Molecular Genetics and Genomics (2020) 295:177–193
https://doi.org/10.1007/s00438-019-01614-3

ORIGINAL ARTICLE

A high-quality cucumber genome assembly enhances computational comparative genomics

Paweł Osipowski[1] · Magdalena Pawełkowicz[1] · Michał Wojcieszek[1] · Agnieszka Skarzyńska[1] · Zbigniew Przybecki[1] · Wojciech Pląder[1]

Received: 19 February 2019 / Accepted: 30 September 2019 / Published online: 16 October 2019

Abstract

Genetic variation is expressed by the presence of polymorphisms in compared genomes of individuals that can be transferred to next generations. The aim of this work was to reveal genome dynamics by predicting polymorphisms among the genomes of three individuals of the highly inbred B10 cucumber (*Cucumis sativus* L.) line. In this study, bioinformatic comparative genomics was used to uncover cucumber genome dynamics (also called real-time evolution). We obtained a new genome draft assembly from long single molecule real-time (SMRT) sequencing reads and used short paired-end read data from three individuals to analyse the polymorphisms. Using this approach, we uncovered differentiation aspects in the genomes of the inbred B10 line. The newly assembled genome sequence (B10v3) has the highest contiguity and quality characteristics among the currently available cucumber genome draft sequences. Standard and newly designed approaches were used to predict single nucleotide and structural variants that were unique among the three individual genomes. Some of the variant predictions spanned protein-coding genes and their promoters, and some were in the neighbourhood of annotated interspersed repetitive elements, indicating that the highly inbred homozygous plants remained genetically dynamic. This is the first bioinformatic comparative genomics study of a single highly inbred plant line. For this project, we developed a polymorphism prediction method with optimized precision parameters, which allowed the effective detection of small nucleotide variants (SNVs). This methodology could significantly improve bioinformatic pipelines for comparative genomics and thus has great practical potential in genomic metadata handling.

Keywords Genome assembly · Variant calling · Comparative genomics · Polymorphism detection · Cucumber · *Cucumis sativus* L

Introduction

The scope of this study was to sequence and assemble the nuclear genome of the highly inbred cucumber (*Cucumis sativus* L.) B10 Borszczagowski line, annotate the gene structure, assign functions to the genes, and characterize the genetic variations among three individual plants of the B10 line. Demonstration of individual variability at such a high level of homozygosity (inbred) will reveal the genome dynamics (also called real-time evolution) and help to confirm the continuous character of evolution. The obtained results allowed us to conclude that plant populations have internal systems that allow them to adapt and survive in changing environmental conditions. It is extremely important to have a very good genome reference sequence because resequencing the lines and aligning them to the reference sequence enables the study of natural evolutionary changes

Communicated by Stefan Hohmann.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-019-01614-3) contains supplementary material, which is available to authorized users.

✉ Magdalena Pawełkowicz
magdalena_pawelkowicz@sggw.pl

✉ Wojciech Pląder
wojciech_plader@sggw.pl

[1] Department of Plant Genetics, Breeding and Biotechnology, Institute of Biology, Warsaw University of Life Sciences - SGGW, 159 Nowoursynowska St, Warsaw, Poland

as well as changes caused by breeding processes. Our aim was to estimate the nature of genomic variation in the real-time evolutionary process. Genomic variations range in size from single base pairs to large chromosomal events, and such variations are common in single organisms and in population development. Comparative genomics can reveal detailed genetic variations in the form of small nucleotide variants (SNVs), larger structural rearrangements as structural variants (SVs), and copy number variations (CNVs). Although there is no clear terminology that is accepted by the research community, SNVs are considered to include single nucleotide polymorphisms (SNPs), multiple nucleotide polymorphisms (MNPs) that are variants in a nucleotide sequence, and insertions/deletions (InDels) of up to 50 nucleotides.

The development of next-generation sequencing (NGS) technologies and computational algorithms that can deal with genomic data makes precise variant predictions possible. Rapid genomic advances together with decreasing costs of NGS provide genuine opportunities to characterise genomes by whole-genome sequencing (WGS) of a wide set of genomes (Gudbjartsson et al. 2015; Zhang et al. 2015), individual genomes (Chen et al. 2012), and even the genomes of single cells (Macaulay and Voet 2014). Nevertheless, the length and quality of NGS reads produced by these technologies still require improvement (Schatz et al. 2010). Second-generation reads have a low error rate ($<0.01\%$) but are too short to overcome many problems posed by long repetitive elements in genomes. Third-generation sequencing technologies can produce single molecule real-time (SMRT) reads up to 100,000 bp long, but have an error rate of at least 0.13 (Goodwin et al. 2016).

Many computational tools for calling SNVs and SVs are based on the alignment of WGS short reads to a reference sequence (resequencing) (Van der Auwera et al. 2013; Hwang et al. 2015; Guan and Sung 2016; Kavak et al. 2017). Long reads also have been used in the alignments (Chaisson et al. 2015; Huddleston et al. 2017; Merker et al. 2018). Considering the bias in sequencing technologies and limitations of the algorithms used for variant detection by resequencing, accurate bioinformatic assessments remain a challenge. SNV calling requires a good understanding of variant call filtering to achieve improved accuracy (Li 2014). However, achieving high accuracy in variant calling is difficult, especially for incomplete draft genomes and/or when limited variant data are available.

SNVs and SVs can change the phenotypes of plants (Springer et al. 2009; Saxena et al. 2014; Zhang et al. 2015) and other organisms (McCarroll and Altshuler 2007; Stankiewicz and Lupski 2010; Raphael 2012; Carvalho and Lupski 2016). In plants, SNP-related research is focused on population genotype scanning for diversity estimation or for desirable traits for breeding purposes (Varshney et al. 2009; Uchida et al. 2011; Lindner et al. 2012; Varshney et al. 2014). This is achieved using technologies that range from whole-genome shallow to medium resequencing to genotyping-by-sequencing. The results can be used for high-throughput marker-assisted genotyping, which is useful in plant genetics and breeding (Torkamaneh et al. 2018). Accurate computational variant calling has been achieved for *Arabidopsis thaliana* (Ossowski et al. 2010; Cao et al. 2011), cucumber (Qi et al. 2013), and some other plants (Torkamaneh et al. 2018). The development of bioinformatic methods for accurate WGS at an individual level will strongly facilitate plant genetics research and provide valuable background data for large-scale genomic studies.

Cucumber is a model plant for sex determination and is considered valuable for omics research (Pawełkowicz et al. 2016, 2019). The cucumber genome is estimated to be 367 Mb long (Arumuganathan and Earle 1991), which is small compared with other model crops such as maize (2300 Mb) (Schnable et al. 2009) and soybean (1115 Mb) (Schmutz et al. 2010). Three cucumber genome drafts are available: Chinese line 9930 (GenBank: GCA_000004075.2) (Huang et al. 2009); American Gy14 (http://wenglab.horticulture.wisc.edu/) (Cavagnaro et al. 2010); and North European B10 line B10v1 (GenBank: GCA_000224045.1), which we sequenced previously (PCC Genomics http://csgenome.sggw.pl). B10 is a highly inbred monoecious accession of ‘Borszczagowski’, an old field cultivar from Poland, from which many breeding lines have been derived through mutagenesis, in vitro regeneration, and transgenesis.

In the present study, we performed an initial bioinformatic assessment of the genome dynamics in the B10 line. The aim of the study was twofold: (1) to enhance genomic sequence information by SMRT read assembly and to detect polymorphisms among three individuals B10 plants that were inbred separately for over 20 generations, and (2) to study the SNV rate, genomic features, and event distribution on the B10 chromosomes. Additionally, we designed a novel analytic SNV calling method that we called reciprocal reference variant calling (RRVC) and an additional false positive filter that we called the reference sample variant filter (RSVF). Initial test of their effectiveness was carried out. We assembled a highly improved genome draft, B10v3 (GenBank: LKUO00000000) from PacBio SMRT reads, which has longer scaffolds and contiguity of contigs as well as better quality gene annotations than the previous drafts. The genome dynamics data across generations of the B10 line also are presented. Three individual B10 genomes were sequenced using Illumina paired-end technology and were scanned for genomic variations using bioinformatics tools. We designed a calling method with improved precision rather than sensitivity that uses WGS to produce consensus data from three individual plants in a way that can contribute to more accurate variant recognition. This is the first study to measure genome dynamics at the level of a single breeding line and with no access to polymorphism databases specific to the target genome.

Materials and methods

Plant material cultivation, genomic DNA preparation, sequencing, and read pre-processing

For PacBio sequencing, cucumber plants from the B10 line were cultivated in a greenhouse during the summer 2014 under a controlled photoperiod of 16-h/8-h day/night. Material for Illumina sequencing and comparative genomics was cultivated in a polytunnel during the summer 2014 with no controlled photoperiod. For both cultivation systems, young leaves were collected, frozen in liquid nitrogen, and stored at – 74 °C. For PacBio sequencing, the leaves were harvested from many plants and pooled. For Illumina sequencing, tissue was collected from three individual plants (P1, P2, and P3). DNA was extracted from 1 g of tissue using a DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. The amount and purity of the DNA were determined using a NanoDrop 2000 spectrophotometer and by electrophoresis quality check on a 1% agarose gel. Long SMRT read sequencing was performed using the PacBio RS II system in two phases: P5C3 chemistry and P6C4 chemistry. Paired-end read sequencing was performed using the Illumina HiSeq system set to 100 bp long reads paired in 300-bp inserts for P1, P2, and P3. These plants had a last common ancestor plant 21 (P1, P2) or 22 (P3) generations ago (Fig. 1). After sequencing, the short read data were trimmed of adapters and low-quality bases using Trimmomatic v0.35 (Bolger et al. 2014) to leave reads no shorter than 50 bp in length. Before and after trimming, the read base quality was checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

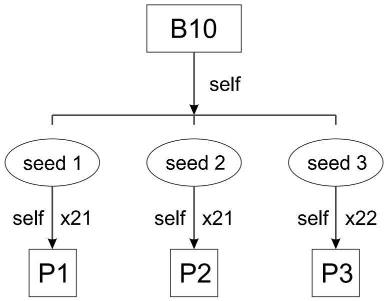


Fig. 1 Pedigree tree of the three cucumber plants from the B10 line (P1, P2, P3) used in this study

Genome reference sequence assembly and correction

A genome sequence draft was assembled from all the obtained PacBio reads using the PBcR pipeline with Celera Assembler v8.3rc2 (Berlin et al. 2015), which is suited to long SMRT reads (Fig. 2, step A). After assembly, the contigs were corrected and quality filtered by long reads genome alignment using PacBio Pbalign v3.0 (https://github.com/PacificBiosciences/pbalign/blob/master/doc/howto.rst) for the SMRT read alignment and the variant calling algorithm Quiver v2.1.0 (https://github.com/PacificBiosciences/GenomicConsensus) for the correction. For the raw draft correction, only the newest 27 × P6C4 PacBio reads were used as input data (Fig. 2, step B). The output draft sequence (B10v2) underwent an NCBI foreign contamination screening process to filter out contigs with mitochondrial, plastidial, and other foreign elements, and was then submitted to GenBank (GCA_001483825.2). The B10v2 sequence was corrected using the short Illumina reads from P1. The BBnorm Ecc Linux shell script from the BBTools v35.82 suite (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/) was used to correct the quality of the P1 Illumina reads by k-mer distribution count-based modification. The P1 processed reads were aligned to the genome draft sequence using Bowtie2 v2.2.9 (Langmead and Salzberg 2012). Samblaster v0.1.24 (Faust and Hall 2014) was used to deduplicate aligned read pairs. The obtained alignment results were input to Pilon v1.20 (Walker et al. 2014) to correct the contig sequences. The genome draft sequence corrected using the P1 data was considered the final version of the genome draft sequence (B10v3) and was used as the reference genome for the subsequent analyses (Fig. 2, step C).

Corrected contigs also were mapped to the cucumber chromosomes using the marker primer pairs from the most recent cucumber consensus map (Yang et al. 2013) to compare the mapping with previous findings (Wóycicki et al. 2011) about rearrangements between investigated B10 cucumber line and the cucumber 9930 line reference genome. First, we excluded all duplicated sequence pairs from the accessible map data file. Then we aligned the remaining marker primer pairs to the genome B10v3 sequence using BWA software v0.7.13 (Li and Durbin 2010). Primer pairs with mapping quality < 40 (Phred) were rejected. The marker positions on the chromosomes and the primer pair B10v3 contig alignment position were used to detect discrepancies in sequence consecutiveness between the two genomes. The occurrences of large intra-chromosomal translocations and inter-chromosomal genome

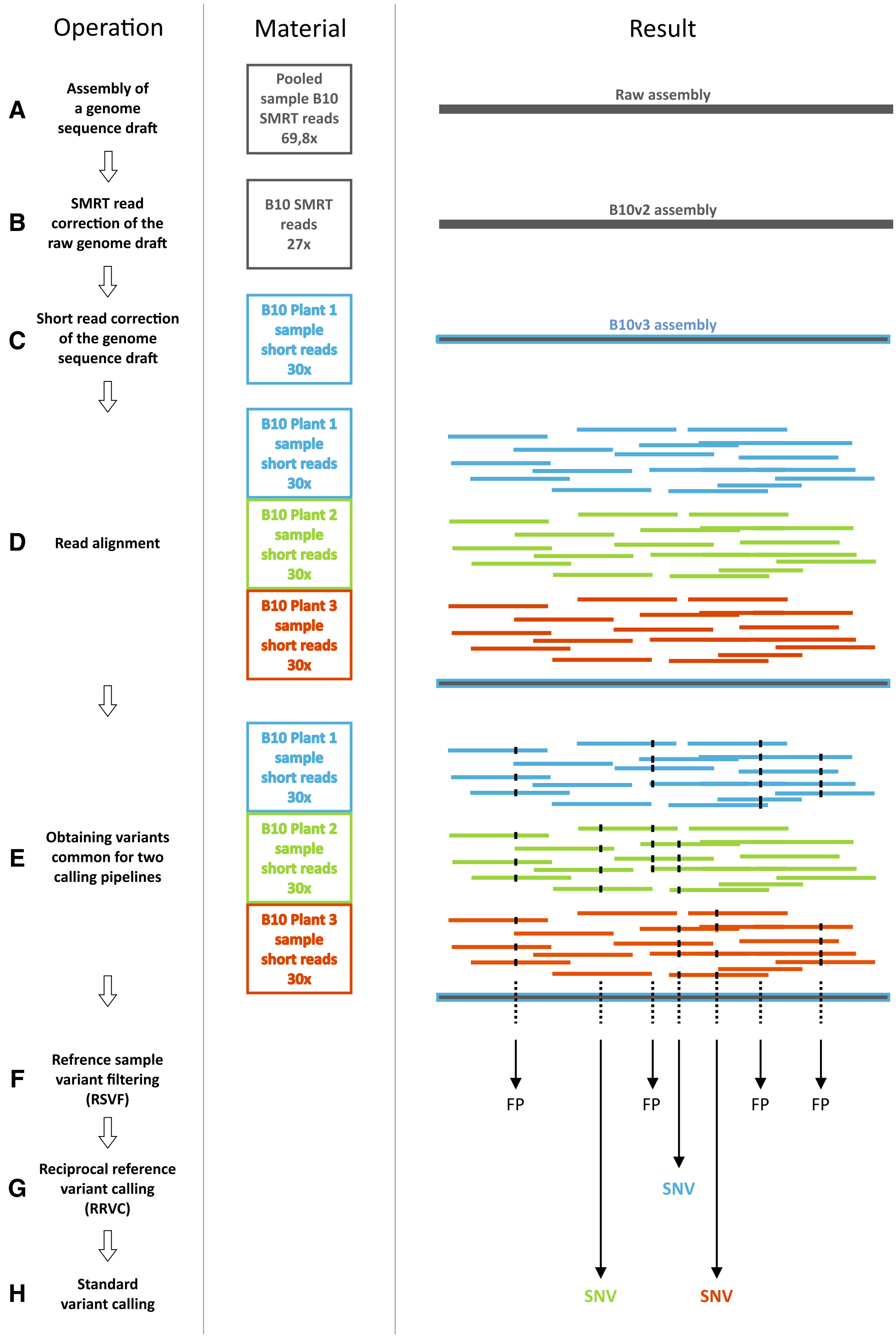

Operation
Material
Result
A
Assembly of a genome sequence draft
Pooled sample B10 SMRT reads 69,8x
Raw assembly
B
SMRT read correction of the raw genome draft
B10 SMRT reads 27x
B10v2 assembly
C
Short read correction of the genome sequence draft
B10 Plant 1 sample short reads 30x
B10v3 assembly
D
Read alignment
B10 Plant 1 sample short reads 30x
B10 Plant 2 sample short reads 30x
B10 Plant 3 sample short reads 30x
E
Obtaining variants common for two calling pipelines
B10 Plant 1 sample short reads 30x
B10 Plant 2 sample short reads 30x
B10 Plant 3 sample short reads 30x
F
Refrence sample variant filtering (RSVF)
FP
FP
FP
FP
G
Reciprocal reference variant calling (RRVC)
SNV
H
Standard variant calling
SNV
SNV

Fig. 2 Genome draft assembly, variant calling, and analysis pathway. The box arrows indicate the order of operations. B10 cucumber line plant samples and derived sequence information are colour-coded: blue (Plant 1), green (Plant 2), and red (Plant 3). Dark grey indicates a pooled DNA sample from many B10 line plants, sequenced by SMRT technology. Long thick horizontal line represents the genome reference sequence. Thin, short lines represent short reads after Illumina sequencing aligned to the reference sequence. Vertical dotted lines crossing the reference line represent the aligned read set of each raw variant called in a certain location. Analysis of the raw variant calling results was performed in three subsequent steps: reference sample variant filtering, reciprocal reference variant calling, and standard variant calling. Black arrows represent the location of final variant call results to raw variant call results. FP, filtered-out false positives; SNV, unique variant called for the specific plant sample (colour figure online)

rearrangements were counted. The marker information was used to determine inter-chromosomal rearrangements when more than two different chromosome sequences were found on one B10v3 contig. Intra-chromosomal rearrangements were determined when the consecutiveness of specific chromosome sequences was deranged on a contig. The results obtained using the rearrangement counting method based on map markers are highly dependent on map density, which does not allow the precise measurement of the sizes of large events between two lines. The mapped contigs were used to compare the genomic distributions of some genomic features. The density of predicted variants, genes, interspersed repetitive elements (IREs), and repeats was assessed by counting the start positions of every feature and event in

200-kb genomic sequence consecutive bins and drawing graphical representations of their relative positions in a whole chromosome context. Additionally, maximum weight match scaffolding of B10v3 contigs was performed with ScaffMatch v0.9 (http://alan.cs.gsu.edu/NGS/?q=content/scaffmatch) using BAC end sequences from the B10 cucumber BAC library characterised by Gutman et al. (2008).

We compared the B10v3 draft genome with the three currently accessible cucumber genomes: Gy14 cucumber line genomic draft v1.0 (Gy14) sequence (downloaded from http://wenglab.horticulture.wisc.edu/), 9930 reference genome sequence (9930) for cucumber (GenBank: GCA_000004075.2), and B10v1 (GenBank: GCA_000224045.1). We applied Perl scripts to compare contigs after modifying the other cucumber genome sequences by splitting them wherever 'N' occurred in the sequences. Draft sequence quality assessment was performed by evaluating the gene structure assembly using the Benchmarking Universal Single-Copy Orthologs (BUSCO) program v3, locating near-universal single-copy orthologues for plants (Simão et al. 2015). B10v2 contigs and B10v3 scaffolds also were assessed for quality parameters with BUSCO by comparing them with the final B10v3 Illumina-corrected contigs. The B10 line genome sequence length and single copy sequence length were estimated using the information about 31-mer frequency distribution in Illumina read sets from individual plant genomes (https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/).

Genome annotation

The B10v3 genome draft was annotated using a variety of bioinformatics tools. CpG islands were detected using CpGcluster (Hackenberg et al. 2006). Barrnap v0.7 (http://www.vicbioinformatics.com/software.barrnap.shtml) was used to predict rRNA genes, and ARAGORN v1.2.26 (Laslett and Canback 2004) was used to predict tRNA genes. Other non-coding RNAs, namely, small nuclear (sn) RNAs, small nucleolar (sno) RNAs, ribozymes, and miscellaneous (misc) RNAs were annotated using Infernal v1.1.2 (Nawrocki and Eddy 2013) and the Rfam database (Nawrocki et al. 2015). RepeatMasker v4.7.0 (http://www.repeatmasker.org/) was used to annotate repetitive elements in the genome, including single sequence repeats (SSRs), low copy repeats (LCRs), and IREs. Two types of IRE databases were used as input: (a) RepeatMasker-implemented databases built from the Dfam consensus HMM database release 20170127 (http://www.dfam-consensus.org) and RepBase release 20170127 (http://www.girinst.org/); and (b) the database of consensus models of putative repetitive elements for B10v3, created using RepeatModeler v1.0.4 (http://repeatmasker.org/RepeatModeler/). For both RepeatMasker and RepeatModeler we used Rmblastn v2.6.0 + (Camacho et al. 2009) for the alignments.

The gene structural annotation of B10v3 was performed in two steps: the main transcriptome-based method and the ab initio unsupervised learning approach. For the transcriptome-based annotation, Illumina paired-end RNA-seq data from 150 transcriptome samples of different organs: leaves, fruits, shoot apex, and floral buds (four developmental stages: 1–2 mm, 3–5 mm, 6–8 mm and 9 mm), extracted from 21 different cucumber lines were applied. The transcriptome data were obtained from the Polish Consortium of Cucumber Genome Sequencing (http://csgenome.sggw.pl/en-us/). Illumina adapters, poor quality regions, and sequences matching cucumber rRNAs were removed from RNA-seq reads using BBDuk2 software from BBTools suite v36.32 (http://jgi.doe.gov/data-and-tools/bbtools). FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to assess the short read quality before and after read correction. De novo RNA-seq assembly was performed using Trinity v2.3.2 (Grabherr et al. 2011) for strand-specific libraries, and rnaSPAdes v3.9.0 (http://cab.spbu.ru/software/rnaspades/) for non-strand-specific libraries. Expression of the assembled transcripts was estimated separately for each sample using Salmon v0.7.2 (Patro et al. 2017). Only transcripts with an expression value of at least 1 transcript per million were kept for further analysis. Assembled transcripts were used to annotate the genome using PASA v2.1.0 (Haas et al. 2003). Only spliced alignments that met the following criteria were used for annotation: minimum 75% of the assembled transcript aligned to the genome, minimum 95% alignment identity, perfect match required for three nucleotides flanking the splice boundaries, minimum intron length of 20 bp, and maximum intron length of 10,000 bp. TransDecoder v3.0.1 (Haas et al. 2013) was used to predict coding regions within transcripts annotated by PASA. For each transcript, only the single best open reading frame was retained, considering the length of the coding region and Blastp/PFAM homology information.

For the complementary unsupervised learning annotation, we used Braker v1.1 pipeline, wrapping GeneMark-ES Suite v4.3.2, and Augustus v3.2.3 algorithms (Hoff et al. 2016). Repetitive elements and rRNA and tRNA genes detected using the main transcriptome-based approach and (secondly) annotated by us as IREs in B10v3 were masked. GeneMark was trained with pooled RNA-seq data from four tissues (leaves, fruits, shoot apex, and floral buds) and two developmental stages (1–2 mm and 3–5 mm) of the B10 line, and mapped using TopHat2 v2.1.1 (Kim et al. 2013). The unsupervised learning approach was used to extract gene structures unannotated by the transcriptome-based method. All annotated genomic structures were compared using Bedtools v2.26.0 (Quinlan 2014). The final structural annotation was compared with the annotated genomes of *A. thaliana*

(GenBank: GCF_000001735.4) and melon (*Cucumis melo*) (GenBank: GCF_000313045.1) by determining the overall genome protein length ratio. The translated protein sequences from the three genomes were aligned by Blastp 2.6.0+(Camacho et al. 2009).

Symbols, descriptions, orthology groups, plant gene ontology (GO) slim terms, and functional domains were assigned to the genes using eggNOG-mapper v1.0.3 (Huerta-Cepas et al. 2016, 2017) and InterProScan v5.30-69.0 (Jones et al. 2014). After defining the protein-coding genes in B10v3, they were compared with the annotated protein-coding genes in the published cucumber genome drafts (9930, Gy14, and B10v1).

Variant calling, comparative genomics, and SNV call verification

To define the dynamics of the B10 genome in terms of polymorphisms, we generated unique SNV, SV, and CNV predictions for each plant sample. Pre-processed plant sample read sets were quality-corrected using the BFC tool r181 (Li 2015a). The reads from each plant were aligned to B10v3 using BWA v0.7.13 (Fig. 2, step D) (Li and Durbin 2010). The alignment output was deduplicated using Samblaster v0.1.24 (Faust and Hall 2014).

To call CNVs, we used CNVnator v0.3.2 (Abyzov et al. 2011) and to call SVs we used FermiKit r178 (Li 2015b), both with default settings. We used two separate pipelines to call SNVs and produce consensus results of variability: FreeBayes v1.1.0-3-g961e5f3 (https://arxiv.org/abs/1207.3907) (https://github.com/ekg/freebayes) and DeepVariant v0.4.1 (https://github.com/google/deepvariant). Outputs for each variant calling software were stored in VCF, GFF, or BED file format. Comparisons and manipulations of the variant data files were performed using Bedtools v2.26.0 (Quinlan 2014) and Bcftools v1.4-6-g5349659 (http://samtools.github.io/bcftools/). The two SNV calling programs were selected because of their strong algorithm diversification and improved chance of obtaining accurate results.

On the basis of the Li 2014 study (Li 2014), we applied the following filters to the results of FreeBayes to increase the accuracy of this pipeline: low-complexity region exclusion, maximum read depth, unbiased double strand coverage, quality filter of minimum variant quality 30 (Phred), and minimum read depth > 30. Mdust script (https://github.com/lh3/mdust) was used to detect repetitive sequences as an input for the low-complexity filtering. For DeepVariant, the raw results that passed the two built-in quality filters were used for further analysis. Separate read sets for each plant sample were used for SNV calling with the exception of the P2 and P3 read alignments in the FreeBayes pipeline, which had a stronger discriminative potential when many samples were used. Both calling pipelines also estimated the genotype for every predicted SNV. Calls that were assigned a homozygous reference genotype or no genotype was specified were filtered out from each result set before subsequent analysis. The final outputs from the two pipelines were used to generate a consensus result by selecting SNV calls that were common to both pipelines for each plant sample (Fig. 2, step E). By comparing the P2 and P3 called variant sets, calls were extracted that were unique for each plant (Fig. 2, step H). If P2 and/or P3 read-derived SNV calls had an equivalent in the P1-derived result, it was considered a false positive (FP) because all variant calling results from the P1 read alignment to B10v3 were assumed to be an overall method bias effect. In this way, we applied the reference sample variant filter, RSVF (Fig. 2, step F). Next, RRVC was performed and common P2 and P3 calls were extracted to determine polymorphisms unique to the P1 genome (Fig. 2, step G). First, all heterozygous calls common to P2 and P3 were removed because B10v3, being a flattened diploid genome representation of one allele from two possible alleles, limits the prediction to homozygous calls. This step can significantly increase the specificity of the method at the cost of sensitivity.

Genes, exons, and up to 1000-bp long upstream sequences from the start codon of the predicted genes (upstream promoters) and 500-bp long downstream sequences from the stop codon (downstream promoters) were checked for overlap with the positions of the SNVs, SVs, CNVs, and IREs. All IRE predictions with no additional filters were included in comparison. To differentiate between exonic sequence derived from primary transcriptome-based approach and IRE we applied a filter for IRE sequences to be no more than 5% diverged from matching consensus sequence used for IRE annotation. The CNV and SV locations also were compared with the IRE locations to predict active transposable elements (TEs). TEs were specified if their location on a genome sequence intersected with that of a rearrangement or if at least one of the ends was ≤ 10 bp from the rearrangement.

From the set of identified unique plant SNVs, 34 randomly selected calls were processed for validation by PCR amplification and Sanger sequencing. Primers for the PCR amplifications were designed based on the B10v3 reference sequence. For every selected SNV location, Sanger sequencing was performed for the DNA sequence unique to the plant with the predicted SNV and for the corresponding DNA sequence from one of the other plant samples. If the unique plant sequence was recognised as modified following the SNV prediction and the other plant sequence matched the reference sequence, the SNV prediction was validated. Precision (positive predictive value, PPV) for each plant SNV set was calculated as a fraction of true positives (TPs) among all the results verified: PPV = TP ÷ (FP + TP). After PPV estimation, the mutation rate per base per generation was assessed for each plant using the newly calculated B10 line genome length estimation and the PPV parameter for result normalisation.

Results

Sequencing and read alignment

PacBio sequencing was performed in two phases using P3C5 and P4C6 chemistry. The P3C5 read set was an average of 3119-bp long (maximum length, 35,190 bp) with approximately 42.8-fold genome depth (42.8 ×) coverage. The P4C6 read set was an average of 6733-bp long (maximum length, 45,973 bp) with 27 × coverage (Table S1). The Illumina read data had 40.79 × coverage for P1, 37 × for P2, and 36 × for P3. Adapter and low-quality base trimming filtered out about 4% of the reads in every set. The mean quality assessment in the Phred scale of the read sets from each sample was 28 for the raw read sets and 33 after the reads were corrected.

Read correction did not affect the overall read numbers, but it reduced the occurrences of unique k-mers in the read sets to approximately 48% with BBtools and 12% with BFC tools. For the raw genome draft sequence correction with P1 short reads, 91.52% of the reads were aligned to the genome draft, with at least one read coverage spanning about 96% of the B10v2 draft sequence. Aligned reads covered an average depth of 40 × for B10v3 bases, and 59.78% of all reads were properly paired after alignment.

For variant calling, 93.50%, 92.35%, and 93.71% of the aligned reads contained 90.85%, 89.73%, and 91.23% properly paired reads of the P1, P2, and P3 samples, respectively, and 91%, 92%, and 91%, of the B10v3 sequence had at least 1 × coverage of aligned reads of P1, P2, and P3, respectively. Aligned P1 reads covered an average depth of 40 × for B10v3 bases, and close to 36 × for the aligned reads of both P2 and P3.

Genome reference sequence assembly

The obtained raw assembly draft (from PacBio P3C5 and P4C6 reads) was 343.58-Mb long and contained 8096 contigs with a maximum length of 12.66 Mb and 842-kb N50 parameter. After the first read correction, none of the main parameters of the new B10v2 assembly differed from those of the raw assembly, except the total contig length increased slightly to 343.85 Mb. Screening of B10v2 identified 61 contigs that were contaminated with foreign sequences (detected by BLAST searches of the NCBI database). Illumina read correction of the P1 sample resulted in the final B10v3 draft, which was 342.29-Mb long (93.27% of the genome length) and contained 8035 contigs with a maximum length of 12.67 Mb and 858 kb N50 parameter. Statistics for the contigs at the three stages, raw genome draft, B10v2 (long read corrected draft), and B10v3 (Illumina-corrected draft), are presented in Table 1.

The completeness assessment with BUSCO software, which evaluated near-universal single-copy orthologue

Table 1 Statistics for the contigs in the raw draft, SMRT-corrected (B10v2) draft, and Illumina-corrected B10v3 reference draft

	Raw genome draft	B10v2	B10v3
No. contig sequences	8096	8096	8035
Total nucleotides in contigs (Mb)	343.58	343.85	342.29
Genome coverage (%)	92	92	93
Max contig length (Mb)	12.66	12.66	12.67
N50 contig length (kb)	842	842	858
L50	56	56	55

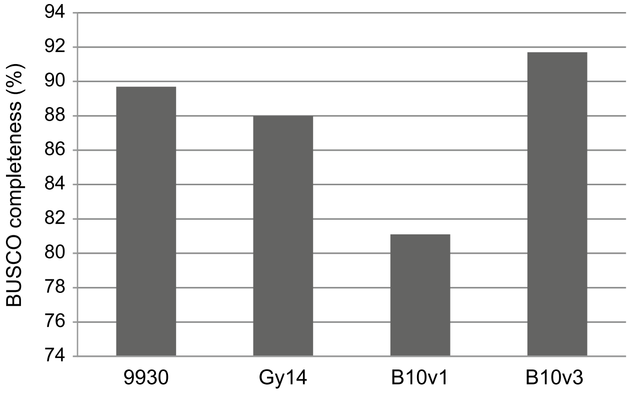


Fig. 3 Completeness assessments using BUSCO v3 of the three assessable cucumber genome drafts and the Illumina-corrected B10v3 reference draft 9930, cucumber reference genome; Gy14, cucumber line genomic draft v1.0; B10v1, published B10 genome draft; B10v3, current B10 genome draft short read corrected

Table 2 Statistics for the contigs in the three accessible cucumber genome drafts

	9930	Gy14	B10v1[a]	B10v3[b]
No. sequences	11,366	13,604	16,454	8035
Total bases (Mb)	195.7	192.3	193.2	342.29
Genome coverage (%)	53	52	52	93
N50 length (kb)	42.3	37.6	23.2	858
L50	1290	1476	2417	55

[a] B10v1 is the first B10 line genome assembly

[b] B10v3 is the Illumina-corrected B10 cucumber line reference genome draft

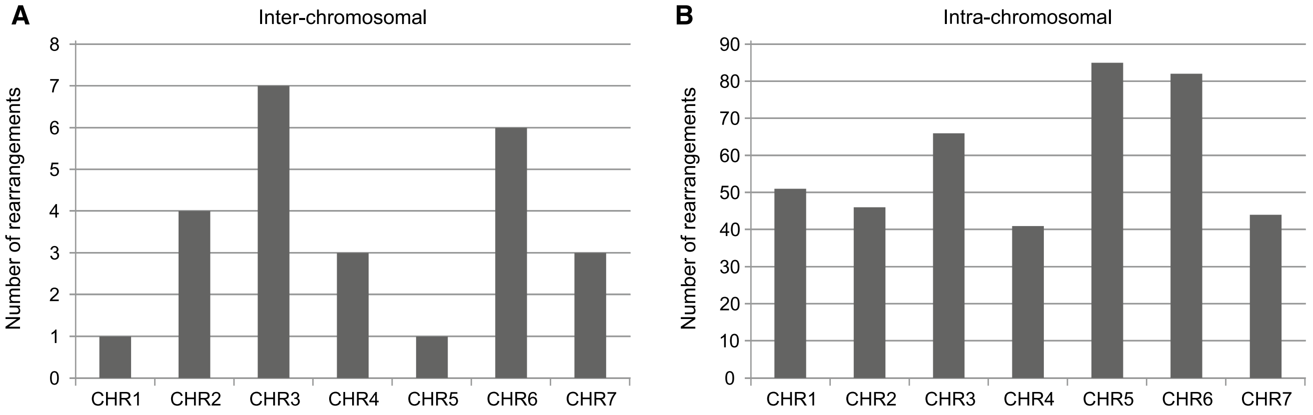


Fig. 4 Chromosomal distribution of inter-chromosomal (**a**) and intra-chromosomal (**b**) events detected between B10 (B10v3) and 9930 cucumber lines by mapping B10 contigs on 9930 chromosomes. One event was counted more than once depending on the number of chromosomes involved

structures present in the B10v3 genome, draft detected 91.7% complete and 1.9% fragmented gene sequences. This result was better than those obtained for the Gy14, 9930, and B10v1 sequences (Fig. 3, Table 2). The B10v3 scaffold completeness check gave a worse result (91.3%) than for the contigs. We also compared B10v3 with B10v2 which had not been short read corrected, and observed identical gene completeness metrics for both sequences. Our results also showed that short read assemblies of 9930 and Gy14 were of relatively high quality, and differed in completeness metrics from B10v3 by only 2% and 3.7%, respectively.

We mapped 119 contigs (1.48%) from a single copy sequence of 196.6 Mb of the B10v3 genome, which made up 57.4% of the reference length (53.6% of the 367-Mb cucumber genome length), onto the seven cucumber chromosomes. Chromosome 3 had 25.3% of the mapped contigs, and the remaining chromosomes had from 4.4% to 19.9% of the total mapped contig length. Out of the 1656 unique marker primer pairs available, 1480 (89.3%) aligned uniquely to B10v3 contigs. A primer alignment quality threshold of 40 (Phred) eliminated the occurrence of primer pairs that aligned with high quality more than once to the reference sequence. About 200 primer pairs were rejected because of low mapping quality, which may indicate sequence diversification between the genomes. For 24 pairs only one primer was mapped with sufficient quality. None of the filtered primer pair alignments was split between two contigs, which confirmed solid Chinese map transition to the B10v3 genome. By comparing marker mapping chromosome positions on B10v3 contigs and the 9930 draft sequence, we found 25 inter-chromosomal events (Fig. 4a) and 415 intra-chromosomal large rearrangements (Fig. 4b). Most intra-chromosomal events were complex, comprising more than one rearrangement. This is consistent with previous findings about diversified intra-species chromosome sequences between North European and Asian cucumber genomes (Wóycicki et al. 2011). By analysing k-mer distribution from three Illumina read sets, the B10 line haploid genome length was estimated to be 413.6 Mb and a single copy sequence of the genome was estimated to be about 196 Mb.

Table 3 Coverage of interspersed repetitive elements (IREs) and transposable element (TE) classes related to the length of the B10v3 draft genome sequence

Transposable element class	No. interspersed elements	Percentage of genome length
DNA TEs	41,280	4%
LINE	23,557	5%
LTR	73,855	21%
Rolling circle	3567	1%
SINE	37	0.00%
Other	7	0.00%
Unknown	96,476	19%

Annotation of the genome draft reference sequence

We identified 141,992 SSRs, 39,510 LCRs, and 29,791 CpG islands in the B10v3 draft sequence. IRE prediction identified 238,779 called structures covering 49.31% of the B10v3 total length. A total of 142,303 sequences were of known TE classes: 41,280 DNA TEs, 3567 rolling-circles, 73,855 long terminal repeats, 23,557 long interspersed nuclear elements, and 37 short interspersed nuclear elements; 96,476 TE sequences were of unknown class (Table 3).

The transcriptome-based gene structure predicted 21,714 genes with an average of 8.48 exons per transcript. The unsupervised learning method assessment resulted in 23,673 genes with an average of 5.41 exons per transcript. A total

Table 4 Numbers of annotated genes in the B10v3 reference genome

Gene prediction method	Genes predicted	Genes with COG assigned	Genes with COG function described	Genes with COG gene assigned	InterProScan match	Genes with GO slim annotation
TB	16,104	15,456	14,240	2919	14,102	11,441
UL	5557	987	806	125	3777	1464
Total	21,661	16,443	15,046	3034	17,879	12,905

TB, transcriptome-based method; UL, unsupervised learning method; COG, clusters of orthologous groups; GO, gene ontology

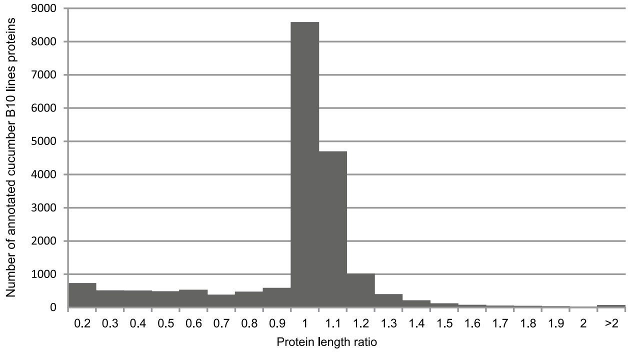


Fig. 5 Distribution of the protein length ratios between the predicted proteins in B10v3 and the top matches in the *Arabidopsis thaliana* and melon (*Cucumis melo*) genomes

of 18,116 gene structures from the unsupervised learning method spanned the same regions as 18,573 structures from the transcriptome-based method. Therefore, the additional unsupervised learning prediction identified 5557 additional structures with 5676 allelic variants and an average of 1.79 exons per transcript. Altogether, 27,271 genes were predicted with an average of 7.09 exons per transcript for all the genes. The longest gene was about 98 kb, 19 genes were ≤ 150 bp, and the average gene length was 4177 bp. In B10v3, fewer genes were specified, but they had a higher number of exons than the genes in the B10v1 draft sequence of Wóycicki et al. (2011). Although the overall number of predicted genes was smaller (by approximately 2500) in the B10v3 reference genome draft then in B10v1, 665 additional genes were annotated.

In addition, 23 miRNAs, 191 snoRNAs, 111 snRNAs, 2543 rRNAs (of which 56 were mitochondrial rRNAs), and 1835 tRNAs were annotated in B10v3. A total of 5610 (approximately 20%) and 197 of the transcribed genes were annotated as long intervening non-coding (linc) RNAs and miscRNAs, respectively. These numbers are likely to be higher if both a larger variety of RNA-seq samples is used for the annotation and software specifically designed for element prediction are used. Nevertheless, our annotations can be used to explore long non-coding RNAs that may be involved in key biological processes in plants (Liu et al. 2012; Zaynab et al. 2018). Of the 21,661 protein coding genes, 83% had an InterProScan match and 76% were assigned to COGs (clusters of orthologous groups) by eggNOG-mapper (Table 4). Among the InterProScan and eggNOG predictions, 568 and 2368, respectively, had hypothetical, putative or expressed function descriptions or no description at all. This is significantly less than the approximately 8000 genes that were annotated as hypothetical function in the B10v1 draft (Wóycicki et al. 2011). GO slim terms were assigned to 11,441 genes from the transcriptome-based method and 1464 from the complementary unsupervised learning method, giving a total of 12,905 GO slim annotated genes (Table 4, Figures S1–S3). In B10v1, approximately 450 fewer genes had GO terms assigned, indicating the GO annotations were similar in both genome drafts. In the other two genome drafts, 23,248 and 21,491 protein-coding genes were predicted in the 9930 line (Li et al. 2011) and in Gy14 (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Csativus), respectively. Herein we predicted 1578 less protein-coding genes than in the 9930 line, but the number of exons per transcript was 1.78 higher in B10 line. Next, we assessed the protein length ratio between our final B10v3 annotations and those of *A. thaliana* and *C. melo*. Blastp searches detected 19,356 proteins from our annotations that matched proteins in the other two genomes. We selected the top matches for comparison. Approximately 69% of the matched B10 line protein sequences had length ratios to 1.0 with the corresponding proteins in the other two genomes. Approximately 10% of the annotated proteins from B10v3 were longer and 18% were shorter than the corresponding proteins in the other genomes (Fig. 5).

A comparison of the B10v3 gene annotations with IRE locations detected 9638 genes, 374 exons, and 16,061 promoters that appeared to have at least one IRE in their sequence (Table 5). Exons of genes annotated using the supervised learning method did not have overlapping IREs because IREs were masked before annotation. Many IRE sequences overlapped with more than one exon location.

Genomic variant analysis and mapping

The DeepVariant and FreeBayes pipelines produced 42, 20, and 33 SNV consensus calls for the P1, P2, and P3 samples,

Table 5 Numbers of predicted genomic structural features that overlapped with interspersed elements in the cucumber B10v3 genome assembly

	Genes	Exons	Promoters	Upstream promoter regions[a]	Downstream promoter regions[b]
No. of features overlapped with IREs	9638	374	16,061	11,492	11,422
No. of IRE overlapped with features	30,789	487	33,141	18,586	18,326
No. of classified TE overlapped with features	2872	46	3580	1833	2004

[a] Upstream promoter region, up to 1000-bp long sequence from the start codon of a predicted gene

[b] Downstream promoter region, 500-bp long sequence from the stop codon of the gene

Table 6 Numbers of called single nucleotide variants (SNVs) by type and presence in genes of P1, P2, and P3

	Total SNVs[a]	SNPs	InDels	MNPs	SNPs in		InDels in	
					Exons	Promoters	Exons	Promoters
P1	42	17	19	6	2	0	2	0
P2	20 (17)	15	2	3	0	2	1	0
P3	33 (29)	30	0	3	1	0	0	0
Total	95 (46)	62	21	12	3	2	3	0

[a] Numbers in brackets are the number of heterozygous genotypes assigned to SNV calls

Table 7 Numbers of predicted copy number variations (CNVs) and structural variants (SVs) unique to each of three plant samples, and numbers of these rearrangements that intersect with gene structures

	CNVs predicted	Genes with CNVs	CNVs in genes	Exons with CNVs	CNVs in exons	Promoters with CNVs	CNVs in promoters
Plant 1	258	115	60	367	60	123	61
Plant 2	240	141	75	477	75	155	78
Plant 3	128	52	35	186	35	52	33
	SVs predicted	Genes with SVs	SVs in genes	Exons with SVs	SVs in exons	Promoters with SVs	SVs in promoters
Plant 1	14	6	8	9	7	3	3
Plant 2	17	6	5	3	1	4	4
Plant 3	30	10	10	34	9	8	7

More than one structural rearrangement (CNV) can occur in features such as genes or exons but it is counted once, even if there are many CNVs within a single feature. Therefore, there might be fewer exons with CNVs than CNVs in exons because some exons have more than one CNV. A CNV also may occur in two different features as features can be found on both forward and reverse DNA strands

respectively (Table S2). Each pipeline produced at least 200 times more separate calls than the final consensus result. The DeepVariant calls were much higher than the FreeBayes calls, which reflects the many yet unchallenged methodological uncertainties in calling SNVs (Table S3). All common results were consistent between the pipelines, although one call of SNP by one method was called as InDel by the other. After further analysis, we defined it as InDel because the DeepVariant results are generally more accurate. The P1 call set contained a significantly higher InDel to SNP ratio (1.12) than the P2 and P3 sets. The P3 call set had no InDels. Eight SNV calls were allocated in genes, six were found in the exons of different genes, and two were in the promoter region (Table 6). Five, four, and one SNV calls were found in functional loci in P1, P2, and P3, respectively. Two of the genes were assigned hypothetical function. The gene with a SNP called in the upstream promoter region was annotated as thioredoxin-like fold glutathione S-transferase and a gene with a SNP called in an exon was annotated as putative homologue of carbon catabolite repressor protein 4. Two ribosomal subunits, one lincRNA, and two genes with unknown function had an SNV predicted in exons (Table S4). Approximately twice as many called SNPs were G:C → A:T transitions rather than A:T → G:C transitions, and the majority of them were outside genes and their promoters.

The SV prediction resulted in 61 SV calls, of which 54 (89%) were translocations, five were deletions, and two were insertions. Most of the SV predictions were for the P3 sample ($n = 30$). About 38% of the SV calls were within gene structures, and approximately 28% intersected with exons. Close to 23% were found in promoter sequences (Table 7). CNV prediction resulted in 626 events, of which 60% were deletions and the remainder were duplications. Nearly half as many unique CNVs were predicted for the P3 sample. Approximately 27% of CNV locations intersected with genes, exons, or promoters, and many CNVs spanned more than one genomic feature (Table 7).

By assessing IRE surroundings for any predicted rearrangements such as CNVs and SVs, we detected all the CNVs that were up to 10 bp away or that intersected with IRE locations. Most of these IREs were not assigned to any known TE class (Table 8). The IREs that were in close proximity to rearrangements affected six predicted gene structures: four in the P1 sample and two in the P2 sample. None of these predicted IREs were within an exon or an assigned TE class.

All predicted genomic features and events were positioned on chromosomes relatively to the mapped contig positions (Table S5). Approximately 23% of predicted events, 57% of all IREs (12% of known class TEs), and 84% of genes were positioned on the chromosomes. Most variants were not assigned to chromosomes and repetitive regions of the genome were difficult to assign precisely to chromosomes by sequence alignment. This indicates that most genomic sequence dynamic activity was in repetitive regions of the B10 line genome. The chromosomal distribution of genomic features showed that most of the gene dense regions overlapped with TE dense regions; however, some regions (up to about 1-Mb long), even regions that were densely populated with genes, had no classified TE or repetitive elements (Fig. 6).

A total of 34 SNV predictions were selected randomly for PCR amplification and Sanger sequencing verification as follows: 16 (nine SNPs and seven InDels), 10 (eight SNPs and two InDels), and eight SNPs from P1, P2, and P3 samples, respectively (Table S6). The PCRs for 10 calls (six SNPs and four InDels) resulted in no product. Sanger sequencing for one SNP called in P3 could not be read precisely, and primers could not be designed for two calls. In the verified sample subset, 12 out of 23 SNVs were validated; two in P1 and three in P2 were FPs. None of the predictions was validated in P3. Thus, a precision of 0.44, 0.8, and 0.52 for SNPs, InDels, and the entire variant set, respectively, was obtained. Precision estimations for the sample subset were 0.78, 0.63, and 0.0 for P1, P2, and P3, respectively (Table 9). The negative results for the P3 sample suggests that the P3 reads may be of lower quality than the reads from the other samples, but this requires advanced scrutiny by, for example, k-mer read set analysis. By normalising SNV calling numbers with a precision parameter, an average mutation rate per generation per base of 1.74×10^{-9} was obtained for the three plants.

Discussion

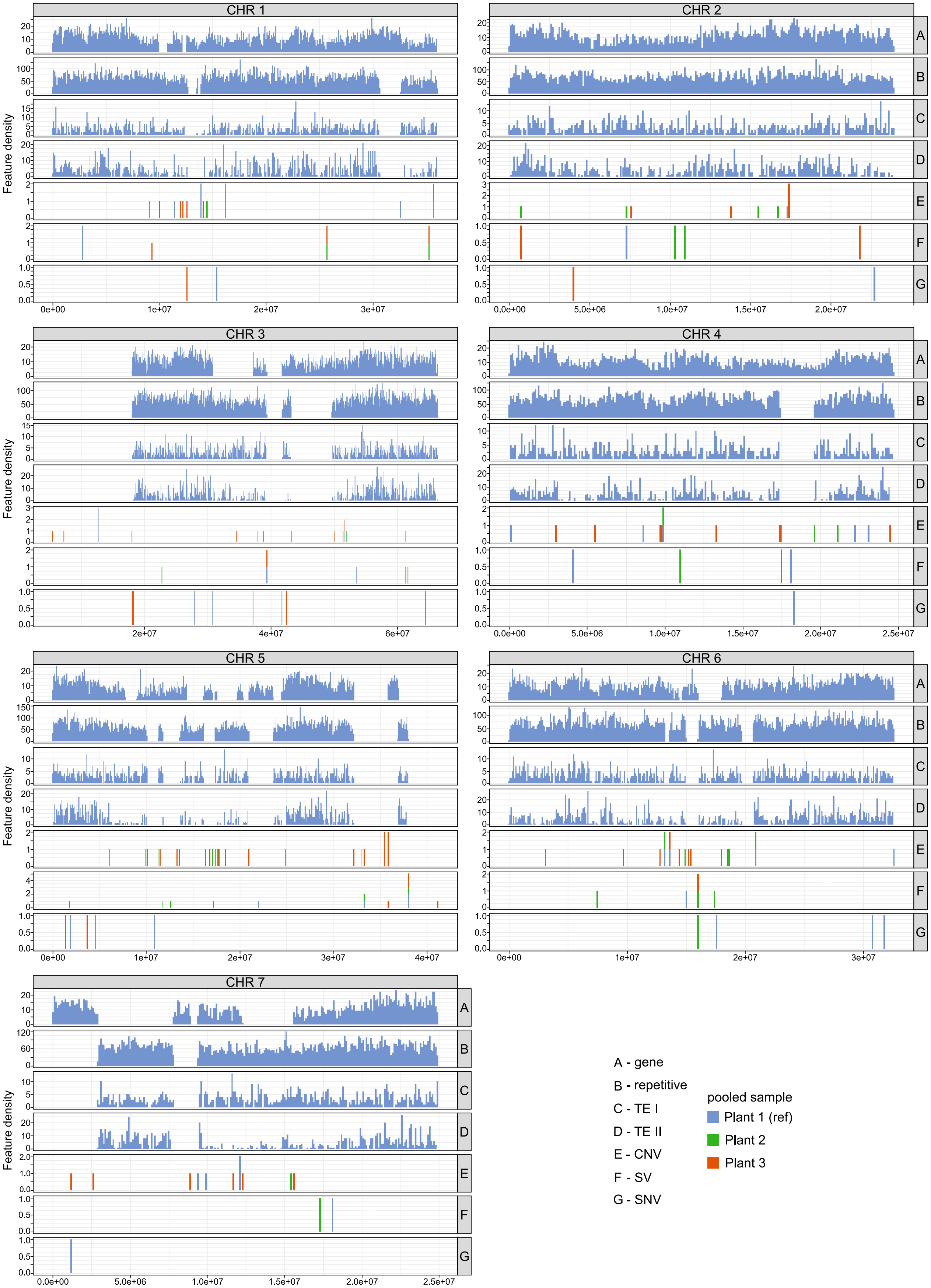
Long SMRT sequencing in genomics projects has contributed to an overall rise in the quality of research. In January 2016, three plant genomes assembled solely using PacBio technology were published in GenBank (Osipowski et al. 2016), compared with 40 current assemblies. New and developing bioinformatic methods can overcome more sequencing errors than previous techniques and SMRT sequencing offers high genome contiguity, which will be particularly beneficial when the cost of such sequencing falls. Nevertheless, the use of higher quality short read data for comparative genomics is currently unavoidable because of the much lower costs and already established bioinformatic pipelines.

Efficient and precise WGS-based variant prediction is used widely in well studied genomes, but its effectiveness for lesser known genome sequences is unclear. Some plant research using solely bioinformatics tools has been carried out recently at the population scale, supported by comprehensive variant databases (Torkamaneh et al. 2017). However, because of a lack of information about genomic variation, tuning pipelines for variant prediction remains a challenge for most organisms. Prediction inaccuracies can lead to results deficiencies, and this together with strong beliefs in the reliability of bioinformatic results have become important matters of scientific debate in plant research (Torkamaneh et al. 2018). Moreover, for certain species, different intra-species genotypes might be sequenced with

Table 8 Numbers of predicted copy number variations (CNVs) and structural variants (SVs) unique to each of three plant samples, and numbers of rearrangements up to 10 bp from or intersecting with interspersed repetitive elements (IREs)

	CNVs	IREs	TE classified[a]
P1	239	1086	83
P2	231	1107	87
P3	120	640	60
	SVs	IREs	TE classified[a]
P1	10	8	0
P2	1	2	1
P3	3	3	1

[a] TE classified, IREs assigned to Class I or Class II transposable elements (TEs)


CHR 1
CHR 2
CHR 3
CHR 4
CHR 5
CHR 6
CHR 7
Feature density
A
B
C
D
E
F
G
0e+00
1e+07
2e+07
3e+07
4e+07
6e+07
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
A - gene
B - repetitive
C - TE I
D - TE II
E - CNV
F - SV
G - SNV
pooled sample
🟦 Plant 1 (ref)
🟩 Plant 2
🟥 Plant 3

◂**Fig. 6** Graphical representation of genomic feature and event distribution in chromosomes by mapping B10 cucumber line contigs using Yang et al. (2013) markers. Density was assessed by counting the start positions of each feature and event in 200-kb consecutive bins. Features are: genes, repetitive (LCRs and SSRs), and transposable elements class I (TE I) and II (TE II). Events are: copy number variants (CNVs), structural variants (SVs), and single nucleotide variants (SNVs). Information specific to the individual plant sample colour coded: blue (P1), green (P2), and red (P3). Genomic features are coloured blue because they were annotated on the B10v3 genome reference sequence corrected by P1-derived short read data. In an unmapped section only contigs containing genes are included (colour figure online)

varying quality, depending on the interests and capabilities of a research group. This poses additional challenges in terms of pooling efforts for more precise genome research and collaborating on the development of universal and integrated species-specific genome databases. Overcoming such obstacles is crucial for plant-specific genomic research.

The multidimensional integration of data can result in optimal quality and contiguity of genome drafts and contribute to correctly designed variant calling pipelines. Achieving this is essential for accurate bioinformatics-based polymorphism prediction as well as for developing efficient breeding strategies. Many crop genomic projects were designed to combine genomic techniques and WGS data to achieve reproducible and accurate bioinformatic variant calling for deep phenomics analysis (D'Agostino and Tripodi 2017). Additionally, plant accession association studies require highly accurace variant calling, and contribute to such studies by database creation (Goodwin et al. 2016; D'Agostino and Tripodi 2017; Torkamaneh et al. 2018). It also is feasible to use WGS data and advanced bioinformatics in smaller genomic projects to improve accuracy.

Genome dynamics

A newly SMRT read-assembled genome draft sequence and three Illumina genomic read sets from individual plants that are 21–22 generations distant from the common ancestor were used in this study. This allowed us to assess genome dynamics such as genomic feature and event distribution and their inter-occurrence. We designed a specific approach to predict unique SNVs among individual genomes, with the focus on accuracy. This approach to measuring variation between individual plants of an inbred crop line is new and could be used for other organisms with available genome draft reference sequences. Specifically, we investigated whether genomic variation could be called within a highly inbred B10 cucumber line, and determined the characteristics of genome dynamics such as the inter-occurrence of genes and variants as well as TEs. Theoretically, the phylogenetic relationship of an individual plant and the configuration of an individual genome of a highly inbred line might translate to more accurate variant predictions. However, the state of the genome draft sequence, read sequence errors, and the lack of comprehensive support for polymorphism data are challenges that need to be taken into account and overcome with reliable results (Li 2014). Our approach addressed these challenges by the systematic enhancement of raw read-derived data. Sequence correction of both the genome reference draft and the read sets should positively influence the accuracy of analysis (Li 2015a). The corrected genome draft gave significantly better results for the P1 reads set after correction than before in terms of the consistency of the read pairs, suggesting the importance of such correction after SMRT assembly.

We assembled a genome reference sequence of significantly better contiguity and quality than the previously published cucumber genome sequences (Huang et al. 2009; Cavagnaro et al. 2010; Wóycicki et al. 2011; Li et al. 2011). The B10v3 sequence helped to broaden the previous studies by including some unassembled coding and non-coding regions in the published assemblies. The B10v3 contigs contiguity was approximately 44% higher than that of the other cucumber genome sequences. The genome k-mer analysis indicated that the B10 genome was 46.7 Mb longer than the cucumber genome measured cytogenetically (Arumuganathan and Earle 1991), but further analysis is required to confirm this. From the B10v3 scaffold quality check, we concluded that the BAC end sequence information previously used for scaffolding was not sufficient to significantly enhance the B10v3 contiguity results. Illumina mate pair read-based scaffolding and B10 line-specific genetic maps could vastly improve all aspects of the present B10v3 draft. Genome quality comparisons showed that short read correction in SMRT assembly quality enhancement was less important in coding regions, but it may be more useful during genome-wide variant calling. Comparison with other cucumber line assemblies suggests that finalisation of a genomic annotation (100% completeness) would require very detailed contiguity work and many more resources than is required for 90% annotation completeness.

IRE occurrence results are in line with general TE class occurrences in other eukaryotes genomes (Grzebelus 2018). The results of the two annotations are consistent with findings that genome drafts with good contiguity have less well-predicted structures but their genes have a higher average exon number (Denton et al. 2014). After filtering out small predicted genes, our results are consistent with recent results for the upgraded *C. melo* genome (Ruggieri et al. 2018) and confirms that high-quality Cucurbitaceae genome sequences tend to have more predicted genes. Our study clearly showed that with extended contiguity, a lower number of annotated genes are expected with an increased number of exons per gene. We have clear evidence that increasing contig contiguity through assembly can help to update protein sequences

Table 9 Numbers of single nucleotide variants (SNVs) called and verified, and precision parameter computed after verification for the P1, P2, and P3 samples

	Total SNV called	SNV randomly chosen	SNP verified	InDel verified	SNV verified	Precision
P1	42	16	5 (4)	4 (3)	9 (7)	0.78
P2	20	10	7 (4)	1 (1)	8 (5)	0.63
P3	33	8	6 (0)	0	6 (0)	0.00
Total	95	34	18 (8)	5 (4)	23 (12)	0.52

True positives in brackets

that were shortened in previous assembly annotations. As a model plant, *A. thaliana* can be presumed to have a high-quality predicted protein set. The *C. melo* genome also has been studied extensively. The annotated protein sets from both of these plants were used as benchmarks for our annotation. We recognised from the length ratio distribution that improvements can still be made in the annotation of proteins that had predicted sequences that were shorter than their orthologs in *A. thaliana* and *C. melo*. However, our annotations did produce many longer protein sequences, which indicates that this and forthcoming B10v3 annotations could significantly improve the protein sequence quality in the Cucurbitaceae genomics field. The numbers of classified TEs detected in exons and promoters (Table 5) indicate some dynamics within coding and regulating elements of genes. Number of IREs detected in exons might seem too high but most of the detected IREs were matches to unknown consensus sequences generated by RepeatModeler approach highly rising sensitivity of the annotation method. Number of classified TEs overlapped with exons seems to be in line with general findings (Nekrutenko and Li 2001). In promoter regions, the number of predicted IREs that overlapped upstream sequences was similar to the number that overlapped downstream sequences, even though they were twice as short. This indicates there were more IREs in the downstream promoter regions.

The SNP transition ratio was similar to that from *A. thaliana* studies and was hypothesised to be caused by the deamination of methylated cytosines and ultraviolet light-induced mutagenesis (Ossowski et al. 2010; Cao et al. 2011). Comparison of the TE and gene structure positions indicated that TE dynamics were high in non-genic regions, but that relatively low activity was expected from IREs in genes. The chromosomal distribution of genomic features indicates the existence of highly conserved genomic regions that may be of crucial biological importance for the plant (or for B10 line phenotypic identity). However, singular rearrangements were predicted in some of these regions. This is the first time the mutation rate has been measured among plant accession individuals. It was approximately 3.5 times lower than the mutation accumulation in *A. thaliana* lines that were 30 generations distant from each other (Ossowski et al. 2010). However, it must be noted that the cucumber genome is about three times longer than the *A. thaliana* genome, and that the SNV calling method was not optimally sensitive. Therefore, the mutation rate in B10 line individuals may be much higher.

In this study, we conducted a genomic comparison of several highly inbred line generations at the full genome scale. It clearly shows that dynamics between these genomes existed. We used a new significantly improved, cucumber B10 line genome sequence as the reference and predicted the variability of the genome.

Predicted SNV calls were selectively verified. If a genome has no database to benchmark a variant calling approach, a method that produces positive results is considered successful and can contribute to the development of comprehensive databases. Focusing on precision at the cost of sensitivity, our method successfully achieved high accuracy for the P1 sample (0.78). The P1 SNV calling was performed using our own methodology, which is very different from the methods used for the other two plants (P2 and P3).

For P2 and P3, SNVs were predicted by a standard method (Fig. 2) and gave worse results. For P3, the results were of much lower quality than those for P1 and P2. This may be because the sequencing quality of P3 seemed to be low, but the read data require more detailed checking to confirm this.

The approximately 30 × coverage per sample used in this study to detect variants via read alignment seems to be minimal for our goals. However, our novel RRVC approach achieved twice that coverage for the P1 sample, which helped increase the accuracy of variant calling. Precision results for each plant clearly reflected the advantage of RRVC. A similar algorithm could be useful for variant calling studies that have a limited sequencing budget. Moreover, the RRVC approach allowed us to positively verify most of the randomly selected P1 SNVs in a manner that was more sensitive to InDel occurrences than the standard methods that were used for P2 and P3 variant calling. This was evident by the much higher ratio of predicted InDels to SNPs in the P1 sample than in the other samples. The InDel ratio obtained by RRVC for the P1 sample might indicate that the approach itself increased InDel calling sensitivity. The design of the analytical reference sample variant FP filter also increased the overall accuracy (Fig. 2).

We concluded that in relatively high-quality genomes there are no obstacles to developing our approach further and to make sample-specific reference corrections for subsequent comparative analyses. However, this approach may be more difficult to implement for highly fragmented genome drafts because each correction would significantly alter contig lengths, making it much harder to compare corrected genome sequences. The applied method developed for this study was more sensitive to InDel detection and more precise in detecting SNPs than the standard methods because we used two sets of DNA with a nominal coverage of 37 × each. This methodology allowed the real-time evolution study of a frequency index of SNV mutations per nucleotide per generation for a single plant breeding line, based on a comparative analysis of individual genomes of individual plants.

In summary, PacBio SMRT reads proved to be of great value in enhancing the overall cucumber B10 line genome contiguity and quality (B10v3). We propose that SMRT reads are the best future solution for sophisticated comparative genomics of lesser known eukaryotic genomes. Importantly, our method could significantly improve bioinformatic pipelines for comparative genomics and thus has great practical potential in genomic metadata handling.

Acknowledgements We thank Prof. Stefan Malepszy for helpful comments. We thank Sarah Williams, PhD, and Margaret Biswas, PhD, from Edanz Group (www.edanzediting.com) for editing drafts of this manuscript.

Authors' contribution PO, MP, ZP, and WP designed and conceived the experiments. PO, MP, MW, and AS performed the experiments. PO contributed the new methodology. PO, MP, MW, and AS analysed the data. PO, MP, and WP wrote the manuscript.

Funding This work has been supported by grants from the National Science Center 2013/11/B/NZ9/00814 and 2011/01/B/NZ2/01631.

Data availability The contigs of the B10v3 cucumber genome draft have been deposited in NCBI databases (http://www.ncbi.nlm.nih.gov) under BioProject PRJNA29678, BioSample SAMN04103735, GenBank accession number LKUO00000000. Additional files are available at PCC Genomics (Polish Consortium of Cucumber Genome Sequencing; csgenome.sggw.pl/en-us/projects/b10v3).

Compliance with ethical standards

Conflict of interest The authors declare no known conflicts of interest and in particular: Paweł Osipowski declare that he has no conflict of interest, Magdalena Pawełkowicz declare that she has no conflict of interest, Michał Wojcieszek declare that he has no conflict of interest, Agnieszka Skarzyńska declare that she has no conflict of interest, Zbigniew Przybecki declare that he has no conflict of interest and Wojciech Pląder declare that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21:974–984

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218

Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM (2015) Corrigendum: assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol 33:1109. https://doi.org/10.1038/nbt1015-1109c

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST + : architecture and applications. BMC Bioinform 10:421. https://doi.org/10.1186/1471-2105-10-421

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43:956–963

Carvalho CMB, Lupski JR (2016) Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 17:224–238

Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, Harkins TT, Kodira CD, Huang S, Huang S, Weng Y (2010) Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). BMC Genom 11:569. https://doi.org/10.1186/1471-2164-11-569

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MH, Korlach J, Eichler EE (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 148:1293–1307

D'Agostino N, Tripodi P (2017) NGS-based genotyping, high-throughput phenotyping and genome-wide association studies laid the foundations for next-generation breeding in horticultural crops. Diversity 9(3):38. https://doi.org/10.3390/d9030038

Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW (2014) Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol 10:e1003998. https://doi.org/10.1371/journal.pcbi.1003998

Faust GG, Hall IM (2014) SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 30:2503–2505

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Grzebelus D (2018) The functional impact of transposable elements on the diversity of plant genomes. Divers (Basel) 10(2):18. https://doi.org/10.3390/d10020018

Guan P, Sung W-K (2016) Structural variation detection using next-generation sequencing data: a comparative technical review. Methods 102:36–49

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Th Sigurdsson G, Stacey SN, FriggeML Holm H, Saemundsdottir J, Th Helgadottir H, Johannsdottir H, Sigfusson G, Thorgeirsson G, Th Sverrisson J, Gretarsdottir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdottir H, Steingrimsdottir T, Gudmundsdottir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdottir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardottir O, Olafsson I, Arnar DO, Th Magnusson O, Kong A, Masson G, Thorsteinsdottir U, Helgason A, Sulem P, Stefansson K (2015) Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 47:435–444

Gutman W, Pawełkowicz M, Woycicki R, Piszczek E, Przybecki Z (2008) The construction and characteristics of a BAC library for *Cucumis sativus* L. "B10". Cell Mol Biol Lett 13:74–91

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning KM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31:5654–5666

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512

Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL (2006) CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinform 7:446. https://doi.org/10.1186/1471-2105-7-446

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32:767–769

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y, Zhu H, Li J, Lin K, Jin W, Fei Z, Li G, Staub J, Kilian A, van der Vossen EAG, Wu Y, Guo J, He J, Jia Z, Ren Y, Tian G, Lu Y, Ruan J, Qian W, Wang M, Huang Q, Li B, Xuan Z, Cao J, Asan WuZ, Zhang J, Cai Q, Bai Y, Zhao B, Han Y, Li Y, Li X, Wang S, Shi Q, Liu S, Cho WK, Kim J-Y, Xu Y, Heller-Uszynska K, Miao H, Cheng Z, Zhang S, Wu J, Yang Y, Kang H, Li M, Liang H, Ren X, Shi Z, Wen M, Jian M, Yang H, Zhang G, Yang Z, Chen R, Liu S, Li J, Ma L, Liu H, Zhou Y, Zhao J, Fang X, Li G, Fang L, Li Y, Liu D, Zheng H, Zhang Y, Qin N, Li Z, Yang G, Yang S, Bolund L, Kristiansen K, Zheng H, Li S, Zhang X, Yang H, Wang J, Sun R, Zhang B, Jiang S, Wang J, Du Y, Li S (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41:1275–1281

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, Peluso P, Boitano M, Chin C-S, Korlach J, Wilson RK, Eichler EE (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res 27:677–685

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 34:2115–2122

Hwang S, Kim E, Lee I, Marcotte EM (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep 5:17875. https://doi.org/10.1038/srep17875

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240

Kavak P, Lin Y-Y, Numanagić I, Asghari H, Güngör T, Alkan C, Hach F (2017) Discovery and genotyping of novel sequence insertions in many sequenced individuals. Bioinformatics 33:i161–i169

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36. https://doi.org/10.1186/gb-2013-14-4-r36

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11–16

Li H (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30:2843–2851

Li H (2015a) BFC: correcting Illumina sequencing errors. Bioinformatics 31:2885–2887

Li H (2015b) FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics 31:3694–3696

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

Li Z, Zhang Z, Yan P, Huang S, Fei Z, Lin K (2011) RNA-Seq improves annotation of protein-coding genes in the cucumber genome. BMC Genom 12:540. https://doi.org/10.1186/1471-2164-12-540

Lindner H, Raissig MT, Sailer C, Shimosato-Asano H, Bruggmann R, Grossniklaus U (2012) SNP-Ratio Mapping (SRM): identifying lethal alleles and mutations in complex genetic backgrounds by next-generation sequencing. Genetics 191:1381–1386

Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua N-H (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell 24:4333–4345

Macaulay IC, Voet T (2014) Single cell genomics: advances and future perspectives. PLoS Genet 10:e1004126. https://doi.org/10.1371/journal.pgen.1004126

McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. Nat Genet 39:S37–S42

Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott L, Utiramerur S, Hou Y, Smith KS, Montgomery SB, Wheeler M, Buchan JG, Lambert CC, Eng KS, Hickey L, Korlach J, Ford J, Ashley EA (2018) Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet Med 20:159–163

Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–2935

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn RD (2015)

Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 43:D7–D130

Nekrutenko A, Li WH (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17:619–621

Osipowski P, Wojcieszek M, Pawełkowicz M, Skarzyńska A, Koren S, Lomsadze A, Wóycicki R, Pląder W, Yagi K, Borodovsky M, Malepszy S, Przybecki Z (2016) Progress in assembling the cucumber (*Cucumis sativus*) Borszczagowski B10 line genome using long single molecule, real-time reads. In: Cucurbitaceae 2016, XIth Eucarpia Meeting on Cucurbit Genetics & Breeding, July 24–28, 2016, Warsaw, Poland, pp 72–74 ref 17. Cucurbitaceae 2016 Organizing Committee

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327:92–94

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14:417–419

Pawełkowicz M, Zieliński K, Zielińska D, Pląder W, Yagi K, Wojcieszek M, Siedlecka E, Bartoszewski G, Skarzyńska A, Przybecki Z (2016) Next generation sequencing and omics in cucumber (*Cucumis sativus* L.) breeding directed research. Plant Sci 242:77–88

Pawełkowicz M, Skarzyńska A, Pląder W, Przybecki Z (2019) Genetic and molecular bases of cucumber (*Cucumis sativus* L.) sex determination. Mol Breed 39(3):50

Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Gu X, Du Y, Li Y, Lin T, Yuan J, Yang X, Chen J, Chen H, Xiong X, Huang K, Fei Z, Mao L, Tian L, Städler T, Renner SS, Kamoun S, Lucas WJ, Zhang Z, Huang S (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet 45:1510–1515

Quinlan AR (2014) BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinform 47(11–12):1–34

Raphael BJ (2012) Chapter 6: structural variation and medical genomics. PLoS Comput Biol 8:e1002821. https://doi.org/10.1371/journal.pcbi.1002821

Ruggieri V, Alexiou KG, Morata J, Argyris J, Argyris J, Pujol M, Yano R, Nonaka S, Ezura H, Latrasse L, Boualem A, Benhamed M, Bendahmane A, Cigliano RA, Sanseverino W, Puigdomènech P, Casacuberta JM, Garcia-Mas J, Garcia-Mas J (2018) An improved assembly and annotation of the melon (Cucumis melo L.) reference genome. Sci Rep 8:10.1038/s41598-018-26416-2

Saxena RK, Edwards D, Varshney RK (2014) Structural variations in plant genomes. Brief Funct Genom 13:296–307

Schatz MC, Delcher AL, Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. Genome Res 20:1165–1173

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-C, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5:e1000734. https://doi.org/10.1371/journal.pgen.1000734

Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61:437–455

Torkamaneh D, Laroche J, Tardivel A, O'Donoughue L, Cober E, Rajcan I, Belzile F (2017) Comprehensive description of genome-wide nucleotide and structural variation in short-season soybean. Plant Biotechnol J 16:749–759

Torkamaneh D, Boyle B, Belzile F (2018) Efficient genome-wide genotyping strategies and data integration in crop plants. Theor Appl Genet 131:499–511

Uchida N, Sakamoto T, Kurata T, Tasaka M (2011) Identification of EMS-induced causal mutations in a non-reference *Arabidopsis thaliana* accession by whole genome sequencing. Plant Cell Physiol 52:716–722

Van der Auwera GA, Carneiro MO, Hartl C, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA (2013) From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinform 11:11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522–530

Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLoS Biol 12:e1001883. https://doi.org/10.1371/journal.pbio.1001883

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9:e112963. https://doi.org/10.1371/journal.pone.0112963

Wóycicki R, Witkowicz J, Gawroński P, Dąbrowska J, Dąbrowska J, Lomsadze A, Pawełkowicz M, Siedlecka E, Yagi K, Pląder W, Seroczyńska A, Śmiech M, Gutman W, Niemirowicz-Szczytt K, Bartoszewski G, Tagashira N, Hoshi H, Borodovsky M, Borodovsky M, Karpiński S, Malepszy S, Malepszy S, Przybecki Z (2011) The genome sequence of the North-European cucumber (*Cucumis sativus* L.) unravels evolutionary adaptation mechanisms in plants. PLoS ONE 6:e22728. https://doi.org/10.1371/journal.pone.0022728

Yang L, Li D, Li Y, Gu X, Huang S, Garcia-Mas J, Weng Y (2013) A 1681-locus consensus genetic map of cultivated cucumber including 67 NB-LRR resistance gene homolog and ten gene loci. BMC Plant Biol 13:53. https://doi.org/10.1186/1471-2229-13-53

Zaynab M, Fatima M, Abbas S, Umair M, Sharif Y, Raza MA (2018) Long non-coding RNAs as molecular players in plant defense against pathogens. Microb Pathog 121:277–282

Zhang Z, Mao L, Chen H, Bu F, Li G, Sun J, Li S, Sun H, Jiao C, Blakely R, Pan J, Cai R, Luo R, Van de Peer Y, Jacobsen E, Fei Z, Huang S (2015) Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. Plant Cell 27:1595–1604