Hulse-Kemp *et al. Horticulture Research* (2018)5:4
DOI 10.1038/s41438-017-0011-0

Horticulture Research

ARTICLE Open Access

Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library

Amanda M. Hulse-Kemp[1,2,3], Shamoni Maheshwari[1], Kevin Stoffel[1], Theresa A. Hill[1], David Jaffe[4], Stephen R. Williams[4], Neil Weisenfeld[4], Srividya Ramakrishnan[5], Vijay Kumar[4], Preyas Shah[4], Michael C. Schatz[5], Deanna M. Church[4] and Allen Van Deynze[1]

Abstract

Linked-Read sequencing technology has recently been employed successfully for *de novo* assembly of human genomes, however, the utility of this technology for complex plant genomes is unproven. We evaluated the technology for this purpose by sequencing the 3.5-gigabase (Gb) diploid pepper (*Capsicum annuum*) genome with a single Linked-Read library. Plant genomes, including pepper, are characterized by long, highly similar repetitive sequences. Accordingly, significant effort is used to ensure that the sequenced plant is highly homozygous and the resulting assembly is a haploid consensus. With a phased assembly approach, we targeted a heterozygous F_1 derived from a wide cross to assess the ability to derive both haplotypes and characterize a pungency gene with a large insertion/deletion. The Supernova software generated a highly ordered, more contiguous sequence assembly than all currently available *C. annuum* reference genomes. Over 83% of the final assembly was anchored and oriented using four publicly available *de novo* linkage maps. A comparison of the annotation of conserved eukaryotic genes indicated the completeness of assembly. The validity of the phased assembly is further demonstrated with the complete recovery of both 2.5-Kb insertion/deletion haplotypes of the *PUN1* locus in the F_1 sample that represents pungent and nonpungent peppers, as well as nearly full recovery of the BUSCO2 gene set within each of the two haplotypes. The most contiguous pepper genome assembly to date has been generated which demonstrates that Linked-Read library technology provides a tool to *de novo* assemble complex highly repetitive heterozygous plant genomes. This technology can provide an opportunity to cost-effectively develop high-quality genome assemblies for other complex plants and compare structural and gene differences through accurate haplotype reconstruction.

Introduction

Pursuing a gold-standard reference genome for each biologically important organism has become a goal of the individual research communities in order to have a tool for answering biologically relevant questions[1–5]. The construction of contiguous genome assemblies has allowed for discovery of genes and gene function, as well as improved our understanding of genomic elements and structure that regulate biological processes in humans, microbes, animals, and plants. These high-quality reference genome assemblies allow for not only complete gene models but also complete promoter regions and more remote regulatory sequences of every gene, as well as true representation of other complex features that are important for trait expression. Having a highly accurate, contiguous, and complete genomic representation allows for unprecedented studies on chromosome-scale evolution, as well as of molecular evolution of polyploid events, gene amplifications, haplotype tracking, and mobile element proliferations. The importance of having complete, well-ordered genomic representations continues to be a

Correspondence: Amanda M. Hulse-Kemp (amanda.hulse-kemp@ars.usda.gov) or Allen Deynze (avandeynze@ucdavis.edu)
[1]Department of Plant Sciences, University of California, Davis, CA, USA
[2]USDA-ARS Genomics and Bioinformatics Research Unit, Raleigh, NC, USA
Full list of author information is available at the end of the article

© The Author(s) 2018
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

significant concern, as it is difficult to attribute a simple explanation for phenotypes and disease, even when overcoming sample size problems; such as in human studies with hundreds of thousands of subjects, where 50–75% of the height phenotype has currently been unable to be explained[6,7].

In plant breeding, the availability of a contiguous genome provides a means to better understand traits and how they interact with their environment in different genetic backgrounds. At the simplest level, it allows for association of genetic markers for selection and introgression of traits across germplasm to enable hypothesis-driven crop improvement by understanding the pathways and development of novel products for consumers. A high-quality genome serves as a tool for more efficient studies with higher statistical power for localization of causal genomic regions and genes responsible for economically important traits.

When the first high-quality plant genomic sequences were achieved using bacteria artificial chromosome (BAC)-based approaches coupled with Sanger sequencing technology[8], it was limited to a few small diploid species, such as Arabidopsis and rice, due to the labor-intensive and expensive protocols, with project costs in the tens to hundreds of millions of dollars for a single genome. Next-generation sequencing technology, such as Illumina sequencing-by-synthesis, has dramatically reduced costs in the past decade and led to the construction of a large number of draft genomes using a combination of paired-end and mate-pair libraries with short reads and high redundancy. However, these draft genomes are usually of low quality and comprised of a large number of contigs (in 100,000 s or more) with scaffold N50s in the hundreds of kilobase pairs (Kb) or less and contig N50s in the tens of Kb, where N50 is the contig or scaffold size at which 50% of the entire assembly is contained. A combination of short-read sequencing technology with physical and genetic maps has led to a drastic improvement in scaffold sizes, but not to contigs, leaving many gaps and misassembled or unassembled regions, especially in repeated regions. Long-read technology was introduced in the last decade and led to dramatic improvements in resolution with N50s over 1 megabase (Mb). However, this technology is considerably more expensive than Illumina sequencing (~10–20 times more expensive or more), making it difficult to implement with large complex plant genomes. This is particularly problematic for crop genomes which range in size from 0.35 Gb in rice to over 23 Gb in loblolly pine, and are further complicated by polyploidy, varying levels of heterozygosity, and large stretches of highly similar repeat sequence, all of which make their sequencing and assembly difficult with standard technologies[9].

To combat the issue of heterozygosity in plants, plant geneticists have taken great concern over choosing a highly homozygous plant for sequencing, such as sequencing highly inbred varieties or by sequencing haploid tissues. This reduced the number of problematic "bubbles", or points in the sequence which the software is unsure of the correct sequence path where there may be multiple choices, during computational assembly generation and produced a haploid consensus sequence assembly. However, in many species, it is not practical to breed homozygous varieties or collect haploid tissue in sufficient quantity for sequencing. Furthermore, generating a haploid consensus sequence will not accurately portray both haplotypes in plants that have varying levels of heterozygous presence/absence variants (PAVs) or insertions/deletions. PAVs are variant types which can quickly lead to alterations in phenotypes when occurring in genic regions, as events which affect sequences outside of a derivative of 3 bps lead to frameshifts which are likely to alter the function of proteins. One economically important example of a change in phenotype due to a PAV is the altering of pungency, or spiciness, in pepper. The *PUN1* gene is a putative acyltransferase in which a 2.5-Kb PAV has been shown to be the causal variant in determining a pepper's distinctive pungent flavor[10]. Having an accurate representation of the PAV regions within an individual line will provide additional power to assess the true biology behind a trait instead of working with a synthetic sequence generated through traditional haploid consensus sequence assembly.

Pepper is a member of the Solanaceae family, which contains several of the most economically important crop species including tomato, potato, eggplant, and tobacco. The pepper genome is a representative complex plant genome; it has one of the largest genome sizes in the Solanaceae family at ~3.5 Gb and is comprised largely of repetitive elements, estimated at 75–80% of the genome[11,12]. The most cultivated pepper species (*C. annnum*) is diploid, and to date has three draft genome assemblies developed using short-read sequencing technology. All three assemblies focused on a different *C. annuum* line, CM334 which is a Mexican landrace hot pepper[11], Zunla-1 which is a widely cultivated accession and Chiltepin which is a wild progenitor of Zunla-1[12], each being mostly homozygous due to their primarily self-pollinating mating type. Similar to most short-read sequencing derived references, the three pepper assemblies are comprised of a large number of small scaffolds with 37,989 scaffolds in CM334[11], 967,017 scaffolds in Zunla-1[12], and 1,973,483 scaffolds in the Chiltepin[12] genomes with the largest scaffold N50 at 2.47 Mb in the CM334 assembly and largest contig N50 of 55 Kb with Zunla-1. Additional genetic resources have also been recently developed in pepper with one study producing two high-quality

manually curated genetic linkage maps with a custom Affymetrix Genechip[13], another study producing a map with an Illumina Infinium array[14], and finally a study producing a map with skim sequencing of a population utilizing the CM334 genome assembly[15]. While multiple short-read assemblies have been implemented in pepper, the large genome size has prohibited sequencing using long-read technology, as such projects would currently have sequencing costs upward of $25–100 K.

Recent advances in library preparation methods have allowed for integration of structural location information of a sequence with the cost-effective short-read sequencing technology. The 10x Chromium technology (10x Genomics, San Francisco, USA) has the potential to strike a big impact in making complex plant genomes more generally accessible by generating long-range information analogous to traditional BAC-by-BAC sequencing technologies but at a tiny fraction of the cost and at high throughput. This technology isolates large DNA fragments (~150 Kb) and creates barcoded Illumina genomic libraries allowing the short reads to be localized by capturing not only sequence, but physical association of the DNA. These libraries are in turn sequenced to about 60× coverage to use for *de novo* assembly using a phased assembly strategy where two individual haplotypes are generated as output[16].

In the current study, we have investigated the Linked-Read technology as a cost-effective resource for sequencing the 3.5-Gb complex pepper plant genome, validated the assembly produced by the technology using four high-density genetic maps, and tested the feasibility of the new assembly for answering biologically relevant questions of the genome structure related to the flavor of the plant with the *PUN1* locus. We show that a significantly improved *de novo* genome assembly can be achieved at a fraction of the time and cost of traditional short-read and long-read assemblies with little input DNA, for a highly repetitive complex heterozygous pepper plant.

Materials and methods

Plant material and DNA extraction

High-molecular-weight DNA was isolated using a modified CTAB protocol[17] in a F_1 of a wide cross (UCD-10X-F_1) between a landrace, Criollos de Morelos 334 (CM334) and a nonpungent blocky pepper-breeding line. The only protocol modification was running the Pippen Pulse gel on the 5–150-Kb setting to perform size selection of fragments over 48 Kb (Supplemental Fig. 1).

Library construction and sequencing

High-molecular-weight DNA (1.25 ng) was loaded onto a Chromium controller chip, along with 10x Chromium reagents and gel beads following manufacturers recommended protocols (https://support.10xgenomics.com/de-novo-assembly/library-prep/doc/user-guide-chromium-genome-reagent-kit-v1-chemistry). Briefly, initial library construction takes place within droplets containing beads with unique barcodes (called GEMs). The library construction incorporates a unique barcode that is adjacent to read one. All molecules within a GEM get tagged with the same barcode, but because of the limiting dilution of the genome (roughly 300 haploid genome equivalents), the probability that two molecules from the same region of the genome are partitioned in the same GEM is very small. Thus, the barcodes can be used to statistically associate short reads with their source long molecule. The resulting library was sequenced on two lanes of an Illumina HiSeq X Ten sequencer to produce 2 × 150 paired-end sequences. The resulting data type is called "Linked-Reads"[18]. Raw data have been uploaded to NCBI's Small Read Archive (SRP117183).

Assembly and assembly verification

The Linked-Read data were assembled using an intermediate version of the Supernova™ assembler[16], between versions 1.0 and 1.1, using the default recommended settings and available on GitHub: https://github.com/10XGenomics/supernova-chili-pepper to produce two "pseudohap" assembly outputs. Molecule size following sequencing of the Chromium library was estimated using the LongRanger tool from 10x Genomics (Supplemental Fig. 2). Genome-wide statistics were calculated on the total number of phase blocks and the N50 of individual phase block sizes in the pseudohap outputs produced in the Supernova 10x assembly. The size distribution of phase blocks was plotted. Corresponding contigs from each pseudohap were compared using LongRanger and the total number of single-nucleotide polymorphisms (SNPs) between pseudohaps was determined. For all of the following analyses, pseudohap1 was utilized, unless differently specified.

To verify the quality of the assembly, we compared the order of contigs to four high-density intraspecific (*C. annuum*) and interspecific (*C. annuum* × *C. frutescens*) genetic maps: three transcriptome-based maps[13,14] and one genomic map[15]. Marker and/or flanking sequences for all maps were aligned as query sequences to the Supernova 10x assembly using GMap v09-14-2016 with default settings[19]. Alignment results were filtered for hits obtaining >90% of the query sequence of at least 200 base pairs (bp) aligning to the target at 98% sequence identity. In the case of the Array map[14], the cutoff for size of query sequence was >100 bp to conform to the assayed sequences on the array. The number of scaffolds that a marker aligned to were examined; after filtering, the majority of the markers were homologous to a single scaffold (Supplemental Fig. 3A–D). For further analyses, markers/ESTs that aligned to more than one 10x scaffold

were removed. Alignment positions (Mb) of the linkage-mapped markers were plotted against their centiMorgan (cM) position in the linkage map, and if markers/ESTs from >1 linkage group aligned to the same 10x scaffold, then, the primary linkage group of that scaffold plotted is that of the majority of corresponding markers. The 10x scaffolds belonging to the same primary linkage group were sorted in order of increasing genetic distance of their aligned ordered markers. Additionally, where previously reported positions of markers on the Kim et al.[11] assembly were available, their positions were plotted against aligned position on the 10x scaffolds, where the 10x scaffolds were first assigned to primary chromosomes using similar logic previously mentioned and then sorted in order of the location of the marker on the Kim et al. genome sequence.

Marker alignment positions on the 10x scaffolds were converted into csv format for input into AllMaps software[20] for all four linkage maps (Supplemental File 1). Interspecific linkage map files were modified to correct for a known translocation between chromosomes 1/8 between *C. annuum* and *C. frutescens*[21]. AllMaps was initially run for sets of markers representing a single chromosome with five different parameter sets for the array map[14], FA map[13], Han map[15], and NM map[13] as follows: (1) equal weights, (2) unequal weights1—1/2/2/2, (3) unequal weights2—1/3/2/4, (4) unequal weights3—1/3/2/3, and (5) unequal weights4—2/3/1/4. The best result of the five-parameter set was determined by the greatest number of anchored and oriented scaffolds.

Pseudomolecule construction

Final pseudomolecules were constructed in AllMaps using the unequal weights2 parameters for a single AllMaps run for the entire genome. The resulting final pseudomolecules were deposited to NCBI under BioProject ID PRJNA376668, designated UCD10X Assembly v1.0. Assembly statistics through each step of the analysis from contigs to scaffolds to pseudomolecules were calculated.

Comparison of assembly with published assemblies

Quast[22] (version 4.1) was utilized to simultaneously compare the UCD10X Assembly v1.0 with three published *C. annuum* sequences, CM334 v1.55[11], Zunla-1 v2.0[12], and Chiltepin v2.0[12]. Default parameters were used except for a minimum contig size of 45 bp and using the scaffold option to also compare broken assemblies. Distributions of contig lengths of the broken assemblies were plotted.

All four pepper assemblies and the *Solanum lycopersicum* version 3.0 (Tomato) were aligned against each other in a pairwise fashion for all pseudochromosome sequences using Mummer version 3.23[23]. The nucmer alignment algorithm was utilized requiring minimum clusters of 100 bp with maximum gap of 500 bp. These results were then filtered for the within-pepper comparisons for alignment lengths of >500 bp at 98% identity and for the pepper-to-tomato comparisons for alignment lengths of >150 bp at 85% identity. These filtered alignments were plotted for visualization using the mummerplot function. Chromosome 2 alignments were extracted from the overall alignments as a highlight. Nucmer –maxmatch algorithm results with minimum alignment length set to 20 bp and minimum length clusters of 100 bp were extracted from the four pepper assemblies aligned to the Tomato and analyzed with Assemblytics software[24].

The benchmarking software BUSCO version 2.0[25] was run against all four pepper assemblies (UCD10X, CM334, Zunla, and Chiltepin), as well as UCD10X-pseudohap2, Tomato—*S. lycopersicum* version 3.00, Potato—*S. tuberosum* version 3, Eggplant—*S. melongena* version 2.5.1, three Tobacco—*Nicotiana tabacum* genomes (K326, TN90, and BX), two *Petunia* genomes (*axillaris* and *inflata*), and Carrot—*Daucus carota* version 2.0 as a related outgroup with dependencies of BLAST 2.6.0, HMMER 3.1b2, and AUGUSTUS 3.2.3[25]. The all-plant ancestry set, embryophyta_odb9, was used as a reference and all runs utilized tomato species parameters through the "–sp tomato" option. The number of complete single copy, complete duplicated, fragmented, and missing BUSCOs were calculated and compared.

Whole-genome sequence of three pepper Recombinant Inbred Lines (RILs) was obtained from the NCBI Small Read Archive (SRR2751915, SRR2751916, and SRR2751917). Sequences were trimmed for quality and TruSeq adapters using CLC Genomics Workbench version 8.5.1 with default parameters except for hard trimming of the first 12 bases and last 5 bases and removing reads shorter than 75 bases. Trimmed reads were aligned to each pepper genome, UCD10X v1.0, CM334 v1.55, Zunla-1 v2.0, and Chiltepin v2.0 using CLC Workbench version 8.5.1 with 0.8-length fraction and 0.9 similarity fraction.

Assessment of *PUN1*

All *PUN1 Capsicum* full-length coding sequences were obtained from the NCBI database (AY819027.1, AY819026.1, AY819029.1, EF104910.1, HM854860.1, GU300812.1, AY819032.1, AY819031.1, and AY819030.1). All sequences were aligned with default BLAST alignment parameters to the UCD10X v1.0 pseudomolecules to identify the PUN1 sequence. The identified region and the corresponding region on the other haplotype sequence were identified and extracted; this was accomplished utilizing the fasta header information which contains the relationship of the scaffolds in the pseudohaps from the Supernova graph assembly structure. A multiple-sequence alignment was

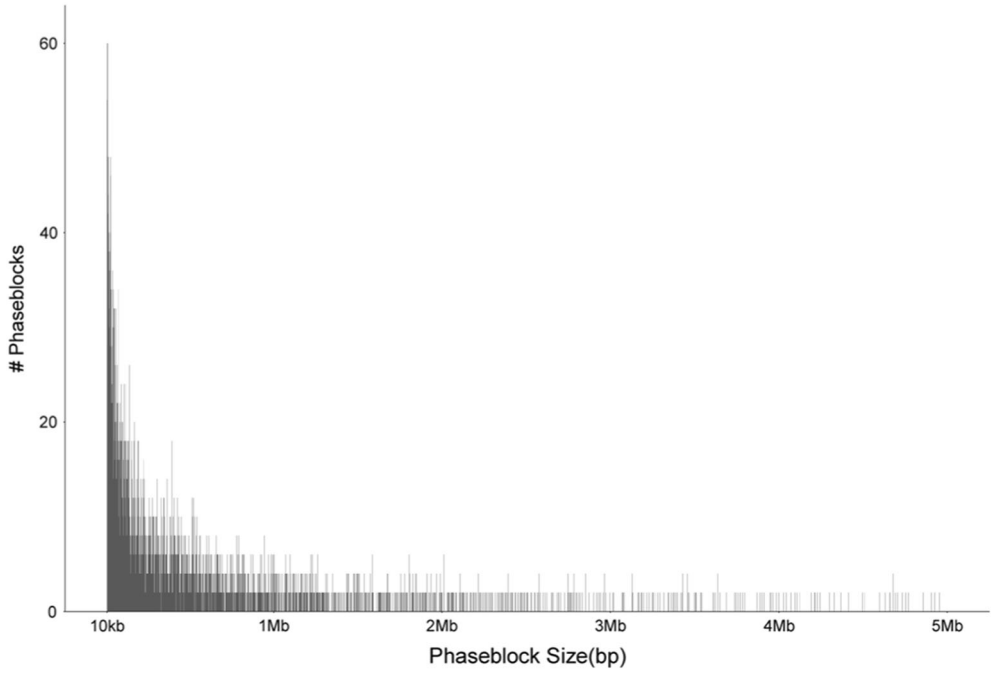

Fig. 1 Size distribution of phase blocks in Supernova pseudohap outputs. Frequency of sizes of all phase blocks in the UCD10X assembly is plotted

generated using MUSCLE software for both extracted sequences and the seven full-length PUN1 sequences from NCBI. The multiple-sequence alignment was analyzed in JalView 2.10.1 and the first three bases of haplotype 2.1 were corrected to positions 24–26. The manually curated alignment was exported in FNA format and visualized.

Results

Sequencing and assembly

A total of ~56-fold read coverage was obtained with paired-end 150-bp reads using 10x Chromium technology sequenced on the Illumina HiSeq X Ten for a single pepper genotype. The sequenced *C. annuum* genotype was an F_1 wide cross hybrid of CM334 and a nonpungent pepper-breeding line. The 56× data were assembled using an intermediate version of the Supernova Assembler[16], between versions 1.0 and 1.1, and available here https://github.com/10XGenomics/supernova-chili-pepper. The assembler utilizes a graph-based assembly approach along with individual molecule barcodes to resolve complex repeats and separate chromosomes based on haplotype information for a phased assembly[16]. Analogous to FALCON-unzip for diploid assembly of PacificBiosciences sequencing[26], the Supernova output produces locally phased haplotype blocks, or pseudohaps, however, full phasing of complete chromosomes is still not completely achieved with the overall output, i.e., full maternal and paternal haplotypes are not deduced. Although the full genome representation is contained in the two resulting pseudohap outputs, it is phased only locally within a phase block (see Fig. 1 in the Supernova paper for an Illustration[16]). The two pseudohaps were compared using LongRanger and a genome-wide SNP rate was found to be 0.4%. The mean distance between the nearest SNP markers was found to be 102 bp with standard deviation of 502 bp, median distance of 22 bp, and largest distance of 393 Kb. For the purpose of generating a reference assembly, a single pseudohap was utilized and the corresponding haplotype for each contig is available from the SOL Genomics Network database (https://solgenomics.net). The pseudohap1 assembly, hereby referred to as the assembly, contained 83,391 scaffold sequences with an N50 of 3.69 Mb for a total assembly size of 3.21 Gb (Supplemental File 2). There were a total of 258,884 phase blocks in the assembly with an N50 size of 1.72 Mb, and the size distribution of the phase blocks is shown in Fig. 1.

We validated the overall structure of this assembly by comparing it to four high-density genetic maps available in pepper and found that the marker orders in the Supernova assembly were highly concordant to three transcriptome-derived maps[13,14] and one genomic-based map[15] (Fig. 2a; Supplemental Figs. S4–7). Physical location of markers on the assembly was also compared with the CM334 Pepper Genome V1.55[11] (Fig. 2b) which showed that physical

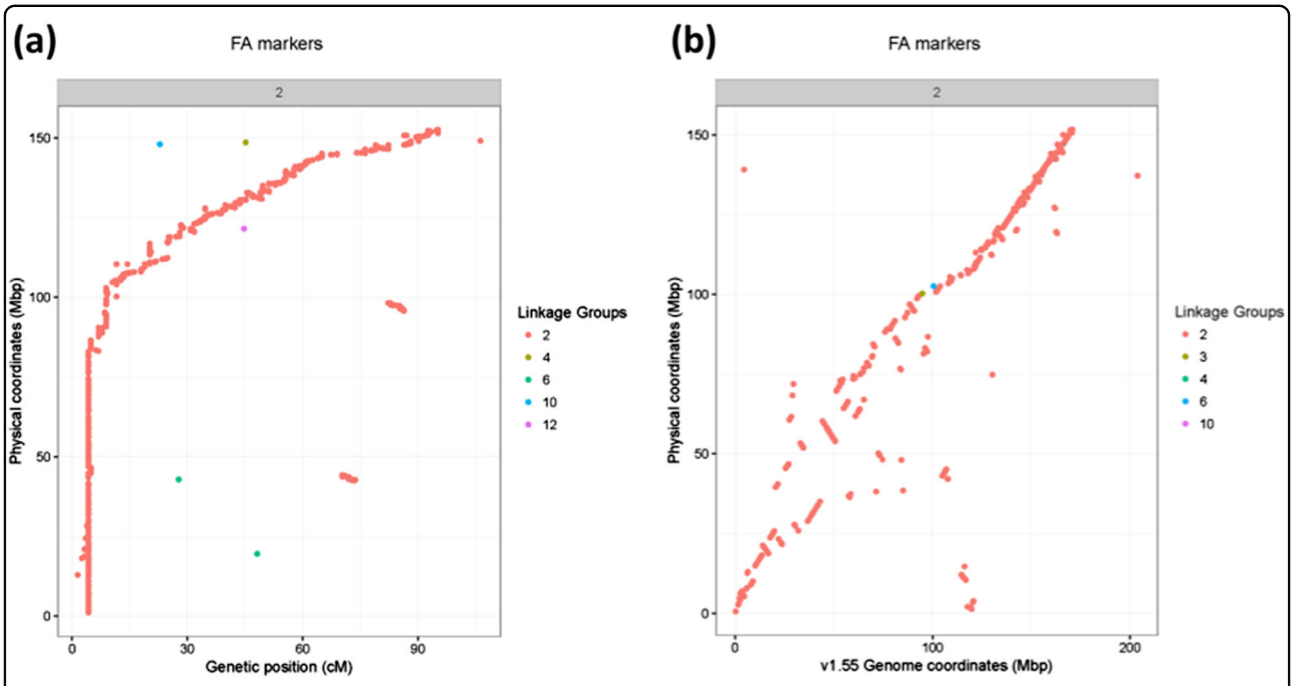

Fig. 2 Assessment of raw 10x assembly contigs compared to Hill et al.[13] *Capsicum frutescens* × *Capsicum annuum* genetic map, shown for Chromosome 2. Assembly scaffolds are ordered based on their primary linkage group, sorted in order of increasing genetic distance. Linkage group colored labels correspond to marker linkage group in Hill et al. **a** Genetic positions of markers (centiMorgans) are shown versus the physical position on the 10x assembly contigs (megabase pair). **b** Physical position on the 10x assembly contigs (Mbp) versus physical position on the *C. annuum* CM334 version 1.55 assembly

location in pericentromeric regions appeared to be more consistent in the 10x assembly positions along the contigs, this corroborates the findings of Hill et al.[13] where some deviations were observed in pericentromeric regions in regard to genetic position on the map.

Filtered alignments of marker sequences for the four maps (Supplemental File 1) were utilized in the AllMaps software[20] using a weighted approach to generate pseudomolecules with currently available genetic map information. The highest weights were applied to the highly manually curated EST maps by Hill et al.[13] and the less manually curated maps were given lower weights[14,15]. Additionally, the intraspecific maps were placed at higher weights, while the interspecific maps were given lower weights. This was done to prioritize the within-species (*C. annuum*) maps as the interspecific maps had to be corrected for a known translocation in the maps due to the structure of *C. frutescens* compared to *C. annuum*[21]. Chromosome-scale pseudomolecules produced 12 major scaffolds corresponding to 12 pepper chromosomes (Fig. 3, Table 1; Supplemental Fig. 8–19). A total of 2.67 Gb was anchored to the 12 chromosomes along with 541 Mb of unplaced sequence for a total assembly of 3.21 Gb designated as UCD10X version 1.0 (available as NCBI bioproject PRJNA376668). The N50 of contigs was 123 Kb, of scaffolds, it was 3.69 Mb, and of pseudomolecules, it was 227.2 Mb. Over 83% of the assembled sequence was anchored into the final assembly (Supplemental File 3).

Comparison to published sequences

UCD10X was compared to the three other publicly available *C. annum* genome sequences: (1) CM334 version 1.55[11], (2) Zunla-1 version 2.0, and (3) Chiltepin version 2.0[12] using QUAST[22]. The GC% of UCD10X was 34.91%, comparable to the other published assemblies, which ranged from 34.97 to 35.09%. The length of sequence anchored to pseudochromosome scaffolds ranked second among the assemblies at 2.67 Gb anchored, compared to CM334 (2.75), Zunla-1 (2.65), and Chiltepin (2.45). The overall size of the assembly (3.21 Gb) was also within the range of the other assemblies.

Although the assemblies at the overall level of pseudochromosomes appeared comparable, the quality within the pseudochromosomes was variable, especially when the assemblies were resolved to their constituent contigs. The UCD10X assembly contained the smallest number of contigs (134,573), with the next most contiguous genome having 32% more contigs (CM334 v1.55—177,870). The UCD10X assembly also has a contig N50 of 123 Kb, 2× greater than

Table 1 Chromosome statistics for UCD10X genome assembly

	Final	Anchored	Oriented			Gaps	Length
Chromosome	#scaffolds	#scaffolds	#scaffolds	bp	bp %	#N's	
LG01	1	134	51	173,703,399	67.96%	13,300	255,602,192
LG02	1	119	36	79,737,851	51.67%	11,800	154,330,250
LG03	1	153	61	155,226,247	57.37%	15,200	270,589,393
LG04	1	108	44	130,088,991	56.19%	10,700	231,497,844
LG05	1	136	44	121,301,069	54.85%	13,500	221,136,838
LG06	1	136	39	99,428,383	43.56%	13,500	228,254,876
LG07	1	139	40	81,805,368	36.01%	13,800	227,195,441
LG08	1	61	25	127,432,512	73.33%	6,000	173,776,113
LG09	1	203	28	75,652,043	34.53%	20,200	219,064,469
LG10	1	121	43	116,352,171	52.48%	12,000	221,721,387
LG11	1	143	35	106,810,631	45.27%	14,200	235,950,708
LG12	1	134	43	122,946,399	52.77%	13,300	232,995,803
Total	12	1,587	487	1,390,485,064	52.04%	157,500	2,672,115,314
Unplaced	81,506	–	–	–	–	–	540,231,928
Including unplaced	81,518	1587	487	1,390,485,064	43.29%	157,500	3,212,347,242

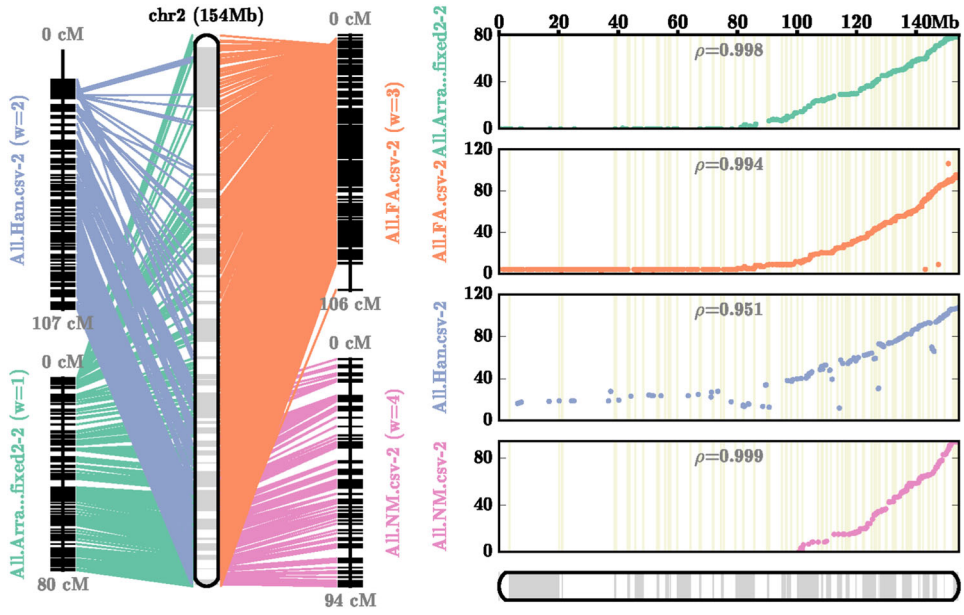

Fig. 3 AllMaps chromosome 2 consensus map for pseudomolecule generation. Order of markers in the four linkage maps compared to the final pseudomolecule generated through consensus map derivation using the AllMaps software with Unequal Weights2 parameters through whole-genome run

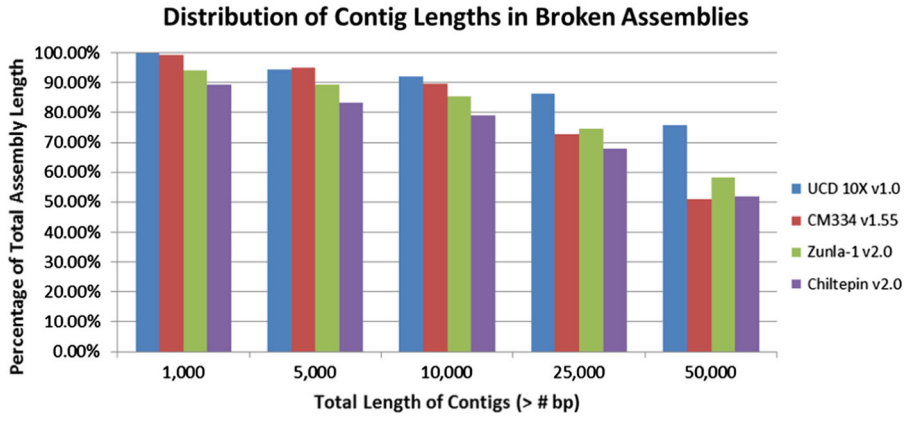

Fig. 4 Distribution of contig lengths in broken assemblies. Proportion of the total assembly length found in individual contigs, based on extracted individual contig sequences whenever there was more than a single "N" character in a row in the scaffolds, for four *Capsicum annuum* genome sequences

the other three genomes. Ultimately, the UCD10X produced the most contiguous assembly with over 75% of the total sequence length in contigs over 50 Kb (Fig. 4).

Gene content of UCD10X-psueodhap1 and the corresponding pseudohap2, the three other Capsicum genomes, as well as assembly sequences for other Solanaceae (Tomato[27], Potato[28], Tobacco[29], Eggplant[30], and Petunia[31]) and an asterid outgroup (Carrot[32]) were assessed using BUSCO2[25], a standard benchmarking software for assessing genome completeness by measuring the number of core genes present and full length in the assemblies. The embryophyta_0db9* standard data set includes 1440 genes that are conserved among 90 representative plants. All runs utilized tomato as the training species for gene model detection and UCD10X-pseudohap1/2 was found to perform comparably to the other pepper genomes containing 1343 (93.26%) and 1345 (93.40%) of the total genes in complete copies, respectively (Fig. 5). The diploid Solanaceae including Potato, Eggplant, and Pepper appeared to all have similar numbers of fully duplicated

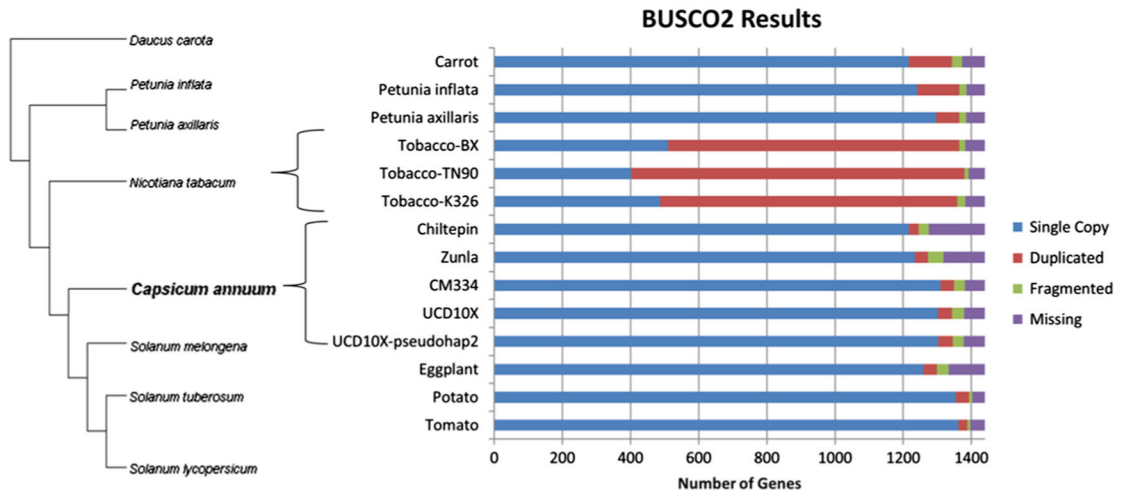

Fig. 5 BUSCO2 conserved gene analysis comparison. BUSCO2 results for the UCD10X assembly and pseudohap2, three other pepper assemblies, and other available Solanaceae family genomes including Tomato, Potato, Tobacco, Eggplant and Petunia, and Carrot as an asteroid outgroup

conserved genes at 38–42, while Petunia had closer to the Carrot outgroup and polyploid Tobacco had high numbers of duplicated genes as expected.

The final pseudomolecules showed high congruence with the other pepper genomes over the euchromatic, noncentromeric regions of the chromosomes, as shown on the long arm of Chromosome 2 (Fig. 6; Supplemental Fig. 20–23). Chromosome 2 is an acrocentric chromosome with a very small short arm compared to the long arm[33]. The heterochromatic region appears to be highly variable between all of the assemblies, except for the comparison between Zunla-1 and Chiltepin. A similar pattern of higher collinearity in euchromatic regions and lower collinearity in centromeric regions is observed for all pepper assemblies when comparing chromosome 2 to the orthologous chromosome 2 in tomato (Fig. 7). Furthermore, when the syntenic region is examined closer, the alignment of the UCD10X sequence appears to be more contiguous and contains better-oriented contigs overall compared to tomato with less switching between plus (red) and minus (blue) orientation in the comparison. As tomato is a closely related Solanaceae member, it is expected that the sequences would show high colinearity. Determining the correct sequence orders between related species is desired to allow for identification of high-resolution syntenic relationships and assessment of gene orthology for comparative studies[34].

We further analyzed the four pepper assemblies reference contigs with the whole-genome MUMmer[23] alignments to the Tomato genome reference with Assemblytics[24] to discover any interspecific structural variants that ranged within 1–10,000 bp. This approach depends on having a high-quality assembly, as it is not

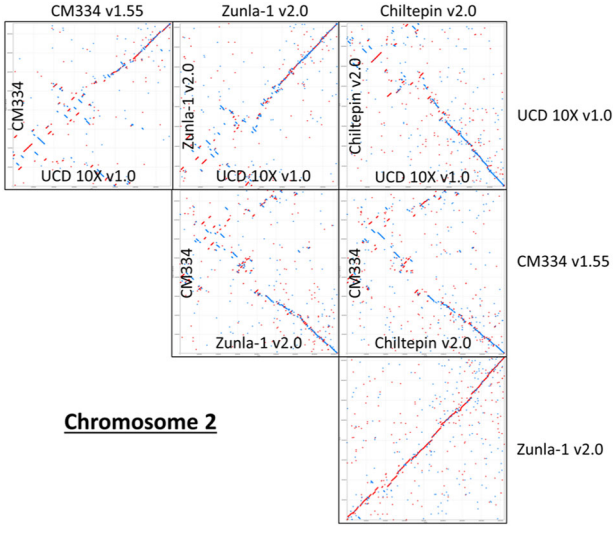

Fig. 6 Comparison of 10x assembly versus other pepper assemblies. Filtered nucmer alignments extracted for Chromosome 2 are shown for alignment lengths of greater than 150 base pairs at greater than 95% sequence identity. Pairwise alignments between the UCD10X version 1.0, CM334 version 1.55, Zunla version 2.0, and Chiltepin version 2.0

possible to find large structural variations if the assembly is highly fragmented. Assemblytics reported that the UCD10X showed the most number of structural variations and most total bases affected by structural variants of all the four pepper assemblies (Supplemental File 4). The UCD10X appears to have captured the most repeat expansions with 6% more events, which covers 20% more in terms of total bases relative to the total bases

Table 2 Comparison of mapping reference utility of four pepper assemblies

	Number Mapped				Number Mapped			
	UCD10X	CM334	Zunla-1	Chiltepin	UCD10X	CM334	Zunla-1	Chiltepin
SRR2751915	22,688,791	22,015,175	21,803,948	21,631,310	1.05	1.02	1.01	1.00
SRR2751916	24,906,044	24,143,895	23,901,098	23,719,715	1.05	1.02	1.01	1.00
SRR2751917	19,652,903	19,098,575	18,927,613	18,812,663	1.04	1.02	1.01	1.00
	Percentage Mapped				Percentage Mapped			
SRR2751915	93.92%	91.13%	90.26%	89.54%	1.05	1.02	1.01	1.00
SRR2751916	93.66%	90.80%	89.88%	89.20%	1.05	1.02	1.01	1.00
SRR2751917	94.05%	91.40%	90.58%	90.03%	1.04	1.02	1.01	1.00
	Non-Perfect Matches				Non-Perfect Matches			
SRR2751915	6,176,100	6,937,196	6,013,706	11,243,193	1.03	1.15	1.00	1.87
SRR2751916	7,056,971	7,818,049	6,878,147	12,370,325	1.03	1.14	1.00	1.80
SRR2751917	5,225,424	6,300,355	5,480,526	9,514,125	1.00	1.21	1.05	1.82
	Reads in Pairs				Reads in Pairs			
SRR2751915	19,725,496	19,085,452	18,602,784	17,452,936	1.13	1.09	1.07	1.00
SRR2751916	21,548,710	20,737,978	20,198,340	19,034,832	1.13	1.09	1.06	1.00
SRR2751917	16,892,830	16,356,714	15,973,206	15,127,852	1.12	1.08	1.06	1.00

Paired-end whole-genome sequences for three pepper recombinant inbred line (RIL) samples were aligned using the CLC Genomics Workbench v8.5.1 to the UCD10X v1.0, CM334 v1.55, Zunla-1 v2.0, and Chiltepin v2.0 genome reference sequences. Number of mapped reads, percentage of mapped reads, number of non-specific matches within mapped reads, and number of reads in pairs are shown in numbers (left) and relative ratios (right). Relative ratio results are highlighted on a spectrum from higherquality (green) to lower quality (red).

identified by Assemblytics than other pepper assemblies. While there are undoubtedly some genuine biological differences between the four pepper varieties, it is possible that these events would lead to the decreased quality of utilization of the developed assembly as a reference for future studies. In order to assess this, we utilized whole-genome paired-end sequences available for three RILs in the NCBI SRA database to compare the functionality of the developed reference compared to the three available references for utility as a reference for mapping sequences for QTL and other studies. The trimmed and QC'ed sequences were aligned to each of the four genomes, and the results of the number and percentage of mapped reads, number of reads mapped in pairs, and the number of non-perfect matches within the alignments are shown in Table 2. Overall, the UCD10X has been found to map a higher percentage of the reads from the individuals and a larger proportion of these reads is found to be mapped in correct pairs. Additionally, the alignments were found to comparatively have a low relative number of non-perfect matches. These findings show an increased benefit in using the UCD10X in the ability to map a higher % of resequencing reads correctly for sequencing-based studies.

Accuracy of haplotype reconstruction

All available full-length *PUN1* gene sequences from NCBI were obtained and aligned to the UCD10X assembly to determine the position of the gene in the assembly. The gene was found to be located on chromosome 2 over positions 135,884,368–135,885,734. This region and the corresponding region in the alternative haplotype were aligned along with the gene sequences obtained from NCBI using MUSCLE software[35]. Sequence alignments showed two distinct sequence types for pungent and nonpungent lines, highlighting the importance of a phased diploid genome assembly (Fig. 7; Supplemental File 5). Specifically, the haplotype sequence from contig 3924 in UCD10X clustered with the genes derived from nonpungent *C. annuum* lines, while the corresponding haplotype sequence from contig 3922 clustered with genes derived from the pungent lines sequenced individually, indicating the complete reconstruction of the 2.57-Kb hemizygous deletion of the *PUN1* gene haplotypes in the sequenced individual (Fig. 8).

Discussion

Our newly assembled pepper genome sequence assembly, UCD10X, was found to have the highest contiguity of any published pepper genome despite being produced with an F_1 individual. This test case has shown that it is possible with the 10x Chromium Linked-Read technology to accurately assemble and recover both parental haplotype sequences while sequencing a single individual, as demonstrated in the *PUN1* region and by nearly complete recovery of genes in each of the two pseudohaps in the BUSCO2 analysis. Recovery of both haplotypes with 10x Chromium Linked-Reads is a powerful advance over standard short-read sequencing, in that

short reads inherently provide low power for discovery and *de novo* reconstruction of genomes, especially in large heterozygous samples. Furthermore, these benefits are generated at a very reasonable cost of data acquisition (~ $6 K for 3.5-Gb genome), and at a fraction of the costs for traditional short-read strategies and long-read sequencing. This opens the exciting possibility to make Linked-Read sequencing and *de novo* assembly a routine operation, on par with regular sequencing efforts, but with much greater power to detect structural variations and haplotype differences.

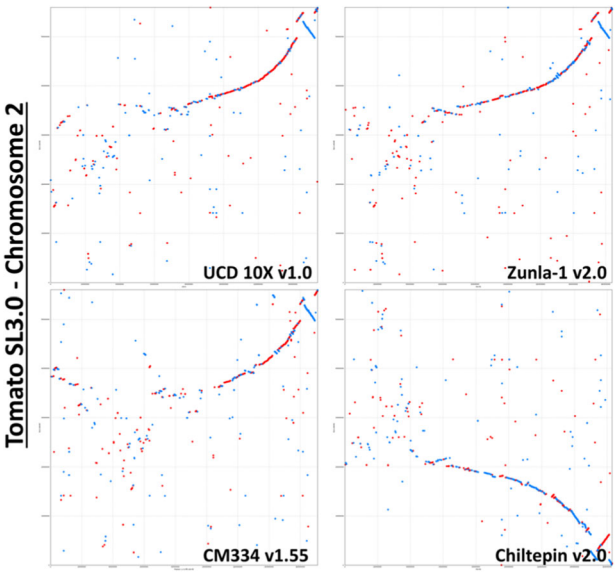

Fig. 7 Comparison of 10x assembly and other pepper assemblies against Tomato. Filtered nucmer alignments that extracted pepper Chromosome 2s and *Solanum lycopersicum* (Tomato version SL3.0) are shown for alignment lengths of greater than 150 base pairs at greater than 85% sequence identity. Pairwise alignments with tomato are shown with UCD10X version 1.0, CM334 version 1.55, Zunla version 2.0, and Chiltepin version 2.0 on the *X* axis and Tomato on the *Y* axis

Although the current assembly is highly contiguous and has resolved the haplotypes over much of the genome, some challenges remain in the sequencing and assembly of complex plant genomes, and some regions may not be able to be accurately assembled with Linked-Reads. Haplotype phasing is still limited to the distance of individual phase blocks within scaffolds, these lengths are primarily determined based on the lengths spanning a heterozygous element such as a SNP and structural variants, so phase block lengths will differ between species as well as individuals. Large portions of complex plant genomes are comprised of repetitive elements. Some of the longest repeats may span distances longer than the individual molecule lengths, which will still cause breaks in contigs across these regions. Consequently, the success with 10x Chromium Linked-Reads or alternative long-read sequencing systems will be dependent on the distribution and length of the repetitive elements in the genome. The repetitive content in pepper constitutes >80% of the genome and was derived by a rapid expansion of retrotransposon elements (70% of the genome), mainly of a single Gypsy element family, after divergence from the other Solanaceae members[11,12]. While these repetitive elements may have caused fragmentation of the genome, the overall product was very contiguous with over 50% of the genome in the *de novo* assembled scaffolds larger than 3.69 Mb. This allowed for most (83%) of the total assembly length to be placed into pseudomolecules using the four available *Capsicum* linkage maps. Only very small scaffolds could not be confidently placed

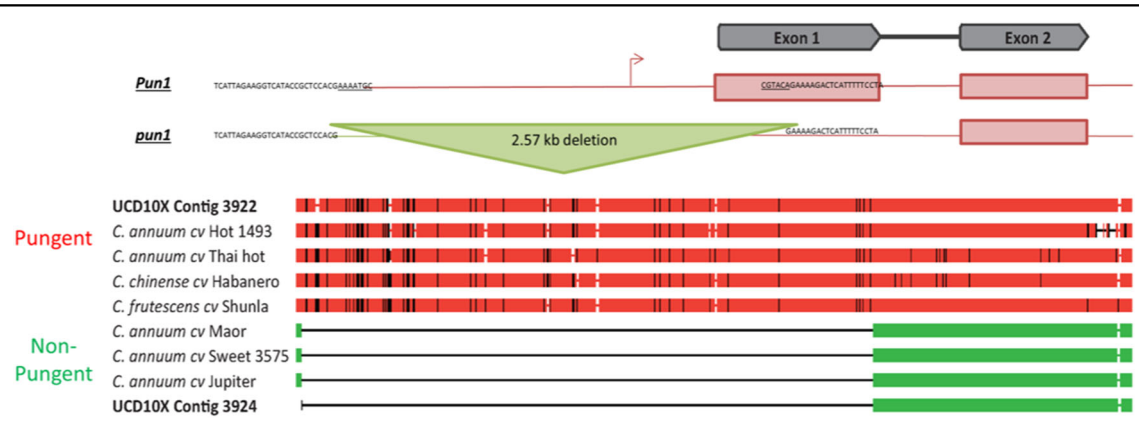

Fig. 8 Analysis of *PUN1* gene assembly sequence. Structure of *PUN1* locus and the corresponding haplotypes of seven *PUN1* gene sequences obtained from NCBI and the loci extracted from both UCD10X assembly haplotypes. Full size of deletion in the alignments is 2574 base pairs. Sequence alignments show high sequence similarity and separation of pungent and nonpungent groups

in the final pseudomolecules as expected (Supplemental Figure 24).

The chromosome sizes determined for UCD10X through use of the four linkage maps are also comparable to the other published genome assemblies (Supplemental Figure 25). The notable differences are with Chromosomes 1 and 8, which are known to have a translocation between *C. annuum* and *C. frutescens*. This is important as these species were used to generate populations with a high-polymorphism rate for genetic mapping[13,14]. The breakpoint in the two interspecific maps in this particular case was pinpointed with manual hand annotation by Hill et al.[13] to ensure that genetic map data were correctly associated with the corresponding *C. annuum* chromosome. It can be seen that this caused UCD10X Chromosome 1 to be slightly smaller than the other three assemblies, while UCD10X Chromosome 8 is larger than the other assemblies and closer in size to the other *C. annuum* chromosomes which would be expected based on the pepper karyotype where most chromosomes are similar in size[36].

Linked-Read genomic library technology paired with short-read sequencing has made it possible to generate long contiguous scaffolds for pepper and moderately sized contigs that were previously not possible through short-read sequencing and at considerably less cost than would be needed for long-read sequencing. This experiment has shown that it is possible to sequence large complex plant genomes such as pepper using the 10x Chromium technology and customized, open-source Supernova assembler. These tools will make it possible to broaden the scope of high-quality draft assemblies in an economically feasible manner for crops which are limited in funding. It also makes it possible to sequence large collections of individual genomes to very high quality, something not tractable with more expensive long-read sequencing. A similar strategy will make molecular-breeding tools more accessible for more crops and advancements to be made at a quicker pace to assist in providing nutritious food for a growing world population.

Conclusions

A highly contiguous assembly for a heterozygous complex genome of pepper has been generated in an economically viable manner through Linked-Read sequencing technology that combines the cost efficiency of short-read sequencing with barcoded sequencing libraries that retain long-range physical information. Importantly, the technology allowed for contiguity across long repeats and pericentromeric regions. We showed that large heterozygous (hemizygous) structural variants can be defined in a single *de novo* assembly, which provides an opportunity to cost-effectively compare structural and gene differences among *de novo* sequence assemblies among genotypes rather than simply mapping reads to a reference genome. This technology greatly enhances researchers abilities to attain new affordable resources for plant breeding; functional analyses of genes and genomic elements; and to improve our understanding of genome evolution across complex organisms.

Availability of data and material

The data sets generated and analyzed during the current study are available in the NCBI database under BioProject ID PRJNA376668 or through the SOL Genomics Network (https://solgenomics.net). All other data generated or analyzed during this study are included in this published article and its supplementary information files.

Acknowledgements

We would like to thank the University of California, Davis Genome Center for maintenance and support of the computational resources utilized for this project. This work was supported by research grants provided by Enza Zaden and Rijk Zwaan. Library preparation and sequencing costs were provided by 10x Genomics. Additional support was from the UC Davis Seed Biotechnology Center and USDA-ARS Project #6066-21310-004-00-D.

Authors' contributions

A.V.D. designed and conceived the project. A.M.H.K., S.M., T.A.H., and D.M.C. designed and performed the analyses. K.S., D.J., S.W., N.W., S.R., V.K., P.S., S.R., and M.C.S. performed the analyses. A.M.H.K. wrote the manuscript. All authors read, edited, and approved the manuscript.

Author details

[1]Department of Plant Sciences, University of California, Davis, CA, USA. [2]USDA-ARS Genomics and Bioinformatics Research Unit, Raleigh, NC, USA. [3]Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, USA. [4]10x Genomics, Inc, 7068 Koll Center Parkway, Suite 401, Pleasanton, CA, USA. [5]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

Competing interests

D.J., S.W., N.W., V.K., P.S., and D.M.C. have competing commercial interests as employees and stockholders of 10x Genomics, which is a commercial company that provides the Linked-Read technology and analysis software. This does not alter the authors' adherence to all of the Genome Biology policies on sharing data and materials. The remaining authors declare that they have no competing interests.

Supplementary information

The online version of this article (https://doi.org/10.1038/s41438-017-0011-0) contains supplementary material.

Received: 5 July 2017 Revised: 13 November 2017 Accepted: 16 November 2017
Published online: 12 January 2018

References

1. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species *J. Hered.* **100**, 659–674 (2009).
2. i5K Consortium. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Hered.* **104**, 595–600 (2013).

3. Zhang, G., Rahbek, C., Graves, G. R., Lei, F., Jarvis, E. D. & Gilbert, M. T. Genomics: bird sequencing project takes off. *Nature* **522**, 34 (2015).
4. Kyrpides, N. C., Hugenholtz, P., Eisen, J. A., Woyke, T., Goker, M. & Parker, C. T. et al. Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biol.* **12**, e1001920 (2014).
5. McCouch, S., Baute, G. J., Bradeen, J., Bramel, P., Bretting, P. K. & Buckler, E. et al. Agriculture: feeding the future. *Nature* **499**, 23–24 (2013).
6. Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R. & Kjaer, T. R. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
7. Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E. & Lee, S. H. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114 (2015).
8. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
9. Bevan, M. W., Uauy, C., Wulff, B. B., Zhou, J., Krasileva, K. & Clark, M. D. Genomic innovation for crop improvement. *Nature* **543**, 346–354 (2017).
10. Stewart, C. Jr., Kang, B. C., Liu, K., Mazourek, M., Moore, S. L. & Yoo, E. Y. et al. The Pun1 gene for pungency in pepper encodes a putative acyltransferase. *Plant J.* **42**, 675–688 (2005).
11. Kim, S., Park, M., Yeom, S. I., Kim, Y. M., Lee, J. M. & Lee, H. A. et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nat. Genet.* **46**, 270–278 (2014).
12. Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L. & Min, J. et al. Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl Acad. Sci. USA* **111**, 5135–5140 (2014).
13. Hill, T., Ashrafi, H., Chin-Wo, S. R., Stoffel, K., Truco, M. J. & Kozik, A. et al. Ultra-high density, transcript-based genetic maps of pepper define recombination in the genome and synteny among related species. *G3* **5**, 2341–2355 (2015).
14. Hulse-Kemp, A. M., Ashrafi, H., Plieske, J., Lemm, J., Stoffel, K. & Hill, T. et al. A HapMap leads to a Capsicum annuum SNP infinium array: a new tool for pepper breeding. *Hortic. Res.* **3**, 16036 (2016).
15. Han, K., Jeong, H. J., Yang, H. B., Kang, S. M., Kwon, J. K. & Kim, S. et al. An ultra-high-density bin map facilitates high-throughput QTL mapping of horticultural traits in pepper (Capsicum annuum). *DNA Res.* **23**, 81–91 (2016).
16. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
17. Stoffel, K., van Leeuwen, H., Kozik, A., Caldwell, D., Ashrafi, H. & Cui, X. et al. Development and application of a 6.5 million feature Affymetrix Genechip(R) for massively parallel discovery of single position polymorphisms in lettuce (Lactuca spp.). *BMC Genomics* **13**, 185 (2012).
18. Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M. & Hindson, C. M. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
19. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
20. Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R. & Schnable, J. C. et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3 (2015).
21. Wu, F., Eannetta, N. T., Xu, Y., Durrett, R., Mazourek, M. & Jahn, M. M. et al. A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus *Capsicum*. *Theor. Appl. Genet.* **118**, 1279–1293 (2009).
22. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
23. Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M. & Antonescu, C. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
24. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
25. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
26. Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T. & Clum, A. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
27. Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K. & Isobe, S. et al. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
28. Xu, X., Pan, S. K., Cheng, S. F., Zhang, B., Mu, D. S. & Ni, P. X. et al. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–U194 (2011).
29. Sierro, N., Battey, J. N., Ouadi, S., Bakaher, N., Bovet, L. & Willig, A. et al. The tobacco genome sequence and its comparison with those of tomato and potato. *Nat. Commun.* **5**, 3833 (2014).
30. Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., Negoro, S. & Ohyama, A. et al. Draft genome sequence of eggplant (Solanum melongena L.): the representative solanum species indigenous to the old world. *DNA Res.* **21**, 649–660 (2014).
31. Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L. & Barry, C. S. et al. Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida. *Nat. Plants* **2**, 16074 (2016).
32. Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P. & Huang, J. et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* **48**, 657–666 (2016).
33. Pickersgill, B. A. Cytogenetics and evolution of Capsicum L. Chromosome engineering in plants: genetics, breeding, evolution, part B. Elsevier, Amsterdam. 1991, pp 139–160.
34. Rinaldi, R., Van Deynze, A., Portis, E., Rotino, G. L., Toppino, L. & Hill, T. et al. New insights on eggplant/tomato/pepper synteny and identification of eggplant and pepper orthologous QTL. *Front Plant Sci.* **7**, 1031 (2016).
35. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
36. Moscone, E. A., Lambrou, M., Hunziker, A. T. & Ehrorfer, F. Giemsa C-banded karyotypes in Capsicum (Solanaceae). *Plant Syst. Evol.* **186**, 213–229 (1993).