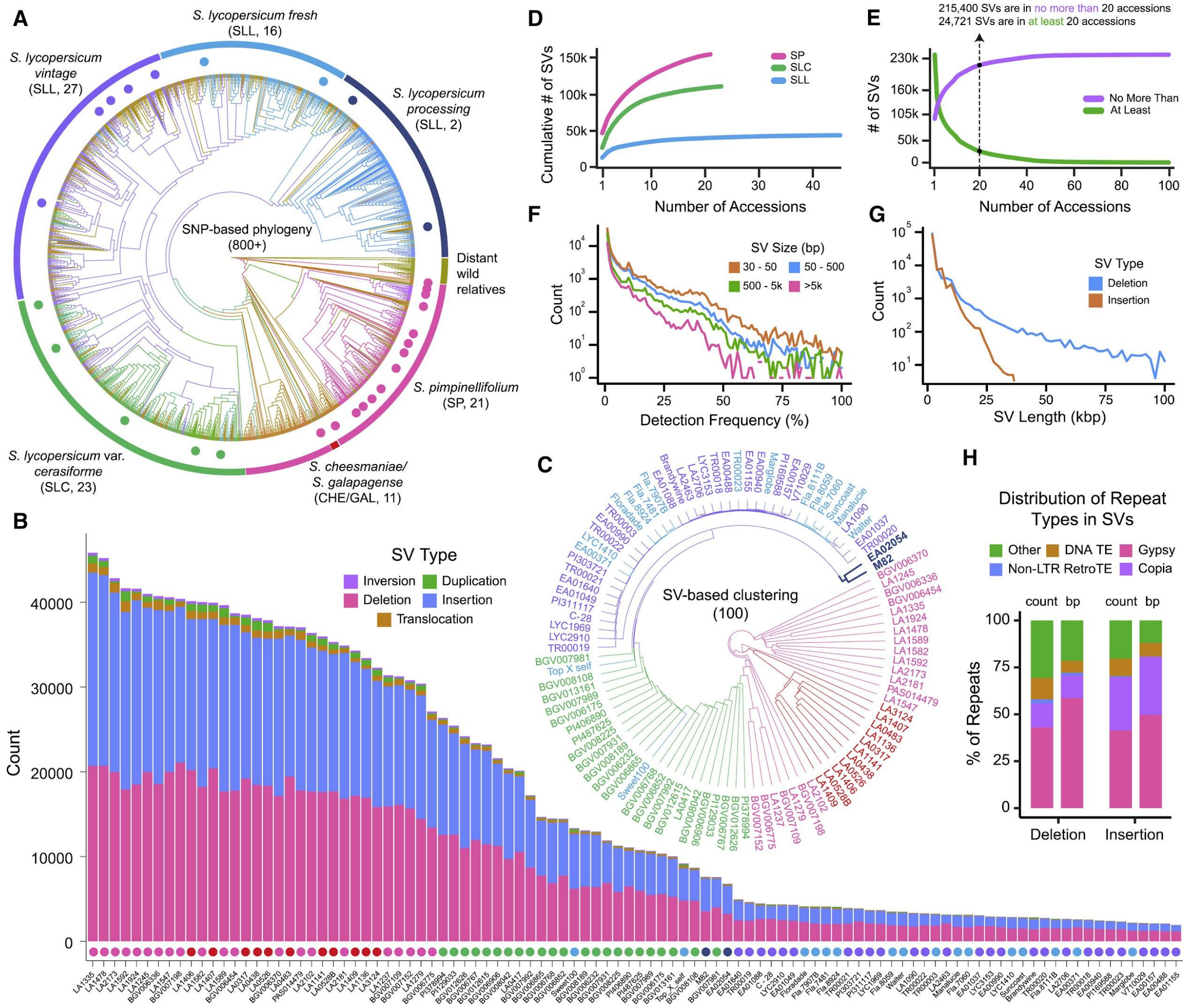


Figure 5.1: The Tomato panSV Genome. (A) SNP-based phylogenetic tree based on short-read sequencing of more than 800 tomato accessions. Major taxonomic groups are marked by colored lines along the circumference. Colored dots indicate a subset of the 100 accessions selected for long-read sequencing. (B) Stacked bar graph showing SV number and type from the 100 accessions. Colored dots indicate the taxonomic group of each accession, corresponding to colors in (A). (C) Hierarchical clustering dendrogram of the SV presence/absence matrix across the 100 accessions, with colors corresponding to (A). Bold branches and names highlight an outgroup of two SLL processing tomato accessions. (D) SVCollector curves of SVs in the three major taxonomic groups. The "greedy" algorithm determines the order of accessions and depicts the cumulative number of SVs as a function of the number of accessions included. (E) Graph showing the number of SVs (y-axis) in "no more than" or "at least" the number of accessions indicated on the x-axis. (F) Histograms of detection frequencies for different SV sizes. (G) Histogram of SV sizes for insertions and deletions. (H) Annotation of the panSV genome. The proportion of repeat types for all insertions and deletions annotations is shown in stacked bar graphs. "Count" shows the proportion of individual repeat annotations, and "bp" shows the proportion of cumulative repeat (not indel) sequence length. "Other" refers to other repeat types. Only indels at least 100 bp in size were considered.

For each of the 100 accessions, we used ONT long-read sequencing to generate a minimum of 40× genome coverage, achieving a total of 7.77 Tb of long-read data with an average read length N50 of 19.6 kbp. Reads were aligned to the recently released SL4.0 reference genome (Heinz 1706, SLL) with NGMLR, and SVs were called with Sniffles [9,11]. We then filtered, sequence resolved, and merged all 100 sets of SV calls, revealing 238,490 total SVs (defined in this study as >30 bp) that comprise the most comprehensive sequence-resolved panSV genome in plants. Importantly, we confirmed that the majority of these variants would not have been revealed using solely short-read sequencing data.

Individual accessions had between 1,928 and 45,840 SVs, with the wild SP, GAL, and CHE accessions harboring the most structural variation relative to the Heinz reference genome (**Figure 5.1B**). Insertions and deletions were the most common SV type, though we also found dozens to hundreds of inversions, duplications, and translocations in all samples. SVs are with respect to the reference genome and do not necessarily reflect the underlying evolutionary context. Clustering of the SV presence/absence matrix revealed a structure that mirrored the larger SNP-based tomato phylogeny, with accessions clustering within their known taxonomic groups (**Figure 5.1C**). Interestingly, the SLL "cherry" variety Sweet100 grouped with the SLCs, and the only two processing cultivars, M82 and EA02054, form a distinct group from the SLLs, suggesting admixture. Comparative analysis of the long-read SVs showed that SP and SLC have more SV diversity compared to SLL, consistent with the loss of genetic variation during the domestication and improvement of tomato (**Figure 5.1D**) [12,13]. This analysis also indicated that even sequencing 100 accessions, many SVs remain to be discovered (**Figure 5.1E**). Consistently, the majority of SVs are singletons, or are otherwise rare, although tens of thousands of SVs are common (>5% detection frequency) (**Figure 5.1F**). We evaluated SV length distribution, which showed that most SVs were relatively small: 30.5%: 30–50 bp; 30.5%: 50–200 bp; and 39%: >200 bp (**Figure 5.1G**). We note that our method has limited ability to detect larger insertions, because, unlike deletion calling, such detection is bounded by read length. SVs are typically composed of, or generated by, transposons and related repeats [14,15], and annotation of our panSV genome showed 84% of deletions and 76% of insertions larger than 100 bp match at least one repeat. Retrotransposon sequences, especially from Gypsy and Copia elements, are the most prevalent among the annotated SVs (**Figure 5.1H**).

Fourteen new high-quality tomato reference genomes

To supplement the panSV genome with additional genomic resources, we selected 14 diverse accessions for genome assembly and annotation. Combining long- and short-read sequencing data, de novo assemblies using the MaSuRCA hybrid assembler yielded an average contig N50 of 1.9 Mbp [16]. Reference-guided scaffolding with RaGOO produced chromosome-scale pseudomolecules that contained, on average, a single copy of 96% of complete benchmarking universal single-copy orthologs (BUSCO) genes [17,18]. Repeats were annotated using REPET, and genes annotations were "lifted over" from reference annotations using geneLift [19]. We used these new reference genomes (referred to as "MAS2.0") to validate SVs in the same 14 accessions, of which 90% were also found in the assemblies. Owing to the diversity of these assemblies, which represent multiple SP, SLC, and SLL accessions, we anchored 22% of recently discovered "pan-genome" genes that are missing from the ITAG reference annotation [20]. These MAS2.0 genomes were critical to link complex SV loci with functional consequences shown below.

SV distribution reveals extensive admixture and introgression

The chromosomal distribution of SVs from our panSV genome revealed several hypervariable genomic regions relative to the Heinz reference shared among subsets of SLL accessions (designated SV "hotspots") (**Figure 5.2A**). Because SP accessions have more structural variants than those of SLL, SV hotspots in SLL could reflect admixture and introgression between wild and domesticated accessions, which was previously partially explored using SNPs [12,21,22]. Introgression is a common practice in tomato breeding, through which disease resistance genes and other desirable traits from wild donors are introduced into SLL breeding germplasm [12]. We found that SV hotspots in SLL correlated with genomic regions that show high similarity with SP and/or SLC based on the Jaccard similarity of SV content between accessions. For example, multiple SV hotspots exist on chromosome 4, including a 2-Mbp region common to all SLL accessions that corresponds to a known unique introgression in the Heinz reference genome (**Figure 5.2A**) [22]. Most SP accessions show a decrease in SV frequency in this region, indicating these accessions are closely related to the introgression donor. We also found a large introgression block shared by five SLLs that occupies two-thirds of the chromosome (**Figure 5.2B**). Notably, two of these accessions are M82 and EA02054, which also carry large introgression blocks that span nearly all of chromosomes 5 and 11

(**Figure 5.2A**), explaining their distinct grouping in SLL and their relatively large number of SVs compared to Heinz 1706, which is also a processing type (**Figures 5.1B and 5.1C**).

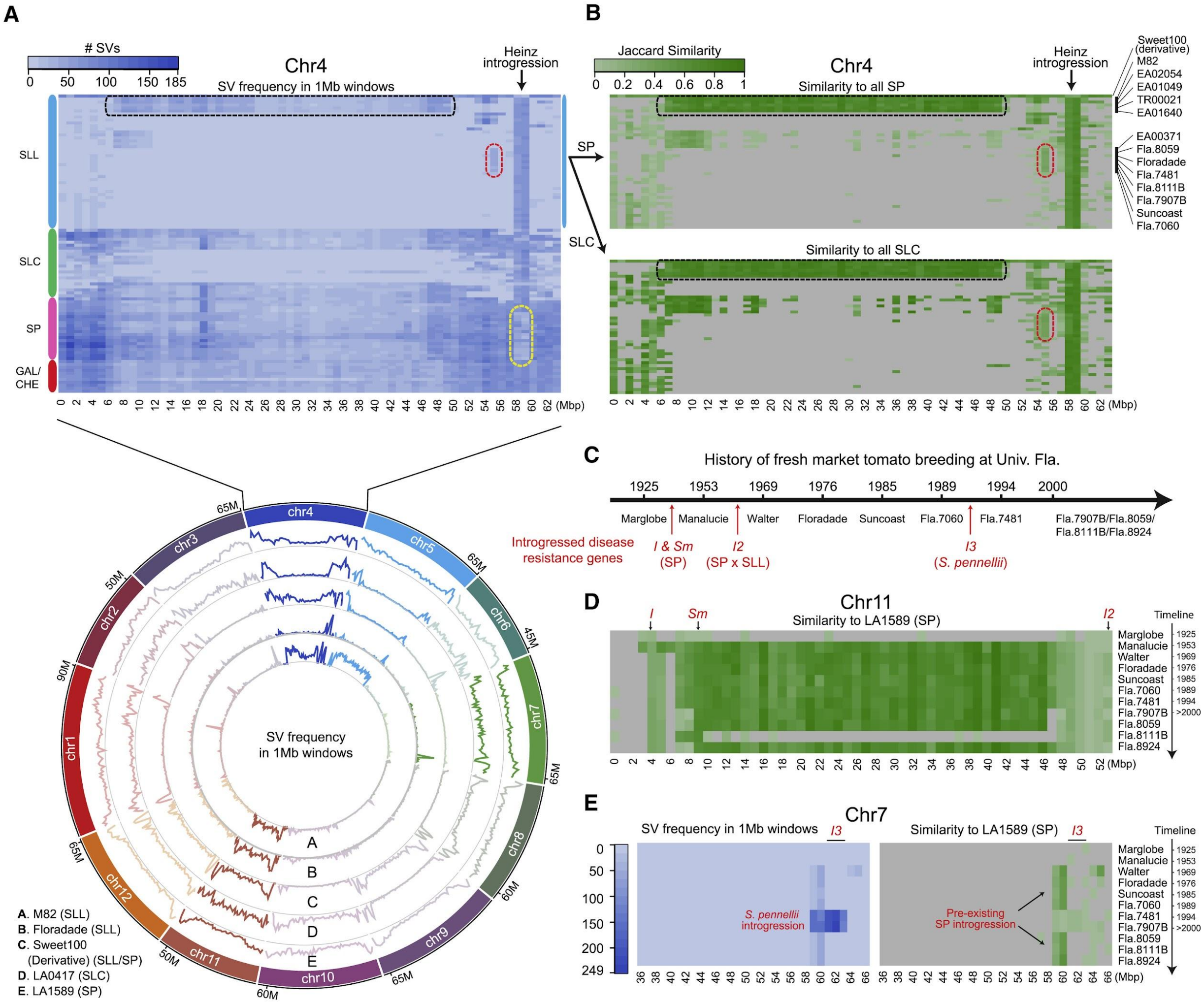


Figure 5.2: SV Distribution Reveals Large-Scale Admixture and Introgression between Wild and Domesticated Genotypes. (A) Heatmap (top) showing SV frequency in 1-Mbp windows (columns) of chromosome 4 relative to the reference genome. Accessions (rows) are grouped by taxonomic group (colored bars). Dotted colored lines mark three notable regions: black, a large SV hotspot for 5 SLLs; red, a small hotspot shared by most UFL SLL lines; and yellow, a SP group with reduced SV frequency, reflecting a small SP introgression in the reference genome. Circos plot (bottom) depicts genome-wide SV frequency for five notable accessions. Rings depict line plots showing the SV number in successive 1-Mbp windows (y-axes are not shared between rings). Chromosomes 4, 5, 7, and 11 are highlighted to show regions of high SV frequency. (B) Heatmaps showing admixture and introgressions on chromosome 4 measured by Jaccard similarity between accessions of SLL and SP (top) and SLC (bottom) in the same row order as (A) (top). For each 1-Mbp window, the SVs for a given SLL accession are compared to the SVs for all SP (top) or SLC (bottom) accessions and the maximum Jaccard similarity is reported. Windows with fewer than 5 SVs in the SLL set are excluded and colored gray. Black and red dotted regions correlate with marked SV hotspots in (A) (top). (C) Timeline of UFL fresh market variety release over the last century. Approximate periods of introgression of key disease-resistance genes are shown in red, along with major donor genotypes for Fusarium wilt (*I*, *I2*, and *I3*) and gray leaf spot (*Sm*). (D) Jaccard similarity for chromosome 11 between the UFL lines (ordered chronologically) and LA1589, the closest SP to this introgression. Locations of *I*, *Sm*, and *I2* are shown in red. (E) The UFL varieties on chromosome 7 showing a small SP introgression in all but two accessions; Fla.7481 and Fla.7907B carry a unique SV hotspot (left) due to introgression of the I3 resistance gene (red) from *S. pennellii*.

Expecting that our panSV genome would illuminate how breeding and introgression have shaped SV content, we examined 11 SLLs included in our 100 genomes from the University of Florida (UFL) tomato breeding program, which has a well-documented history of disease resistance gene introgression [23]. The devastating fungal disease Fusarium wilt first emerged in the 1930s, and the resistance genes *I* and *I2* (from SP donors) and *I3* (from *S. pennellii*) against three races of this disease were successively introduced into UFL breeding material between the 1930s and 1980s (**Figure 5.2C**) [24–26]. Furthermore, the *Sm* resistance gene against Grey leaf spot was introduced in the 1950s [27]. Molecular mapping and gene cloning have shown that *I* and *Sm* are located on the opposite arms from *I2* on chromosome 11. The variants from our panSV genome demonstrated overlapping introgressions from multiple donors, including those contributing resistance to other diseases [28], accounting for the large introgression block in the UFL accessions (**Figure 5.2D**). Interestingly, the modern breeding line Fla.8111B carries the *I*, *I2*, and *Sm* resistance genes but lacks a large portion of this introgression, suggesting this region was later purged during selection.

The *I3* introgression on chromosome 7 was introduced in the 1980s (**Figure 5.2C**). The modern breeding lines Fla.7481 and Fla.7907B that carry *I3* resistance show a 5-Mbp SV hotspot with low similarity to SP and SLC at the *I3* locus, consistent with the donor being the distant green-fruited wild species *S. pennellii* (**Figure 5.2E**). Interestingly, UFL lines lacking *I3* resistance have a 2-Mbp introgression from SP or SLC that first appeared in the 1960s and overlaps the *I3* introgression. The *I3* introgression is negatively implicated with several horticultural characteristics, including reduced fruit size and increased sensitivity to bacterial spot [29–31]. The earlier introduced SP introgression may have provided tolerance to bacterial spot or benefitted other traits, as is likely for many other putative SP or SLC introgressions revealed by our panSV genome. The large number of SVs from wild species introduced in breeding could have broad functional consequences.

<u>*SVs associated with genes have widespread impacts on expression*</u>

SVs may influence the expression of nearby genes by altering the sequence or copy number of a gene or by changing the composition or position of *cis*-regulatory sequences [32,33]. We explored this relationship with the comprehensive catalog of SVs across our tomato panSV genome. Candidate SVs that could potentially impact gene expression were abundant in our collection. Nearly 50% (112,114) of SVs overlap genes and/or flanking regulatory sequences (±5 kbp of coding sequence), and among 34,075 annotated genes, 95% have at least one SV within 5 kbp of coding sequences across the 100 genomes, with the majority found in *cis*-regulatory regions (**Figures 5.3A and 5.3B**). To explore the impact of SVs on gene expression, we performed 3′ RNA sequencing (RNA-seq) on three tissues (cotyledons, roots, and apical meristems) for 23 accessions that capture 44,358 gene-associated SVs. We evaluated a total of 21,156 SV-gene pairs and found hundreds of significant expression changes (**Figure 5.3C**). Nearly half of the SVs affecting coding sequences (deletions of coding sequence [CDS] start, deletions of exons, and duplications) are significantly associated with differences in expression, with many substantially reducing or eliminating expression. In regulatory regions, 1,534 SV-gene pairs (7.3%) showed significant differential expression across all tissues, and overall, these differences were subtler compared to SVs in coding regions (mean log2 fold change 1.36 and 2.47, respectively).

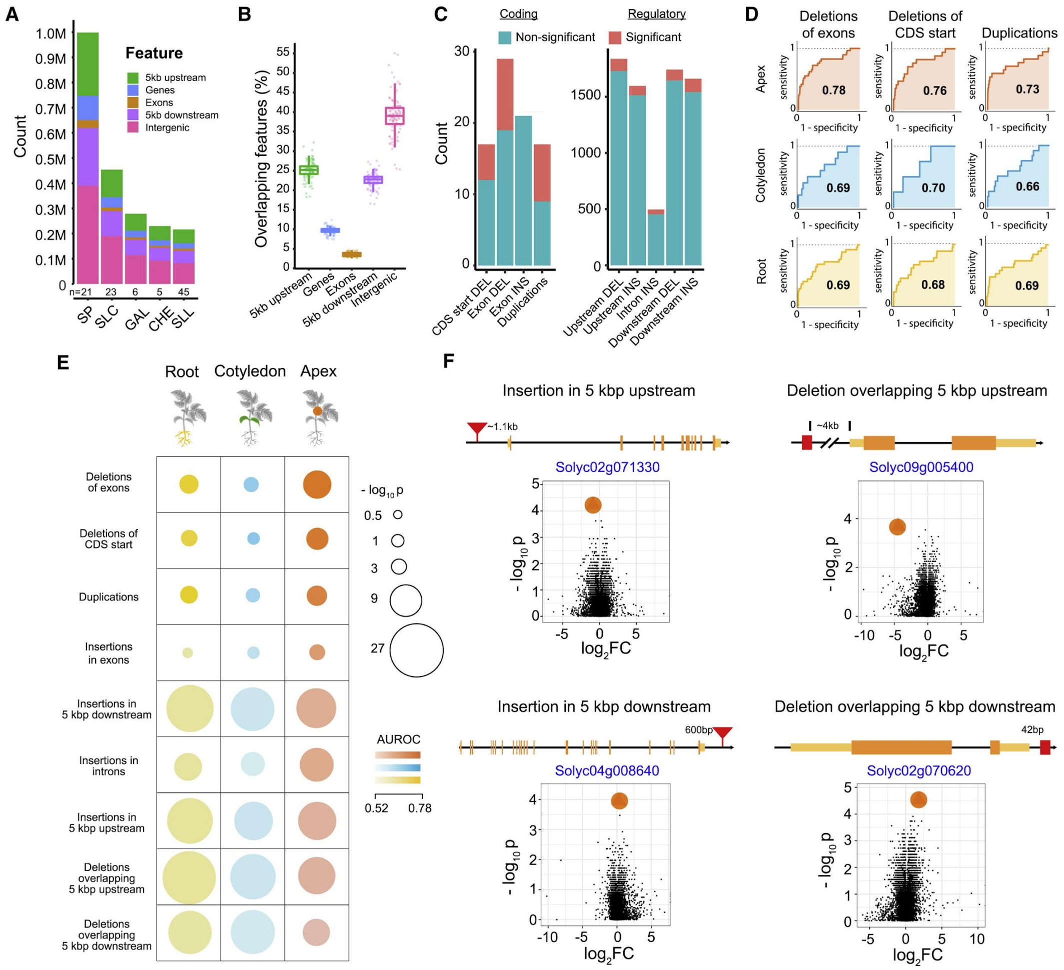


Figure 5.3: Gene-Associated SVs Impact Expression. (A) Stacked bar chart showing total counts of SVs overlapping different genomic features in major taxonomic groups. N represents the number of accessions in each taxonomic group. (B) Percentage of SVs overlapping different genomic features in 100 accessions. Each point is one sample. Fewer SVs are found within genes compared to surrounding regulatory regions. (C) Stacked bar charts showing numbers of differentially expressed genes affected by insertion, deletion, and duplication SVs overlapping coding sequences (left) and regulatory regions (right; significance is defined as an adjusted p < 0.05). Differential expression was tested on common SVs in the 23 accessions used for RNA sequencing (frequency between 0.2 and 0.8; see STAR Methods). (D) ROC curves for the top three SV annotation types, with high AUROC (area under the receiver operating characteristics) scores across the three tissues demonstrating the ability to identify genes containing SVs using changes in expression across the accession split. The AUROC is specified within the ROC curve in each case. The steep rise of the curves in the top panel corresponds to a near-perfect identification of a large fraction of the genes containing SVs based on differential expression. CDS, coding sequence. (E) Differential expression significantly predicts genes with SVs. Overall performance of using "SV splits" and differential expression to predict associated gene(s) is shown. Analyses are broken down into 9 categories across three tissues. Each category is defined based on SV type and relative position to genes. Circle sizes and colors represent the significance of performance (−log10 p-value) and the magnitude of AUROC, respectively. SV categories are ranked in decreasing order of average AUC (area under the curve) across the three tissues. Note that the significance of performance for each SV type is enhanced by the number of annotated SV-gene pairs (for example, p < 1 × 10−4 for duplications, although p < 1 × 10−8 for insertions in introns). (F) Volcano plots for four regulatory SV-gene pair examples with the highest AUROC score highlight the extent of differential expression of SV-containing genes (marked in orange circles), compared to all expressed genes (black dots). p values and expression fold changes are computed across two groups of accessions (with and without the indicated SV). Data shown are for apex tissue. Exons (orange), UTRs (yellow), and SVs (red) are not drawn to scale. Distances between genes and SVs are shown.

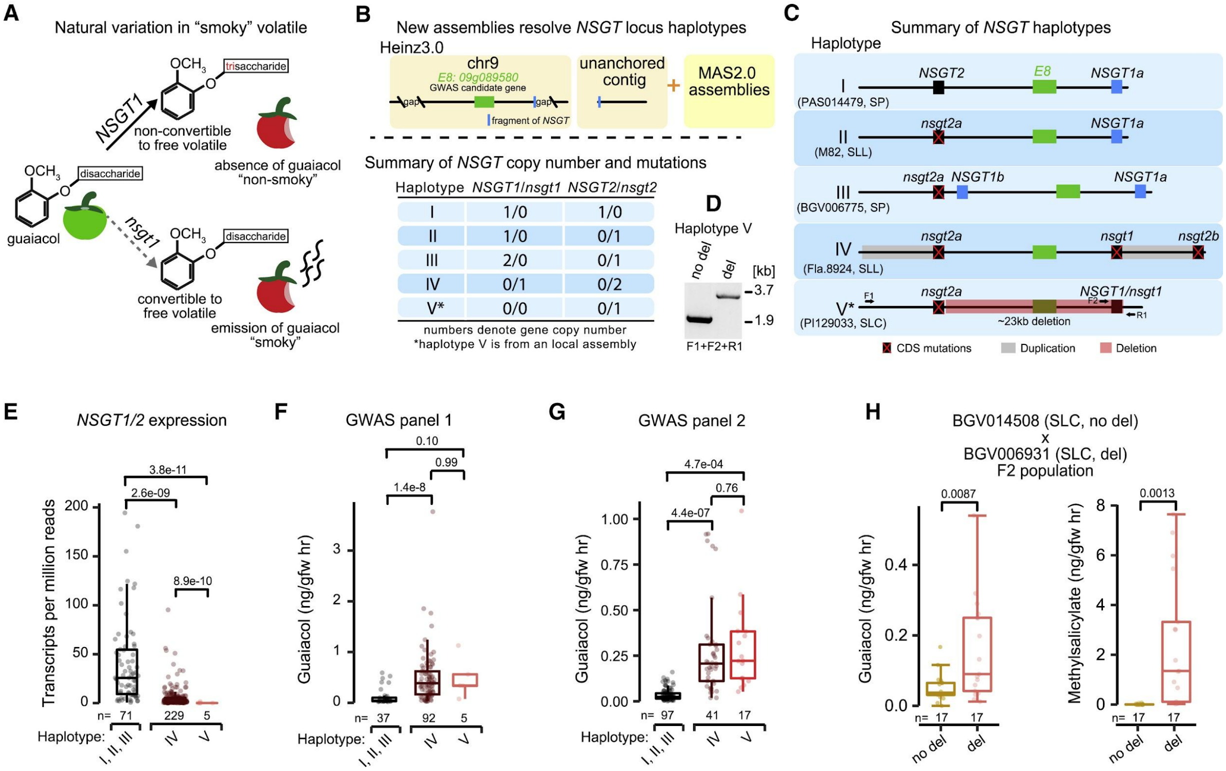
Knowing that a substantial fraction of population-scale expression variation is explained by *cis*-eQTL [34,35], we next formulated a classification task that uses changes in gene expression to predict the presence of a nearby SV. This classifier complements standard fold-change measurements among known SV-gene pairs, and its performance can quantify the extent to which global expression changes are associated with SVs. Notably, this test is robust to population structure because global changes in expression or confounding variants can only serve to weaken any one-to-one relationship between gene expression variation and the existence of a nearby variant.

Broadly, differential expression significantly predicts genes with associated SVs (**Figures 5.3D and 5.3E**). As expected, this classifier performs best on the coding sequence SVs (e.g., deletions of exons, apex tissue expression, area under the receiver operating characteristics [AUROC] > 0.78, and false discovery rate [FDR] < 0.05), as reflected by the sharp initial rise in receiver operating characteristic (ROC) curves (**Figure 5.3D**). The strength of this signature indicates that indirect effects (e.g., *trans* regulation) do not dominate the observed relationship and also demonstrates the high accuracy of our variant calls. Importantly, we also observe subtle but significant effects of regulatory SVs on gene expression (e.g., deletions overlapping 3′ flanking sequence, apex tissue expression, AUROC > 0.53, and FDR < 0.05). The AUROC captures the individual *cis*-regulatory effect size, which is small on a per variant basis. However, in aggregate, these variants have a large impact on expression variation (**Figure 5.3E**), suggesting they globally shape expression profiles. Overall, our results show that SVs can impact gene expression in both substantial and subtle ways and that many such variants in our panSV genome may be functionally relevant (**Figure 5.3F**).

New reference genomes resolve multiple haplotypes for the smoky volatile locus

Our panSV genome, new MAS2.0 assemblies, and expression dataset could help to reveal genes and variants underlying quantitative trait variation that has been masked by hidden genomic complexity. Many fruit aroma volatile QTLs that contribute to flavor have been identified through genome-wide association study (GWAS), but only a few have been functionally characterized [36,37]. One such QTL involves the metabolically linked volatiles guaiacol and methyl salicyate, whose "smoky" or "medicinal" flavors negatively influence consumer appeal. A previous GWAS identified a candidate gene *E8* (*Solyc09g089580*), encoding a putative negative regulator of ethylene biosynthesis involved in fruit ripening [36]. Although transcriptional knockdown of *E8* resulted in accumulation of guaiacol and methyl salicylate, other volatiles were also modified. Furthermore, no causal mutations were identified, likely due to two large gaps flanking *E8* in the reference genome at the time (SL3.0).

A separate study found that mutations in the *NON-SMOKY GLYCOSYLTRANSFERASE1* (*NSGT1*) and *NSGT2* paralogous genes, which are physically close on chromosome 9, cause an accumulation of guaiacol (**Figure 5.4A**) [38]. Whereas *NSGT2* shows little expression and is believed to be non-functional, upregulation of *NSGT1* during ripening converts guaiacol to non-cleavable triglycosides, preventing guaiacol volatilization [38]. To investigate whether *NSGT* genes could be linked to the smoky QTL, we inspected the previous reference genome SL3.0 and found a partial sequence of *NSGT1* near the gap at the chromosome 9 GWAS locus and another *NSGT1* fragment at a second GWAS peak on an unanchored contig (**Figure 5.4B**) [36]. Consistently, a recent short-read k-mer-based analysis also linked the two smoky GWAS peaks and suggested hidden structural complexity [39]. However, all these studies failed to resolve this locus. Importantly, our new MAS2.0 assemblies not only filled the gaps flanking *E8* with these two *NSGT* paralogs but also further revealed coding sequence variants and SVs that are resolved into five haplotypes (**Figures 5.4B and 5.4C**).



Haplotype	NSGT1/nsgt1	NSGT2/nsgt2
I	1/0	1/0
II	1/0	0/1
III	2/0	0/1
IV	0/1	0/2
V*	0/0	0/1

numbers denote gene copy number
*haplotype V is from an local assembly

Figure 5.4: New Reference Genomes Anchor Candidate Genes and Resolve Multiple SV and Coding Sequence Haplotypes for the "Smoky" Volatile GWAS Locus. (A) Schematic showing a key step of the metabolic pathway underlying the "smoky" aroma trait. During fruit ripening, activation of glycosyltransferase *NSGT1* prevents release of smoky-related volatiles by converting them into non-cleavable triglycosides (top). *nsgt1* mutations result in the release of the smoky volatile guaiacol. (B) Genomic resources used to resolve the GWAS locus for guaiacol (top) and summary of haplotypes (bottom). The published locus mapped to a region of chromosome 9 with one candidate gene and multiple gaps and also to an unanchored contig with a fragment of an *NSGT* gene (top). MAS2.0 assemblies revealed multiple haplotypes that include copy number variation for the *NSGT1* and *NSGT2* paralogs and loss-of-function mutations (bottom). A local assembly revealed haplotype V (asterisk). (C) Schematics depicting the five resolved haplotypes. The assemblies and major taxonomic groups from which the haplotypes were identified are shown below. Red "X"s mark coding sequence (CDS) mutations. Grey bars mark duplication in haplotype IV. The red rectangle marks a large deletion in haplotype V. (D) PCR confirmation of the deletion in haplotype V. Primers (F1, F2, R1) are shown in (C). (E) Quantification of *NSGT1*/2 expression by RNA-sequencing. Haplotypes are grouped according to functional *NSGT1* (I, II, III), *nsgt1* CDS mutation (IV), and *nsgt1* deletion (V). Expression data are from pericarp tissue of ripe fruit [37]. (F and G) Guaiacol content of fruits from a previous GWAS study (F) [36] and a new GWAS analysis using a collection of 155 SP and SLC accessions (G). Mutations in *NSGT1* are associated with guaiacol accumulation. Accessions are grouped as in (E). (H) Quantification of guaiacol and methylsalicylate content in an SLC x SLC F_2 population segregating for the haplotype V 23 kbp deletion. In (E–H), n represents sample size in each group. All p values are based on two-tailed, two-sample t tests.

Haplotype I is likely ancestral with the *NSGT1* and *NSGT2* genes flanking *E8*. Although an *NSGT2* coding sequence mutation is found in all other haplotypes, haplotypes II and III have intact *NSGT1*, with the latter carrying two copies of *NSGT1* (**Figure 5.4C**). Finally, copy number and functional variation are extended in haplotypes IV and V; haplotype IV has a 7-kbp duplication, including mutant *nsgt2* that disrupted *NSGT1*, rendering it non-functional, and haplotype V has a large 23-kbp deletion that removes both *NSGT1* and *E8*, leaving only a single mutated copy of *nsgt2* (**Figure 5.4D**).

These haplotypes, along with the previous characterization of *NSGT1* [38], suggest that multiple mutant alleles of *nsgt1* are responsible for natural variation in guaiacol (and methyl salicylate) accumulation and the smoky flavor. Using gene expression and metabolite data from fruits of more than 300 accessions [36,37], we tested associations between functional (I, II, and III), coding sequence non-functional (IV), and deletion non-functional (V) *NSGT1* haplotypes and guaiacol accumulation. Accessions carrying the mutant haplotypes IV and V, which emerged early in domestication in the SLCs, exhibited lower combined *NSGT1/2* expression levels compared to accessions with functional haplotypes, with no *NSGT1/2* expression detected in the five accessions carrying the haplotype V deletion (**Figure 5.4E**). Consistently, both mutant haplotypes accumulated more guaiacol, though the effect from the rare haplotype V showed weak statistical significance (**Figure 5.4F**). We validated these findings using a new GWAS panel of 155 accessions comprised primarily of SP and SLC genotypes [40]. Again, both *nsgt1* coding and deletion mutation haplotypes accumulate significantly more guaiacol than functional haplotypes (**Figure 5.4G**). Finally, we generated an F_2 population between two SLCs segregating for haplotype V and functional *NSGT1*, which confirmed the deletion, lacking both *NSGT1* and *E8*, is associated with accumulation of both guaiacol and methyl salicylate (**Figure 5.4H**). Together, our results anchored two *NSGT* genes to the smoky GWAS QTL and show that multiple *nsgt1* mutations largely explain natural variations of the smoky flavor. This example demonstrates how our high-quality long-read genome assemblies can resolve complex haplotypes and reveal causative variants for poorly understood QTLs.

The fruit weight QTL fw3.2 resulted from a tandem duplication of a cytochrome P450 gene

A substantial increase in fruit weight was a major feature of tomato domestication [41]. The genes underlying five major fruit weight QTL have been identified, with the responsible mutations being either SVs or SNPs [42–46]. Among these is *fw3.2*, which is strongly associated with a SNP in the promoter of the cytochrome P450 gene *SlKLUH*, a known regulator of organ size in multiple species [42,47,48]. The promoter SNP was proposed to account for higher (2- to 3-fold) *SlKLUH* expression (**Figure 5.5A**), and transcriptional knockdown of this gene results in smaller fruits, though a causative role for the SNP was unclear.

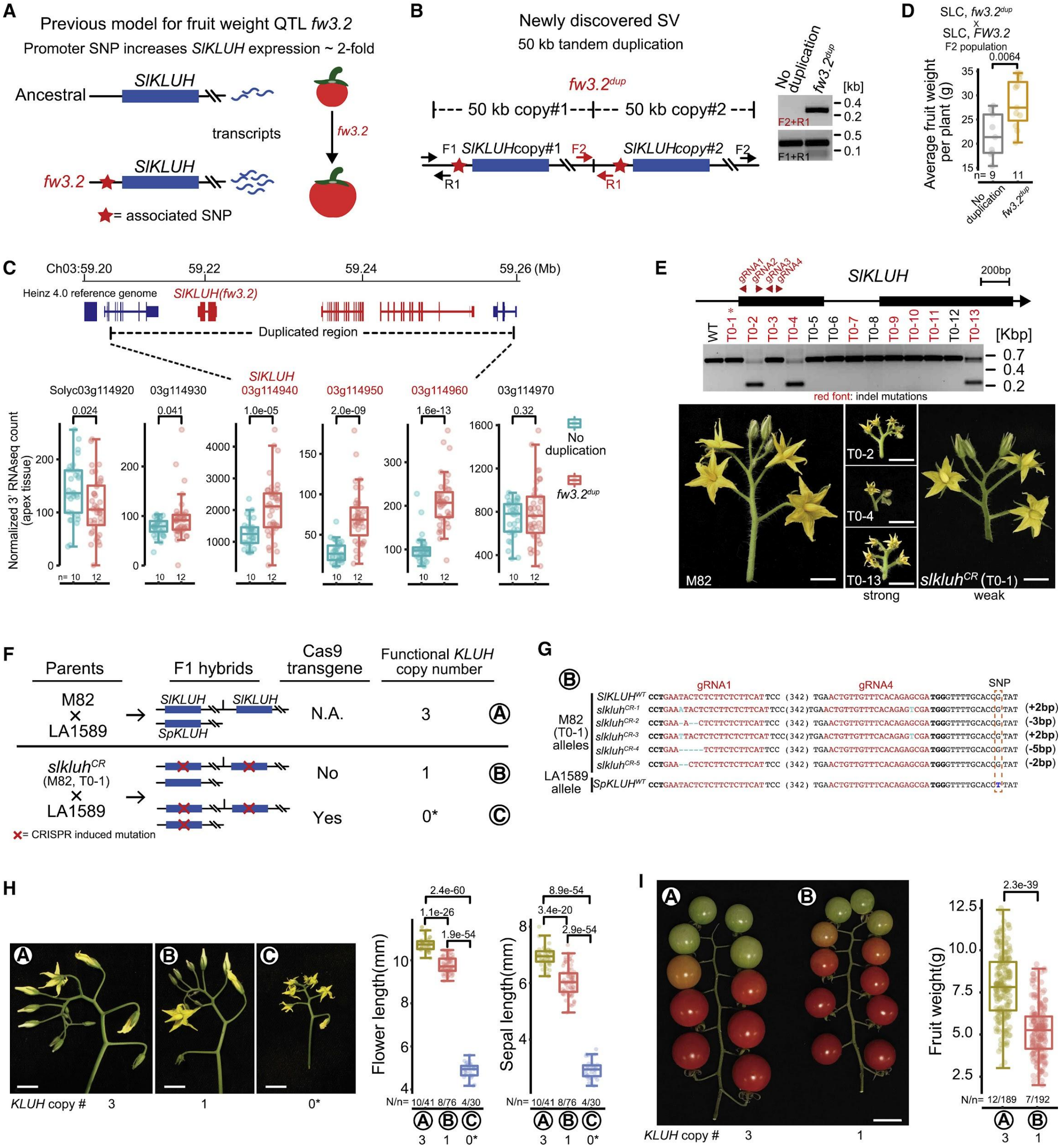


Figure 5.5: The Fruit Weight QTL *fw3.2* Resulted from a Tandem Duplication that Increased Expression of a Cytochrome P450 Gene. (A) Published mechanism for *fw3.2* positing that a SNP in the promoter of the cytochrome P450 gene *SlKLUH* increased expression ~2-fold, resulting in larger fruits. (B) SV analyses revealed a 50-kb tandem duplication at the *fw3.2* locus that included *SlKLUH* (left). PCR validation of the duplication (right) is shown. Primers (F1, F2, and R1) are labeled on the left. "No duplication" refers to the accession without this duplication, and "*fw3.2^dup*" refers to the accession that carries the duplicated copy of *fw3.2* as shown by the PCR product across the duplication junction (F2 + R1). (C) Expressions of genes within the *fw3.2* duplication are ~2-fold higher. Gene coordinates and the duplication region (top) and RNA-seq boxplots of duplicated and flanking genes (bottom) are shown. Each point is one biological replicate from one accession. n, number of accessions. (D) An SLC × SLC F_2 population segregating for the *fw3.2* duplication but fixed for the promoter SNP. Increased fruit weight is associated with the duplication. (E) CRISPR-Cas9 mutagenesis of *SlKLUH* in the M82 background. *SlKLUH* gene model with gRNA targets (top), PCR genotyping (middle), and representative inflorescences (bottom) of *slkluh^CR* T0 plants is shown. The three *slkluh^CR* T0 plants shown have mutations in all four copies of *SlKLUH* and exhibit similar tiny inflorescences, suggesting a null phenotype. Strong phenotypes were also observed for other T0 plants with sequenced indels (red font), except T0-1, which showed a weaker phenotype and was fertile, allowing a genetic test of dosage. (F) Altering tomato *KLUH* gene dosage shows that copy number variation explains *fw3.2*. Schematic shows the M82/M82^{CR} *slkluh* T0-1 (SL) × LA1589 (SP) crossing scheme used to test the phenotypic effects of altering tomato *KLUH* functional copy number in an F_1 hybrid isogenic background. Genotypic groups A and B are isogenic for M82 × LA1589 genome-wide heterozygosity and differ only in having 3 or 1 functional copies of tomato *KLUH*, respectively. Genotypic group C effectively has 0 functional copies due to inheritance of the single insertion Cas9 transgene that targets the single *SpKLUH* allele in *trans*. (G) Mutated *slkluh* alleles and the *SpKLUH* allele in genotypic group B. Red font, guide RNA targets; cyan font, mutations. An LA1589 SNP (blue font) permits distinction of *KLUH* allele parent of origin. All *SpKLUH* sequences in genotypic group B are wild type. (H) Decreasing tomato *KLUH* functional copy number reduces flower organ size. Representative inflorescences (left) and quantifications of flower and sepal length (right) from all three genotypic groups are shown. (I) Decreasing tomato *KLUH* functional copy number reduces fruit weight. Representative fruits (left) and fruit weight quantification (right) from genotypic groups A and B are shown. Reducing tomato *KLUH* copy number from three to one reduces fruit size by 30%. Genotypic group C plants with mutated *SpKLUH* alleles fail to produce fruits. Scale bars represent 1 cm in (E) and (H) and 2 cm in (I). In (H) and (I), N indicates plant number and n indicates flower/fruit number. All p values are based on two-tailed, two-sample t tests.

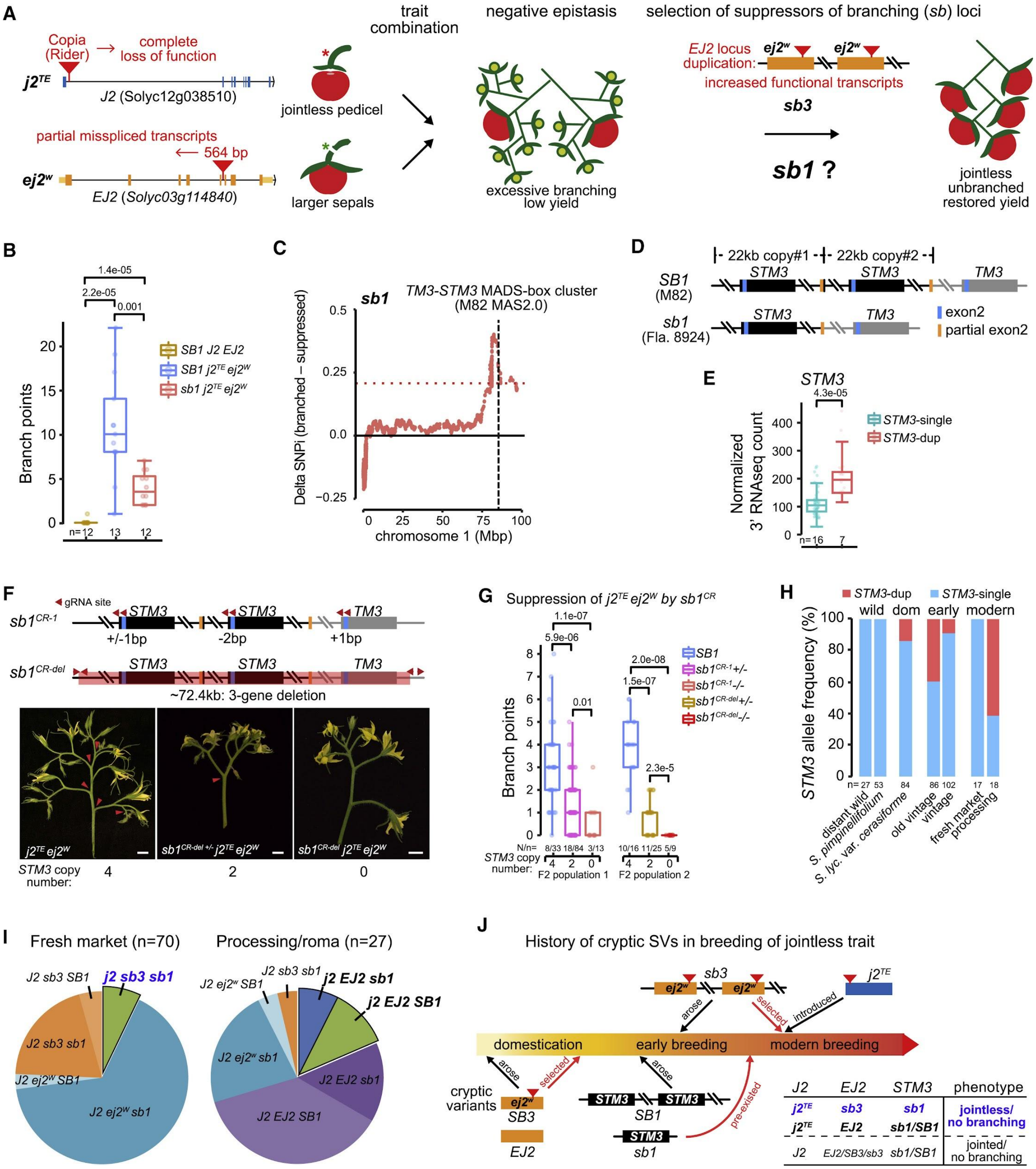
Our panSV genome revealed a ~50-kbp tandem duplication at the *fw3.2* locus containing three genes, including two identical copies of *SlKLUH* (designated *fw3.2^dup*) (**Figure 5.5B**). Although SNPs in promoters can affect expression by modifying *cis*-regulatory elements, we explored whether *fw3.2^dup* is the causative variant, with the hypothesis that an increase in gene copy number explains the higher expression. In support of this, our expression analyses showed that all three intact genes within the duplication are expressed approximately 2-fold higher in accessions carrying *fw3.2^dup* (**Figure 5.5C**). To disentangle the effects of these variants on fruit weight, we generated F_2 populations segregating for *fw3.2^dup* but fixed for the promoter SNP and other known fruit weight QTLs. Higher fruit weight co-segregated with the duplication allele (**Figure 5.5D**). In contrast, there was no association between the promoter SNP and fruit weight in F_2 populations segregating only for the SNP.

Our results suggested that the duplication carrying *SlKLUH* could explain *fw3.2* due to an increase in gene copy number and therefore dosage. We tested this by CRISPR-Cas9 targeting *SlKLUH* in the processing cultivar M82 (carrying *fw3.2^dup* and therefore four functional copies of *SlKLUH*) with multiple guide RNAs (gRNAs). PCR genotyping and sequencing of independent T0 plants showed large deletions and small indels in the target sites. The majority of these plants, including three confirmed to lack wild-type (WT) alleles, were much smaller than control plants and had tiny inflorescences and flowers that were infertile (**Figure 5.5E**).

Fortuitously, one fertile plant (*slkluh*^{CR} T0-1) showed a weaker phenotype from having both WT and mutant alleles, allowing us to directly test how changes in *SlKLUH* dosage affect fruit weight. To work in an isogenic background with uniform cherry type fruits that allows for a robust assessment of fruit size, we crossed the *slkluh*^{CR} T0-1 with the SP accession LA1589. As LA1589 has only two copies of *SpKLUH* (**Figure 5.5F**), the M82 × LA1589 F$_1$ isogenic hybrids have three gene copies of *KLUH* (2 copies *SlKLUH* and 1 copy *SpKLUH*). These control F$_1$ hybrids (group A) were compared with F$_1$ progeny resulting from the cross between *slkluh*^{CR} T0-1 and LA1589. Several F$_1$ hybrid plants that inherited the Cas9 transgene produced small organs and were infertile (group C), which we confirmed was due to inheritance of mutated and further *trans* targeting of all *KLUH* copies. Among F$_1$ plants lacking the Cas9 transgene, a subset inherited two mutated alleles of *SlKLUH* and a single functional allele of *SpKLUH* (group B) (**Figures 5.5F and 5.5G**). Notably, these group B plants produced 15% smaller flowers and 30% smaller fruits compared to group A plants (1 versus 3 functional alleles of *KLUH*) (**Figures 5.5H and 5.5I**). Thus, our panSV genome and functional genetic dissection using CRISPR-Cas9 genome editing show that the duplication including *KLUH*, and the corresponding increase in gene dosage and expression, underlies *fw3.2*.

<u>*Genetic interactions involving four SVs allowed jointless breeding*</u>

We revealed thousands of genes with expression variation that could be caused by SVs. These variants might have little or no phenotypic consequences; however, many may be “cryptic,” having little or no effect on their own but causing phenotypic changes in the context of other variants [49,50]. The “jointless” fruit pedicel is an important tomato harvesting trait that originated from different mutations from wild and domesticated accessions [51]. The jointless trait allows complete separation of fruits from other floral parts and is caused by a transposon insertion that eliminates functional transcripts of the MADS-box transcription factor gene *JOINTLESS2* (*J2*). A cryptic insertion in the related ENHANCER OF *J2* (*EJ2*) gene reduces functional transcripts and causes excessive inflorescence branching with reduced fruit production following introduction of the jointless trait (**Figure 5.6A**). Breeders overcame this negative interaction and restored normal inflorescences by exploiting two natural "suppressor of branching" (*sb*) QTLs that we designated *sb1* and *sb3* [52]. We recently showed that *sb3* is an 83-kbp duplication that includes *ej2*^{w}, which causes a dose-dependent increase of weak allele expression that compensates for the reduced functional transcripts (**Figure 5.6A**).



J2	EJ2	STM3	phenotype
j2TE	sb3	sb1	jointless/ no branching
j2TE	EJ2	sb1/SB1	
J2	EJ2/SB3/sb3	sb1/SB1	jointed/ no branching

Figure 5.6: Four SVs in Three MADS-Box Genes Were Required to Breed for the Jointless Trait. (A) Genetic suppressors were selected to overcome a negative epistatic interaction on yield caused by mutations in two MADS-box genes. The SV mutation *j2*^{TE} causes a desirable jointless pedicel that facilitates harvesting. Introducing *j2*^{TE} in backgrounds carrying the cryptic SV mutation *ej2w* results in excessive inflorescence branching and low fertility. The *sb1* and *sb3* QTLs were selected to suppress *j2*^{TE} *ej2*^{w} negative epistasis. *sb3* is an 83-kb duplication harboring *ej2w*. *sb1* is cloned in this study. (B) Quantification of *sb1* partial suppression of branching in the *j2*^{TE} *ej2*^{w} background. The *SB1 j2*^{TE} *ej2*^{W} and *sb1 j2*^{TE} *ej2*^{W} genotypes were derived from F_3 families. Each data point is one inflorescence from F4 plants (n). (C) Delta SNP index (deltaSNPi, QTL-seq) plot shows the *sb1* locus contains the *TM3-STM3* MADS-box gene cluster. (D) Schematic of the *TM3-STM3* locus in the SLL genotypes M82 and Fla.8924, with M82 having an ~22-kb tandem duplication (designated *SB1*) containing *STM3*. (E) RNA-seq showing increased expression of *STM3* from the *SB1* duplication compared to *sb1*. (F) CRISPR-Cas9 mutagenesis of the *TM3-STM3* cluster (*sb1*^{CR}) suppresses branching in the *j2*^{TE} *ej2*^{w} background. Schematics at top depict two CRISPR lines with indel mutations in the *STM3* and *TM3* genes (*sb1*$^{CR\text{-}1}$) and a large deletion spanning all three genes (*sb1*$^{CR\text{-}del}$; top). Representative inflorescences from the indicated genotypes (bottom) are shown. Arrowheads mark branch points. (G) Quantification and comparison of suppression of inflorescence branching by homozygous and heterozygous *sb1*$^{CR\text{-}1}$ and *sb1*$^{CR\text{-}del}$ mutations in the background of *j2*^{TE} *ej2*^{w}. Genotypes were derived from F_2 populations. N, plant number; n, inflorescence number. (H) *STM3* duplication allele frequency in wild tomato species (distant relatives and SP), early domesticates and cultivars (SLC and SLL vintage), and modern cultivars (SLL fresh market and processing). (I) Distribution of *J2 EJ2 SB1* genotypes in fresh-market and processing/roma tomato types. All *j2* fresh-market genotypes carry *sb1* and *sb3*, whereas processing/roma genotypes have *SB1* or *sb1*, because *EJ2* is functional. (J) Schematic showing the history of breeding for the jointless trait, including when SVs in *EJ2* and *STM3* arose. The pre-existing *sb1* cryptic variant (single-copy *STM3*) mitigated the severity of branching caused by introduction of *j2*^{TE} in varieties carrying the cryptic variant *ej2*^{w}. Selection of the *sb3* cryptic variant (two copies of *ej2*^{w}) resulted in the complete suppression of branching and restoration of normal yield. Gradient colored bar represents timeline. The table summarizes genotypic combinations. Blue and black bold fonts indicate solutions for jointless breeding in fresh market and processing/roma types, respectively (I and J). In (B), (E), (H), and (I), n represents sample size. p values in (B) and (G) are based on two-tailed, two-sample t tests.

The cryptic *sb1* locus is a partial suppressor of branching, and our previous QTL mapping positioned *sb1* to a 6-Mbp interval on chromosome 1 (**Figures 5.6B and 5.6C**). We searched for candidate genes and focused on two neighboring MADS-box paralogs, *TM3* (*Solyc01g093965*) and SISTER OF *TM3* (*STM3*) (*Solyc01g092950*). Notably, *STM3* showed approximately 2-fold higher expression in the branched parental line (M82 *j2*^{TE} *ej2*^{W}) compared to the suppressed parent (Fla.8924 *j2*^{TE} *ej2*^{W}). There were no obvious coding or regulatory mutations in this gene; however, the Heinz 4.0 reference genome has gaps in that area. Our MAS2.0 assemblies filled the gaps and revealed copy number variation for *STM3*, with an extra copy of the gene in the branched parent due to a near-perfect 22-kbp tandem duplication (**Figure 5.6D**). Consistently, genotypes with four copies of *STM3* showed 2-fold higher expression compared to two copy genotypes (**Figure 5.6E**).

To test whether lower dosage and expression from a single *STM3* gene is responsible for the *sb1* QTL, we used CRISPR-Cas9 to generate mutant alleles disrupting the complex *STM3-TM3* locus. A CRISPR construct with two gRNAs gave small indel mutations in all copies of the identical *TM3*/*STM3* exon 2 (*sb1*$^{CR\text{-}1}$), although a second construct with four gRNAs deleted the entire locus (*sb1*$^{CR\text{-}del}$) (**Figure 5.6F**). Both *sb1*$^{CR\text{-}1}$ and *sb1*$^{CR\text{-}del}$ plants were slightly late flowering, but their inflorescences were normal. We then introduced each allele into the highly branched M82 *j2*^{TE} *ej2*^{w} double mutants and identified *j2*^{TE} *ej2*^{w} *sb1*$^{CR\text{-}1}$ and *j2*^{TE} *ej2*^{w} *sb1*$^{CR\text{-}del}$ triple mutants from segregating F_2 populations. Importantly, all of these plants (0 functional copies of *STM3*) showed practically complete suppression of branching compared to *j2*^{TE} *ej2*^{w} double mutants (4 functional copies of *STM3*) (**Figures 5.6F and 5.6G**). Moreover, *j2*^{TE} *ej2*^{w} plants that were heterozygous for the CRISPR alleles (2 functional copies of *STM3*) showed partial suppression of inflorescence branching, mimicking the effect of *sb1* (e.g., Fla.8924; 2 functional copies of *STM3*) (**Figures 5.6F and 5.6G**). Thus, a single-copy *STM3*, and the corresponding lower gene expression, explains *sb1*.

Short-read-based genotyping of more than 500 accessions spanning tomato taxonomic groups showed that the duplication of *STM3* arose early in domestication, but the ancestral single gene has remained common in tomato germplasm (**Figure 5.6H**). In fact, the majority of vintage and modern fresh-market accessions have single-copy *STM3*, indicating that a lower dosage and expression level provided partial suppression of branching upon the introduction of *j2*^{TE} into lines carrying *ej2*^{w}. The duplication of *ej2*^{w}, and the resulting increased expression of this weak allele, arose later and was likely selected to achieve complete suppression of branching. In support, all jointless fresh-market accessions carry both *sb1* (single-copy *STM3*) and *sb3* (duplicated *ej2*^{w}) (**Figure 5.6I**). In contrast, breeding for jointless in processing tomato accessions was achieved by selecting against *ej2*^{w} (**Figure 5.6I**). Consistent with this, *sb1* and *SB1* (duplicated *STM3*) are present at equal frequencies in processing tomato accessions, maintaining cryptic variation in the context of inflorescence development (**Figures 5.6I and 5.6J**). Our analysis reveals *STM3* as a new regulator of tomato inflorescence development, and the dissection of *sb1* shows that the path of jointless breeding depended on four SVs affecting the expression levels of three MADS-box genes and further illustrates how functional consequences of structural variation can remain hidden.

5.4 DISCUSSION

Raising the curtain on structural variation

Advancements in genome-sequencing technologies continue to revolutionize biology by providing an increasingly comprehensive view of the genetic changes underlying phenotypic diversity. The recent development of high-throughput Oxford Nanopore long-read sequencing has provided the opportunity to rapidly reveal the breadth and depth of previously hidden SVs in complex genomes and across populations [53]. Taking advantage of the expansive genetic diversity of wild and domesticated tomatoes, we sequenced a collection of 100 accessions and resolved hundreds of thousands of SVs. These SVs were shaped predominately by transposons [54], are abundant across all chromosomes, frequently reside within or in close proximity to genes, are often associated with expression, and likely contribute to phenotypic variation. Integrating our panSV genome, de novo assemblies, and expression data with genome-editing enabled us to resolve and functionally link SVs to three major domestication and breeding traits. The smoky and *sb1* loci, in particular, demonstrate how these resources were essential to resolve complex haplotypes underlying QTLs where previous assemblies were thwarted by repeats, especially highly similar long and local duplications. Moreover, our analyses of the smoky and *fw3.2* loci show that presumed causative variation may be incomplete or incorrect. More broadly, most QTLs discovered by GWAS in model and crop plants reside in regions with multiple candidate genes and variants. In addition to improving GWAS statistical power, long-read-based discovery of abundant, sometimes complex SVs may immediately pinpoint high-confidence candidate genes and variants for functional analyses. Similar progress in understanding the functional impacts of SVs will likely emerge from generating population-scale panSV genomes in other species [33,55–58].

Duplications, gene copy number variation, and dose-dependent phenotypes

Our panSV genome revealed that *fw3.2* and *sb1* were both associated with previously hidden duplications. In both plants and animals, duplications that alter copy number and expression of dosage-sensitive genes were found to modify phenotypic diversity, including traits important in domestication and breeding [7]. Large, tandem, recent duplications are one of the most challenging SVs to resolve, and even when a strong candidate gene is present, as with *SlKLUH* in the *fw3.2* duplication, directly testing how modified gene dosage and expression impacts quantitative variation is challenging.

Enabled by CRISPR-Cas9 genome editing, we generated plants with different gene copy numbers, and therefore dosages, for *SlKLUH* and *STM3* in the *fw3.2* and *SB1* duplications, respectively. Establishing a dosage series of isogenic genotypes not only confirmed the causality of the duplications and the specific genes but also directly demonstrated their quantitative impact. In particular, heterozygotes of *sb1*^{CR} alleles (2 copies of *STM3* on 1 chromosome) suppressed inflorescence branching of *j2*^{TE} *ej2*^{W} plants to a similar degree as the natural dosage effect from single-copy *STM3* (1 copy of *STM3* on each chromosome). Similarly, reducing functional *KLUH* copy number from three to one recapitulated the natural quantitative effect on fruit size of having four or two copies. Manipulating gene copy number by genome editing now provides a way to systematically interrogate and explore dosage to phenotype relationships [59], which will be important for guiding the design and engineering of specific dosages for crop improvement.

cis-regulatory SVs and quantitative variation

Our panSV genome showed that the majority of gene-associated SVs are in *cis*-regulatory regions, and many are associated with subtle changes in expression. Expanding long-read sequencing and expression analyses to a wider population will reveal even more such SVs. This raises the question to what extent *cis*-regulatory SVs affect phenotypes. For genes that are dosage sensitive, such as those encoding components of molecular complexes or involved in signaling networks, a subtle change in expression could alter phenotype [59]. However, the magnitude of the phenotypic effect may depend on a threshold change in expression and could be weak, making detection challenging in population genetics studies where other mutations and alleles influence trait variation. Genome editing could be used to study the effects of gene-associated SVs by recreating specific mutations or mimicking the expression effects of natural *cis*-regulatory SVs in isogenic backgrounds. Our previous work characterizing collections of CRISPR-Cas9-engineered promoter alleles in multiple developmental genes showed that deletion and inversion SVs can affect expression and phenotypic outputs in various, often unpredictable ways [4]. As SVs could be cryptic, a more powerful and informative approach would therefore be to sensitize the locus or genome, by combining natural *cis*-regulatory SVs with engineered SVs in the same promoter or with engineered mutations in related, potentially redundant genes. Resolving the functional impacts of SVs, particularly those whose effects are subtle or cryptic, will advance our understanding of genotype-to-phenotype relationships and facilitate the exploitation of natural and engineered SVs in crop improvement.

5.5 METHODS

Plant material and growth conditions

A hundred tomato accessions were collected from TGRC (Tomato Genetics Resource Center), USDA (United States Department of Agriculture), University of Florida, EU-SOL (The European Union-Solanaceae project), INRA (The National Institute for Agricultural Research), IVF-CAAS (The Institute of Vegetables and Flowers, Chinese Academy of Agricultural Science) and our stocks. The landrace collection (*S. lycopersicum var. cerasiforme*) was from the seed stocks of E. van der Knaap. Seeds of *S. pimpinellifolium* (LA1589), *S. lycopersicum* cv. M82 (LA3475), and $j2^{TE}$ $ej2^{w}$ mutant are from Lippman lab.

Seeds were either germinated on moistened filter paper at 28 °C in the dark or directly sown in soil in 96-cell plastic flats. Plants were grown under long-day conditions (16-h light/8-h dark) in a greenhouse under natural light supplemented with artificial light from high-pressure sodium bulbs (~250 μmol m-2 s-1). Daytime and nighttime temperatures were 26–28 °C and 18–20 °C, respectively, with a relative humidity of 40%–60%.

Quantification of fruit guaiacol and methylsalicylate contents in this study was conducted from plants grown in North Florida Research and Education Center-Suwannee Valley near Live Oak, Florida. Analyses of fruit weight in F_2 segregation populations were conducted on plants grown at the University of Georgia (Athens, GA). Analyses of floral organ size, fruit weight of F_1 hybrid plants and inflorescence branching in F_4 generation were conducted on plants grown in the fields at Cold Spring Harbor Laboratory (CSHL), Cold Spring Harbor, NY. Seeds were germinated in 96-cell flats and grown for 32 d in the greenhouse before being transplanted to the field. Plants were grown under drip irrigation and standard fertilizer regimes. Analyses of inflorescence branching in two sb^{CR} $j2^{TE}$ $ej2^{W}$ F_2 populations were conducted on plants grown in the greenhouses at CSHL and Weizmann Institute of Science, Israel.

Short-read structural variant calling and sample selection

Publicly available short-read data came from a total of four sources [12,13,36,37]. Phylogenetic trees derived from some of these data have been adapted from their original publication and are shown in **Figure 5.1A** [40,52]. Phylogenetic classifications (branch coloring) were manually curated according to these previous phylogenetic studies and based on knowledge of tomato types and breeding classes. First, the raw reads were trimmed with Trimmomatic (v0.32, LEADING:30 TRAILING:30 MINLEN:75 TOPHRED33) [60]. Reads we aligned to the SL4.0 reference genome with bwa mem (v0.7.10-r789, -M) [11,61]. Alignments were then compressed, sorted, and indexed with samtools view, sort, and index respectively (v0.1.19-44428cd) [62]. Next, PCR duplicates were marked with Picard (v1.126) (https://broadinstitute.github.io/picard/). We removed any samples that had less than 5X alignment coverage or any samples that had a duplication rate > = 20%. If a given accession had more than one associated BAM file, they were merged with samtools.

An ensemble approach was used to call SVs from these short-read alignments. We and others have found that a consensus among multiple short-read SV callers can achieve higher precision without substantially decreasing sensitivity [63]. We used 3 independent tools to call SVs: Delly (v0.7.3, -q 20), Lumpy (v0.2.13, -mw 4 -tt 0.0) and Manta (v1.0.3, -j 15 -m local -g 30) [64–66]. For each accession, SV call sets from Delly, Lumpy, and Manta were then merged with SURVIVOR (v1.0.7, minimum distance of 1kbp, types must match, and a minimum length of 10bp) [67]. Only SVs called by at least 2 of the 3 tools were retained. In total, we produced short-read SV calls for 847 accessions.

We then used SVCollector to select our first set of accessions for long-read sequencing [10]. For SVCollector, we further filtered short-read SV calls to only include SVs that intersect genes (+/− 5 kbp of flanking sequence). These filtered SVs were then used as input into SVCollector (greedy), and the top-ranked SLL (29) and SLC (22) accessions for which we had available seeds were selected. Aside from these 51 accessions selected with SVCollector, we selected an additional 49 accessions for long-read sequencing. These included SLL, SP, GAL, and CHE accessions which were not included in the short-read SV analysis.

Tissue collection and high molecular weight DNA extraction

For extraction of high molecular weight DNA, young leaves were collected from 21-day-old light-grown seedlings. Before tissue collection, seedlings were etiolated in complete darkness for 48 h. Flash-frozen plant tissue was ground using a mortar and pestle and extracted in four volumes of ice-cold extraction buffer 1 (0.4 M sucrose, 10 mM Tris-HCl pH 8, 10 mM MgCl2, and 5 mM 2-mercaptoethanol). Extracts were briefly vortexed, incubated on ice for 15 min, and filtered twice through a single layer of Miracloth (Millipore Sigma). Filtrates were centrifuged at 4000 rpm for 20 min at 4 °C, and pellets were gently resuspended in 1 ml of extraction buffer 2 (0.25 M sucrose, 10 mM Tris-HCl pH 8, 10 mM MgCl2, 1% Triton X-100, and 5 mM 2-mercaptoetanol). Crude nuclear pellets were collected by centrifugation at 12,000g for 10 min at 4 °C and washed by resuspension in 1 ml of extraction buffer 2 followed by centrifugation at 12,000g for 10 min at 4 °C. Nuclear pellets were re-suspended in 500 µl of extraction buffer 3 (1.7 M sucrose, 10 mM Tris-HCl pH 8, 0.15% Triton X-100, 2 mM MgCl2, and 5 mM 2-mercaptoethanol), layered over 500 µl extraction buffer 3, and centrifuged for 30 min at 16,000g at 4 °C. The nuclei were resuspended in 2.5 ml of nuclei lysis buffer (0.2 M Tris pH 7.5, 2 M NaCl, 50 mM EDTA, and 55 mM CTAB) and 1 ml of 5% Sarkosyl solution and incubated at 60 °C for 30 min. To extract DNA, nuclear extracts were gently mixed with 8.5 ml of chloroform/isoamyl alcohol solution (24:1) and slowly rotated for 15 min. After centrifugation at 4000 rpm for 20 min, ~3 ml of aqueous phase was transferred to new tubes and mixed with 300 µl of 3 M NaOAC and 6.6 ml of ice-cold ethanol. Precipitated DNA strands were transferred to new 1.5 ml tubes and washed twice with ice-cold 80% ethanol. Dried DNA strands were dissolved in 100 µl of elution buffer (10 mM Tris-HCl, pH 8.5) overnight at 4 °C. Quality, quantity, and molecular size of DNA samples were assessed using Nanodrop (Thermofisher), Qbit (Thermofisher), and pulsed-field gel electrophoresis (CHEF Mapper XA System, Biorad) according to the manufacturer's instructions.

Short-read DNA sequencing

Aside from the publicly available data used for short-read-based SV calling, we produced additional short-read data in-house for use in genome assembly for all but 2 (M82 and Fla.8924) MAS2.0 accessions. Short-read sequencing was performed according to a previous paper [52]. In brief, libraries were prepared with the Illumina TruSeq DNA PCR-free prep kit from 2 μg genomic DNA sheared to 550 bp insert size. DNA libraries were sequenced on an Illumina NextSeq500 platform at the Cold Spring Harbor Laboratory Genome Center.

Long-read DNA sequencing

Libraries for Oxford Nanopore genome sequencing were constructed using high-quality HMW DNA. DNA was sheared to ~20 kb using Covaris g-tubes or ~75 kb using Megarupter (Diagenode) and purified with a 1 × AMPure XP bead cleanup. Next, DNA size selection was performed using the Short Read Eliminator kit (Circulomics). Library preparation was performed with 1.5 μg of size-selected HMW DNA, using the Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies) following manufacturer's guidelines. Libraries were loaded on MinION or PromethION flow cells and sequenced according to standard protocols. Runs were basecalled with either Albacore v2.3 or with Guppy v2.1 through 3.2. Basecalling was performed using the PromethION r9.4.1 model, with recommended settings for the SQK-LSK109 kit and the FLO-PRO001 or FLO-PRO002 flowcells. At least 40G of data with mean read quality above or equal to Q7 were produced for each sample.

Long-read structural variant calling, filtering, and merging

For each of our 100 accessions selected for long-read sequencing, we aligned a maximum of 60X coverage to the SL4.0 reference genome. The SL4.0 reference genome is a recently published preprint that improves to the previous (SL3.0) tomato reference genome [11]. This PacBio long-reads assembled genome is the most complete and accurate representation of the Heinz 1706 reference genome to date. ITAG4.0, the reference gene models used in this study, are the accompanying reference gene annotation set. To call SVs relative to this reference, we aligned reads with NGMLR (v0.2.7, -x ont–bam-fix) and called SVs with Sniffles (v1.0.11)(–cluster–min_homo_af 0.7 -n 1000) [9]. As is convention, SV labels (insertions, deletions, duplications, inversions, and translocations) are defined with respect to this single reference genome and do not necessarily define the underlying mutations causing the genetic variation. We further note that long insertions are somewhat underrepresented since Sniffles' power to call insertions is bounded by read-length. For read sets exceeding 60X coverage, the longest set of reads achieving 60X was used. We then filtered SVs to remove potentially spurious calls. First, we identified regions of the reference genome prone to producing false SV calls and removed any SVs intersecting these regions (a total of 2,961,888 bp of the SL4.0 reference genome). To define these regions, we simulated ONT reads using SURVIVOR from the SL4.0 reference genome and called SVs with Sniffles. We performed this simulation a total of 9 times and merged the 9 VCF files with SURVIVOR (minimum distance of 1kbp, types must match, and a minimum length of 50bp). We then masked any region of the reference implicated in any SV from this simulation, including 2.5 kbp of flanking sequence. Next, we removed any SVs mapping to the ambiguous reference "chromosome 0" (SL4.0ch00). We also removed SVs larger than 100 kbp or SVs with a "0/0" genotype.

Using this same process described above, we also aligned Heinz 1706 PacBio reads to the SL4.0 reference genome to assess the propensity of the reference genome to produce false positives [11]. We called only 75 from these alignments, suggesting that spurious false positives due to reference bias in our panSV-genome are rare.

For some accessions, duplications were filtered by observing short-read coverage across putative duplications. To do this, we wrote a custom tool similar to CNVnator's genotyping functionality [68]. First, for each accession, we calculated short-read coverage in non-overlapping 200bp windows of the reference genome using bedtools [69]. The same reads and alignments as described in "Short-Read Structural Variant Calling and Sample Selection" were used here. Coverage was then corrected for GC bias using a custom version of the algorithm outlined in Yoon et al. [70]. The global mean coverage was calculated by first removing outliers (using the 1.5 x IQR rule) then fitting a Gaussian distribution to the coverages using SciPy (stats.norm.fit) [71]. Finally, to verify a duplication, we required that the coverage roughly spanning the duplication boundaries must be greater than 1.75X the global mean coverage. Only duplications at least 1 kbp in size were considered. To calculate the coverage of the duplicated region, adjacent 200 bp windows were merged via averaging to obtain 1 window close to the true duplication size. The coverage for this window, aligned to the original duplication coordinates (rounded to the nearest 200bp interval) was then compared to the global mean coverage. The above duplication filtering was only performed on samples for which we had short-read data available. The source code for duplication filtering can be found on GitHub (https://github.com/malonge/DupCheck).

By default, Sniffles provides supporting reads for each insertion call but reports the insertion sequence from a single noisy read. To associate each insertion with an accurate sequence, we used Iris (v1.0.1)(https://github.com/mkirsche/Iris). Iris extracts the reads supporting the insertion sequencing using samtools, computes their consensus using Racon [72], and then replaces the original insertion sequence with the polished consensus. Finally, we used Jasmine to merge SVs across all accessions (v1.0.1, min_support = 1 max_dist = 500 k_jaccard = 8 min_seq_id = 0.25 spec_len = 30)(see "Merging SVs with Jasmine" below). We used the default distance metric for merging, which is Euclidean distance. Briefly, 2-dimensional coordinates for each SV are given by (SV start position, SV length). SVs may be candidates for merging if their Euclidean distance between these 2D points is ≤ 500. The primary SV set was merged across all 100 accessions, though we also produced group-specific merged call sets for SLL, SLC, and SP using the same parameters.

Merging SVs with Jasmine

We developed a new SV merging tool called Jasmine, which is available open-source on GitHub (https://github.com/mkirsche/Jasmine). Jasmine constructs a graph G in which nodes represent SVs from individual samples. Edges connect pairs of SVs that may be merged based on criteria such as the distance between their breakpoints, and in the case of insertions, their sequence similarity. Next, the variants are partitioned based on the reference sequence, SV type, and strand. To compute the best possible set of SV merges for a given group, Jasmine computes a forest on the graph which has a few key properties: 1) The edges in the forest are a subset of the edges in G, 2) No tree in the forest contains multiple nodes representing SVs from the same sample, 3) There are no unused edges in G which can be added to the forest while maintaining the previous properties, and 4) The sum of the breakpoint distances of edges in the forest is minimized. To do this, Jasmine uses a variant of Kruskal's algorithm for computing minimum spanning trees. By considering the edges in non-decreasing order of edge weight, Jasmine greedily adds edges to the forest if they will not violate any of the required properties. To avoid storing this potentially very large network in memory, the network is computed dynamically by finding low-weight edges for each node with a KD-tree. Initially, a small constant number of edges incident to each node is stored, and as these are processed in increasing order of edge weight, new edges to process are added to the set by finding the next nearest neighbors for each node. As a result of this optimization, Jasmine is efficient in terms of both memory and runtime and can merge the entire set of over 1.7 million tomato SV calls in less than ten minutes on a single thread of a laptop.

We tested the efficacy of Jasmine on a simulated dataset. In this experiment, we use our merged tomato panSV-genome as our "ground truth." This provides us with a realistic distribution of allele frequencies, SV types, and SV genomic positions. From this merged SV set, we then derived 100 individual SV sets, essentially reversing the merging process. When assigning variants to their original individual set, we added noise to the SV genomic position. The noise was modeled with a uniform distribution centered at 50 bp for both the start positions and lengths. In addition, the sequences of insertions were changed to model 10% sequencing error. Then, we reran Jasmine (using the same parameters as those used for our panSV-genome) on these noisy individual call sets and compared the results to the original merging. 98.98% of the 19.4 million variant pairs which were merged initially were also merged in the simulated results, while only 0.93% of the merged pairs from the simulation were unmerged in the original dataset. We also found that of the 238k variant calls which originally consisted of merged variants from multiple samples, 97.78% of them contained exactly the same sets of variants after the simulation. The added noise to the variant boundaries caused some previously merged variants to exceed the distance threshold. Also, some originally close variants in the same sample traded places during the merging process. This analysis shows that the method is highly robust to variation in the positions and lengths of structural variants across samples.

MAS2.0 genome assembly

We established de novo genome assemblies and associated gene and repeat annotations for a subset of the 100 accessions sequenced for SV analysis. This included the PAS014479 (SP), BGV006775 (SP), BGV006865(SLC), BGV007989 (SLC), BGV007931 (SLC), PI303721 (SLL), PI169588 (SLL), EA00990 (SLL), LYC1410 (SLL), Floradade (SLL), EA00371 (SLL), M82 (SLL), Fla.8924 (SLL), and Brandywine (SLL) accessions. Collectively, we refer to these assemblies and annotations as "MAS2.0," and they are freely available to download at the Sol Genomics Network (https://solgenomics.net/projects/tomato100).

A hybrid assembly was performed for each accession using the MaSuRCA assembler (v3.3.3 or v3.3.4) [16]. Sequencing data used for assembly are described in "Short-read DNA sequencing" and "Long-read DNA sequencing". M82 and Fla.8924 were not sequenced in-house for this study, but rather come from a previous publication [17]. As is recommended by the MaSuRCA documentation, no preprocessing was done on any of the sequencing data. For the ONT reads, we used the longest 35X coverage of reads with an average Phred quality score of at least 7. Library insert sizes for all Illumina data were set to 500 ± 50. All assemblies employed the Flye unitigger during the final stage of MaSuRCA, except M82, which used default unitigging settings. All other MaSuRCA parameters were set to default values.

Each set of initial draft contigs underwent two rounds of short-read polishing with POLCA (MaSuRCA v3.3.4) [73]. As input for each of the two rounds of polishing, we used seqtk to randomly sample ⅔ of the Illumina data used during assembly (https://github.com/lh3/seqtk). After polishing, we screened each set of contigs for bacterial contamination by aligning them to the tomato SL4.0 reference and a bacterial reference genome. Every RefSeq bacterial genome downloaded on October 1st, 2019, comprised our bacterial reference. Contigs were mapped to both references with Minimap2 (-k19 -w19) [74]. Any contig covered more by bacterial alignments than by tomato alignments were deemed contaminated and removed from the assembly. Only the BGV006865 and PI303721 accessions contained contaminated contigs. Finally, polished and screened contigs were scaffolded according to the SL4.0 reference genome using RaGOO (v1.1) (-T corr) [17]. The MaSuRCA mega-reads associated with the initial assemblies were used for misassembly correction. "Chromosome 0" of the SL4.0 was not considered during RaGOO scaffolding (-e). We generated dotplots for each assembly by aligning the final pseudomolecules to the SL4.0 reference genome using nucmer (-l 100 -c 500) and finally plotting with mummerplot

(–fat–layout) [75]. Finally, we used BUSCO to assess genome completeness (v3.0.2, -l solanaceae_odb10 -m genome -c 10 -sp tomato) [18].

To observe SV concordance between our panSV-genome and the MAS2.0 assemblies, we called SVs from the assemblies using two techniques. First, we aligned the MAS2.0 assemblies to the SL4.0 reference genome using Nucmer (v3.1, -maxmatch -l 100 -c 500) and called SVs with Assemblytics (unique_length_required = 500 min_size = 15, max_size = 100500) [76]. Additionally, we simulated 60X coverage of perfect 25 kbp reads from the MAS2.0 assemblies and called SVs with NGMLR (v0.2.7, -x ont –bam-fix) and Sniffles (v1.0.11, -s 2 -l 15 –cluster –min_homo_af 0.7 -n 1000) with respect to the SL4.0 reference genome. Combining the Assemblytics and Sniffles MAS2.0 SV sets, we observed the pairwise SV concordance with the corresponding 14 accessions in our panSV-genome. The % SV overlap for each of the 14 accessions is as follows: BGV006775: 95.5571, BGV006865: 94.5002, BGV007931: 95.8251, BGV007989: 91.8735, Brandywine: 91.1921, EA00371: 87.8088, EA00990: 86.9073, Fla.8924: 89.4226, Floradade: 84.7832, LYC1410: 93.3863, M82: 90.3600, PAS014479: 92.8686, PI169588: 88.5430, PI303721: 70.9839.

We note that we do not expect perfect overlap between the read-mapping and assembly-based SV calls, since both have unique fallibilities and biases. For example, larger variants found with one approach may be broken into multiple smaller variants found by the other approach. Or, the exact position of variants may shift within genomic repetitive elements. Also, SVs in regions of the genome that fail to assemble may still be detected by aligning reads to a reference genome. Furthermore, expected variability in nanopore sequencing, along with other factors, likely contributes to the between accession variation that we observe. Broadly, an average overlap of 90% is a positive indication of SV accuracy and data quality.

MAS2.0 gene annotation

We used a “lift-over” approach to annotating the MAS2.0 assemblies with gene models. Along with the tomato reference ITAG4.0 gene models, our reference gene model set included previously published “pan-genome” genes which may be missing from ITAG4.0 but present in our assemblies [20]. Gene models were lifted-over onto each of the 14 MAS2.0 assemblies with geneLift (v1.1, -c 90 -i 95) (https://github.com/srividya22/geneLift). Briefly, geneLift maps reference cDNA sequences to target assemblies using GMAP and Minimap2 and retains alignments with at least 90% coverage and 95% identity [77]. The remaining non-overlapping GMAP alignments constitute the initial gene models, which are then supplemented by Minimap2 alignments to unannotated regions providing additional non-redundant gene models. Gene IDs reported by geneLift match the reference gene IDs and any gene duplications reported have an added suffix "-c" followed by the respective copy number of the gene to make them unique. Annotated "pan-genome" genes can be distinguished by a "TomatoPan" gene ID prefix.

MAS 2.0 and SV repeat annotation

We used REPET to annotate MAS2.0 assemblies and panSV-genome insertion/deletion sequences with repeats [19]. From each MAS2.0 genome assembly, we built a sub-genome by selecting the longest contigs up to a cumulative size ranging 360-380 Mbp. This allowed us to sample a large portion of the genome while keeping the downstream computation tractable [78]. Each sub-genome was used to generate libraries of consensus sequences that are representative of repeats present therein using the TEdenovo pipeline from the REPET package v2.4 (parameters were set to consider repeats with at least 5 copies). The libraries produced were filtered to keep only those sequences that are found at least once as a full-length copy in the respective sub-genomes. Each resulting library of consensus sequences was then used as query for annotation of respective whole genomes using the TEannot pipeline from the REPET package v2.4. The library of consensus sequences was classified using PASTEC followed by semi-manual curation [79].

For the annotation of insertions and deletions, the filtered consensus libraries obtained from ten of the 14 MAS2.0 assemblies (the first 10 to be completed) were pooled and appended to those from SL4.0 which were generated previously using the protocol described above. This combined library was then used as a query for whole genome annotation by TEannot using default settings.

PI129033 NSGT local assembly

None of our 14 MAS2.0 assemblies contained the NSGT deletion allele described in "New Reference Genomes Resolve Multiple Haplotypes for the "Smoky" Volatile Locus." Therefore, we performed a local assembly of the NSGT locus in PI129033, a sample known to carry this deletion allele. Using the same long-read alignments as described in "Long-read Structural Variant Calling, Filtering, and Merging," we extracted PI129033 reads that aligned to the NSGT locus (SL4.0ch09:65168601-65653800) using samtools view. These reads were then error corrected with Canu (corOutCoverage = 999, genomeSize = 475k) and assembled with Flye (–nano-corr,–genome-size 475k) [80,81]. Flye produced a single contig 534,847 bp in length representing the NSGT locus in PI129033. We next sought to polish this contig with short reads to produce an accurate representation of the locus. To do this, we first placed the contig into the SL4.0 reference genome in order to provide a suitable reference genome for short-read mapping. This avoids the potential poor quality of mapping when aligning WGS reads to a small segment of the genome. To create this pseudo-reference genome, we first started with the SL4.0 genome and replaced the NSGT locus (SL4.0ch09:65168601-65653800) with our local assembly. We also added 100bp gaps to the flanks of the inserted contig so that we could identify and retrieve it after polishing. We aligned short reads to this pseudo-reference using bwa and performed two rounds of short-read polishing with Racon (-u). Finally, we removed the local assembly from the pseudo-reference using samtools faidx and aligned it with Minimap2 (-ax asm5) to the SL4.0 reference genome to precisely define the deletion coordinates.

SV hotspot and introgression analysis

For each accession, we counted the number of SVs in non-overlapping 1Mpb windows of the reference genome. Bins with a relatively large number of SVs are informally referred to as "SV hotspots". SV frequency, shown in heatmap and circos form, is depicted in **Figure 5.2A** (http://omgenomics.com/circa/). Our observation of "hotspots" usually results from visual interpretation of these plots. SV hotspot heatmap rows are ordered within each phylogenetic group (GAL, CHE, SP, SLC, SLL) by the R "heatmap.2" default row ordering. These ordered groups were then concatenated to produce the final heatmap.

Since we hypothesized that introgression from wild donors could account for many of the observed SLL hotspots, we developed a technique to compare accessions to look for genomic regions of SV similarity. The custom Python code used for this task can be found in a GitHub repository (https://github.com/malonge/CallIntrogressions). The script "get_distances.py" compares SLL accessions to one or many accessions from any other "comparison" group (SP, SLC, GAL, or CHE). The algorithm considers successive 1Mpb windows of the reference genome. For each SLL accession, its set of SVs in a given window is compared to the set of SVs in all accessions in the comparison group in the same window. To compare two sets of SVs, we calculate the Jaccard similarity, requiring at least 5 SVs in both SV sets. The script then outputs, for each 1 Mpb window and for each SLL accession, the maximum Jaccard Similarity with any other comparison accession. If all comparisons for a given window had fewer than 5 SVs in either SV set, an "NA" value is reported.

We calculated similarity for all 45 SLL accessions at the same time by comparing each accession to each non-SLL accession. Comparisons against GAL and CHE did not yield any candidate introgressions from these groups, so we did not display those results. In **Figures 5.2D and 5.2E**, we also show an instance where we compare SLL accessions against a single SP comparison accession (LA1589).

SV genomic feature annotation

Throughout the manuscript, we describe various relationships between SVs and other genomic features such as genes. Generally, we annotated our panSV-genome with genomic features using vcfanno [82]. We define an "annotation" as the association of a particular SV with particular feature IDs (such as a gene ID) based on some relationship. vcfanno annotates SVs by finding their intersection (overlap) with genomic feature intervals. Accordingly, some of the annotations reported in the manuscript can be directly interpreted from vcfanno, such as "Insertions in exons," or "Deletions overlapping 5 kbp upstream," since these can be directly interpreted from feature intersection. Other annotations, such as SV containment of genes, required some combination of intersection calculations. For example, to detect genes contained by SVs, we first checked if the gene start and end positions intersected a given SV. If that SV intersected both the start and end of a gene, it contains that gene.

We ultimately produced many SV/feature annotation classes which are explained in more detail here. In any applicable annotation, "upstream" or "downstream" refers to the 5′ or 3′ flanking regions of genes, respectively. In supplemental material, these "upstream" and "downstream" regions may also be referred to as "5′ UTR" and "3′ UTR" respectively. "Insertions in exons," "Insertions in introns," "Insertions in 5 kbp downstream," "Insertions in 5 kbp upstream," "Deletions overlapping 5 kbp upstream," and "Deletions overlapping 5 kbp downstream" are self-explanatory. "Duplications" are duplications that contain entire genes. "Deletions of exons" are deletions that delete at least one entire CDS exon of a gene, but do not delete the entire gene. Finally, "Deletions of CDS start" are deletions that contain 50 bp upstream and downstream of a CDS start site.

The impact of SVs on gene expression

Data analysis was performed in R using custom scripts. In each tissue (apex, cotyledon, and root), gene expression was averaged over the biological replicates in each accession (23 accessions with 3 replicates each in apex and root, and 22 accessions with 4 replicates each in cotyledon), and the genes with average expression count of at least 1 across the accessions were retained for further analysis. We averaged read counts across replicates to effectively treat the replicate expression as estimating a fixed effect. These gene expression averages within each accession/tissue were ranked and standardized so that the values were constrained between 0 and 1. While most of our analyses operate on these rank data, in order to provide estimates of fold change, we used the average expression profiles across replicates directly. These values were normalized by division of total read count of each accession and then fold changes were calculated across these normalized values between accessions with and without the SV.

Are SV-associated genes differentially expressed?

We first defined a list of SV-gene pairs based on SV annotations (see SV Genomic Feature Annotation). We filtered this list to only include SV-gene pairs that had the SV present in at least 5 and absent in at least 5 of the accessions for which we had RNA-seq data. For each of the SV-gene pairs, the accessions were split into two groups: with and without the SV. The extent of differential expression of the associated gene was calculated using a two-sided Mann-Whitney U test across the accession split. The Mann Whitney U test is a rank-based test that is very robust to underlying distributions in the expression values. The p values among a specific annotation and tissue type were adjusted by applying Benjamini-Hochberg procedure [83]. The adjusted p values for each annotation and tissue type were aggregated using two methods: Fisher's method and a harmonic mean estimate [84,85].

At least half of the SV-associated genes in each SV type were common to all three tissues, exhibiting different levels of differential expression across the same accession split. To determine an average differential expression across the tissues, we used Fisher's method to aggregate p values across the three tissues for each SV-associated gene and subsequently applied Benjamini-Hochberg method to limit the number of false positives.

Can we predict SV-associated genes from their differential expression?

For this analysis, we formulated a prediction task: Using the SV annotations as a "ground truth" labeled feature set (the gene associated with the SV is positively labeled and all other genes are negatively labeled), we measured how well we could predict the presence of an associated SV (positive label) given differential expression. We used AUROC (Area under the ROC) scores as a measure of the performance of this task, which is calculated as follows: For each SV of a given annotation type, the p values corresponding to the differential expression across the accession split (with or without the SV) was calculated for all genes in a given tissue via a two-sided Mann-Whitney U test, and the list of p values was ranked (highest rank corresponds to the most significant p-value). For each SV, AUROC scores were analytically calculated by determining the positively labeled gene's position in the ranked list of all gene p values (high AUROC score corresponds to a near-perfect identification of the SV-associated gene). In other words, genes are predicted to be associated with a variant if they exhibit excess differential expression when comparing accessions with versus without the SV. Conceptually, this can also be described as our classifier choosing a series of cutoff positions in this list, generating a ROC curve (and associated AUROC) by calculating the true and false positive rate associated with each cutoff. Since all genes are affected by the underlying phylogenetic structure in the data, successful prediction of the true SV-associated gene in the list of all genes only occurs when predictions are robust to confounding population structure.

We have thus far described our prediction task when considering a single SV-gene pair. To assess the broad impact of SVs on expression, we combined all SV-gene pairs in a given annotation and tissue type. This is conceptually the same as for single SV-gene pairs, except the gene labels are combined into an aggregated labeled set where there is one positive gene label for each SV-gene pair. The resulting ROC curve and associated AUROC effectively measures the average performance of the classifier over all SV-gene pairs. A high AUROC would indicate SVs globally have a significant impact on associated gene expression.

Our aggregated classifier's performance can be measured by computing an overall p-value as follows. For a given variant and tissue type, the ranks of p values of all SV-associated genes are removed from the list of sequential ranks of all expressed genes in a given tissue (for example, the ranks of 17 genes associated with duplications in apex tissue are removed from the sequence of ranks 1:20029 of the 20029 expressed apex genes). A One-tailed Mann-Whitney U test was performed to evaluate if the median of the ranks of SV-gene pair p values was lower than the median of ranks of p values of all other expressed genes. The resulting p-value is depicted by the size of the circle in **Figure 5.3E**. It is important to note that the overall p values (circle size) are influenced by the number of SV-associated genes used in classification in each case, as well as the fold change in expression. For instance, duplications in apex have a larger p-value (with 17 variants used in classification) than insertions in 5 kbp downstream (with 1129 variants used in classification).

Plant phenotyping

To quantify floral organ size, lengths of sepals and anther cones of closed yellow flower buds just before opening were measured. Inflorescence complexity was measured by counting the number of branching events per inflorescence. Flowering time was quantified by counting the number of leaves before the first inflorescence.

NSGT haplotype analyses

Thirteen of the fourteen MAS2.0 genome assemblies filled the gaps at the chromosome 9 “guaiacol” GWAS locus. To annotate this region, the full-length protein sequence of NSGT1 was used for BLAST search against the Heinz SL4.0 reference genome and the 14 MAS2.0 assemblies. We used the protein sequence as the query for BLAST to achieve more sensitive and more contiguous alignments while still allowing for the discrimination of NSGT alleles. Based on the BLAST results and sequence differences, four coding sequence variants including NSGT1, NST2, nsgt1, and nsgt2 are annotated in these genomes [38]. We observed several accessions missing sequencing coverage at this locus, suggesting a deletion. We selected one such accession (PI129033) for a local assembly of the deletion haplotype (see “PI129033 NSGT Local Assembly”). The local assembly revealed the large deletion haplotype V.

Short-read based genotyping

NSGT locus coding sequence variants genotyping

From short-read alignments to the SL4.0 reference genome, we extracted reads overlapping with NGST locus (SL4.0ch09:65390765-65417476) using samtools view. In addition, we included previously unmapped reads. These mapped and unmapped read sets were converted back to fastq files using samtools bam2fq. Subsequently, the reads were mapped to the unique portion of nsgt1 (117bp, GTTAGGTTTTAGGGTTTCAATTATGCTTGGAAATTTGGAagaagccatttgaaaggcttgaataaggtttaggtaccATCTTTAACAACTACCTCCAAAATTATAAACCTTTTTCTT), nsgt2 (86bp, CCAATACTTGAATGgttcaaaattagactttgtactttcaagaaaaccttgtGGAACCATTTCTTCAATTGTTTTGTTCACCCCTT), NSGT1 (100bp, ATATAATAGCTTCAACAACTTTTTAACCCCTTcatcaatagctttcaattttatcttctcactcaattgCATTGCCTTCAAATGAATTTGTTTCCTAGGC) and NSGT2(123bp, CAAAGGCTTTCTCATCGCGTGGTTTTATTGGTTTCATATCTAATTTCTTGatctcatagtcatgaagaaaaggAAAAGATGTAAGGCTTGAACTCCCATAAAGAAATTGGTGGTAAAGGTAGG) simultaneously using bwa mem (-M). After mapping, reads with edit distance (NM tag) smaller than 15 and a minimum mapping quality of 20 were extracted.

We used samtools depth to compute the coverage of the filtered reads across only the core of the unique regions (lower case sequences above) for nsgt1, nsgt2, NSGT1, and NSGT2. If more than 4 core bp had 0 coverage, we discarded the total mapped read counts for the sequence. If there was read count support for any of the nsgt1, nsgt2, NSGT1, or NSGT2 haplotypes, we report them as “presence.” Since the “unique” sequence of NSGT1 is also present in nsgt1, if both nsgt1 and NSGT1 were genotyped as “presence,” we only labeled nsgt1 as “presence.” This is based on the observation that no sequencing resolved haplotypes have both nsgt1 and NSGT1 together. This genotyping was consistent with the observed haplotypes in our MAS2.0 assemblies.

NSGT locus deletion variant genotyping

From the short-read alignments to SL4.0, we counted the reads with a mapping quality of at least 20 in the middle region of the haplotype V deletion: SL4.0ch09:65401889-65404136. Accessions with less than 5 mapped reads were genotyped as “deletion.” The pipeline was benchmarked against PCR genotyped samples including 138 accessions with no deletion and 17 accessions with deletions. Results from our pipeline were 100% consistent with PCR genotyping results.

sb1 duplication genotyping

From the short-read alignments to SL4.0, we extracted the reads mapped to a broad region that contained the *sb1* duplication locus: SL4.0ch01:77727550-77765153. For each sample, we also extracted the unmapped reads. Mapped and unmapped read sets were converted to fastq files using samtools. Subsequently, we aligned the extracted reads to a portion of the *sb1* locus (SL4.0ch01:77737550-77745153), which avoided high copy number TEs and represented a unique sequence of this locus. This was done with bwa mem (-M). We counted the number of reads mapped to this locus using samtools idxstats. The raw counts were normalized based on the total number of reads mapped for each sample. We manually checked the read alignments to SL4.0 and verified 22 single-copy accessions and eight duplication accessions. Accessions with normalized coverage lower than mean (verified single-copy accessions) – 1 standard deviation were genotyped as “single-copy” and accessions with normalized coverage greater than mean (verified duplication accessions) + 1 standard deviation were genotyped as “duplication.”

Tissue collection, RNA extraction and quantification

For 3′ RNA-sequencing (3′ RNA-seq), seeds were treated with 50% bleach for 20 minutes to homogenize germination and were germinated in Petri dishes with moistened filter paper in the dark at 28 °C. Whole root tissues were collected 3 days after germination with a mixture of several seedlings as one biological replicate and three such replicates for each of a total of 23 accessions. For cotyledon tissues, seedlings after germination at similar stages were transplanted to soil in 96-cell flats and grown in the greenhouse. Cotyledons of seedlings were collected when two true leaves start to visibly emerge (10~11 days after sowing). Four biological replicates each with several seedlings combined for each of a total of 22 accessions were collected. For apex tissue, seedlings after germination at similar stages were transplanted to soil in 96-cell flats and grown in the greenhouse. For apex tissue collection, seeds were germinated, and seedlings were transplanted as above. Vegetative apical meristem together with the two youngest/smallest leaf primordia were collected 4 days after transplanting [86]. Eight to twelve apices were combined as one biological replicate and three replicates were collected for each of a total of 23 accessions. Total RNA was extracted using the RNeasy Plant Mini Kit (QIAGEN) and treated with the RNase Free DNase Set (QIAGEN) according to the manufacturer's instructions. Total RNA samples were sent to the Genomic Diversity Facility at Cornell University for high-throughput 3′ RNA (single-end, read length = 75bp) as described [87].

For quantitative RT-PCR, seeds were germinated on moistened filter paper at 28°C in dark. After germination, seedlings at similar stages were transferred to soil in 96-cell plastic flats and grown in the greenhouse. Shoot apices were collected at the transition and floral meristem stage of meristem maturation [86], and immediately flash-frozen in liquid nitrogen. Total RNA was extracted as described above. 100 ng to 1 μg of total RNA was used for cDNA synthesis using the SuperScript III First-Strand Synthesis System (Invitrogen). qPCR was performed with gene-specific primers using the iQ SYBR Green SuperMix (Bio-Rad) reaction system on the CFX96 Real-Time system (Bio-Rad).

NSGT1/2 expression analysis

Published RNA-seq data of tomato fruit pericarp tissue from 405 accessions were downloaded from SRA PRJNA396272. Reads were trimmed by quality using Trimmomatic (ILLUMINACLIP:TruSeq3-PE-2.fa:2:40:15:1:FALSE LEADING:30 TRAILING:30 MINLEN:100) and aligned to the cDNA annotation of reference genome sequence of tomato (SL4.0) using kallisto quant [88]. The output of kallisto generates normalized transcripts per million reads (TPM) which was used for quantifying NSGT1/2 expression. Because only one copy of NSGT1/2 is annotated in the SL4.0 and sequences of NSGT1 and NSGT2 are highly similar, we used the TPM of the annotated copy of NSGT (*Solyc09g089585*) to represent the expression level of both NSGT1 and NSGT2.

Metabolite profiling

Published fruit guaiacol contents were obtained from Tieman et al. (2017). To minimize environmental effects, only data from one field season (2015) were used. Fruit guaiacol and methylsalicylate contents in our new GWAS panel were quantified as previously described [36]. Briefly, at least six fruits (two fruits for each replicate) of red ripe stage were collected from each variety. Volatile compound identification was determined by gas chromatography-mass spectrometry and co-elution with known standards (Sigma-Aldrich, St. Louis MO).

3′ RNA-seq data processing and gene expression analysis for individual duplication locus

3′ RNA-seq reads were trimmed by quality using Trimmomatic (v0.36, ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 LEADING:30 TRAILING:30 MINLEN:30 HEADCROP:12) and mapped to SL4.0 reference genome using STAR with default parameters [89]. Bam files generated by STAR were sorted by read name and gene expression was quantified as uniquely mapped reads to annotated gene features in the ITAG4.0 reference annotation using HTSeq-count (–format = bam–order = name–stranded = no–type = exon–idattr = Parent) [90]. Gene counts were processed in R for visualization. First, we filtered expressed genes by only keeping genes with the sum of counts across all samples greater than the sum of replicates. Then the count table was imported into R package “DESeq2” [91] and normalized counts were used for making boxplots.

Generation of F_2 populations segregating for the fw3.2 duplication or promoter SNP

The *fw3.2* duplication and the derived allele of the promoter SNP are highly, but not completely associated. From our collection of accessions, we carefully selected four pairs of accessions carrying either single or double copies of *fw3.2* but fixed at the promoter SNP (M9) of *KLUH* and all other known fruit weight QTL genes. Four bi-parental F_2 populations were developed from each pair of accessions so that the duplication of *fw3.2* would segregate. We genotyped the F_2 plants by *fw3.2* duplication markers and markers flanking the entire duplicated region. Similarly, six bi-parental F_2 populations that segregated for the promoter SNP but fixed as the single-copy of *fw3.2* and other known fruit weight QTL genes were developed. We genotyped F_2 plants using M9 markers. In each population, ten homozygous F_2 plants carrying each of the contrasting genotypes were grown in the field. At harvest, we selected 15 to 20 large fruits after mature green stage and recorded their average weight to represent the potential of largest fruit from a single plant. Poor fruit setting was observed in population 19S313 so only about 10 representative fruits were used for each plant. In extreme cases, the fruit weight of three plants were represented by less than 5 fruits.

CRISPR-Cas9 mutagenesis, plant transformation, and selection of mutant alleles

CRISPR-Cas9 mutagenesis and generation of transgenic tomato was performed following our standard protocol [92]. Briefly, guide RNAs (gRNAs) were designed using the CRISPRdirect tool (https://crispr.dbcls.jp/) [93]. Binary vectors for gRNAs and Cas9 were assembled using the Golden Gate cloning system as described [4,51,94]. Final binary vectors were transformed into the tomato cultivar M82 by Agrobacterium tumefaciens-mediated transformation through tissue culture [95]. Transplanting of first generation transgenic (T0) plants and genotyping of CRISPR-generated mutations were performed as Soyk et al. [51]. Briefly, CRISPR-targeted region was PCR amplified, and wild-type (WT) size products were sequenced for T0 plants and those with mutations were selfed or crossed to WT M82 plants for further characterization of mutant alleles.

Generation of hybrid plants for different KLUH dosages

To test the dosage-dependent effect of *KLUH* in an isogenic background with uniform "cherry" fruit type, the fertile T0 plant with CRISPR-Cas9 targeting *SlKLUH* (*slkluh*^{CR} T0-1) was crossed with the SP accession LA1589. About half of F_1 plants carried the Cas9 transgene (1:1 segregation of transgene). Analyses were focused on F_1 plants that did not inherit the Cas9 transgene, because they are a fixed, uniform genotype. In contrast, plants with the Cas9 transgene would be genetically intractable for dosage analyses, because of the random chimerism that occurs within individual plants carrying the Cas9 transgene. From eight individual F_1 plants without the Cas9 transgene (genotypic group B), *KLUH* gene PCR products were cloned and eight individual clones were sequenced. All eight plants were confirmed to have only mutant *slkluh* alleles and a WT *SpKLUH* allele. Sepal length, flower length and fruit weight were quantified from these plants. Most of the F_1 plants with the Cas9 transgene showed slightly smaller floral organs, and several of these plants had extremely small floral organs and no fruit set. From four individual F_1 plants with the Cas9 transgene that showed tiny floral organs (genotypic group C), sepal length and flower size were quantified. To determine whether this effect was due to trans-targeting of *SpKLUH*, two plants with extremely small floral organs were randomly selected and sequenced for multiple PCR-cloned *KLUH* alleles. Consistently, sequencing of the two plants showed only mutant alleles for *SlKLUH* and *SpKLUH*, consistent with the CRISPR-Cas9 trans-targeting the *SpKLUH* gene copy. WT M82 was crossed with LA1589 and the F_1 plants were used as controls.

STM3 Phylogenetic analyses and sequence analyses

Sequences of homologous proteins of *STM3* and *TM3* were obtained from tomato and Arabidopsis genome and aligned using the ClustalW2.1 program in Geneious 11.1.5. Phylogenetic tree was constructed using "Geneious Tree Builder" with Jukes-Cantor genetic distance model and Neighbor-Joining method with 1,000 bootstrap replicates. *STM3* and *TM3* fell in the same clade with Arabidopsis flowering time regulator SOC1 [96].

Delta SNP index plot and genome coverage plot

Mapping of genomic position of *sb1* was reported in Soyk et al. [52]. Briefly, F_2 segregation population was generated from crosses between a branched M82 *j2*^{TE} *ej2*^{W} double mutant with an unbranched *j2*^{TE} *ej2*^{W} double mutant (Fla.8924). A group of excessively branched inflorescences (6–36 branches) and a group of clearly suppressed plants (1–4 branches) were selected. An equal amount of tissue from each plant (~0.2 g) was pooled for DNA extraction for the two groups using standard protocols. Libraries were prepared with the Illumina TruSeq DNA PCR-free prep kit from 2 µg genomic DNA sheared to 550 bp insert size and sequenced on an Illumina NextSeq platform at the CSHL Genome Center. After aligning reads to reference genome (SL3.0), SNPs were called with samtools/bcftools [62,97] using read alignments for the two genomic DNA sequencing pools in addition to the M82 [98] and Fla.8924 [99] parents. Called SNPs were then filtered for bi-allelic high-quality SNPs at least 100 bp from a called indel using bcftools [97]. Read depth for each allele at segregating bi-allelic SNPs in 100-kb sliding windows (by 10 kb) was summed for the various sequencing pools and allele frequencies were calculated. Finally, the difference in allele frequency (SNP index) between the branched and unbranched pools was calculated and plotted across the 12 tomato chromosomes. One of the two regions that exceeded a genome-wide 95% cut-off in SNP index was located on chromosomes 1 and was named *sb1*. The candidate interval based on SL3.0 is SL3.0ch01:80006250-86570024.

To show the genome coverages at the *sb1* locus in M82, M82 *j2*^{TE} *ej2*^{W}, Fla.8924, and *S. pimpinellifolium*, we calculated the coverage from Illumina data using bedtools multicov only counting properly paired reads in 10-kb windows across chromosome 1. Depths in the four genotypes were normalized by dividing by the average depth using R.

Generation of F_2 populations segregating for sb1 CRISPR alleles, j2TE and ej2W

Homozygous *sb1*$^{CR\text{-}1}$ and *sb1*$^{CR\text{-}del}$ plants were each crossed with M82 *j2*^{TE} *ej2*^{W}, respectively, to construct two F_2 populations segregating at those three loci. In the F_2 generation, plants were first genotyped for *j2*^{TE} and *ej2*^{W} mutations at seedling stage in flats. All double mutants were transplanted and further genotyped for CRISPR alleles and quantified for inflorescence complexity/branching.

Quantification and statistical analysis

"n" is defined in all relevant figure legends. All statistical tests were performed in R. Significance is only ever defined for the SV differential expression analysis (**Figure 5.3C**) and it is defined as a p-value less than 0.05. Two-sided Mann-Whitney U tests were used for analysis in **Figures 5.3C–5.3F**. The Mann-Whitney U test provides a robust estimate to compute the significance of the expression change that does not depend on any assumption of underlying distributions. The p values for these tests underwent FDR correction with the Benjamini-Hochberg procedure. Adjusted p values were aggregated using Fisher's method and a harmonic mean estimate. Detailed methods for these analyses can be found in "The Impact of SVs on Gene Expression." For expression analysis in **Figures 5.4E, 5.5C, and 5.6E**, numbers of accessions for each genotype are presented in the figures, and differences between groups were compared using two-tailed, two-sample t tests. Fruit guaiacol and methylsalicylate contents were compared between genotypes using two-tailed, two-sample t tests. For quantitative analysis in sepal length, flower length, fruit weight, and inflorescence complexity n = number of flowers and inflorescences quantified was used for two-tailed, two-sample t tests. The number of plants (n =) used for each genotype is also labeled in the figures. For the above analysis, all data points were plotted as single dots in the boxplots. For expression analysis with qRT-PCR, three biological replicates of pooled meristems were used for each genotype and two technical replicates were performed for each biological replicate. Mean values of normalized expression were compared using two-tailed, two sample t tests. For flowering time quantification, number of plants of each genotype is labeled in the figure. Means ± s.d. were shown and mean values between groups were compared by two-sample t tests.

5.6 ACKNOWLEDGEMENTS

We thank members of the Schatz and Lippman labs at JHU and CSHL for helpful discussions. We thank Aleksey Zimin at JHU for helpful discussions about genome assembly. We thank T. Mulligan, A. Krainer, S. Qiao, and K. Schlecht from CSHL for assistance with plant care. We thank S. Muller, R. Wappel, S. Mavruk-Eskipehlivan, and E. Ghiban from the CSHL Genome Center for sequencing support. We thank Sanwen Huang at the Chinese Academy of Agricultural Sciences and Dani Zamir at the Hebrew University for sharing sequencing data and germplasm. We thank Prashant Hosmani, Susan

Strickler, Naama Menda, and Lukas Mueller at the Boyce Thompson Institute for Plant Research for providing early access to the SL4.0 reference genome and hosting data on the Solanaceae Genomics Network. We thank Surya Saha and Lukas Mueller for early access to Heinz 1706 PacBio data. This work was supported by the Howard Hughes Medical Institute and grant support from the US National Institutes of Health for their support of the CSHL Cancer Center Next Generation Shared Resource (5P30CA045508-31) and (R01LM012736 and R01MH113005) to J.G. and (UM1 HG008898) to F.J.S., the USDA National Institute of Food and Agriculture (AFRI no. 2018-67013-27896; SCRI no. 2015-51181-24312) to S.H. and (AFRI no. 2016-67013-24452) to S.H. and Z.B.L., the BARD (BARD US-Israel Agricultural Research and Development Fund) (IS-5120-18C) to Y.E. and Z.B.L., the ISF (Israel Science Foundation) (1913/19) to Y.E., and the National Science Foundation Plant Genome Research Program (IOS 1855585) to D.M.T. and H.K.; (IOS 1564366) to A.L.C., D.M.T., and E.v.d.K.; (IOS-1350041) to M.C.S.; and (IOS-1732253) to J.V.E., E.v.d.K., M.C.S., and Z.B.L.

5.7 REFERENCES

1. Meyer RS, Purugganan MD. Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet. 2013;14:840–52.

2. Olsen KM, Wendel JF. A bountiful harvest: genomic insights into crop domestication phenotypes. Annu Rev Plant Biol. 2013;64:47–70.

3. Chen K, Wang Y, Zhang R, Zhang H, Gao C. CRISPR/Cas Genome Editing and Precision Plant Breeding in Agriculture. Annu Rev Plant Biol. 2019;70:667–97.

4. Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB. Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. Cell. 2017;171:470–80.e8.

5. Wallace JG, Rodgers-Melnick E, Buckler ES. On the Road to Breeding 4.0: Unraveling the Good, the Bad, and the Boring of Crop Quantitative Genomics. Annu Rev Genet. 2018;52:421–44.

6. Coster WD, De Coster W, Van Broeckhoven C. Newest Methods for Detecting Structural Variations. Trends Biotechnol. 2019;37:973–82.

7. Lye ZN, Purugganan MD. Copy Number Variation in Domestication. Trends Plant Sci. 2019;24:352–65.

8. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020;21:171–89.

9. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. nature.com; 2018;15:461–8.

10. Sedlazeck FJ, Lemmon Z, Soyk S, Salerno WJ, Lippman Z, Schatz MC. SVCollector: Optimized sample selection for validating and long-read resequencing of structural variants. bioRxiv. biorxiv.org; 2018;342386.

11. Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, et al. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. bioRxiv. biorxiv.org; 2019;767764.

12. 100 Tomato Genome Sequencing Consortium, Aflitos S, Schijlen E, de Jong H, de Ridder D, Smit S, et al. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant J. Wiley; 2014;80:136–48.

13. Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, et al. Genomic analyses provide insights into the history of tomato breeding. Nat Genet. 2014;46:1220–6.

14. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. Cell. 2019;176:663–75.e19.

15. Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, et al. Structural variants in 3000 rice genomes. Genome Res. genome.cshlp.org; 2019;29:870–80.

16. Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, et al. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. genome.cshlp.org; 2017;27:787–92.

17. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate

reference-guided scaffolding of draft genomes. Genome Biol. 2019;20:224.

18. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. academic.oup.com; 2015;31:3210–2.

19. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering transposable element diversification in de novo annotation approaches. PLoS One. 2011;6:e16526.

20. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet. 2019;51:1044–51.

21. Aflitos SA, Sanchez-Perez G, de Ridder D, Fransz P, Schranz ME, de Jong H, et al. Introgression browser: high-throughput whole-genome SNP visualization. Plant J. Wiley Online Library; 2015;82:174–82.

22. Consortium TTG, The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485:635–41.

23. Scott JW. University of Florida tomato breeding accomplishments and future directions. Soil Crop Sci Soc Fla Proc 58. tombreeding.ifas.ufl.edu; 1998;8–11.

24. Bohn GW, Tucker CM. IMMUNITY TO FUSARIUM WILT IN THE TOMATO. Science. cabdirect.org; 1939;89:603–4.

25. Scott JW, Jones JP. Monogenic resistance in tomato to Fusarium oxysporum f. sp. lycopersici race 3. Euphytica. Springer Science and Business Media LLC; 1989;40:49–53.

26. Strobel JW, Hayslip NC, Burgis DS, Everett PH. Walter: A determinate tomato resistant to races 1 and 2 of the fusarium wilt pathogen. Fla Agr Exp Sta Circ S [Internet]. agris.fao.org; 1969; Available from: https://agris.fao.org/agris-search/search.do?recordID=US201301188332

27. Walter JM, Kelbert DGA. Manalucie: A Tomato with Distinctive New Features. University of Florida, Agricultural Experiment Stations; 1953;

28. Foolad MR, Panthee DR. Marker-Assisted Selection in Tomato Breeding. CRC Crit Rev Plant Sci. Taylor & Francis; 2012;31:93–123.

29. Hutton SF, Scott JW, Vallad GE. Association of the Fusarium wilt race 3 resistance gene, I-3, on chromosome 7 with increased susceptibility to bacterial spot race T4 in tomato. J Am Soc Hortic Sci. American Society for Horticultural Science; 2014;139:282–9.

30. Li J, Chitwood J, Menda N, Mueller L, Hutton SF. Linkage between the I-3 gene for resistance to Fusarium wilt race 3 and increased sensitivity to bacterial spot in tomato. Theor Appl Genet. 2018;131:145–55.

31. Scott JA. 83.43 An interesting infinite series. The Mathematical Gazette. Cambridge University Press; 1999;83:305–7.

32. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nat Genet. 2017;49:692–9.

33. Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet. 2019;51:1052–9.

34. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund,

NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. europepmc.org; 2017;550:204–13.

35. Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell. 2016;166:492–505.

36. Tieman D, Zhu G, Resende MFR Jr, Lin T, Nguyen C, Bies D, et al. A chemical genetic roadmap to improved tomato flavor. Science. 2017;355:391–4.

37. Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C, et al. Rewiring of the Fruit Metabolome in Tomato Breeding. Cell. 2018;172:249–61.e12.

38. Tikunov YM, Molthoff J, de Vos RCH, Beekwilder J, van Houwelingen A, van der Hooft JJJ, et al. Non-smoky glycosyltransferase1 prevents the release of smoky aroma from tomato fruit. Plant Cell. 2013;25:3067–78.

39. Voichek Y, Weigel D. Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nat Genet. 2020;52:534–40.

40. Razifard H, Ramos A, Della Valle AL, Bodary C, Goetz E, Manser EJ, et al. Genomic Evidence for Complex Domestication History of the Cultivated Tomato in Latin America. Mol Biol Evol. 2020;37:1118–32.

41. van der Knaap E, Chakrabarti M, Chu YH, Clevenger JP, Illa-Berenguer E, Huang Z, et al. What lies beyond the eye: the molecular mechanisms regulating tomato fruit weight and shape. Front Plant Sci. 2014;5:227.

42. Chakrabarti M, Zhang N, Sauvage C, Muños S, Blanca J, Cañizares J, et al. A cytochrome P450 regulates a domestication trait in cultivated tomato. Proc Natl Acad Sci U S A. 2013;110:17125–30.

43. Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J, et al. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science. 2000;289:85–8.

44. Mu Q, Huang Z, Chakrabarti M, Illa-Berenguer E, Liu X, Wang Y, et al. Fruit weight is controlled by Cell Size Regulator encoding a novel protein that is expressed in maturing tomato fruits. PLoS Genet. 2017;13:e1006930.

45. Muños S, Ranc N, Botton E, Bérard A, Rolland S, Duffé P, et al. Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near WUSCHEL. Plant Physiol. 2011;156:2244–54.

46. Xu C, Liberatore KL, MacAlister CA, Huang Z, Chu Y-H, Jiang K, et al. A cascade of arabinosyltransferases controls shoot meristem size in tomato. Nat Genet. 2015;47:784–92.

47. Anastasiou E, Kenz S, Gerstung M, MacLean D, Timmer J, Fleck C, et al. Control of plant organ size by KLUH/CYP78A5-dependent intercellular signaling. Dev Cell. 2007;13:843–56.

48. Miyoshi K, Ahn B-O, Kawakatsu T, Ito Y, Itoh J-I, Nagato Y, et al. PLASTOCHRON1, a timekeeper of leaf initiation in rice, encodes cytochrome P450. Proc Natl Acad Sci U S A. 2004;101:875–80.

49. Paaby AB, Rockman MV. Cryptic genetic variation: evolution's hidden substrate. Nat Rev Genet. 2014;15:247–58.

50. Sackton TB, Hartl DL. Genotypic Context and Epistasis in Individuals and Populations. Cell. 2016;166:279–87.

51. Soyk S, Lemmon ZH, Oved M, Fisher J, Liberatore KL, Park SJ, et al. Bypassing Negative Epistasis on Yield in Tomato Imposed by a Domestication Gene. Cell. 2017;169:1142–55.e12.

52. Soyk S, Lemmon ZH, Sedlazeck FJ, Jiménez-Gómez JM, Alonge M, Hutton SF, et al. Duplication of a domestication locus neutralized a cryptic variant that caused a breeding barrier in tomato. Nat Plants. 2019;5:471–9.

53. Beyter D, Ingimundardottir H, Eggertsson HP, Bjornsson E, Kristmundsdottir S, Mehringer S, et al. Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease. bioRxiv. 2019;848366.

54. Domínguez M, Dugas E, Benchouaia M, Leduque B, Jiménez-Gómez JM, Colot V, et al. The impact of transposable elements on tomato diversity. Nat Commun. 2020;11:4058.

55. Danilevicz MF, Tay Fernandez CG, Marsh JI, Bayer PE, Edwards D. Plant pangenomics: approaches, applications and advancements. Curr Opin Plant Biol. Elsevier; 2020;54:18–25.

56. Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants. 2020;6:34–45.

57. Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet. 2018;50:1289–95.

58. Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, et al. The population genetics of structural variants in grapevine domestication. Nat Plants. 2019;5:965–79.

59. Veitia RA, Bottani S, Birchler JA. Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. Trends Genet. 2013;29:385–93.

60. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

61. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

63. Zarate S, Carroll A, Krashenina O, Sedlazeck FJ, Jun G, Salerno W, et al. Parliament2: Fast Structural Variant Calling Using Optimized Combinations of Callers. bioRxiv. biorxiv.org; 2018;424267.

64. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32:1220–2.

65. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15:R84.

66. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333–9.

67. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. nature.com; 2017;8:14061.

68. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974–84.

69. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

70. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. genome.cshlp.org; 2009;19:1586–92.

71. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72.

72. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27:737–46.

73. Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLoS Comput Biol. 2020;16:e1007981.

74. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

75. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

76. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016;32:3021–3.

77. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.

78. Jouffroy O, Saha S, Mueller L, Quesneville H, Maumus F. Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. BMC Genomics. bmcgenomics.biomedcentral.com; 2016;17:624.

79. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: an automatic transposable element classification tool. PLoS One. 2014;9:e91929.

80. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6.

81. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. genome.cshlp.org; 2017;27:722–36.

82. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. Springer; 2016;17:118.

83. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc. Wiley; 1995;57:289–300.

84. Sitgreaves R, Haggard EA. Intraclass correlation and the analysis of variance. J Am Stat Assoc. JSTOR; 1960;55:384.

85. Wilson DJ. The harmonic mean p-value for combining dependent tests. Proc Natl Acad Sci U S A. 2019;116:1195–200.

86. Park SJ, Jiang K, Schatz MC, Lippman ZB. Rate of meristem maturation determines inflorescence architecture in tomato. Proc Natl Acad Sci U S A. 2012;109:639–44.

87. Kremling KAG, Chen S-Y, Su M-H, Lepak NK, Romay MC, Swarts KL, et al. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature. 2018;555:520–3.

88. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.

89. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

90. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. Oxford University Press (OUP); 2015;31:166–9.

91. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. genomebiology.biomedcentral.com; 2014;15:550.

92. Brooks C, Nekrasov V, Lippman ZB, Van Eck J. Efficient gene editing in tomato in the first generation using the clustered regularly interspaced short palindromic repeats/CRISPR-associated9 system. Plant Physiol. academic.oup.com; 2014;166:1292–7.

93. Naito Y, Hino K, Bono H, Ui-Tei K. CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. Bioinformatics. academic.oup.com; 2015;31:1120–3.

94. Werner S, Engler C, Weber E, Gruetzner R, Marillonnet S. Fast track assembly of multigene constructs using Golden Gate cloning and the MoClo system. Bioeng Bugs. Taylor & Francis; 2012;3:38–43.

95. Gupta S, Van Eck J. Modification of plant regeneration medium decreases the time for recovery of Solanum lycopersicum cultivar M82 stable transgenic lines. Plant Cell Tissue Organ Cult. Springer Nature; 2016;127:417–23.

96. Lee J, Lee I. Regulation and function of SOC1, a flowering pathway integrator. J Exp Bot. academic.oup.com; 2010;61:2247–54.

97. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. academic.oup.com; 2011;27:2987–93.

98. Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, et al. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat Genet. 2014;46:1034–8.

99. Lee TG, Shekasteband R, Menda N, Mueller LA, Hutton SF. Molecular markers to select for the j-2–mediated jointless pedicel in tomato. HortScience. American Society for Horticultural Science; 2018;53:153–8.

As genome assemblies become more accurate, the standards by which we judge them must increase as well, with the ultimate goal of perfectly representing genomes.

6

Conclusion

THE PRECEDING CHAPTERS OUTLINE MY CONTRIBUTIONS to the field of plant genomics which primarily involve characterizing plant genotypes. Chapter 0 introduces how long DNA sequencing reads and associated mapping technologies can produce genome assemblies that completely represent genotypes. Chapters 1 and 2 describe RaGOO and RagTag, genome assembly scaffolding and improvement methods that facilitate chromosome-scale sequence-resolved physical maps. RaGOO and RagTag use genome assemblies to improve other genome assemblies — a timely and convenient methodological solution to a growing demand for assembly scaffolding. Importantly, the exceptional accuracy of modern reference and draft genome assemblies mitigates RaGOO/RagTag scaffolding errors, even when the genome assemblies represent distinct genotypes. While draft assemblies will become more accurate and contiguous, they will also be more numerous, and many will still require scaffolding. Therefore, RagTag's homology-based patching and scaffolding will continue to be a cheap and convenient solution for accurate scaffolding at scale.

As genome assemblies become more accurate, the standards by which we judge them must increase as well, with the ultimate goal of perfectly representing genomes [1,2]. While most future genome assemblies are expected to be chromosome-scale, researchers must still devise ways to automatically polish, finish, and validate raw assemblies [3,4]. This is exemplified in chapter 4, where we developed specialized techniques to locally reassemble telomeres, polish centromeres, and generally validate the COL-Cen reference genome. Beyond this specific example, many new so-called "gapless" or "Telomere-2-Telomere" (T2T) genome assemblies still omit or misassemble rDNAs, telomeres, segmental duplications, and satellite repeats. Assemblies will likely continue to imperfectly separate haplotypes and/or subgenomes, and as we showed in our evaluation of the recent human T2T assembly, they will likely harbor systematic technology-specific sequencing errors [1,3]. Correcting such errors is currently a laborious and manual process and a major bottleneck for genome assembly projects, and a new generation of computational tools is needed to automate this process.

Even if eukaryotic genome assembly eventually becomes a facile process, researchers need tools to store and organize pan-genomes and their metadata in order to actualize their utility. Already, researchers plan to assemble the genomes of thousands of humans and other researchers are planning to assemble thousands of fungi, animals, and other eukaryotic genomes. In plants, genome assembly will be applied to new cultivars, wild accessions, mapping populations, model genotypes, and mutant lines. To date, there is a lack of tools to organize and store such vast amounts of genome assemblies, with most research groups using internal *ad hoc* systems. To address this problem, we have created the pan-sol specification, a set of guidelines for storing and organizing sample and associated genome resource data (https://github.com/pan-sol/pan-sol-spec). The pan-sol specification is influenced by the Vertebrate Genomes Project, which similarly provides genome assembly organizational guidelines [1]. Additional work is needed to improve these guidelines, build accompanying databases and software, and foster participation from scientific communities.

Beyond storing foundational sample and genome assembly information, the genomics community must create new ways to store large databases and annotations that depend on specific assemblies. If this challenge is not addressed, plant genomics communities will suffer the same fate as human genomics which is dependent on legacy reference builds due to the cost of transitioning large databases to new and improved references [5]. Advances in cloud computing and "lift over" approaches show promise as solutions to reannotate new genome assemblies [5–9]. Doing so will enable any individual genome assembly within a pan-genome to serve independently as a robust reference genome.

Ultimately, the purpose of producing and organizing pan-genomes is to use them to study biological processes and to link important phenotypes to underlying genotypes. Chapter 5 describes our use of long-reads to catalog SVs in 100 diverse tomato accessions, ultimately uncovering broad and specific examples of SVs influencing important phenotypes. Researchers must continue to sequence more plant genomes to uncover SVs in larger populations. This will enable the discovery of important rare SVs, and it will also enable more robust genotyping of known SVs. Deeper catalogs of SVs and accompanying phenotype data will facilitate a broader understanding of the function of natural SV alleles in diverse genetic backgrounds via robust association mapping, machine learning, and probabilistic modeling. With rich functional annotation of natural SVs, researchers will be able to more precisely engineer SV alleles to efficiently produce specific phenotypes [10,11].

The natural successor to within-species pan-genomics is between-species pan-genomics. Such analysis in plants can reveal how genome evolution influences domestication [12,13], which can in turn be combined with genome editing for *de novo* domestication of minor crops [14–16]. Researchers can also leverage pan-genus genomes to study the relationship of gene orthologs and paralogs as well as the evolution of gene and transposable element families [17–21]. While robust genome assemblies facilitate cross-species comparisons, new specialized techniques are required for sensitive comparison of diverged species, including at the intergenic, gene, transcript, coding sequence, and protein level [17,22–25]. Comparing accurate genome assemblies both within and between species will provide a broad understanding of genome structure, function, and evolution, further yielding important applications for agriculture.

REFERENCES

1. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46.

2. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV. The complete sequence of a human genome. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.05.26.445798v1.abstract

3. Mc Cartney AM, Shafin K, Alonge M, Bzikadze AV. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. biorxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.07.02.450803.abstract

4. Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, et al. Significantly improving the quality of genome assemblies through curation. Gigascience [Internet]. academic.oup.com; 2021;10. Available from: http://dx.doi.org/10.1093/gigascience/giaa153

5. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation [Internet]. bioRxiv. 2021 [cited 2021 Sep 23]. p. 2021.07.12.452063. Available from: https://www.biorxiv.org/content/10.1101/2021.07.12.452063v1.abstract

6. Schatz MC, Langmead B, Salzberg SL. Cloud computing and the DNA data race. Nat Biotechnol. nature.com; 2010;28:691–3.

7. Zarate S, Carroll A, Mahmoud M, Krasheninina O, Jun G, Salerno WJ, et al. Parliament2: Accurate structural variant calling at scale. Gigascience [Internet]. academic.oup.com; 2020;9. Available from: http://dx.doi.org/10.1093/gigascience/giaa145

8. Mun T, Chen N-C, Langmead B. LevioSAM: Fast lift-over of variant-aware reference alignments. Bioinformatics [Internet]. academic.oup.com; 2021; Available from: http://dx.doi.org/10.1093/bioinformatics/btab396

9. Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. academic.oup.com; 2014;30:1006–7.

10. Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB. Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. Cell. 2017;171:470–80.e8.

11. Wang X, Aguirre L, Rodríguez-Leal D, Hendelman A, Benoit M, Lippman ZB. Dissecting cis-regulatory control of quantitative trait variation in a plant stem cell circuit. Nat Plants. 2021;7:419–27.

12. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell. 2020;182:145–61.e23.

13. Wang X, Gao L, Jiao C, Stravoravdis S, Hosmani PS, Saha S, et al. Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding. Nat Commun. nature.com; 2020;11:5817.

14. Lemmon ZH, Reem NT, Dalrymple J, Soyk S, Swartwood KE, Rodriguez-Leal D, et al. Rapid improvement of domestication traits in an orphan crop by genome editing. Nat Plants. nature.com; 2018;4:766–70.

15. Zsögön A, Čermák T, Naves ER, Notini MM, Edel KH, Weinl S, et al. De novo domestication of wild tomato using genome editing. Nat Biotechnol [Internet]. nature.com; 2018; Available from: http://dx.doi.org/10.1038/nbt.4272

16. Li T, Yang X, Yu Y, Si X, Zhai X, Zhang H, et al. Domestication of wild tomato is accelerated by genome editing. Nat

Biotechnol [Internet]. nature.com; 2018; Available from: http://dx.doi.org/10.1038/nbt.4273

17. Hendelman A, Zebell S, Rodriguez-Leal D, Dukler N, Robitaille G, Wu X, et al. Conserved pleiotropy of an ancient plant homeobox gene uncovered by cis-regulatory dissection. Cell. 2021;184:1724–39.e16.

18. Glover N, Dessimoz C, Ebersberger I, Forslund SK, Gabaldón T, Huerta-Cepas J, et al. Advances and Applications in the Quest for Orthologs. Mol Biol Evol. academic.oup.com; 2019;36:2157–64.

19. Wei C, Wang Z, Wang J, Teng J, Shen S, Xiao Q, et al. Conversion between 100-million-year-old duplicated genes contributes to rice subspecies divergence. BMC Genomics. bmcgenomics.biomedcentral.com; 2021;22:460.

20. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of Gene Duplication in Plants. Plant Physiol. academic.oup.com; 2016;171:2294–316.

21. International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators:, Appels R, Eversole K, Feuillet C, Keller B, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science [Internet]. science.sciencemag.org; 2018;361. Available from: http://dx.doi.org/10.1126/science.aar7191

22. Wu Y, Johnson L, Song B, Romay C, Stitzer M, Siepel A, et al. A multiple genome alignment workflow shows the impact of repeat masking and parameter tuning on alignment of functional regions in plants [Internet]. bioRxiv. 2021 [cited 2021 Sep 24]. p. 2021.06.01.446647. Available from: https://www.biorxiv.org/content/10.1101/2021.06.01.446647v1.abstract

23. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587:246–51.

24. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. Springer; 2019;20:238.

25. Gilchrist CLM, Chooi Y-H. Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. Bioinformatics [Internet]. academic.oup.com; 2021; Available from: http://dx.doi.org/10.1093/bioinformatics/btab007