DNA Research, 2023, **30**, 1–14
https://doi.org/10.1093/dnares/dsac033
Advance access publication 8 October 2022
Resource Article: Genomes Explored

Resource Article: Genomes Explored

A contiguous *de novo* genome assembly of sugar beet EL10 (*Beta vulgaris* L.)

J. Mitchell McGrath[1,*], Andrew Funk[2], Paul Galewski[2], Shujun Ou[2], Belinda Townsend[3], Karen Davenport[4], Hajnalka Daligault[4], Shannon Johnson[4], Joyce Lee[5], Alex Hastie[5], Aude Darracq[6], Glenda Willems[6], Steve Barnes[6], Ivan Liachko[7], Shawn Sullivan[7], Sergey Koren[8], Adam Phillippy[8], Jie Wang[9], Tiffany Liu[9], Jane Pulman[9], Kevin Childs[9], Shengqiang Shu[10], Anastasia Yocum[11], Damian Fermin[11], Effie Mutasa-Göttgens[12], Piergiorgio Stevanato[13], Kazunori Taguchi[14], Rachel Naegele[1], and Kevin M. Dorn[15,*]

[1]USDA-ARS Sugarbeet and Bean Research Unit, Michigan State University, 1066 Bogue St., East Lansing, MI 48824, USA
[2]Plant Breeding, Genetics, and Biotechnology Program, Michigan State University, East Lansing, MI 48824, USA
[3]Department of Plant Sciences, Rothamsted Research, West Common, Harpenden, Hertfordshire AL5 2JQ, UK
[4]Los Alamos Nat'l Lab, Biosecurity and Public Health, Los Alamos, NM 87545, USA
[5]BioNano Genomics, 9640 Towne Centre Drive, San Diego, CA 92121, USA
[6]SESVANDERHAVE N.V., Industriepark Soldatenplein Zone 2 Nr 15, 3300 Tienen, Belgium
[7]Phase Genomics, 4000 Mason Road, Suite 225, Seattle, WA 98195, USA
[8]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, MD, USA
[9]Center for Genomics-Enabled Plant Science, Plant Biology Department, Michigan State University, East Lansing, MI 48824, USA
[10]United States Department of Energy, Joint Genome Institute, Berkeley, CA, USA
[11]A2IDEA, 674 S. Wagner Rd., Ann Arbor, MI 48103, USA
[12]University of Hertfordshire, Division of Biosciences, Hatfield, Hertfordshire AL10 9AB, UK
[13]DAFNAE, University of Padova, Viale Università 16, 35020 Legnaro (PD), Italy
[14]Hokkaido Agricultural Research Center, National Agriculture and Food Research Organization, Shinsei Memuro, Hokkaido 082-0081, Japan
[15]USDA-ARS Soil Management and Sugarbeet Research Unit, Crops Research Laboratory, 1701 Centre Ave, Fort Collins, CO 80526, USA

[*]Corresponding author: Email: mitchmcg@msu.edu (J.M.M.); kevin.dorn@usda.gov (K.M.D.)

Abstract

A contiguous assembly of the inbred 'EL10' sugar beet (*Beta vulgaris* ssp. *vulgaris*) genome was constructed using PacBio long-read sequencing, BioNano optical mapping, Hi-C scaffolding, and Illumina short-read error correction. The EL10.1 assembly was 540 Mb, of which 96.2% was contained in nine chromosome-sized pseudomolecules with lengths from 52 to 65 Mb, and 31 contigs with a median size of 282 kb that remained unassembled. Gene annotation incorporating RNA-seq data and curated sequences via the MAKER annotation pipeline generated 24,255 gene models. Results indicated that the EL10.1 genome assembly is a contiguous genome assembly highly congruent with the published sugar beet reference genome. Gross duplicate gene analyses of EL10.1 revealed little large-scale intra-genome duplication. Reduced gene copy number for well-annotated gene families relative to other core eudicots was observed, especially for transcription factors. Variation in genome size in *B. vulgaris* was investigated by flow cytometry among 50 individuals producing estimates from 633 to 875 Mb/1C. Read-depth mapping with short-read whole-genome sequences from other sugar beet germplasm suggested that relatively few regions of the sugar beet genome appeared associated with high-copy number variation.

Key words: Beta vulgaris, beet, genome assembly, genome annotation, comparative genomics

1. Introduction

Humans have used beet or chard (*Beta vulgaris* spp. *vulgaris* L.) as early as the late Mesolithic, initially as a leafy pot herb and for medicinal uses.[1] It was not until the Middle Ages that the enlarged taproot became used as a vegetable. The origin of the enlarged taproot is not clear, but by the 18th century beets were widely used as fodder and fuelled the prelude to the Industrial Revolution in Europe. Sugar beet was selected from lower sucrose fodder beets (6–8% sucrose fresh weight) in the late 1700s, with the first true sugar beet commercial varieties available by 1860.[2] Since then, improvements in sucrose content and processing quality have been continuous, resulting in an industry average in the USA and Europe approaching 19% sucrose fresh weight (~75% dry weight). Public sector sugar beet breeding today focuses generally on

Received 11 April 2022; Revised 26 August 2022; Accepted 12 September 2022

crop protection traits.[3,4] The EL10.1 genome summarized here was recently interrogated for resistance gene signatures[5] and crop-type attributes.[6] An alternate annotated assembly, EL10.2, is also available.

Beta vulgaris is a basal eudicot in the family Amaranthaceae (Caryophyllales).[7] Wild forms are native to European and Mediterranean coastlines and collectively classified as subspecies *maritima*.[1,8,9] There are no known barriers to cross-fertilization among beet crop and wild types, and the genomes of crop wild relatives are beginning to be described in detail.[10] Most *B. vulgaris* types, and all characterized *maritima* types, are diploid. Chromosomes are morphologically similar at mitotic metaphase, and highly repetitive DNA sequences comprise ~60% or more of the beet genome.[11,12] Each chromosome shows different patterns of repeat-sequence distribution[13,14] supporting the notion that sugar beet genomes are true diploids.[12,15] An ancient genome triplication appears to be shared with the basal asterid and rosid eudicot clades.[16] A uniform linkage group nomenclature was derived from Schondelmaier and Jung's[17] linkage group assignments and made more portable with SSR markers.[18] Extensive marker technologies remain proprietary within the commercial sugar beet breeding sector who supply hybrid seed to growers worldwide.

We seek to fill knowledge gaps in understanding of sugar beet traits by completing a genome framework for beet and incorporating crop genetic trait information into the framework, focussing on crop quality and preservation traits. Creating highly contiguous genome assemblies is challenging in plants due to the generally high-repetitive nature of large portions of their genomes. Genome annotation is perhaps more challenging as expressed gene functions are generally predicted from relatively few physiologically verified protein functions derived from unrelated plant taxa on the basis of nucleotide and amino acid sequence similarity. Improved approaches are becoming available and more commonly used to develop high-quality assemblies and annotations for plant species previously considered too complex or costly.[19] The EL10 genome assemblies described here include long-read length technologies to span longer low-complexity repeat regions, optical mapping to create larger scaffolds from long-read contig assemblies, and Hi-C to link together scaffolds across the genome into chromosome-sized scaffolds. Highly contiguous assemblies exhibit the full organization of hereditary material and minimize the uncertainty of position and distribution of genetic markers to allow closer focus on any region of the genome.

Scaffolds of the EL10 assemblies show high concordance with genetic maps and the RefBeet genome sequence,[12] which is an excellent but fragmented genome sequence assembled using first- and second-generation sequencing technologies. The EL10.1 genome has been used to anchor other assemblies including those used for beet curly top virus resistance identification.[20] The current work, presented here, provides a more comprehensive picture of genome size variability of sugar beet, and global changes in repeat-sequence depth and coverage between sugar beet inbreds and breeding populations. Further, gene number in beet appears to be diminished relative to other eudicots, at least for gene classifiers that are shared among representative angiosperm genomes. Contiguous genome assemblies will allow routine inter-cultivar comparisons between accessions that vary for important traits, and help deduce causal from associated genomic features influencing a trait of interest or general performance.

2. Materials and methods

2.1. Plant material

USDA-ARS germplasm release EL10 (PI 689015) was derived by single seed descent from C869 (PI 628755) by self-pollinating over six generations. C869 is a biennial sugar beet conditioned by the self-fertile (S^fS^f) allele and is segregating for nuclear male sterility (*Aa*), with resistance to several diseases.[21] The initial selfing occurred from one self-fertile C869 CMS plant (EL-A013483) in 2002. Seed was field grown at the Bean and Beet Farm (Saginaw, MI) in 2005, roots were harvested, potted into fibre pots (5 L, Stock # ITMLFNP08090RBRD040TW, BFG Supply, Burton, OH), vernalized for 16 weeks, and grown in the greenhouse until flowering. Flowers were inspected for visible pollen, and when present, a #16 white grocery bag (Duro Bag, Novolex, Hartsville, SC) was placed over the bolting stem to effect self-pollination. Seed harvested from a single plant (EL-A018880) was considered the S1 generation, and subsequent generations were derived by single seed descent using field grown mother roots and selfing with the same methods. The S2 generation (EL-A022144) was obtained in 2007, and the S3 (EL-A025943) in 2010. Nine individuals of this population were genotyped with 69 SESVanderhave proprietary SNP markers evenly spaced across the beet linkage map, and a single homozygous individual (#17) of this population was sequenced for a preliminary assembly (named C869_UK[22]). A sibling of this line (EL-A026195) with good field performance in the 2011 Michigan field (Saginaw Valley Research and Extension Center, SVREC, Richville, MI) was selfed in the same manner to yield the S4, while S5 (EL-A13-03870) and S6 generations were produced solely under greenhouse conditions in 2013 and 2015, respectively. Sixteen S6 individuals were genotyped with 24 SSR markers,[18] and 6 individuals (EL-A15-01096, EL-A15-01098, EL-A15-01099, EL-A15-01101, EL-A15-01102, and EL-A15-01103) were chosen as sequencing candidates based on marker homozygosity and similar growth habit and appearance, and pooled for long-read sequencing. One of these (EL-A15-01101) provided the sole tissue source for Illumina sequencing and nuclear DNA content estimation, and seed was named and released as EL10. Seed of EL10 was increased and deposited in the National Plant Germplasm System repository as a genetic stock (PI 689015).

Additional taxa were used, depending on the availability of materials, for the assessment of genome size, cytometric estimates. Material included progeny of EL-A15-01101 whose genome was assembled here, advanced progeny of table beet W357B (a self-fertile parental line graciously provide by Dr. Irwin Goldman) which were inbred by single seed descent for five generations (accession EL-A1400766), an East Lansing open-pollinated self-sterile sugar beet breeding population (termed '5B'), and an open-pollinated USDA-ARS release used for a disease nursery check entry (F1042, PI 674103).

2.2. Genome sequencing, assembly, and finishing

High-molecular-weight DNA for PacBio sequencing isolated nuclei using the HMW preparation protocols suitable for BAC library construction by Amplicon Express (Pullman, WA). PacBio RSII sequencing was performed at the Los Alamos National Laboratory (Los Alamos, NM), in 86 single-molecule, real-time cells using P6-C4 chemistry. PacBio reads greater than 6 kb were assembled with the Falcon Assembler

(version 0.2.2), resulting in 938 primary contigs (SBJ_80X assembly, Supplementary Table S1). Optical mapping was performed using the BioNano Irys sequential hybrid protocol with enzymes *Bss*SI and *Bsp*QI (SBJ_80X_BN assembly, Supplementary Table S1). For the EL10.1 assembly, scaffolding was accomplished using Proximity Guided Assembly and Hi-C reads by Phase Genomics (Seattle, WA). Resulting scaffolds were polished and gap-filled using PBJelly, Arrow, and Pilon, following Bickhart *et al.*[23] Briefly, PBJelly from PBSuite v15.8.24 was run using the Protocol.xml (https://gembox.cbcb.umd.edu/shared/Protocol.xml) with default parameters and minimum gap size set to 3 as: Jelly.py setup Protocol.xml --minGap=3, Jelly.py mapping Protocol.xml, Jelly.py support Protocol.xml, Jelly.py extraction Protocol.xml, Jelly.py assembly Protocol.xml, and finally Jelly.py output Protocol.xml. Pilon v1.13 was run using --fix local bases and the pipeline at https://github.com/skoren/PilonGrid. Arrow v2.0.0 was run using the pipeline available at https://github.com/skoren/ArrowGrid. Pilon v1.21 was run using --fix indels using the pipeline at https://github.com/skoren/PilonGrid.

To generate the EL10.2 assembly, the SBJ_80X_BN assembly was further scaffolded using 462 million DoveTail Genomics Hi-C read pairs (Chicago and Dovetail Hi-C technologies) using the HiRise algorithm as described[24] (Putman *et al.*, 2015). The resulting scaffold-level assembly was subsequently polished with POLCA,[25] utilizing high depth Illumina paired-end reads (Table 1). Assembly metrics for RefBeet 1.2, EL10.1, and EL10.2 were assessed using the stats.sh tool from the BBTools software package (https://jgi.doe.gov/data-and-tools/software-tools/bbtools/) using the default parameters.

2.3. Whole-genome alignment

Whole-genome alignment of the EL10.1 assembly (as reference) and the RefBeet 1.2 assembly (as query) was conducted using modules from MUMmer v.4.0.0beta2. Initial alignments were created with the nucmer module, with options --mum --minmatch 30 (uses only anchor matches that are unique in both the reference and the query, and sets the minimum length of a single exact match to 30 bp). The resulting delta alignment was filtered using the delta-filter module with options -1 -i 70 -l 5000 (to use only 1-to-1 alignments, with a minimum 70% sequence identity, and minimum alignment length of 5,000 bp). Summary reports were created using dnadiff, and plots were created from the filtered delta file using mummerplot with options --png --fat -r (with output image as png, and using layout sequences using fattest alignment only).

2.4. Annotation

The EL10.1 assembly was annotated using the MAKER pipeline.[26] A custom repeat library for EL10 was created and used for repeat masking.[27] Protein and transcript evidence were used to aid gene model prediction. Protein evidence was obtained from the following species or databases: *Arabidopsis thaliana* proteins from Araport11,[28] *Solanum lycopersicum* proteins from IPTG 2.4,[29] *Populus trichocarpa* proteins from Phytozome genome v3.0,[30] and curated plant proteins from UniProt release 2017_03.[31] Transcript evidence was derived from 25 RNA-seq read sets (BioProject PRJNA450098, Illumina 2500, 150 bp paired-end) using StringTie v1.3.3b[32] and TransDecoder v5.0.1 (Haas and Papanicolaou *et al.*, manuscript in prep., http://transdecoder.github.io).

Gene prediction programmes AUGUSTUS[33] and SNAP[34] were trained using the transcript sequences generated by StringTie (above), and both AUGUSTUS and SNAP were used to predicted gene models within the MAKER pipeline.[26] When AUGUSTUS and SNAP predicted genes at the same locus, MAKER chose the gene model that was the most concordant with the transcript and protein evidence, and that model was retained at that locus. HMMER v 3.1[35] was used to determine the presence of Pfam-A protein domains in the initial predicted protein sequences. Gene models supported either by protein or transcript evidence or by the presence of a Pfam domain were collected as high-quality gene models for the final genome annotation. Both transcript and protein sequences were searched against the SwissProt and UniRef databases using BLAST.[36] HMMER v3.1[35] identified PfamA

Table 1. Sequence inputs and metrics used in construction of EL10.1

Technology	Library		Coverage[a]
		PacBio passed reads	
PacBio long reads	RS II, P6-C4 chemistry	6,540,795	79.3
	Mean length = 9,096 nt (std. dev. = 6,528)		
	>40 kb initial mapping and pre-assembly	5,176	0.38
		BioNano passed labels	
Optical physical map	BioNano Genomics *Bss*SI–*Bsp*PQ1 Hybrid Scaffold	121 Gb	161.3
	*Bsp*PQ1 (7.6 labels/100 kb)	40 Gb	
	*Bss*SI (10 labels/100 kb)	81 Gb	
		Illumina passed reads	
Paired-end short reads Cross-linked *in vivo*	HiSeq 2500, TruSeq Libraries, 125 bp PE	447,211,041	149.0
	Phase Genomics Hi-C library, HiSeq 2500, TruSeq Libraries (EL10.1)	355,892,798	118.6
	Dovetail Genomics Hi-C library, HiSeq 10X, TruSeq Libraries (EL10.2)	927,545,984	183.3

[a]Using genome size of 758 Mb.

domains within predicted protein sequences. Signal peptide and transmembrane domains were predicted using SignalP v4.1[37] and TMHMM v2.0,[38] respectively. Searches and predicted results were parsed and combined in the final functional annotation.

The online sequence functional classification and annotation tool Mercator4 ver. 2.0[39] were supplied with the EL10.1 MAKER predicted protein fasta file using default settings. Four gene models were excluded from analysis due to their short length (<5 amino acids) (e.g. EL10Ac2g04429.1, EL10Ac8g20093.1, EL10Ac1g00658.1, and EL10Ac7g16947.1). Comparisons were made with Mercator4-supplied representatives of the Tracheophyta (i.e. *Oryza sativa*, *Brachypodium distachyon*, *A. thaliana*, *S. lycopersicum*, and *Manihot esculenta*). The EL10.2 assembly was later annotated by JGI (annotation version EL10.2_2) and is available on Phytozome.

2.5. LTR annotation

De novo identification of intact long terminal repeats (LTR) retrotransposons were performed using LTR_Retriever v1.6 with default parameters.[40] The insertion time of each intact LTR-RT is estimated by LTR_retriever based on $T = K/2\mu$ where K is the divergence between an LTR pair and μ is the mutation rate of 1.3×10^{-8} per bp per year. Whole-genome LTR sequence annotations were achieved using the non-redundant LTR library generated by LTR_Retriever and RepeatMasker v4.0.0 (www.repeatmasker.org).

2.6. LTR Assembly Index estimation

The assembly continuity of repeat space was assessed using the LTR Assembly Index (LAI) deployed in the LTR_retriever package (v1.6).[40] LAI was calculated based on either 3 Mb sliding windows or the whole assembly using raw_LAI = (Intact LTR-RT length × 100)/Total LTR-RT length. For the sliding window estimation, a step of 300 kb was used (-step 300000 -window 3000000). The estimation of LAI was adjusted using the mean identity of LTR sequences in the genome based on all-versus-all BLAST.

2.7. Tandem repeats

Tandem Repeats Finder Program Version 4.09 was used to characterized tandemly duplicated sequences. using the default Alignment Parameters (e.g. match = 2, mismatch = 7, indels = 7, PM = 80, PI = 10, minimum alignment score = 50, maximum period size = 500).[41]

2.8. Self-synteny

CoGe SynMap[42] was used, inputting *B. vulgaris* (vEL10.1.0, id37197) and EL10.1 MAKER annotation gff files. Coding sequences were compared using LAST[43] and DAGChainer[44] (with input settings maximum distance between two matches = 20 genes, minimum number of aligned genes = 5). Kn/Ks ratios[45] were calculated using default parameters on CoGe (genomevolution.org/wiki/index.php/SynMap).

2.9. Genome size variation

Four *B. vulgaris* populations were evaluated for nuclear DNA content as described.[46] Briefly, young and healthy true leaf tissues from greenhouse grown seedlings were placed in between moist paper towels in zip-lock bags and shipped to the Flow Cytometry Lab at Benaroya Research Institute at Virginia Mason (Seattle, WA) for next day delivery. 50 mg of leaf tissue from each sample was finely chopped using a razor edge to release intact nuclei for flow cytometric analysis. Chicken erythrocyte nuclei (2.50 pg/2C) were used as an internal standard. A value of 978 Mb per pg was used for genome size conversion.[47] Statistical analyses were performed with JMP Pro version 14 (SAS, Cary, NC).

2.10. Read count mapping

Reads from five Illumina paired-end sequencing datasets were trimmed and subsampled to produce sets of 25 Gb for normalized mapping to the EL10.1 assembly. These were the single sequenced EL10 plant, a single plant two generations less inbred than EL10 (i.e. C869_UK), a pool of 25 individual from the parental population from which EL10 was derived (C869_25), the doubled haploid from which RefBeet was generated (KWS2320), and a single plant of a Japanese O-type breeding line (NK-388mm-O) (each accessible at NCBI BioProject PRJNA563463). Four samples of KWS2320 genomic reads (SRR869628, SRR869631, SRR869632, and SRR869633) were obtained from the NCBI SRA and pooled prior to filtering. FASTQ reads from the 5 mapping samples were filtered for a minimum FASTQ quality of 6 and minimum length of 80 bp after trimming. The reads that passed the filter were randomly subsampled to obtain 25 Gb of reads per sample. Each pool of 25 Gb was independently mapped to the EL10 assembly using BBMap v. 36.67.[48] Read mapping was done with default parameters and kmer length = 13 with the addition of 'local=t' to allow soft-clipping the ends of alignments and 'ambiguous=random' to randomly assign reads with multiple best matches among all best sites, to facilitate mapping of repetitive sequences evenly across the genome. For plotting read depth, 5 kb bins were created across each chromosome and the read coverage per base pair was calculated for each bin. The 'basecov' and 'covstats' outputs of BBMap were used to determine read depths and their standard deviations.

2.11. Multispecies synteny

The analysis of synteny was accomplished by plotting collinear blocks relative to beet chromosomes. Collinear blocks were defined using the program MCScanX using default recommendations.[49] Protein sets for *A. thaliana*, *V. vinifera*, *Spinacea oleraceae*, and *A. hypocondriacus* were downloaded from phytozome (https://phytozome.jgi.-doe.gov/pz/portal.html) with their corresponding gff files. Quinoa data were downloaded from chenopodiumdb (www.cbrc.kaust.edu.sa/chenopodiumdb/) and the *B. vulgaris* proteins and gff files were developed for this report.

2.12. Accession numbers

Sequence data from this article can be found in the EMBL/GenBank data libraries. The EL10 sugar beet whole-genome project has been deposited in NCBI under the accession PCNB00000000. EL10.1 is version PCNB01000000. Associated NCBI database pointers are BioSample SAMN07736104, BioProject PRJNA413079; Assembly GCA_002917755.1, and WGS Project PCNB01. All raw reads used in EL10 genome assemblies are deposited in the short-read archive (SRA): Illumina reads SRR6305245; PacBio Reads SRR6301225; and Hi-C Library reads SRR10011257 (Phase Genomics) and SRR12507442 and SRR12507443 (Dovetail Genomics). BioNano Maps are located at SAMN08939661 (*Bsp*Q1) and SAMN08939667 (*Bss*S1).

Table 2. Assembly metrics for sugar beet genome versions

Assembly name	Assembly input and method	# Contigs	Assembly size (contigs)	Contig N/L50 (# fragments/length)	# Scaffolds	Assembly size (scaffolds)	Scaffold N/L50 (# fragments/length (Mb)	%N (Scaffold assembly)	% assembly in scaffolds >50 kb	Pseudochromosome assembly size (n = 9 scaffolds)
RefBeet 1.2	RefBeet 1.2[12]	61,805	517,837,822	3,863/39.1 kb	40,508	566,571,340	72/2.013	8.60	91.94	NA
EL10.1	PacBio + BioNano + Phase Hi-C + polishing	363	540,479,261	64/2.701 Mb	40	540,537,112	5/57.939 Mb	0.01	100	520,115,771
EL10.2	PacBio + BioNano + DoveTail Hi-C + polishing	3,098	534,762,237	119/1.287 Mb	18	568,751,015	5/61.987 Mb	5.99	100	564,173,179

Read mapping accessions are deposited under BioProject PRJNA563463, and BioSamples SAMN12674955 (C869_UK), SAMN12674956 (C869_25), and SAMN12674957 (NK-388mm-O). The EL10.1 genome assemblies and annotations can be viewed and downloaded via the CoGe Genome Browser available at genomevolution.org/coge/, both EL10.1 (Genome ID = 54615) and Phytozome (phytozome-next.jgi.doe.gov/info/Bvulgaris_EL10_1_0). The EL10.2 assembly and JGI annotation (EL10.2_2) is available on Phytozome under genome ID 782 (https://phytozome-next.jgi.doe.gov/info/Bvulgarisssp_vulgaris_EL10_2_2).

Genome browsing and file resources including transcript assemblies are available at http://sugarbeets.msu.edu and beetbase.scinet.usda.gov. Transcript assemblies were constructed from root development and leaf RNA-seq reads derived from C869 (the EL10 progenitor) from 3 to 10 weeks post-emergence[50] [3-week-old root (SRR10039097), 4-week-old root (SRR10039086), 5-week-old root (SRR10039081), 6-week-old root (SRR10039080), 7-week-old root (SRR10039079), 10-week-old root (SRR10039098), and mature leaf (SRR10037935)]. Also included were RNA-seq sets of 96-h germinated seedlings from other germplasm germinated under aqueous stress conditions,[51] including 150 mM NaCl, 0.3% hydrogen peroxide, and biologically extreme temperatures (10 and 41°C) (SRR10039075, SRR10039076, SRR10039077,SRR10039078,SRR10039082,SRR10039083, SRR10039084,SRR10039085,SRR10039087,SRR10039088, SRR10039089,SRR10039090,SRR10039091,SRR10039092, SRR10039093, SRR10039094, SRR10039095, and SRR10039096). The transcript assemblies are located at http://sugarbeets.msu.edu/data.html.

3. Results

A five-generation inbred genome of the sugar beet 'C869' (PI 628755) was released as a genetic stock 'EL10' (PI 689015). C869 is the common seed parent for East Lansing recombinant inbred populations previously described.[52] To purify sufficient high-molecular-weight DNA, five plants from one inbred family showing no gross phenotypic differences and no polymorphism among 24 selected unlinked SSR markers[18] were chosen for nuclei isolation, long-read sequencing, and assembly. The resulting assembly, using only one of the five plants, was scaffolded via opto-physical mapping, and the two assemblies described here share this common backbone. Two different chromatin conformation-sequencing technologies were used to construct two independent assemblies (EL10.1 vs. EL10.2), with the goal of evaluating the effect of the two technologies and Hi-C library sequencing depth on contiguity. Holistically, the two Hi-C technologies and scaffolding pipelines greatly reduced the number of scaffolds to the haploid chromosome number in beet (n = 9), with assembly EL10.2 slightly improved in contiguity over assembly EL10.1 (Table 2). In-depth analyses, including annotation, repeat content, genetic map co-linearity focussed on the EL10.1 assembly below due to EL10.1's substantially better contig-level statistics and percent ambiguous bases, with only slightly worse scaffold-level statistics. As EL10.1 has been used in at least two publications,[5,6] it is therefore important to document both genomes. Insights described below pertain to EL10.1 unless otherwise noted.

3.1. Sequencing and assembly

High-molecular-weight DNA was isolated from intact gel-embedded nuclei of true leaves from young seedlings and pooled from the five inbred plants for long-read sequencing using standard protocols for BAC library construction (Amplicon Express, Pullman, WA). Eighty-six PacBio SMRT cells yielded 79.3-fold coverage (58,655 Mb) of the *circa* 750 Mb *B. vulgaris* genome size (see below). The Falcon Assembler (version 0.2.2) was used to assemble long reads (Table 1), initialized with reads exceeding 40 kb in length. The Falcon assembly resulted in 938 primary contigs, 70.9% with a length greater than 100,000 nucleotides and a total length of 562.76 Mb. G + C content was similar between EL10.1 and RefBeet contigs (35.8% vs. 36.1%, respectively).

Scaffolding the Falcon assembly with a BioNano two-enzyme (*Bsp*QI and *Bss*SI) sequential hybrid optical (physical) map resulted in substantial improvement. The *Bsp*QI optical map was generated from 141,462 molecules with an average length of 285 kb and labelled to an average density of 11.8 sites kb^{-1}, and the *Bss*SI optical map was generated from 270,071 molecules, also with an average length of 285 kb, labelled to a density of 7.7 sites kb^{-1}. Optical maps were aligned to PacBio Falcon contigs and the resulting *Bsp*QI and *Bss*SI map lengths were 628 and 590 Mb with N50 contig sizes of 1.99 and 1.21 Mb, respectively. After merging PacBio, *Bsp*QI, and *Bss*SI contigs, the final hybrid genome map consisted of 86 scaffolds with a total length of 566.8 Mb, and an N50 of 12.5 Mb (SBJ_80X_BN, Supplementary Table S1).

Long range scaffolding was carried out using Proximity Guided Assembly with 118× coverage of Phase Genomics Hi-C reads. The Phase Genomics Hi-C reads were used to scaffold the SBJ_80X_BN PacBio/BioNano using self-pollinated progeny of the individual that was optically mapped. The resulting Hi-C assembly was subsequently polished and gap-filled using a combination of approaches (PBJelly, Arrow, and Pilon; following Bickhart *et al.*[23]). The resulting 540.5 Mb assembly consisted of 9 chromosome-sized scaffolds, numbered via Butterfass chromosome nomenclature,[53] and 31 unscaffolded contigs. These comprise the genome assembly version EL10.1. The 9 chromosome-sized scaffolds (designated Chromosomes below) were relatively similar in size (mean = 57.8 Mb, std. dev. = 3.9 Mb) (Table 2). Unanchored-contigs (n = 31) represented 3.9% of the final EL10.1.

With the goal of generating the most contiguous sugar beet genome possible from EL10 background, a second assembly (EL10.2) was created using an alternative Hi-C library technology from DoveTail Genomics with higher sequencing coverage (183×). The SBJ_80X_BN PacBio-BioNano assembly was scaffolded with these DoveTail Genomics Hi-C reads using the Dovetail Genomics HiRise assembly pipeline. Chromosomes were numbered according to the nomenclature by.[17] The EL10.2 pseudo-chromosomes (9 largest scaffolds) encompass 564 Mb, just shy of the total assembly size of the 40,508 RefBeet 1.2 scaffolds (566 Mb). The EL10.2 assembly appeared to resolve the major assembly associated inversions on Chromosomes 7 and 9 in EL10.1 (see below), as well as placed the unlinked 31 scaffolds into the larger whole-genome chromosome context (Supplementary Fig. S1), many of which appeared to be placed within the context of Chromosome 5 in EL10.2.

3.2. Assessment

No complete chloroplast or mitochondrial genomes were incorporated into the EL10.1 assembly, although fragments of both plastid genomes were detected in the EL10.1 assembly. The position of RefBeet 1.2 scaffolds were determined for EL10.1 Chromosomes (Fig. 1). Contigs >5 kb in length were largely co-linear between the two assemblies. Two small inverted-orientation contigs were evident on Chromosome 7, as were small inverted (e.g. Chromosome 6) and misplaced segments (e.g. Chromosomes 3 and 7). RefBeet 1.2 was anchored with genetic markers,[12] and 345 of these with 100% match identity across 75 nt or greater were placed in concordant order on the EL10.1 assembly. In addition, 3,279 proprietary SNP markers from the SESVanderhave (Tienen, Belgium) molecular marker genetic map were placed to the EL10.1 assembly. Most marker orders were highly concordant. However, a third of the mapped markers were inverted on Chromosome 9, and a complex rearrangement involving 40% of markers was evident on Chromosome 7 (mapped inversions; Supplementary Table S2). Genetic markers also added nine scaffolds to five chromosomes (mapped integrations; Supplementary Table S3). Genetic markers used to orient the cytogenetic map[14] also aligned with the EL10.1 assembly. Chromosomes 1 and 3 were cytogenetically congruent with their North–South orientation, and the rest were reversed relative to the orientations given in that publication. Scaffold 5 was located to the South end of cytogenetic Chromosome 5 (Supplementary Table S4), consistent with SESVanderhave marker data (Supplementary Table S3).

3.3. Annotation

The EL10.1 assembly contained the entire first linkage group described in beet,[54] the *R–Y–B* linkage group on Chromosome 2. Each of these genes has been recently cloned (*R*, for the red alkaloid betalain synthesis by a novel cytochrome P450,[55] *Y*, a Myb transcription factor required for production of red colour,[56] and *B* for the bolting gene which determines annual or biennial life habit.[57,58] Both the direction and the distance agree with published genetic map intervals, and the EL10.1 assembly indicates that the bolting gene is physically located proximal towards the centromere and the colour genes are more distal (Supplementary Table S5).

Results from the MAKER annotation pipeline[26] conservatively predicted 24,255 protein-coding gene models, numerically 88.5% of the 27,421 predicted in RefBeet.[12] To annotate the MAKER-derived gene models with predicted function descriptions, three sources of evidence were used, in the priority: (i) UniProt, (ii) Pfam-A, and (iii) Uniref90. If no homologous proteins were found in these three highly curated sets, protein-coding gene models were assigned to the functional class of 'hypothetical' proteins. Gene model completeness was checked using BUSCO v4.06 (Supplementary Table S6).[59] A higher proportion of missing BUSCOs was seen in EL10.2.2 than either EL10.1 or RefBeet 1.1. Overall, protein-coding gene predictions covered a relatively small proportion of the assembled EL10.1 genome (39,161,207 nt; 7.2%). GC content of predicted coding genes was marginally higher than that of the whole genome (41.1% vs. 35.8%, respectively). Predicted proteins were named using the underlined characters in the key: EL10 / annotation version A / *c*hromosome or scaffold number / genomic in origin / a sequential number / and appended with .1 to signify that only one isoform was considered at this level of analysis (e.g. EL10Ac7g16740.1). Of the 24,255 genes identified in EL10.1, only 86 were not able to be assigned to the EL10.2 genome using LIFT (Supplementary Table S7).

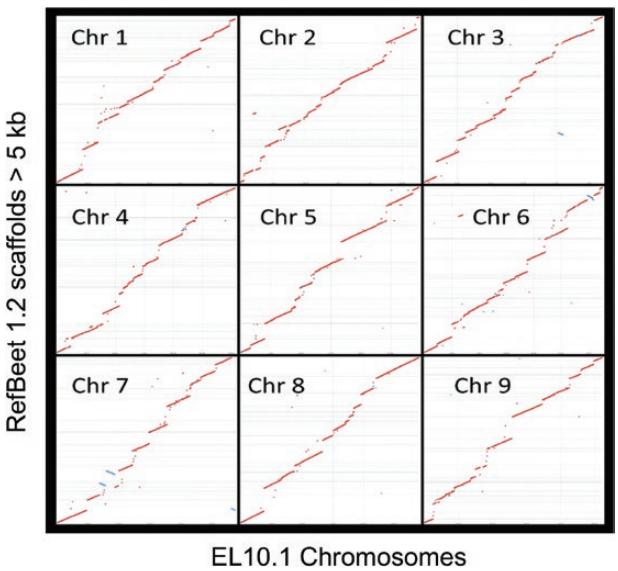

Figure 1. Chromosome alignment of the EL10.1 assembly (x-axis) versus RefBeet 1.2 assembly (y-axis) by EL10.1 Chromosome. Alignments less than 5 kb in length were removed before plotting. Alignment blocks with positive upwards slope indicate matching orientation between the two assemblies. Unassembled RefBeet regions are indicated by gaps.

The number of MAKER annotations ascribed across Chromosomes of EL10.1 was relatively consistent (mean = 2,559, std. dev. = 173.8), but highly variable between scaffolds (mean = 44.0, std. dev. = 47.6) (excluding Scaffolds 23, 29, 30, and 31 for which no gene models were predicted) (Supplementary Table S8). A total of 3,940 gene models had no functional annotation among curated comparative databases (and thus were designated hypothetical), and were evenly distributed among chromosomes but not scaffolds (Supplementary Table S8). Approximately 55% of gene models were considered unique or single copy based on assignment of curated-database proteins (Supplementary Table S8), more than 45% of predicted genes could be members of gene families.

Self-synteny of MAKER gene models with the EL10.1 genome sequence was explored using the CoGe SynMap platform.[42] Few internal syntenies were detected. Mean copy number of the 2,327 discovered tandem gene models was 2.82 (std. dev. = 1.96), and 65.8% of these tandem duplications were two copies. For syntenic regions with at least 5 matches in a span of 20 gene models (encompassing 1,858 genes in 268 synteny blocks), average Kn/Ks values were all less than 1, suggesting stabilizing selection for genes in these blocks. For individual gene pairs, five gene pairs had Kn/Ks values >1 (suggesting diversifying selection) but only two of these pairs had interpretable annotations. EL10Ac6g14284.1 and EL10Ac9g20883.1 were predicted as Clathrin heavy chain 2 genes (i.e. vesicle trafficking) and EL10Ac1g01568.1 and EL10Ac5g12109.1 were predicted SET Domain Protein genes (i.e. chromatin structure modulation).

A comparative gene annotation perspective was gained using the MapMan4 ontology of plant proteomes.[39] EL10.1 MAKER gene models were placed in 99.6% of 4,145 ontologies assigned to one of 28 'bins' (infrequently allowing for assignment to more than one bin), organized in a hierarchal, conceptual, plant-specific context (e.g. Photosynthesis, Cell cycle, Hormones, etc.). Where possible, each bin resolves to a gene from a high-quality genome assembly in the Mercator4 web implementation of MapMan4. Specific comparisons for each of the 4,127 EL10.1 occupied terminal, termed 'leaf', bins were made with five other angiosperms (e.g. *A. thaliana*, *O. sativa*, *B. distachyon*, *S. lycopersicum*, and *M. esculenta*). Most EL10.1 predicted proteins in the found set were placed in one (or more) MapMan4 leaf bins (Supplementary Table S9). Since the MapMan4 ontology is hierarchal, the number of genes in each leaf bin was averaged for all five angiosperms, and compared with EL10.1 (Supplementary Table S9).

Enrichment analysis can shed light on biological processes that may have assumed greater or lesser importance in the evolutionary success of a lineage. Given the general reduced gene copy number in EL10.1, genes whose copy number equalled or exceeded the mean of five angiosperms (grape, *Arabidopsis*, spinach, amaranth, and quinoa) were tentatively considered as enriched, and those that were substantially lower than the overall mean of EL10.1 were considered as reduced. EL10.1 appeared particularly depauperate in at least two top-level ontologies: Cellular respiration (Bincode 2) and Phytohormones (Bincode 11) (Supplementary Table S9). Equal or over-represented ontologies included DNA Damage Response (Bincode 14) and Coenzyme metabolism (Bincode 7) (Supplementary Table S9).

Proteome content of the five averaged angiosperms relative to EL10.1 was gauged for missing members, which could suggest regions in EL10.1 that were not assembled, genes that were not annotated, or perhaps reflect biological divergence or biochemical alternatives that beet followed during its evolution. Not detected in EL10.1 were 154 genes that were present in at least one copy in each of the five angiosperms. Missing annotations were assignable across all 28 top-level bins, with the exception of Bincode 8 (Polyamine metabolism) (Supplementary Table S10). Among the five taxa being used, mean copy number was low (1.6 genes per leaf bin), and failure to assemble or annotate low-copy number genes in EL10.1 was possible. However, in 12 cases, each of the five other plants had small gene families (mean copy number = 3.7 genes per family) but no EL10.1 homologue was annotated. It seems less probable that all members of these gene families would have been missed during assembly and annotation, thus their functions in beet may have been dispensable, their genes sufficiently diverged, or their functions assumed by other non-homologous genes.

Under-represented genes in 'Cell wall' (Bincode 21) included those involved with hemicellulose, lignin, cutin, and suberin metabolism, as might be expected from selection for a mechanically sliced root crop for sucrose extraction (e.g. less knife wear during processing, which is a trait that has not necessarily been under conscious selection) (Supplementary Table S11). Phytohormone representation was low across all second-level categories, especially salicylic acid (Bincode 11.8). External stimuli response (Bincode 26) was rich in drought response but poor in biotic stress response genes. Multi-process regulation (e.g. integration of development with response-to-environment) was over-represented by the TOR (Target of Rapamycin) signaling pathway (Bincode 27.2) and under-represented in the SnRK1 metabolic regulator system (Bincode 27.3). RNA biosynthesis (Bincode 15) was generally over-represented, however Bincode 15.8 (transcriptional repression) was greatly under-represented. Overall, 138 leaf bins were similar or over-represented and 447 were under-represented in EL10.1.

Transcription factor genes (Bincode 15.7) were under-represented overall in the EL10.1 annotation. On average, there were ~10 fewer genes in EL10.1 transcription factor classes than the average of five other angiosperms. Transcription factor classes with a >50 gene deficiency between the angiosperm average and EL10.1 included MADS box, NAC, MYB, and bHLH transcription factors (Supplementary Table S12). Most of the transcription factor classes showing larger deficiency in copy number were members of large common gene families. Few transcription factor classes were equally or over-represented, and most of these were from gene families characterized by lower copy number (Supplementary Table S12). However, the FAR1 (in *Arabidopsis*, transposon-derived transcription factors associated with far red-light response) transcription factor class was abundant in EL10.1, and highly variable in the group of five other angiosperms (Supplementary Table S12). It is likely that each of these differences in transcription factor copy number has potential to impact plant phenotype, development, and/or response to the environment.

3.4. Genome size

Discrepancies between reported genome sizes (714–758 Mb[46]; derived from estimates for one plant each of table and sugar beet, respectively) and assembled genome sizes of sugar beet

(~540.5–566.6 Mb, Table 2) may be explained by failure to assemble repetitive sequence arrays completely. To better assess genome size as a gauge of the completeness of assemblies in *B. vulgaris*, an additional 50 independent cytometrically determined nuclear DNA content estimates were obtained from four unrelated germplasm accessions: two traditional out-crossing progenies and two from progeny of deeply inbred accessions of EL10 and an inbred table beet derived from germplasm 'W357B'. Nuclear DNA content estimates of these materials ranged from 633 to 875 Mb, as estimated from at least four biological replicates from each accession (at least 20 from inbreds) with four technical replicates performed per biological replicate (Supplementary Table S13). Overall, genome size between crop types was not statistically different (sugar beet, n = 120, mean = 729.0 Mb/1C, std. dev. = 51.2; table beet, n = 80, mean = 742.3, std. dev. = 52.8; P = 0.079). Average genome size differences of each sugar beet accession were significantly different from one another ($P < 0.001$, means and dispersion values are presented in Supplementary Table S13), and only the difference between sugar beet '5B sugar breeding population' and Inbred Table beet was not significantly different than the other two sugar beet accessions. Inbreds showed a statistically significant smaller average genome size (Supplementary Table S13: inbreds, mean = 728.5 Mb/1C, out crossed, mean = 764.9 Mb/1C, P = 0.0002), and at least 2-fold higher variation than out-crossers (Supplementary Table S13). The average cytometrically determined genome size of all tested accessions was 734.3 Mb (std. dev. = 50.3 Mb).

3.5. Repetitive element content estimation

Plant genomes are characterized by high-repetitive sequence content, found either as tandem arrays or as multiple copies distributed throughout the genome.[60] More than 180,000 named repetitive elements (identified by RepeatMasker) were placed on the EL10.1 assembly (Supplementary Table S14). DNA class transposable elements were the most frequent (58.1%), in contrast with RefBeet,[12] and LTR elements the next most frequent class (36.0%) of annotated transposable elements (Supplementary Table S14). Numbers and types of LTR elements were estimated similarly using RepeatMasker and LTR_Retriever.[40] However, distribution of the filtered high-confidence intact LTR_Retriever-predicted Gypsy and Copia elements (Supplementary Table S14) showed Copia elements generally were more frequently found the ends of Chromosomes and Gypsy elements biased towards centromeric regions (Fig. 2).

Repeats associated with centromeric histone variants have been characterized in beets,[61] and these consist of the Gypsy element Beetle7 as well the pBV class of major satellites (Supplementary Table S14). High-similarity Beetle7 sequences (90% identity over 1,000 nt or better) were located on all chromosomes and eight of the scaffolds. The 35S and 5S ribosomal RNA genes are also tandemly arrayed in beets.[14] The 35S arrays in EL10.1 were localized to Chromosome 2, as expected, and also to Scaffolds 7 and 19. The 5S array localized to Chromosome 4, also as expected, and to Scaffold 11. Only one canonical plant telomere array $(TTTAGGG)_n$ greater than three tandem copies was found in the EL10.1 assembly, near the end of Scaffold 5. However, terminal repeat arrays defined by the major satellite class pAV[62] were found near the ends of most chromosomes, except at one end each of Chromosomes 1, 5, 7, and 9 (Supplementary Table S14). pAV arrays were seen on each of these except Chromosome 1, where the South terminus appeared absent. Evidence suggests Chromosome 5 South is Scaffold 5, Chromosome 9 may have a pericentric inversion or an assembly artefact that misplaced Chromosome 9 South, and complex inversions in Chromosome 7 may have failed to accurately assemble the North terminal repeat region (these appeared to have been resolved in the EL10.2_2 assembly). Notably, interstitial pAV arrays were evident in both Chromosomes 5 and 7 (Supplementary Table S14D).

Tandem repeats (unit length 500 nt or less assessed with Tandem Repeat Finder) were evenly spread across the EL10.1 assembly (Supplementary Table S15), with an average of 630.4 repeats Mb^{-1} (std. dev. = 19.3) across chromosomes, and similar for scaffolds but with 25-fold higher variation (mean 661.0 repeats Mb^{-1}, std. dev. = 460.8). Shorter repeats were more frequent, and the most frequent size class was 21 nt (23,163 instances). Size classes of tandem repeats may reflect the predominant repeat unit size for centromeric sequence in a species,[63] and for EL10.1, the most frequent repeat size above 100 nt was 160 nt (781 copies), followed by 170 nt (382 copies). Relatively high numbers of repeats (67–134 copies) in the 314–325 nt repeat unit size range were evident, as might be consistent with a heterodimeric model of centromere repeats.[63]

Assembly continuity was assessed using the LAI.[40] After adjusting for the amplification time of LTR-RTs, the whole-genome LAI of the EL10 assembly was estimated to be 13.3, which is considered reference quality and improves upon the RefBeet assembly (LAI = 6.7) (Supplementary Fig. S2). Thus, the EL10.1 sugar beet genome assembly appeared to be largely complete with respect to repetitive element landmarks and assembled in a largely congruent fashion with respect to genetic markers.

3.6. Read count mapping

Read-depth variation provided a means to compare accessions using readily available and deeper coverage short reads. Low variation in read depth suggests relatively even distribution of coverage across assembly coordinates, while higher variation suggests regions of low sequence complexity that may not have assembled in a consistent fashion, perhaps contributing to differences in genome size between cytometry and assembly estimates. Five independent Illumina-derived read sets were read mapped to the EL10.1 genome assembly, one from EL10 and one each from four other sugar beet germplasms (including two EL10 relatives and two unrelated germplasms). Overall, more than 99.6% of EL10's cleaned reads mapped to the EL10.1 assembly, with relatively even coverage (e.g. ~36 reads per assembled nucleotide), but scaffold coverage was slightly less and the standard deviation was 22-fold higher. Similar results were evident in the other four germplasms (Supplementary Table S16). There appeared no 'degree-of-relatedness' discrimination between disparate germplasm at this level of analysis, as EL10 relatives showed as much overall difference in read-depth variation as individuals drawn from unrelated populations.

High read-depth locations were localized using a conservative, computationally facile, and relatively crude sequence-independent approach. High read-depth locations were defined as a region of 5 kb with average per-base read mapping depth above 500 in one or more of the five tested germplasms (indexed from the lower nucleotide position of

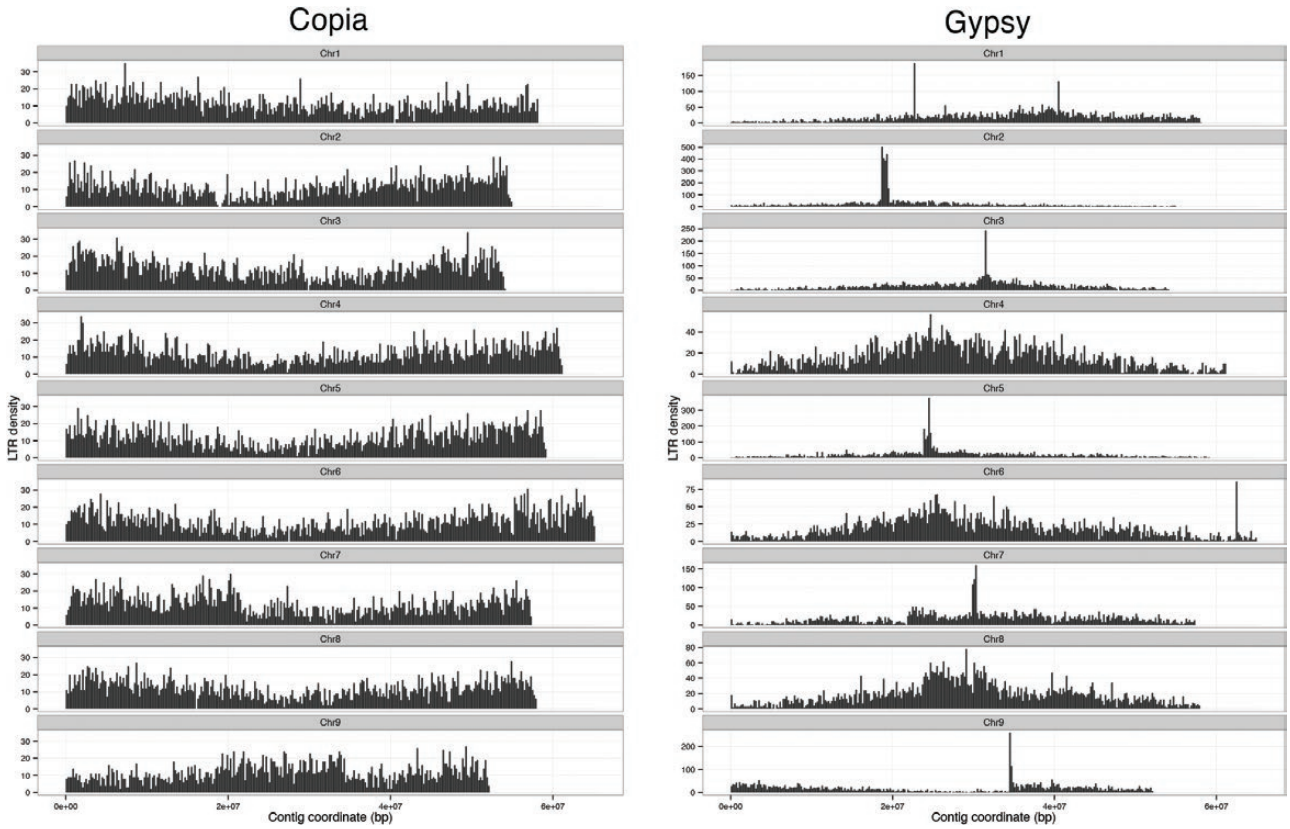

Figure 2. Distribution of LTR Copia and Gypsy retrotransposon elements across the EL10.1 Chromosomes.

the EL10.1 assembly). This binning approach is conservative in the sense that most highly repetitive elements are shorter than the 5 kb window size used, but provided a computational advantage for an initial assessment whether changes in genome size could crudely be restricted to specific genomic bins, or were otherwise more or less independently distributed across the genome. Along Chromosome 1, 47 bins were flagged as differential between C869_25 (i.e. the base genotype for EL10 and C869_UK) and each other accession. Each flagged bin in each of the five germplasms occurred predominantly in the same places on Chromosome 1. Most of these bins were occupied by Gypsy or Copia LTRs, however Bin 44,615,000 was occupied by chloroplast sequence (Sequence ID: KR230391.1) and Bins 8,100,000, 22,360,000, and 22,365,000 were occupied with mitochondrial sequences (Sequence ID: FP885845.1). It is not unusual to find plastid sequences within plant genomes,[64] and plastid sequence read depths are likely subject to external influences (e.g. plant growth and DNA isolation methods). The large differences in the remaining read-depth estimates at specific sites suggest that copy number changed since a last common ancestor. These sites have the potential to contribute to intra-specific genome size variation. Further evaluation of such sites across the genome in a more precise sequence-specific fashion (e.g. not binned) may help deduce special features related to their lability and whether changes in genome size at this level of resolution have phenotypic effects.

3.7. Broader synteny

Caryophyllales members spinach (*S. oleracea*), grain amaranth (*Amaranthus hypochondriacus*), and quinoa (*Chenopodium quinoa*) have annotated genome assemblies that were used to compare with EL10.1,[65–67] respectively; note that quinoa and amaranth are each amphidiploid). Chromosome 4 synteny appeared maintained in chromosome-sized blocks among Caryophyllales, as well as *Vitis* to a lesser extent, but not *A. thaliana*, as outgroup representatives of the Rosids (Fig. 3). Chromosome 1 synteny also appeared relatively conserved in chromosome-sized blocks among the Caryophyllales, with the exception of the spinach assembly version used here, which will likely improve in the future. Elements of Chromosomes 2, 6, and 9 were found in extended blocks in quinoa and amaranth, but also not spinach. Extended synteny for Chromosomes 5 and 8 were evident in quinoa but were not as extended in amaranth, while extended blocks for Chromosomes 3 and 7 were present in amaranth but not as well maintained in quinoa. Genome evolution within the Caryophyllales produced significant genomic variation in chromosome number, number of syntenic regions, and size of syntenic regions relative to beet (Table 3).

4. Discussion

A high-quality *de novo* assembly of the sugar beet genome was created. The EL10.1 assembly contains most of the 'EL10' genome organized into 9 linkage groups plus 31 extra unplaced scaffolds. Most scaffolds contain predicted genes, and many were able to be placed in context of the larger chromosome-sized assemblies using genetic markers. Ends of chromosomes were captured to some degree, however additional work and sequencing experiments will be required to finish the EL10 genome assembly to exacting standards. The EL10.2 assembly appears to resolve at least the major assembly/scaffolding-induced inversions evident

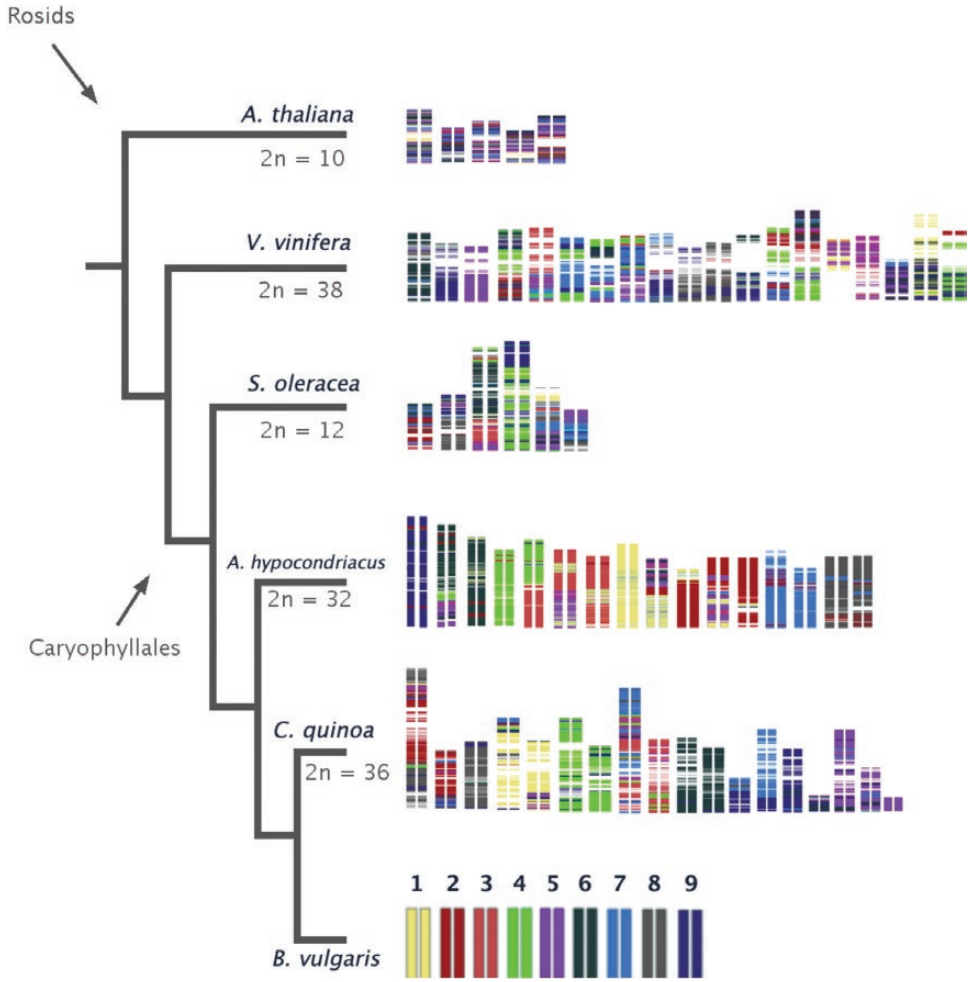

Figure 3. Visualization of syntenic blocks among Caryophyllales genomes relative to *B. vulgaris* EL10.1 Chromosomes compared with two representative Rosid species, colour coded by EL10.1 Chromosome.

Table 3. Proportion and metrics of synteny (co-linear blocks of MAKER beet gene predictions) shared among five species

Species	*C. quinoa*	*A. hypocondriacus*	*S. oleracea*	*V. vinefera*	*A. thaliana*
Common name	Quinoa	Amaranth	Spinach	Grape	*Arabidopsis*
Chromosome number	$2n = 4x = 36$	$2n = 4x = 32$	$2n = 2x = 12$	$2n = 2x = 38$	$2n = 2x = 10$
Number of synteny blocks	854	599	410	547	734
Number of genes in blocks	25,832	14,519	8,437	10,711	9,228
Mean number of genes per block	30.2	24.2	20.6	19.6	12.6
Std. dev.	51.4	31.2	27.6	22.4	8.7
Range	5–490	5–245	6–261	6–223	6–74

in Chromosomes 7 and 9. Genome assembly is fraught with uncertainty. In most cases, there is no *a priori* information to gauge the completeness and correctness of an assembly, particularly when genetic map(s) are not available for the background being sequenced. In this case, the fortuitous availability of a published sugar beet assembly[12] allows for comparisons, however EL10 provides an independent perspective on the organization of an inbred beet genome compared with that of a doubled haploid. The major difference between the two studies is that the sequencing technologies have improved to provide longer range scaffolding resulting in a substantially improved contiguity of sugar beet genome assembly. Such improvements also presumably better reflect copy number variations and gene content.

Long reads alone are currently insufficient for a high-quality assembly of a plant genome of moderate to large genome size. Beet might be considered a moderately sized plant genome. The addition of an opto-physical map to long reads alone provided a ~10-fold reduction in the number of contigs, as well as set an upper bound for the size of the sequenced EL10 genome assembly (628 Mb). However, this also was insufficient to achieve chromosome-level assembly. Further addition of Hi-C data, where intact nuclei are cross-linked *in vivo* and where the native genome organization is presumably

preserved, provided the means to map chromosome-level associations. The sequential application of at least four independent technologies (i.e. short- and long-read sequencing, physical/optical maps, Hi-C chromatin conformation capture, and genetic maps) helped to overcome many limitations spanning low-complexity regions of a genome over previous technologies in creating contiguous *de novo* genome assemblies of moderate to large plant genomes. Future applications of higher-quality long-read data types (e.g. PacBio HiFi) coupled with endonuclease-free chromatin conformation capture (e.g. DoveTail Omni-C) should provide additional improvements to genome completeness (maximizing actual genome size) and chromosome-level contiguity and accuracy, as has been recently seen in other plant genomes.[68]

A reduction in gene copy number in beet (relative to annotated protein genes generated for comparative purposes, e.g. MapMan4) was observed. No clear evidence of gene copy number amplification was observed among the EL10.1 predicted protein set for most of the gene families. Clear reductions in gene copy number were detected across multiple gene classes and for transcription factors in particular, also observed by[12] and more recently.[69] Exceptions to the transcription factor reduction observed in EL10.1 included the FAR1 class of transcription factors, which may be anciently derived from Mutator-like transposons and co-opted in *Arabidopsis* for red-light perception and signaling.[70,71] The role for this class of sequences remains unknown in beets, and copy number variation was high for FAR1 between the five other angiosperms considered. The lower overall gene copy number in beets may be suggestive of a basal gene copy number in dicots where beet numbers (or Caryophyllids in general) approximate a baseline condition, while other dicots may have increased copy numbers and diversified gene families.

Genome size estimates of the cultivated beets examined here were quite variable, ranging from 633.0 to 875.5 Mb per haploid genome. Genome size estimates of 21 wild *B. vulgaris* spp. *maritima* genotypes from Portugal ranged from 660.1 to 753.1 Mb,[72] thus variability in genome size is known to occur in the species. The range of estimates was 2.6 times higher in the cultivated beets relative to the wild types. This was also observed relative to the breeding system of the cultigens, where the range in genome size among the out-crossers was 2.7 times lower than that of the inbreds (e.g. Supplementary Table S13). Variation in read-depth coverage may be useful for tracking genome size changes.[73] Areas of high variation are intriguing from a chromosome biology and evolution perspective, as well as their potential effect on phenotype and on the origin of novel variation. Many plant genomes are large because of their highly repetitive nature, and many classes of repetitive elements are known to vary across kingdoms, often with little in common other than size, the fact they are repetitive, and characteristic footprints (target site duplications, terminal repeats, etc.).[60] Speciation seems to favour whole-scale sequence replacement of repeat elements while retaining their size, however inter-specific amplified repeats seem to be present at low copy number in related genomes.[74] Exactly how, and in particular when and what effects the efficiency, distribution, and specificity of divergent repeat amplification, is not as easy to investigate. Additional sugar beet genomes sequenced and assembled using high-quality long reads technologies should provide new insights on the repetitive regions of the sugar beet genome that were previously inaccessible with older technologies.

Beet is naturally a wind-pollinated out-crosser, which means that genetic diversity is partitioned within populations rather than between populations. Inbreeding depression is high, and inbred beets are not necessarily representative of the genomic landscape of hybrids. Each of the germplasm examined here, with the exception of C869_25, was highly inbred, using one of three different breeding methods. Both C869_UK and EL10 were derived from C869_25 through single seed descent, for three and five generations, respectively. RefBeet (aka KWS2320) was derived as a doubled haploid, and NK-388mm-O is a seed parent for hybrids inbred through conventional sib-mating.[75,76] The method used to generate the inbred seems not to relate to generation of read-depth differences. However, each germplasm had a set of read enrichment events specific to their own lineage, and others that were shared among two, three, or all germplasms. For instance, NK-388mm-O was enriched in depth at EL10.1 Chromosome 1 positions 53,310,000 to 53,325,000 Mb, KWS2320 was depauperate at positions 40,540,000 to 44,615,000, and C869_25 over-represented from 22,735,000 to 22,7750,000. Responsible sequences underlying these regions have not yet been investigated, except where wide differences in chloroplast content and mitochondria were particularly rich in NK-388mm-O. While these read-depth differences may be artefacts of assembly, it is equally likely that they are by-products of artificially induced inbreds. In other highly heterozygous organisms like grape, large variation in genome composition (structure, size, and genes) have been found in sequenced hybrids compared with the original sequenced doubled haploid grape PN40024.[77–79]

Exploration of synteny between species is accessible from a contiguous well-annotated genome sequence. For EL10.1, annotations were conservatively estimated from well curated plant gene resources, which likely improved confidence in assessing similarity between well-known plant genes. Following the syntenic organization of such genes across phylogenetic groups showed that closely related species retained higher levels of synteny than more distantly related species, as expected. Also expected, was that recombination and schism of synteny blocks increased with increasing phylogenetic distance. Perhaps unexpected was differential synteny conservation by individual chromosomes. However, relatively few plant genomes are available that are highly contiguous, and this caveat limits interpreting results.

5. Conclusion

Here, we present a contiguous, hybrid genome assembly of a sugar beet line, EL10.1, consisting of 540 Mb, of which 96.2% was contained in nine chromosomes. Compared with the previous genome assembly (RefBeet 1.2), our hybrid approach improved chromosome resolution and coverage of highly repetitive regions. A total of 24,255 proteins were predicted using MAKER, covering an estimated 7% of the EL10.1 genome. Compared with sequenced core angiosperms, the sugar beet genome had a reduction in gene number for certain gene families and groups including transcription factors. Genome size variation was also explored for this species using flow cytometry, with an estimated size range of 633–875 Mb depending on the use-type and pedigree. The genomic data presented here will enable further molecular studies of sugar beet and other Caryophyllales members.

Funding

J.M.M.: funding provided by USDA-ARS CRIS 3635-21000-011-00D and the Beet Sugar Development Foundation, Denver, CO, USA. S.K. and A.P. were supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

Acknowledgements

We thank Safa Alzohairy for isolating RNA from germinating sugar beet seedlings under various germination regimes. We also thank K. Arumuganathan, Director of the Flow Cytometry Lab of the Benaroya Research Institute at Virginia Mason, Seattle, WA for contracting the genome size estimates, and the staff at Dovetail Genomics (Santa Cruz, CA) for their assistance with the EL10.2.2 assembly.

Authors' contributions

J.M.M., B.T., E.M.-G., and K.D.: conceived and organized the work; J.M.M., B.T., E.M.-G., K.D., R.N., and P.G.: wrote the manuscript; A.F., P.G., and S.O.: characterized EL10.1 assembly sequence organization; K.D., H.D., and S.J.: created PacBio resources; J.L. and A.H.: created BioNano resources; I.L., S.S., S.K., and A.P.: conducted Hi-C assembly and finishing of EL10.1; A.D., G.W., S.B., P.S., and K.T.: applied proprietary genetic markers and materials to assess integrity of the EL10 assemblies; J.W., T.L., J.P., and K.C.: provided MAKER gene annotations for EL10.1; S.S.: created gene models for EL10.2 for Phytozome integration; A.Y. and D.F.: created RNA-seq transcriptome assemblies used in the EL10.1 annotation.

Supplementary data

Supplementary data are available at *DNARES* online.

Supplementary Table S1. Assembly statistics of intermediate EL10 assemblies.

Supplementary Table S2. Inversions in the EL10.1 genome assembly assessed using genetic markers.

Supplementary Table S3. Co-locations of scaffolds and chromosomes deduced by genetically mapped markers.

Supplementary Table S4. Orientation of EL10.1 Chromosomes relative to the cytogenetic map of Paesold *et al.*[14]

Supplementary Table S5. The Y–R–B linkage group in the EL10.1 genome assembly.

Supplementary Table S6. Gene models detected via Benchmarking Universal Single-Copy Orthologs (BUSCO).

Supplementary Table S7. Distribution of MAKER annotations across the EL10.1 genome assembly.

Supplementary Table S8. Comparison of MapMan4 first-order functional classifications for EL10.1 and four other Tracheophytes [*Oryza sativa* (*Os*), *Brachypodium distachyon* (*Bd*), *Arabidopsis thaliana* (*At*), *Solanum lycopersicum* (*Sl*), and *Manihot esculenta* (*Me*)].

Supplementary Table S9. Comparison of MapMan4 leaf bins where no gene was predicted by MAKER in EL10.1 and five Tracheophytes [*Oryza sativa* (*Os*), *Brachypodium distachyon* (*Bd*), *Arabidopsis thaliana* (*At*), *Solanum lycopersicum* (*Sl*), and *Manihot esculenta* (*Me*)].

Supplementary Table S10. Comparison of over- and under-represented MapMan4 second-order functional classifications for EL10.1 and four other Tracheophytes [*Oryza sativa* (*Os*), *Brachypodium distachyon* (*Bd*), *Arabidopsis thaliana* (*At*), *Solanum lycopersicum* (*Sl*), and *Manihot esculenta* (*Me*)].

Supplementary Table S11. Comparison of transcription factor classes for EL10.1 and four other Tracheophytes [*Oryza sativa* (*Os*), *Brachypodium distachyon* (*Bd*), *Arabidopsis thaliana* (*At*), *Solanum lycopersicum* (*Sl*), and *Manihot esculenta* (*Me*)].

Supplementary Table S12. Beet genome size estimates obtained by flow cytometry.

Supplementary Table S13. Frequency of transposable element (TE) classes in the EL10.1 genome assembly. (A) RepeatMasker-derived annotations. (B) Complete LTR retrotransposons. (C) Interstitial repeat classes (from Kowar *et al.*[61]). (D) Terminal repeat locations (from Dechyeva and Schmidt[62]) integrated with cytogenetic orientation (parentheses indicate reversed orientation relative to Paesold *et al.*[14]).

Supplementary Table S14. Characteristics of tandem repeats in the EL10.1 genome assembly.

Supplementary Table S15. Read count mapping of short reads from EL10 and four other germplasms to the EL10.1 genome assembly.

Supplementary Figure S1. Comparison of contiguity between EL10.1 and EL10.2_2 genome assemblies. Alignments with matching orientation are shown in red, inversions are show in blue.

Supplementary Figure S2. LTR Assembly Index (LAI) of the RefBeet assembly (A) and EL10 assembly (B) of the sugar beet genome. *X*-axes denote pseudochromosomes of the two assemblies. Each dot represents regional LAI in a 3 Mb window. Red-dotted lines indicate the LAI cutoff of the reference genome quality (LAI = 10). Blue-dotted lines indicate the mean LAI.

References

1. Biancardi, E., Panella, L.W. and Lewellen, R.T. 2012, *Beta maritima: the origin of beets*. Springer: New York.
2. Galon, J. and Zallen, D.T. 1998, The role of the Vilmorin Company in the promotion and diffusion of the experimental science of heredity in France, 1840–1920, *J. Hist. Biol.*, **31**, 241–62.
3. Panella, L., Campbell, L.G., Eujayl, I.A., Lewellen, R.T. and McGrath, J.M. 2015, USDA-ARS sugar beet releases and breeding over the past 20 years, *J. Sugar Beet Res.*, **52**, 22–67.
4. Panella, L., Lewellen, R.T. and Hanson, L.E. 2008, Breeding for multiple disease resistance in sugar beet: registration of FC220 and FC221, *J. Plant Reg.*, **2**, 146–55.
5. Funk, A., Galewski, P. and McGrath, J.M. 2018, Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome, *Plant J.*, **95**, 659–71.
6. Galewski, P. and McGrath, J.M. 2020, Genetic diversity among cultivated beets (*Beta vulgaris*) assessed via population-based whole genome sequences, *BMC Genomics*, **21**, 189.
7. Yang, Y., Moore, M.J., Brockington, S.F., et al. 2015, Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing, *Mol. Biol. Evol.*, **32**, 2001–14.
8. Andrello, M., Henry, K., Devaux, P., Desprez, B. and Manel, S. 2016, Taxonomic, spatial and adaptive genetic variation of *Beta* section *Beta*, *Theor. Appl. Genet.*, **129**, 257–71.
9. Andrello, M., Henry, K., Devaux, P., et al. 2017, Insights into the genetic relationships among plants of *Beta* section *Beta* using SNP markers, *Theor. Appl. Genet.*, **130**, 1857–66.
10. Del Rio, A.R., Minoche, A.E., Zwickl, N.F., et al. 2019, Genomes of the wild beets *Beta patula* and *Beta vulgaris* ssp. *maritima*, *Plant J.*, **99**, 1242–53. doi:10.1111/tpj.14413

11. Flavell, R.B., Bennet, M.D. and Smith, J.B. 1974, Genome size and the proportion of repeated nucleotide sequence DNA in plants, *Biochem. Genet.*, **12**, 257–69.
12. Dohm, J.C., Minoche, A.E., Holtgräwe, D., et al. 2014, The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*), *Nature*, **505**, 546–9.
13. Schmidt, T. and Heslop-Harrison, J.S. 1998, Genomes, genes and junk: the large-scale organization of plant chromosomes, *Trends Plant Sci.*, **3**, 195–9.
14. Paesold, S., Borchart, D., Schmidt, T. and Dechyeva, D. 2012, A sugar beet (*Beta vulgaris* L.) reference FISH karyotype for chromosome and chromosome-arm identification, integration of genetic linkage groups and analysis of major repeat family distribution, *Plant J.*, **72**, 600–11.
15. Halldén, C., Ahrén, D., Hjerdin, A., Säll, T. and Nilsson, N.O. 1998, No conserved homoeologous regions found in the sugar beet genome, *J. Sugar Beet Res.*, **35**, 1–13.
16. Dohm, J.C., Lange, C., Holtgräwe, D., et al. 2012, Palaeohexaploid ancestry for Caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (*Beta vulgaris*), *Plant J.*, **70**, 528–40.
17. Schondelmaier, J. and Jung, C. 1997, Chromosomal assignment of the nine linkage groups of sugar beet (*Beta vulgaris* L.) using primary trisomics, *Theor. Appl. Genet.*, **95**, 590–6.
18. McGrath, J.M., Trebbi, D., Fenwick, A., et al. 2007, An open-source first-generation molecular genetic map from a sugarbeet × table beet cross and its extension to physical mapping, *Crop Sci.*, **47**, S-27–44.
19. Jung, H., Winefield, C., Bombarely, A., Prentis, P. and Waterhouse, P. 2019, Tools and strategies for long-read sequencing and *de novo* assembly of plant genomes, *Trends Plant Sci.*, **24**, 700–24.
20. Galewski, P. and Eujayl, I. 2022, A roadmap to durable BCTV resistance using long-read genome assembly of genetic stock KDH13, *Plant Mol. Biol. Rep.*, **40**, 176–87.
21. Lewellen, R.T. 2004, Registration of Rhizomania Resistant, Monogerm Populations C869 and C869CMS Sugarbeet, *Crop Sci.*, **44**, 357–58.
22. McGrath, J.M., Drou, N., Waite, D., et al. 2013, The 'C869' sugar beet genome: a draft assembly. https://pag.confex.com/pag/xxi/webprogram/Paper5768.html (29 August 2020, date last accessed).
23. Bickhart, D.M., Rosen, B.D., Koren, S., et al. 2017, Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome, *Nat. Genet.*, **49**, 643–50.
24. Putnam NH, O'Connell B, Stites JC, et al. 2016, Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome. Res.*, **26**, 342–350.
25. Zimin, A.V. and Salzberg, S.L. 2020, The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies, *PLoS Comput. Biol.*, **16**, e1007981.
26. Holt, C. and Yandell, M. 2011, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, *BMC Bioinf.*, **12**, 491.
27. Campbell, M.S., Law, M., Holt, C., et al. 2014, MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations, *Plant Physiol.*, **164**, 513–24.
28. Cheng, C.-Y., Krishnakumar, V., Chan, A.P., et al. 2017, Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome, *Plant J.*, **89**, 789–804.
29. Fernandez-Pozo, N., Menda, N., Edwards, J.D., et al. 2015, The Sol Genomics Network (SGN)—from genotype to phenotype to breeding, *Nucleic Acids Res.*, **43**, D1036–41.
30. Tuskan, G.A., DiFazio, S., Jansson, S., et al. 2006, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science*, **313**, 1596–604.
31. The UniProt Consortium. 2017, UniProt: the universal protein knowledgebase, *Nucleic Acids Res.*, **45**, D158–69.
32. Pertea, M., Pertea, G.M., Antonescu, C.M., et al. 2015, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, *Nat. Biotechnol.*, **33**, 290–5.
33. Stanke, M. and Waack, S. 2003, Gene prediction with a hidden Markov model and a new intron submodel, *Bioinformatics*, **19**, ii215–25.
34. Korf, I. 2004, Gene finding in novel genomes, *BMC Bioinf.*, **5**, 59.
35. Finn, R.D., Clements J. and Eddy, S.R. 2011, HMMER web server: interactive sequence similarity searching, *Nucleic Acids Res.*, **39**, W29–37.
36. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–10.
37. Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H. 2011, SignalP 4.0: discriminating signal peptides from transmembrane regions, *Nat. Methods*, **8**, 785–6.
38. Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L.L. 2001, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J. Mol. Biol.*, **305**, 567–80.
39. Schwacke, R., Ponce-Soto, G.Y., Krause, K., et al. 2019, MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis, *Mol. Plant*, **12**, 879–92.
40. Ou, S. and Jiang, N. 2018, LTR_retriever: a highly accurate and sensitive program for identification of long terminal-repeat retrotransposons, *Plant Physiol.*, **176**, 1410–22.
41. Benson, G. 1999, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acid Res.*, **27**, 573–80.
42. Lyons, E., Pedersen, B., Kane, J., et al. 2008, Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar and grape: CoGe with rosids, *Plant Phys.*, **148**, 1772–81.
43. Kielbasa, S.M., Wan, R., Sato, K., Horton, P. and Frith, M.C. 2011, Adaptive seeds tame genomic sequence comparison, *Genome Res.*, **21**, 487–93.
44. Haas, B.J., Delcher, A.L., Wortman, J.R. and Salzberg, S.L. 2004, DAGchainer: a tool for mining segmental genome duplications and synteny, *Bioinformatics*, **20**, 3643–6.
45. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.*, **24**, 1586–91.
46. Arumuganathan, K. and Earle, E.D. 1991, Nuclear DNA content of some important plant species, *Plant Mol. Biol. Rep.*, **9**, 208–18.
47. Doležel, J., Bartoš, J., Voglmayr, H. and Greilhuber, J. 2003, Letter to the editor, *Cytometry*, **51A**, 127–8.
48. Bushnell, B. 2014, *BBMap: a fast, accurate, splice-aware aligner*. Report Number: LBNL-7065E.
49. Wang, Y., Tang, H., DeBarry, J.D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res.*, **40**, e49.
50. Trebbi, D. and McGrath, J.M. 2009, Functional differentiation of the sugar beet root system as indicator of developmental phase change, *Physiol. Plant.*, **135**, 84–97.
51. McGrath, J.M., Derrico, C.A., Morales, M., Copeland, L.O. and Christenson, D.R. 2000, Germination of sugar beet (*Beta vulgaris* L.) seed submerged in hydrogen peroxide and water as a means to discriminate cultivar and seedlot vigor, *Seed Sci. Technol.*, **28**, 607–20.
52. McGrath, J.M., Koppin, T.K. and Duckert, T.M. 2005, Breeding for genetics: development of Recombinant Inbred Lines (RILs) for gene discovery and deployment. In: Proceedings of the American Society of Sugar Beet Technologists, pp. 124–32. www.bsdf-assbt.org/wp-content/uploads/2017/04/PASSBTAgp124to132BreedingforGeneticsDevelopmentofRecombinantInbredLinesforGeneDiscoveryandDeployment.pdf (16 June 2019, date last accessed).
53. Butterfass, T. 1964, Die chloroplastenzahlen in verschiedenartigen zel- len trisomer zuckerruben (*Beta vulgaris* L.), *Z. Bot.*, **52**, 46–77.
54. Keller, W. 1936, Inheritance of some major color types in beets, *J. Agric. Res.*, **52**, 27–38.
55. Hatlestad, G.J., Sunnadeniya, R.M., Akhavan, N.A., et al. 2012, The beet *R* locus encodes a new cytochrome P450 required for red betalain production, *Nat. Genet.*, **44**, 816–20.
56. Hatlestad, G.J., Akhavan, N.A., Sunnadeniya, R.M., et al. 2014, The beet *Y* locus is a co-opted anthocyanin MYB that regulates betalain pathway structural genes, *Nat. Genet.*, **47**, 92–6.

57. Pin, P.A., Benlloch, R., Bonnet, D., et al. 2010, An antagonistic pair of FT homologs mediates the control of flowering time in sugar beet, *Science*, **330**, 1397–400.
58. Pin, P.A., Zhang, W., Vogt, S.H., et al. 2012, The role of a pseudo-response regulator gene in life cycle adaptation and domestication of beet, *Curr. Biol.*, **22**, 1095–101.
59. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2. published online 9 June 2015, doi:10.1093/bioinformatics/btv351
60. Bennetzen, J.L. and Wang, H. 2014, The contributions of transposable elements to the structure, function, and evolution of plant genomes, *Annu. Rev. Plant Biol.*, **65**, 505–30.
61. Kowar, T., Zakrzewski, F., Macas, J., et al. 2016, Repeat composition of CenH3-chromatin and H3K9me2-marked heterochromatin in sugar beet (*Beta vulgaris*), *BMC Plant Biol.*, **16**, 120.
62. Dechyeva, D. and Schmidt, T. 2006, Molecular organization of terminal repetitive DNA in *Beta* species, *Chromosome Res.*, **214**, 881–97.
63. Melters, D.P., Bradnam, K.R., Young, H.A., et al. 2013, Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution, *Genome Biol.*, **2013**, R10.
64. Pichersky, E., Logsdon, J.M., McGrath, J.M. and Stasys, R.A. 1991, Fragments of plastid DNA in the nuclear genome of tomato: prevalence, chromosomal location and possible mechanism of integration, *Mol. Gen. Genet.*, **225**, 453–8.
65. Yang, W.-D., Tan, H.-W. and Zhu, W.-M. 2016, SpinachDB: a well-characterized genomic database for gene family classification and SNP information of spinach, *PLoS One*, **11**, e0152706.
66. Lightfoot, D.J., Jarvis, D.E., Ramaraj, T., et al. 2017, Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution, *BMC Biol.*, **15**, 74.
67. Jarvis, D.E., Ho, Y.S., Lightfoot, D.J., et al. 2017, The genome of *Chenopodium quinoa*, *Nature*, **542**, 307–12.
68. Sato, K., Abe, F., Mascher, M., et al. 2021, Chromosome-scale genome assembly of the transformation-amenable common wheat cultivar 'Fielder', *DNA Res.*, **28**, dsab008.
69. Hamdi, J., Kmeli, N. and Bouktila, D. 2021, Genome-wide survey of sugar beet (*Beta vulgaris* subsp. *vulgaris*) Dof transcription factors reveals structural diversity, evolutionary expansion and involvement in taproot development and biotic stress resistance, *Biologia*, **76**, 2421–36.
70. Hosoda, K., Imamura, A., Katoh, E., et al. 2002, Molecular structure of the GARP family of plant Myb-related DNA binding motifs of the *Arabidopsis* response regulators, *Plant Cell*, **14**, 2015–29.
71. Mason, M.G., Mathews, D.E., Argyros D.A., et al. 2005, Multiple type-B response regulators mediate cytokinin signal transduction in *Arabidopsis*, *Plant Cell*, **17**, 3007–18.
72. Castro, S., Romeiras, M.M., Castro, M., Duarte, M.C. and Loureiro, J. 2013, Hidden diversity in wild *Beta* taxa from Portugal: insights from genome size and ploidy level estimations using flow cytometry, *Plant Sci.*, **207**, 72–8.
73. Pucker, B. 2019, Mapping-based genome size estimation, *bioRxiv*, doi:10.1101/607390, 13 April 2019, preprint: not peer reviewed. https://www.biorxiv.org/content/10.1101/607390v1.article-info
74. Schmidt, T., Jung, C. and Metzlaff, M. 1991, Distribution and evolution of two satellite DNAs in the genus *Beta*. *Theor. Appl. Genet.*, **82**, 793–9.
75. Taguchi, K. 2014, Genetics and breeding studies on *Aphanomyces* root rot resistance of sugar beet, covers from the discovery of genetic resources to development of new varieties, *Breed. Res.*, **16**, 186–91.
76. Taguchi, K., Kuroda, Y., Okazaki, K. and Yamasaki, M. 2019, Genetic and phenotypic assessment of sugar beet (*Beta vulgaris* L. subsp. *Vulgaris*) elite inbred lines selected in Japan during the past 50 years, *Breed. Sci.*, **69**, 255–65.
77. Minio, A., Massonnet, M., Figueroa-Balderas, R., Castro, A. and Cantu, D. 2019, Diploid genome assembly of the wine grape Carménère, *G3*, **9**, 1331–7.
78. French-Italian Public Consortium for Grapevine Genome Characterization. 2007, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, *Nature*, **449**, 463–7.
79. Wang, Y., Xin, H., Fan, P., et al. 2020, The genome of Shanputao (*Vitis amurensis*) provides a new insight into cold tolerance of grapevine, *Plant J.*, **105**, 1495–506.