South African Journal of Botany 175 (2024) 241–252

Contents lists available at ScienceDirect

South African Journal of Botany

journal homepage: www.elsevier.com/locate/sajb

Identification and characterization of sequence variants from a *de novo*-assembled partial pan-genome of cactus pear (*Opuntia* L.)

Mathabatha F. Maleka[a,*], Tsholofelo J. Modise[a], Morné G. Du Plessis[a], Gesine M. Coetzer[b]

[a] *Department of Genetics, University of the Free State, Bloemfontein. 9301, South Africa*
[b] *Department of Soil, Crop and Climate Sciences, University of the Free State, Bloemfontein, 9301, South Africa*

ARTICLE INFO

Article History:
Received 3 April 2024
Revised 28 August 2024
Accepted 9 October 2024
Available online xxx

Edited by: Dr I. Risenga

Keywords:
Opuntia ficus-indica
Opuntia robusta
Illumina sequencing
De novo sequence assembly
Sequence annotation
Simple sequence repeats (SSRs)
Single nucleotide polymorphisms (SNPs)

ABSTRACT

Members of the genus *Opuntia* exhibit extensive morphological variation that is seemingly linked to their propensity for occupying extreme environments all over the world. The plants are very important to the agriculture industry as they have many health-promoting nutritional and bioactive compounds. Also, as succulents, they are interesting because they develop peculiar stems – which, however, fulfil the role of leaves – that enable the storage of copious amounts of water. Yet, little is known about the genomes of *Opuntia* species although having such data can give insight into the genetic diversity that regulates the observed morphological variation as well as the molecular processes responsible for the development of succulent leaves. The current study aimed to sequence and *de novo* assemble a partial reference genome of *Opuntia* that would be used to find sequence variants for differentiating species and cultivars of this important agricultural crop. Illumina sequencing was performed on 10 cultivars that represented two species (*Opuntia ficus-indica* and *O. robusta*) within the South African *Opuntia* germplasm. Sequence assembly of 214 million filtered high-quality reads generated a 657 Mbp partial pan-genome that represented ~30 % of a predicted *O. ficus-indica* genome. Functional annotation of the assembled sequences at different bioinformatic databases revealed many genes relating to diverse developmental pathways as well as water storage. Further, sequence variants including some 60,000 simple sequence repeats (SSRs) and 118,000 biallelic single nucleotide polymorphisms (SNPs) were discovered by mapping reads of each cultivar against the assembled partial reference pan-genome of *Opuntia*. While the identified number of SNPs is significant it, however, corresponded to moderate and higher frequency variants only (minor allele frequency [MAF] > 0.2). Also, a random selection of 1,000 SNPs genotyped across all specimens were able to differentiate the 10 cultivars and two species studied herein. Overall, the current study reports the first partial reference pan-genome of *Opuntia* specimens. The generated data should serve as a valuable genetic resource for future studies seeking to analyze diversity-related phenomena or elucidate the molecular development and growth of *Opuntia* specimens.

1. Introduction

Opuntia ficus-indica, also known as cactus pear, is a drought tolerant plant that occurs mostly in semi-arid and arid environments all over the world (Russell and Felker, 1987). The plant has many uses in human and animal lives, and these include being consumed as fruit, vegetables, or livestock feed. In addition to the agriculture industry, cactus pear is also a rich source of many additives that are used in the cosmetic, pharmaceutical and civil construction industries, whereas others apply the additives to alternative fuels (Martins et al., 2023; Stintzing and Carle, 2005). In commercial plantations, cactus pear is commonly propagated asexually via terminal cladodes due to the ease of growing these organs and the resultant plantlets will tend to have shorter juvenile phases and higher survival rates (Wessels et al., 1997). Over time, such plantlets may present uniform phenotypes and maintain 'true-to-type' phenotypic traits (Peña-Valdivia et al., 2008). Such uniformity, however, can make it challenging for novices to accurately differentiate *Opuntia* species and cultivars. The need to accurately differentiate *Opuntia* taxa draws from their intrinsic performance differences when grown under different climatic conditions (Kumar et al., 2023; Neupane et al., 2021). Hence, other methods are needed to distinguish members of *Opuntia* effectively.

Molecular DNA markers have long been established as genomic identifiers that can distinguish plant cultivars and species accurately. Markers that have been generally used for this purpose include single nucleotide polymorphisms (SNPs, Chen et al. 2009; Oliveira et al. 2014), simple sequence repeats or short tandem repeats (SSRs/STRs, Bandelj et al. 2002; Choudhary et al. 2012), amplified fragment length

[*] Corresponding author.
E-mail address: malekamf@ufs.ac.za (M.F. Maleka).

https://doi.org/10.1016/j.sajb.2024.10.023

polymorphisms (AFLPs, Aranzana et al. 2003; Parks and Moyer, 2004) and randomly amplified polymorphic DNA (RAPD, Forapani et al. 2001; Diederichsen and Fu, 2006). An increase in genome sequencing projects due to the accessibility of next-generation sequencing (NGS) platforms in recent years has encouraged genome-wide marker identification and applications, particularly in non-model species. Consequently, genetic diversity estimates are now being performed at the genome level while also identifying and functionally annotating genes found in those genomes (Shalev et al., 2022; Wang et al., 2021; Xu et al., 2017). In all, NGS has not only aided the sequencing and functional annotation of genes in non-model genomes, but also the discovery of genetic markers that can be used to evaluate levels of diversity and differentiation of taxa.

Despite the significant progress made in the genome sequencing of non-model plant species until now (Marks et al., 2021; Sun et al., 2022), the sequencing and characterization of cacti has, however, been very limited. Prior to this, only three studies had previously reported the (nuclear) genome sequencing of cacti within the Cactaceae (Franco et al., 2022) – the plant family into which the genus *Opuntia* is classified (Wallace and Gibson, 2002). Several studies had, however, sequenced cacti plastomes for assessing phylogenomic and evolutionary relationships, whereas others sequenced and analyzed cacti transcriptomes to study the effect of diverse stresses toward ecological and environmental adaptations (Franco et al., 2022). In addition, the development and application of genetic markers in cacti is well established, with most studies, however, using variants in the internal transcribed spacer (ITS) region and/or SSRs as markers of choice (Franco et al., 2022; Omar et al., 2021). While these markers have been effective, there is still a need to grow the collection of genetic markers that can be used to study other aspects of cacti biology. In South Africa, especially, efforts to find and develop more markers will be significant to the world-renowned collection of cultivars known as the South African *Opuntia* germplasm (Chapman et al., 2002). The germplasm has been acclimatizing to the local conditions for more than a century (Potgieter and Mashope, 2009). During this period, the cultivars were sustained at several sites across the country and also bred to serve different purposes in the agricultural industry (Mashope et al., 2011). Overall, accessibility of NGS platforms and the exclusive *Opuntia* germplasm present an opportunity for South Africa to disseminate knowledge on the genetics of cacti.

Herein, the aims were to sequence and assemble a partial *Opuntia* reference genome that will be used for finding sequence variants that can differentiate species and cultivars within this taxonomic group. Notably, the reference genome was projected to be partial due to the limited sequencing done on each cultivar. Also, the reference genome will be assembled using DNA sequenced from different cultivars and species within the genus *Opuntia* so as to make it representative – a model that is currently termed a ‘pan-genome’ (Della Coletta et al., 2021; Li et al., 2022). While partial, the current study pioneered the sequencing and annotation of an *Opuntia* genome while discovering thousands of SNPs and SSRs that may be efficient in distinguishing cultivars and species of *Opuntia*. The generated data will surely expand the pool of genetic markers that can be used to study *Opuntia* specimens (Franco et al., 2022). Moreover, the current *Opuntia* reference pan-genome will provide a great resource for future studies seeking to unravel the molecular genetics of development in this key agricultural succulent plant.

2. Materials and methods

2.1. Plant material and DNA extraction

One-year old cladodes of *Opuntia* cultivars were harvested from a local plantation established at the University of the Free State (UFS) campus in Bloemfontein, Free State Province, Republic of South Africa (RSA). The cultivars were mainly selected for study due to their importance in the commercial cactus pear industry, whereby they are used for fruit production or as livestock feed (Mashope et al., 2011), or they displayed potential for being valuable in the industry (Coetzer et al.; unpublished data). Cladode tissues were cut up (5 × 5 cm^2) to gather material for extracting genomic DNA. All cultivars, except one, corresponded to the species *Opuntia ficus-indica* and these included ‘Algerian’, ‘Berg x Mexican’, ‘Direkteur’, ‘Fusicaulis’, ‘Gymno Carpo’, ‘Meyers’, ‘Morado’, ‘Santa Rosa’ and ‘Sicillian Indian fig’. The cultivar ‘Robusta’ represented the species *O. robusta*. Genomic DNA was isolated from each sample through the NucleoSpin® Plant II Kit (Macherey-Nagel GmbH & Co. KG, Düren, Germany). However, the protocol was adjusted due to the mucilage that is abundant in cladode tissues of cactus pear plants. For this purpose, the incubation time during cell lysis was changed to 60 min, whereas centrifugation of crude lysates was done for 7 min. Also, the centrifugation time for binding DNA to columns was set to 5 min and columns were washed twice. The extracted DNA was measured by spectrophotometry (NanoDrop ND-1000, Thermo Fisher Sci., Waltham, MA, USA), while quality was analysed by gel electrophoresis using 1 % (w/v) TAE agarose gels that were stained with GelRed nucleic acid dye (Biotium Inc., Fremont, CA, USA). Gels were pictured under UV light using a G:Box Gel Documentation system (Syngene, Cambridge, UK).

2.2. NGS, adapter removal and quality trimming of raw reads

The extracted *Opuntia* genomic DNA was subjected to NGS through a TruSeq Kit (Illumina Inc., San Diego, CA, USA) that generated 125 bp paired-ends (PE) reads. NGS services were performed at the Agricultural Research Council Biotechnology Platform (ARC-BP Pretoria, RSA; https://www.arc.agric.za/pages/btp.aspx) through an Illumina HiSeq 2500 Sequencing System. The statistics and quality assessment of raw reads were analysed with the software FastQC (v.0.12.1; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Read trimming and the removal of adapter sequences as well as low-quality reads was performed with the software BBDuk, which forms part of the BBTools package (https://sourceforge.net/projects/bbmap/). Reads were length trimmed on either or both ends, whereas quality trimming was performed along read lengths with average base quality set to 15 (trimq = 15). The kmer size was set to 23 (k = 23) and permitted one mismatch (hdist = 1) between matched reads. All these procedures were applied equally on PE reads. Where reads were filtered, but their matching mates passed the filtering criteria; the latter were saved for subsequent analysis procedures.

2.3. De novo *sequence assembly and sequence annotation*

Before assembling the *Opuntia* pan-genome, all trimmed high-quality reads were mapped against the chloroplast and mitochondrial genomes of *Opuntia quimilo* (MN114084.1; https://www.ncbi.nlm.nih.gov/) and cultivated beet (*Beta vulgaris* subsp. *vulgaris*; NC_002511.2), respectively. Read mapping was performed via the CLC Genomics Workbench platform (v21.0.5; https://digitalinsights.qiagen.com/) using default settings while saving all unaligned (both the PE and unpaired) reads. The unaligned PE reads were assembled *de novo* using the multi-kmer software SPAdes (v3.15.5; Bankevich et al. 2012) and IDBA-UD (v1.1.3; Peng et al. 2012). Also, another assembly was made using the built-in *de novo* assembler of the CLC Genomics Workbench. The assembler was set to automatically choose a suitable kmer based on input data and subsequently mapped the input PE reads against the assembly for error correction. The three assemblies were individually extended by mapping the unpaired high-quality reads that were kept earlier and this was done with the tool Tadpole (v38.90; BBTools suite). An assembly with the best parameters was verified using a web application of QUAST (http://

cab.cc.spbu.ru/quast/; Gurevich et al. 2013) and, henceforth, considered to be an *Opuntia* pan-genome.

Putative protein-coding genes in the *Opuntia* pan-genome were predicted using the software package AUGUSTUS (v3.5.0; Stanke et al. 2004). Given that genome annotation is still lacking in most plants, the putative *Opuntia* genes were predicted based on gene models in the *Arabidopsis thaliana* genome as applied in the AUGUSTUS workflow. The predicted *Opuntia* proteins were annotated by searching them against proteins at the NCBI non-redundant (NR release 255, https://ftp.ncbi.nlm.nih.gov/blast/db/) and UniProtKB/SwissProt (release 2023_03; UniProt, 2023) databases using BLASTp algorithms (Altschul et al., 1990). The searches were performed through the high-throughput alignment platform DIAMOND (v2.1.6.160; Buchfink et al. 2015), with significant alignments set to exhibit E-values ≤ 1e^{-5} and percentage identity ≥ 50 %. Following that, accession numbers of protein subjects were used to describe the functions of predicted genes based on the gene ontology (GO; Ashburner et al. 2000), plant ontology (PO; Avraham et al. 2008) and plant reactome (PR; Naithani et al. 2017) vocabularies. Enzyme-coding genes were characterized by mapping the sequences against the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto 2000; Kanehisa et al. 2015) pathways as applied at the Plant Gene Set Annotation Database (PlantGSAD; Ma et al. 2022). Throughout these analyses, enriched annotation pathways were verified by hypergeometric distribution analysis along with false discovery rate (FDR) correction ($p \leq 0.05$). All the gene annotation analyses were performed with the graphical web application ShinyGO (v0.77; Ge et al. 2020).

2.4. Discovery and characterization of sequence variants

The identification of sequence variants in the 10 *Opuntia* cultivars was preceded by indexing the *de novo*-assembled pan-genome using the software Burrows-Wheeler Alignment (BWA v0.7.17-r1188; Li and Durbin, 2009). Subsequently, the high-quality PE reads of each sample were separately mapped against the indexed genome, with quality threshold of trimming reads prior to alignments set to 10 (q = 10). The ensuing mapping files were pooled, used to call sequence variants and filtered through the packages SAMtools (Li et al., 2009), BCFtools (Li, 2011) and VCFtools (Danecek et al., 2011). The filtering criteria included retaining biallelic SNP sites only, in addition to deleting base positions with low call quality (QUAL ≤ 20), sequence depth (DP < 10) and minor allele frequency (MAF < 0.2). The data was further filtered to only include sites that were genotyped across all 10 *Opuntia* samples. A subset of the genotyped positions were used to perform multidimensional scaling (MDS) analysis through the software PLINK (v1.90; Purcell et al. 2007) and the results where plotted via the R-based *ggplot2* package (Wickham, 2016).

2.5. Identification of SSRs in the Opuntia pan-genome

The *Opuntia* pan-genome was searched to find and characterize SSRs using the software programme MISA (Thiel et al., 2003). The searches were based on loci containing a tandem array of no less than eight repeats for 2-mers, seven repeats for 3-mers, six repeats for 4-mers, five repeats for 5-mers and four repeats for each of the 6- and 7-mers. The 8- to 10-mers, on the contrary, were each set to consist of at least three repeats. When encountered, composite SSRs – e.g., $(CT)_n(GA)_n$, whereby n indicates the number of repeats per motif – were also recorded if they occurred at distances of 6 – 100 bp apart. Imperfect SSRs – e.g., $(CT)_n$NNNNN$(GA)_n$, whereby N = A, T, G or C – were also documented, so long as adjacent loci were separated by no more than 5 bp of non-motif nucleotides.

2.6. Primer design and polymerase chain reaction (PCR) amplification of selected SSR loci

Oligonucleotide primers were designed with the software package Geneious Prime (Biomatters Ltd., Auckland, New Zealand) to have the following features: sequence length of 18 – 27 bp, a GC content of 40 – 70 % and a melting temperature (T_m) of about 60 °C. Moreover, each primer pair was designed to have a maximum T_m difference of 5 °C and yield amplification products of 100 – 400 bp in length. PCR amplification of SSR loci was performed using the EmeraldAmp® GT PCR Master Mix (Takara Bio. Inc., Shiga, Japan). Each reaction comprised 40 – 50 ng of template DNA, 0.2 μM of each primer, 1 × EmeraldAmp® GT PCR Master Mix and filled to 10 uL with nuclease-free water. Thereafter, the following PCR conditions were completed on an Applied Biosystems 2720 Thermal Cycler (Applied Biosystems, Foster City, CA, USA): initial denaturation at 98 °C for 3 min, followed by 30 cycles of denaturation at 98 °C for 10 s, primer annealing at locus-specific temperatures for 20 s and primer extension at 72 °C for 20 s. Final extension was done at 72 °C for 1 min. Following PCR, amplicons were stained with GelRed (Biotium Inc.) and electrophoresed on 3 % (w/v) TAE agarose gels loaded with a PCR DNA ladder (Ampliqon A/S, Odense M, Denmark) for determining amplicon sizes. Gels were visualized under UV light (Syngene G:Box).

3. Results

3.1. NGS, adapter removal and quality trimming of raw reads

Illumina sequencing of the 10 *Opuntia* cultivars generated >118.5 million raw PE reads that totalled 28.1 Gbp (Table 1). The reads ranged in length from 70 – 125 bp and showed an overall GC content of 39 %. After read trimming, just over 94 % of the PE reads remained along with six million unpaired high-quality reads that were stored (Table 2). Aligning PE reads against the chloroplast and mitochondrial genomes from related taxa returned over 214 million

Table 1
Statistics of the raw paired-end (PE) reads sequenced from the 10 *Opuntia* cultivars.

Cultivar	Raw PE reads	%GC	Length (bp)
Algerian	5,440,685	37	70 – 125
Berg x Mexican	15,768,448	38	125
Direkteur	48,906,938	39	125
Fucicaulis	7,167,350	38	70 – 125
Gymno Carpo	7,566,508	39	70 – 125
Meyers	7,918,839	38	70 – 125
Morado	6,336,871	38	70 – 125
Robusta	6,612,292	38	70 – 125
Santa Rosa	6,065,206	38	70 – 125
Sicillian Indian fig	6,762,034	38	70 – 125
Total	**118,545,171**	**39**	**70 – 125**

Table 2
Statistics on the quality filtering and mapping of reads against the plant organellar genomes (i.e., chloroplast and mitochondria). Mapped reads included both trimmed paired-end (PE) and unpaired reads.

NGS read data	Number of reads	Total bases (bp)
Total raw PE reads	237,090,342	28,177,618,483
Filtered PE reads	223,533,516	21,672,283,943
Filtered unpaired reads	6,001,712	594,447,203
cpDNA-mapped PE and unpaired reads	7,329,909	692,076,208
mtDNA-mapped PE and unpaired reads	949,338	83,602,795
De novo-assembled PE reads	214,746,510	20,852,660,752

Table 3
Statistics for the *de novo* assembly and extension of the filtered high-quality reads using three different sequence assemblers.

Assembly parameter	SPAdes		IDBA-UD		CLC	
	Number of sequences	Total length (bp)	Number of sequences	Total length (bp)	Number of sequences	Total length (bp)
Scaffolds	235,184	523,444,959	335,539	657,330,882	320,774	408,424,949
Scaffolds ≥ 50 Kbp	18	1,005,714	21	1,198,338	0	N/A
Scaffolds ≥ 25 Kbp	941	29,345,610	874	27,487,674	15	459,690
Scaffolds ≥ 10 Kbp	11,628	183,935,985	11,498	180,927,769	1,881	24,224,500
Scaffolds ≥ 5 Kbp	30,333	315,110,338	138,506	327,113,118	13,104	99,364,827
Scaffolds ≥ 1 Kbp	97,905	474,301,111	138,506	559,794,161	110,585	306,108,691
Scaffold N50		6,942		5,028		2,382
Scaffold L50	20,377		32,221		44,112	
Longest scaffold		67,530		79,713		47,115
Mean scaffold length		2,225		1,959		1,273
N's per 100 kbp	24.62		2.22		909.52	

unaligned sequences that should correspond to the *Opuntia* nuclear genomic DNA. *De novo* read assembly followed by extension with unpaired reads that were saved earlier yielded assemblies of 408 Mbp (CLC), 523 Mbp (SPAdes) and 657 Mbp (IDBA) (Table 3). The latter presented the best parameters among the three assemblies and, as such, was used in subsequent analyses as a partial *Opuntia* pan-genome.

3.2. *Annotation of the partial* Opuntia *pan-genome*

The prediction of protein-coding genes based on gene models from the *Arabidopsis thaliana* genome yielded 135,243 protein sequences that ranged from 30 to 4,736 (with an average of 245) amino acids in length (data not shown). Searching the protein sequences against proteins at the NR and SwissProt databases revealed that around 58 % (78,318) of the queries had matches at the NR database, but only 17 % (22,977) matched subjects at the SwissProt database (data not shown). Most subjects at the NR database originated from *Carnegiea gigantea* (32,494; 41.49 %), *Beta vulgaris* subsp. *vulgaris* (4,992; 6.37 %), *Chenopodium quinoa* (2,987; 3.81 %) and *Spinacia oleracea* (2,612; 3.34 %). On the contrary, most subjects at the SwissProt database came from *Arabidopsis thaliana* (15,569; 67.76 %), followed by *Oryza sativa* Japonica Group (1,082; 4.71 %), *Oenothera berteroana* (1,019; 4.43 %) and *Nicotiana tabacum* (590; 2.57 %). Notably, some subjects at the SwissProt database were recorded as transposon sequences from the common fruit fly (*Drosophila melanogaster*; 157 sequences, 0.68 %) and budding yeast (*Saccharomyces cerevisiae S288C*; 83 sequences, 0.36 %). Nonetheless, given that most hits at the NR database featured vague identifiers including ‘uncharacterised’, ‘hypothetical’, ‘putative’, or ‘unnamed’, protein and enzyme functions encoded by the *Opuntia* pan-genome were classified using the SwissProt subjects only.

Sequence analysis based on GO term assignments revealed that the most gene-enriched pathway in the subgroup Biological Process (BP) was ‘shoot system development’, with 492 of the 964 pathway genes (FDR $< 6.1 \times 10^{-21}$) having been detected (Fig. 1A). Other pathways with many identified genes included ‘developmental process’ (1,597 genes; FDR $< 3.3 \times 10^{-53}$), ‘anatomical structure development’ (1,553 genes; FDR $< 3.8 \times 10^{-55}$), ‘cellular component organization or biogenesis’ (1,525 genes; FDR $< 4.4 \times 10^{-32}$) and ‘multicellular organismal process’ (1,510 genes; FDR $< 2.4 \times 10^{-46}$). In the subgroup Cellular Component (CC), pathways with many identified genes included ‘plasma membrane’ (1,443 genes; FDR $< 1.4 \times 10^{-08}$), ‘organelle membrane’ (1,352 genes; FDR $< 1.4 \times 10^{-30}$) and ‘plastid’ (1,282 genes; FDR $< 4.8 \times 10^{-12}$). The pathways ‘late endosome’ and ‘endosome membrane’ were, however, the most gene-enriched with 70 of 109 genes (FDR $< 3.2 \times 10^{-07}$) and 118 of 190 genes (FDR $< 1.1 \times 10^{-10}$) identified, respectively (Fig. 1B). The subgroup Molecular Function (MF) presented several pathways of which many genes have been found and these included ‘small molecule binding’ (1,476 genes; FDR $< 7.1 \times 10^{-23}$), ‘anion binding’ (1,421 genes; FDR $< 1.4 \times 10^{-21}$) along with ‘nucleotide binding’ and ‘nucleotide phosphate binding’ (1,390 genes; FDR $< 4.9 \times 10^{-22}$) (Fig. 1C). Regarding gene-enrichment, three pathways stood out as highly enriched and these included ‘protein serine kinase activity’ (358 of 673 pathway genes; FDR $< 1.1 \times 10^{-20}$), ‘ATP hydrolysis activity’ (278 of 513 pathway genes; FDR $< 1.6 \times 10^{-17}$) and ‘ATP-dependent activity’ (438 of 828 pathway genes; FDR $< 4.2 \times 10^{-24}$).

Regarding pathway mapping at the KEGG database, many sequences were mapped under ‘metabolic pathways’ (530 genes; FDR $< 2.8 \times 10^{-71}$), while the pathway ‘aminoacyl-tRNA biosynthesis’ was the most gene-enriched with 39 of the 48 pathway genes having been found (FDR $< 1.2 \times 10^{-15}$) (Fig. 2A). Other notable pathways included the ‘biosynthesis of plant hormones’ (114 genes; FDR $< 9.0 \times 10^{-08}$), ‘oxidative phosphorylation’ (62 genes; FDR $< 6.1 \times 10^{-07}$), ‘spliceosome’ (57 genes; FDR $< 2.9 \times 10^{-10}$) and ‘ubiquitin mediated proteolysis’ (52 genes; FDR $< 2.8 \times 10^{-09}$). Last, the pathways ‘glycerolipid metabolism’ (20 of 30 pathway genes; FDR $< 2.9 \times 10^{-06}$) and ‘glycerophospholipid metabolism’ (32 of 49 pathway genes; FDR $< 5.1 \times 10^{-09}$) were, interestingly, among the most gene-enriched.

Annotation of the predicted *Opuntia* proteins at plant-specific databases yielded many gene-enriched pathways of interest. At the PO database, such pathways included ‘radicle’ (14 of 22 pathway genes; FDR $< 2.0 \times 10^{-03}$), ‘adult vascular leaf’ (22 of 40 pathway genes; FDR $< 7.6 \times 10^{-04}$), ‘style’ (19 of 35 pathway genes; FDR $< 2.5 \times 10^{-03}$), ‘apical meristem’ (20 of 37 pathway genes; FDR $< 2.0 \times 10^{-03}$) and ‘leaf trichome’ (40 of 76 pathway genes; FDR $< 7.4 \times 10^{-06}$) (Fig. 2B). However, the majority of genes were associated with the pathway ‘pollen tube cell’ (1,787 genes; FDR $< 9.7 \times 10^{-85}$), with ‘fruit’ (339 genes; FDR $< 1.2 \times 10^{-13}$), ‘rosette leaf’ (199 genes; FDR $< 2.5 \times 10^{-21}$) and ‘inflorescence’ (112 genes; FDR $< 2.3 \times 10^{-08}$) also being notable. The PR database revealed the pathway ‘biotin biosynthesis II’ as the most gene-enriched, with eight of the 12 pathway genes having been detected (FDR $< 5.8 \times 10^{-06}$) (Fig. 2C). Other gene-enriched pathways of significance included ‘cellulose biosynthesis’ (six of nine pathway genes; FDR $< 6.4 \times 10^{-05}$) and ‘sphingolipid metabolism’ (10 of 17 pathway genes; FDR $< 1.1 \times 10^{-06}$), whereas ‘brassinosteroid signalling’ (22 genes; FDR $< 7.3 \times 10^{-14}$) and ‘auxin signalling’ (19 genes; FDR $< 3.7 \times 10^{-10}$) stood out as containing the most identified genes.

3.3. *Discovery and characterization of sequence variants*

Initially, analysis of sequence data from the 10 *Opuntia* cultivars revealed over 15.3 million variants in the form of single nucleotide substitutions (i.e., SNPs), insertions and/or deletions (i.e., indels) as well as SSRs (data not shown). Data filtering based on the criteria

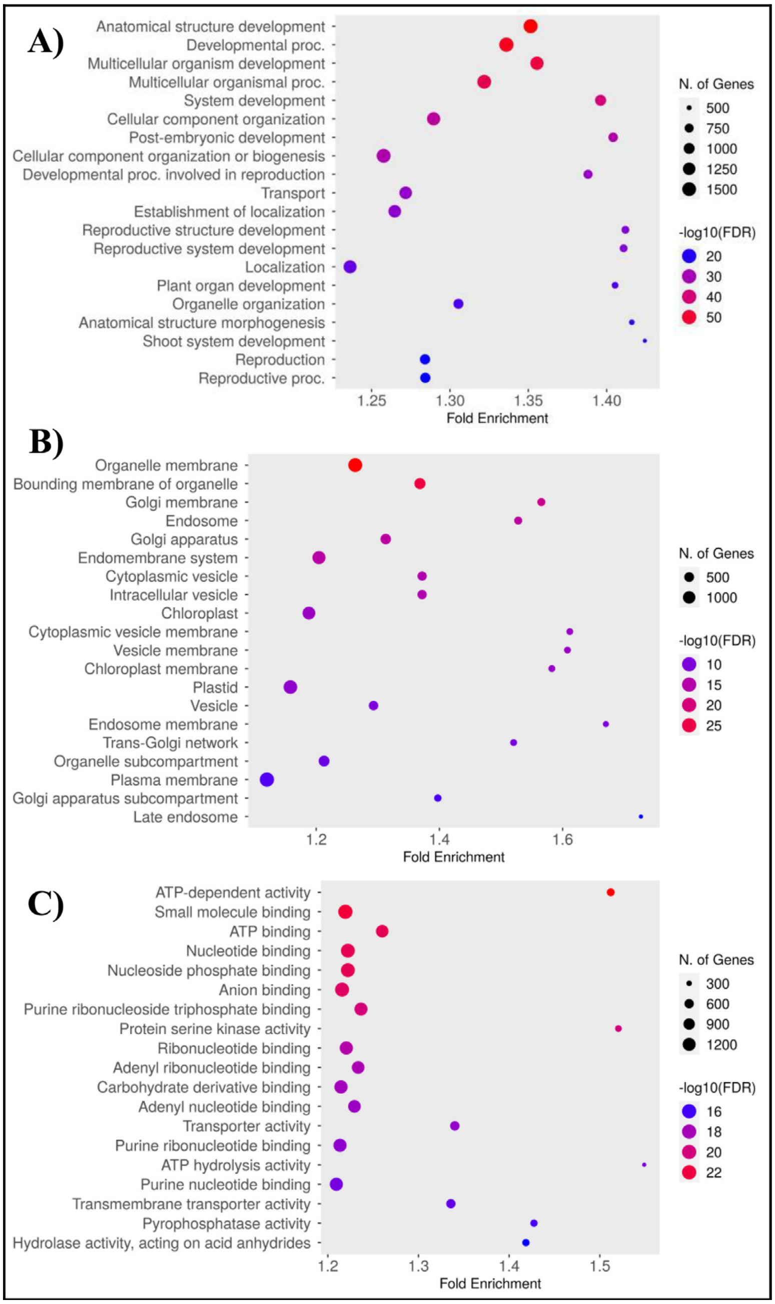

Fig. 1. Annotation of the partial *Opuntia* pan-genome based on gene ontology (GO; Ashburner et al. 2000) terms. The three GO subcategories are presented as follows: A) biological process (BP), B) cellular component (CC) and C) molecular function (MF). The graphics show the top 20 annotations only, and the numbers of annotated genes (N. of Genes) are plotted to scale as illustrated by different sizes of dotplots. 'Fold Enrichment' refers to the number of genes identified per annotation pathway, whereas FDR = false discovery rate ($p \leq 0.05$).

defined earlier (Section 2.4) reduced the number of variants to 2,598,481 biallelic SNPs. Further filtering involving the removal of variants with partial genotypes returned 118,010 high-quality biallelic SNPs, of which 1,000 were selected at random and used to create a MDS plot. The MDS plot revealed complete differentiation of the cultivars, with the cultivars forming three distinct clusters (Fig. 3). In addition, the cultivar 'Robusta' (*Opuntia robusta*) appeared to be fairly separated from the other nine cultivars of the species *O. ficus-indica*.

3.4. Identification of SSR loci in the Opuntia pan-genome

Screening the 335,539 sequences constituting the partial *Opuntia* pan-genome produced 59,963 SSR loci distributed on 38,807 sequences (11.57 %) (Table 4). One-third (34.31 %) of these sequences had more than one SSR locus each, whereas 8,502 sequences (21.91 %) contained 11,131 composite SSR loci. Imperfect SSR loci appeared on 368 of the 8,502 sequences (4.33 %). The majority of SSR loci involved 2-mers (38,840; 64.77 %), followed by 3-mers (12,176; 30.31 %), 4-mers (3,782; 6.31 %) and 6-mers (2,535; 4.23 %) (Table 4). Interestingly, 8-mers (775; 1.30 %) were equivalent to 5-mers (887; 1.48 %). Also, most loci had motifs that were repeated eight- (10,171; 16.96 %), nine- (7,238; 12.07 %) or ten-fold (5,384; 8.98 %). Motifs repeated five-fold were, on the contrary, underrepresented (1,113; 1.86 %) in the current data (Table 4).

The most frequent motif sequence in the current study was the dinucleotide AT (Table 5). It occurred a total of 7,434 instances that corresponded to an overall rate of 19.14 % among 2-mers. Other popular motifs included the 4-mer TATG and 10-mer AGATGTAAGA that

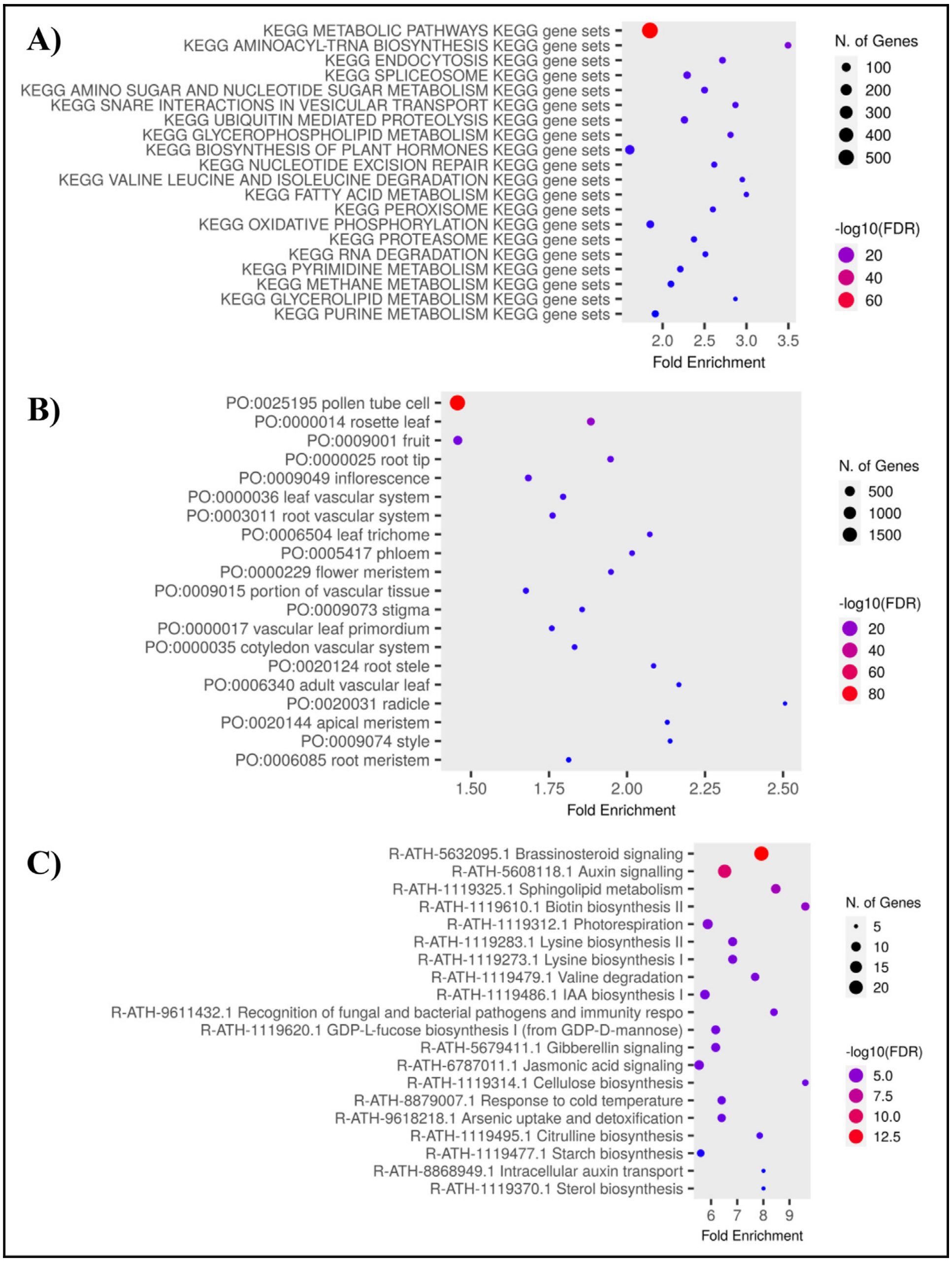

Fig. 2. Annotation of the partial *Opuntia* pan-genome using the plant-specific A) PlantGSAD (Ma et al., 2022), B) Plant Ontology (PO, Avraham et al. 2008), and C) Plant Reactome (PR, Naithani et al. 2017) databases. Each graphic displays the top 20 annotations only, and the numbers of annotated genes (N. of Genes) are plotted to scale as indicated by different sizes of dotplots. 'Fold Enrichment' refers to the number of genes identified per annotation pathway, whereas FDR = false discovery rate ($p \leq 0.05$).

each occurred at frequencies greater than 17 % within their motif categories. Considering paired sequence motifs, the AT/TA dinucleotides were also the most popular with 13,495 occurrences that represented a 34.75 % rate within the 2-mers. Other frequent motif pairs included the 4-mer ACAT/ATGT (67.85 %), the 3-mer AAT/ATT (56.03 %) and the 10-mer AAGAAGATGT/ACATCTTCTT, which occurred at a remarkably high rate of 35.04 % (Table 5).

3.5. Primer design and PCR amplification of selected SSR loci

Oligonucleotide primers were designed to PCR amplify a random set of 10 SSR loci identified in the *Opuntia* pan-genome (Table 6). Chosen loci included two 2-mer motifs as well as four each of the 3- and 4-mer motifs. After PCR, optimal annealing temperatures for the loci ranged from 50.0 – 65.7 °C. PCR validation of the primers in selected *Opuntia* cultivars revealed that all but two loci appeared to be polymorphic (Fig. 4). Therefore, the polymorphic loci can be effective towards distinguishing and assessing the levels of genetic diversity in *Opuntia* samples.

4. Discussion

The plant family Cactaceae (order Caryophyllales) contains numerous agricultural crops that are good sources of nutritional and bioactive compounds (de Araújo et al., 2021). Some members in the family are commercially exploited all over the world (e.g., *Opuntia ficus-indica* and several *Hylocereus* sp.), yet others have been largely underutilised (de Araújo et al., 2021). Until now, the genomic landscapes of seven species in the Cactaceae have been sequenced and characterized to some degree (Amaral et al., 2021; Copetti et al.,

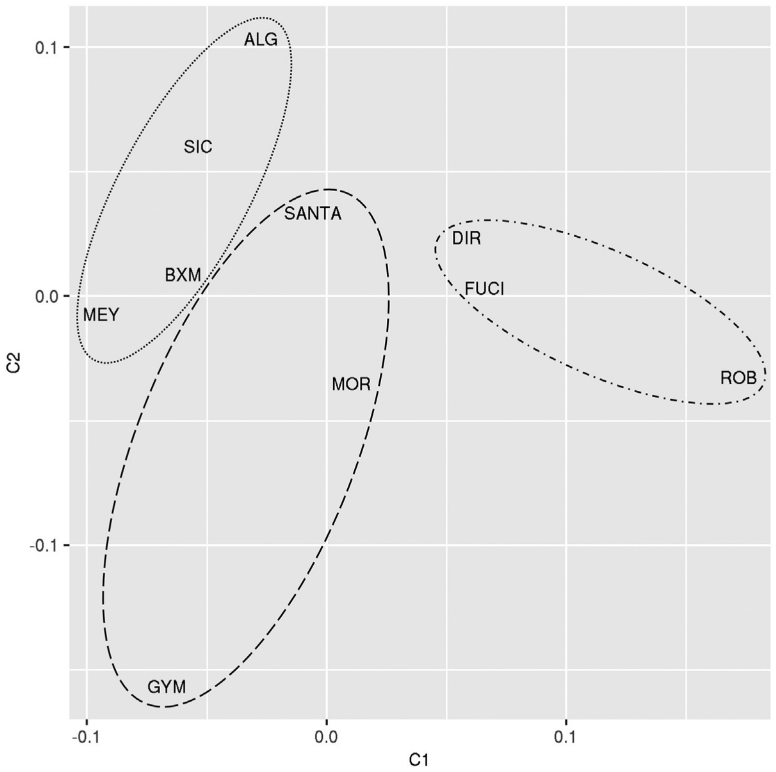

Fig. 3. A multidimensional (MDS) plot indicating the differentiation of the 10 *Opuntia* cultivars based on 1,000 random, biallelic SNPs. Cultivar names are abbreviated as follows: ALG (Algerian), BXM (Berg x Mexican), DIR (Direkteur), FUCI (Fusicaulis), GYM (Gymno Carpo), MEY (Meyers), MOR (Morado), ROB (Robusta), SANTA (Santa Rosa) and SIC (Sicillian Indian fig). All cultivars, except for ROB (*Opuntia robusta*), belong to the species *Opuntia ficus-indica*. The cultivars clustered in line with the three clades generated previously based on SSR data (Modise et al., 2024).

Table 4
The characterization of SSR loci identified in the partial *Opuntia* pan-genome.

Motifs	Number of sequences within repeat category									Total	%Total
	3	4	5	6	7	8	9	10	≥11		
2-mer						8,210	5,891	4,304	20,435	38,840	***64.77***
3-mer					2,333	1,388	906	791	6,758	12,176	***20.31***
4-mer				1,172	692	510	401	264	743	3,782	***6.31***
5-mer			635	170	43	16	8	4	11	887	***1.48***
6-mer		1,752	426	127	53	47	31	20	79	2,535	***4.23***
7-mer		301	43	11	1	0	1	1	45	403	***0.67***
8-mer	718	44	8	4	1	0	0	0	0	775	***1.29***
9-mer	200	12	1	1	0	0	0	0	0	214	***0.36***
10-mer	346	5	0	0	0	0	0	0	0	351	***0.59***
Total	**1,264**	**2,114**	**1,113**	**1,485**	**3,123**	**10,171**	**7,238**	**5,384**	**28,071**	**59,963**	
%Total	***2.11***	***3.53***	***1.86***	***2.48***	***5.21***	***16.96***	***12.07***	***8.98***	***46.81***		

Table 5
The abundance of frequent motifs per motif category in the partial *Opuntia* pan-genome.

Motif	Sequence	Occurrences	% per class	Sequence	Occurrences	% per class
2-mer	AT	7,434	19.14	AT/AT	13,495	34.75
3-mer	AAT	1,595	13.10	AAT/ATT	6,820	56.03
4-mer	TATG	645	17.05	ACAT/ATGT	2,566	67.85
5-mer	AAAAG	91	10.26	AAAAG/CTTTT	239	26.94
6-mer	TATATG	79	3.12	ACATAT/ATATGT	529	20.87
7-mer	AACAGCT	38	9.43	AAACCCT/AGGGTTT	112	27.79
8-mer	AAAAATTT	43	5.55	AAAAAAAT/ATTTTTTT	137	17.68
9-mer	TTTTTAAAT	8	3.74	AAAAAATTT/AAATTTTTT	32	14.95
10-mer	AGATGTAAGA	61	17.38	AAGAAGATGT/ACATCTTCTT	123	35.04

Table 6
Information on primers designed to amplify 10 random SSR loci in the *Opuntia* cultivars.

Sequence ID	Reference for primers	SSR motif sequence and repeats	Primer sequences (5′ → 3′)	Primer length	%GC	T_a
Scaffold_4377	Locus L5354	(TTTC)$_{10}$	For: TTAGGACTCGCCAATCTTCTGG	22	50.0	62.5
	(Modise et al., 2024)		Rev: TTCCATTAGCCTCCTCCATCAA	22	45.5	
Scaffold_40838	Locus L10464	(ACAT)$_7$	For: GGTGGCTCAAGGAAGTGTATGG	22	54.5	62.5
	(Modise et al., 2024)		Rev: TGCATGTTTGGCAGCTCAGTAT	22	45.5	
Scaffold_571	This study	(TC)$_{10}$	For: GCCATCGGAAGTGCTCTCA	19	57.9	55.7
			Rev: TGCAGCCCAACCCAAACATA	20	50.0	
Scaffold_25280	Locus L23031	(GAG)$_{12}$	For: GAGTGATGAATGTTGGTGGTGCT	23	47.8	62.5
	(Modise et al., 2024)		Rev: ACCATCTCCTCCTTCTGGTTGAC	23	52.2	
Scaffold_31908	Locus L28477	(ATT)$_{11}$	For: TGCGAGGTAGACGTGTTGGA	20	55.0	62.5
	(Modise et al., 2024)		Rev: TGCTCTCGACTCTCCCCACT	20	60.0	
Scaffold_15531	Locus L37320	(AAAG)$_8$	For: CAGCCAACGACCCAACATCTAT	22	50.0	62.5
	(Modise et al., 2024)		Rev: CAATGTCCTCCCTTCCACTGTC	22	54.5	
Scaffold_30887	Locus L38909	(TTCT)$_5$	For: AGGTCCGTCATAGTCCCTCC	20	60.0	55.7
	(Modise et al., 2024)		Rev: TCATCGGTTGAGAATGGGCT	20	50.0	
Scaffold_64939	Locus L86067	(ATT)$_{13}$	For: CCACATCACCATGCAAACCATT	22	45.5	62.5
	(Mokoboki et al., 2009)		Rev: TGTTGTTGCGCCTGCTCTATG	21	52.4	
Scaffold_1150	This study	(TA)$_{14}$	For: TGCTGATTATCCACTGAGCGT	21	47.6	50.0
			Rev: CAGATGGTTGCCTCACTCAA	20	50.0	
Scaffold_127085	Locus L161092	(TGT)$_{10}$	For: TGCAAGGGTAAGACTGCCTAC	21	52.4	55.7
	(Modise et al., 2024)		Rev: AAAGCCGACTCAAGCACGA	19	52.6	

2017; Zheng et al., 2021). Yet, the species *O. ficus-indica* has been afforded little attention, though it is one of the key species in the family. Importantly, genome sequences offer an opportunity to gain insight into the biology and evolution of taxa (Henry, 2022; Marks et al., 2021). For instance, comparative genomics in Cactaceae revealed extensive hemiplasy that is said to produce phenotypic trait parallelisms that are extensive among cacti (Copetti et al., 2017). Further, Zheng et al. (2021) described the chromosomal co-localization of Caryophyllales-restricted betacyanin genes in dragon fruit, whereas Amaral et al. (2021) identified signals for the positive selection of metabolic genes that influence drought stress and nutrient absorption in the cactus *Cereus fernambucensis* Lem. (Cereeae) from South America. On the contrary, this study aimed to sequence and assemble a partial reference *Opuntia* pan-genome for the purpose of identifying sequence variants (in the form of SNPs and SSRs) that can be used to distinguish and assess genetic diversity in different cultivars of *Opuntia*. In the end, our sequencing efforts were expectedly quite inadequate to assemble a complete reference genome, yet the assembled partial pan-genome revealed thousands of high-quality sequence variants and should serve as a valuable resource to aid molecular studies in *Opuntia* species.

The lack of a complete nuclear genome from *Opuntia* species, thus far, is most likely because of their very large sizes. Nuclear DNA content in *O. ficus-indica* was determined through flow-cytometry to be 2C = 4.90 pg, and, thus, 1C = 2.45 pg (Segura et al., 2007). Considering this amount, the genome size of *O. ficus-indica* can, therefore, be estimated via the equation: genome size (bp) = (0.978 × 10^9) x DNA content (pg) (Dolezel et al., 2003), to be ~2.4 Gbp. This size is noticeably larger than genomes of all Cactaceae species sequenced to date (~0.98 – 1.68 Gbp; Copetti et al. 2017; Amaral et al. 2021; Zheng et al. 2021). That the current study assembled data totalling only 657 Mbp means roughly 30 % of the predicted *O. ficus-indica* genome was

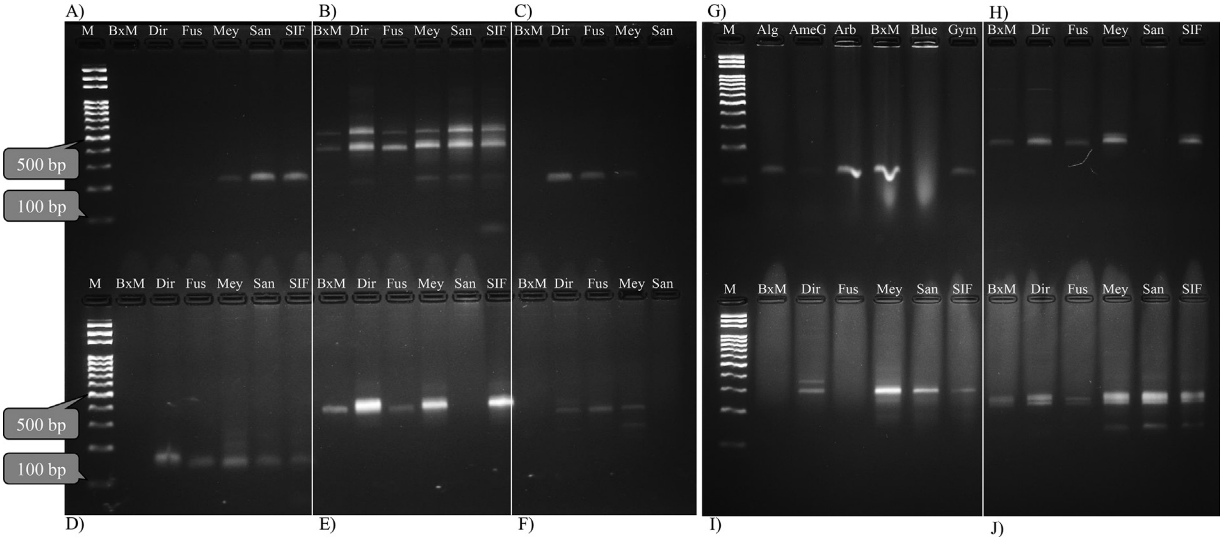

Fig. 4. Agarose gel electrophoresis of the 10 SSR loci amplified from selected *Opuntia* cultivars. The loci occur on *Opuntia* genomic sequences with the following tags: A) Scaffold_4377, B) Scaffold_40838, C) Scaffold_25280, D) Scaffold_31908, E) Scaffold_15531, F) Scaffold_64939, G) Scaffold_ 571, H) Scaffold_30887, I) Scaffold_1150 and J) Scaffold_127085. Cultivar names were abbreviated as follows: Alg = 'Algerian', AmeG = 'American Giant', Arb = 'Arbiter', BxM = 'Berg x Mexican', Blue = 'Blue Motto', Dir = 'Direkteur', Fus = 'Fusicaulis', Gym = 'Gymno Carpo', Mey = 'Meyers', San = 'Santa Rosa' and SIF = 'Sicillian Indian fig'. Lane M contains the PCR DNA ladder.

sequenced and much more is still missing. Indeed, it was anticipated that the generated sequence data would not be enough to assemble an entire pan-genome. Accordingly, additional efforts involving both short- and long-read sequencing (Henry, 2022; Li and Harkess, 2018) are required to sequence an entire *Opuntia* reference genome and improve the current pan-genome.

To our knowledge, the current study is the first to sequence and assemble nuclear genomic sequences of *Opuntia* species. Lately, there have been a few NGS studies in *Opuntia*, however, and they included reports on chloroplast genomes of different species (Chen et al., 2022; Köhler et al., 2020; Majure et al., 2023), the analysis of microRNA expression during fruit development in *O. robusta* (Guerrero-Garibay et al., 2023) as well as the analyses of differential gene expression in fruit and cladode tissues of three *Opuntia* species (Valadez-Moctezuma et al., 2023). The present study, in contrast, generated much more data even if it was still inadequate to assemble a complete *Opuntia* pan-genome. Precisely, the 28.1 Gbp of data represents sequence coverage of around 12-fold based on the estimated genome size of *O. ficus-indica* being ~2.4 Gbp. For the *de novo* assembly of genomes, however, it is advised that sequence coverage should be above 50-fold because high-coverage would help resolve the highly repetitive genomic regions that are generally difficult to sequence and order (Li et al., 2017; Miller et al., 2010). Further, that the *de novo* assembler IDBA-UD yielded a better assembly was not surprising because it seems to be good at assembling genome sequences having large regions with heterozygous positions (Lischer and Shimizu, 2017), as in the current data. Thus, although it is still short, the current data is the largest produced so far for *Opuntia* species, and the assembled partial pan-genome is likely the most optimal that can be created using the available sequence data.

Despite the relatively small amount of genome sequence assembled herein, a large portion of the data could not be annotated at the NCBI NR or UniProtKB/SwissProt protein databases. Having said that, the annotated *Opuntia* sequences generally matched protein sequences from other plant species, with subjects at the NR database mainly representing taxa from different members of Caryophyllales – the same taxonomic order to which *Opuntia* belongs (Simpson, 2010). Nonetheless, Valadez-Moctezuma et al. (2023) likewise achieved a low annotation rate (23 %) for the 16,068 *Opuntia* transcripts queried at the SwissProt database. In this study, the low annotation rate can be reasoned by three causes. First, the *Opuntia* protein-coding gene regions were predicted based only on gene models from *Arabidopsis thaliana*, as this is the best characterized plant genome. However, the enormous genomes of *Opuntia* taxa imply that there is potential for novel Caryophyllales- or *Opuntia*-specific genes occurring in the genome, which would not be identified using this tactic. Also, *ab initio* gene prediction based on large draft genomes can be quite an error-prone process (Scalzitti et al., 2020). As such, the protein-coding genes predicted in the partial *Opuntia* pan-genome must be validated in future, after additional efforts of genome sequencing. Nonetheless, the current data should still provide a good preliminary resource for gene-based studies in *Opuntia*.

Second, the unannotated sequences may correspond to non-coding genic regions including promoters and introns. In addition, some sequences may represent intergenic regions, and these can be hundreds of kilobases long in certain plant genomes (Gottlieb et al., 2013). The non-coding genic and intergenic sequences do not occur in protein databases, and so, cannot be annotated via the current plan. Notably, genome collinearity has been used as a key concept toward the annotation of intergenic sequences in different grass genomes (Gottlieb et al., 2013). Accordingly, further characterization of the currently sequenced Cactaceae genomes could help annotate non-coding sequences in the partial *Opuntia* pan-genome.

Third, the unannotated sequences may correspond to sequence misassemblies that are inherent to *de novo* assembly methods (Sohn and Nam, 2018). Correcting such misassemblies may involve the use of new software tools and suitable reference genomes (Chawla et al., 2016; Zhu et al., 2015) – a plan that is currently not viable for the *Opuntia* pan-genome because available putative reference genomes are not only smaller and but also annotated partially. Another issue is that the larger *Opuntia* genomes may comprise lineage-specific sequences that will be challenging to assemble and annotate accurately especially as read coverage in this study was limited. Perhaps the method of reference-guided *de novo* assembly, which involves using genomes from species that are closely related to the study organism, may offer a better result for correcting sequence misassemblies (Lischer and Shimizu, 2017). Overall, the current partial *Opuntia* pan-genome requires improvement via additional sequencing and analysis with appropriate tools that can resolve assembly challenges including misassemblies.

Annotation of sequences in the partial *Opuntia* pan-genome showed that they were enriched for genes involved in the formation of new shoots and developmental structures including vegetative and flower tissues. Such functions are expected in actively growing plant tissues (Gaillochet and Lohmann, 2015; Srivastava, 2002) and these results correlate with the sequenced *Opuntia* specimens originating from young, actively growing tissues. Other genes were associated with the 'protein serine kinase activity', which effects cellular functions including hormone-mediated signalling and cell-cycle control during plant development (Laurie and Halford, 2001; Srivastava, 2002). Indeed, genes in the brassinosteroid and auxin signalling pathways – whose end-products are key hormones that enable plant growth and development (Gaillochet and Lohmann, 2015; Srivastava, 2002; Tian et al., 2017) – were similarly enriched in the partial *Opuntia* pan-genomic sequences. Moreover, that the glycerophospholipid and glycerolipid pathway genes were also enriched in the current data is reasonable since their metabolic products aid survival in water-stressed environments (Xu et al., 2020). The genus *Opuntia* comprises succulent plants that can endure harsh temperatures and water-stressed conditions (Granados-Aguilar et al., 2022; Snyman et al., 2007). Valadez-Moctezuma et al. (2023), however, found that the *Opuntia* tissues were enriched with transcripts corresponding to peculiar GO functions, and these were linked to varying climatic conditions endured by the plants around the time they were harvested. In any case, our results correlate with the fact that the sequenced samples originated from actively growing tissues. Genes identified herein should benefit future efforts aiming to study the molecular genetics of growth and development in *Opuntia* specimens. Such information can help understand the genetic control of leaf development in other succulent plants (Heyduk, 2021).

An underlying aim of this study was to identify SNP and SSR loci that can be used to distinguish and assess the levels of genetic diversity in *Opuntia* samples. The discovery of such variants in non-model plant genomes is now common due to the increased accessibility of NGS platforms. Recent examples on the genome-wide discovery of SNPs from NGS data include studies in avocado (*Persea americana* Mill.; Talavera et al. 2019), African oil palm (*Elaeis guineensis* L.; Xia et al. 2019), selected pumpkin species (*Cucurbita* spp.; Nguyen et al. 2020), Oriental melon (*Cucumis melo* L. var. *makuwa*; Kishor et al. 2021) and lettuce (*Lactuca sativa* L.; Park et al. 2022). Because different studies apply different threshold criteria to locate sequence variants, the identified variants ranged from 5,640 biallelic SNPs genotyped across 48 melon varieties (Kishor et al., 2021) to >1.2 million loci showing MAF > 0.05 across 200 oil palm individuals (Xia et al., 2019). In this study, the 10 *Opuntia* samples yielded over 118,000 biallelic SNPs that each occurred at a relatively higher level of frequency (MAF > 0.2). A higher MAF cutoff applied in this study means that the identified variants were of moderate to high frequency – implying that the studied cultivars harbour moderate to high levels of genetic diversity. Just as a small selection of the SNPs were adequate to distinguish the 10 *Opuntia* cultivars from each other, future studies can extract SNPs from the collection 118,000 and develop a core dataset for widespread cultivar identification (Park et al., 2022; Yuan et al., 2022) or even perform trait-marker association analyses (Hou et al., 2018; Kishor et al., 2021). In *Opuntia*, fruit quality has become one of the most desirable traits (Coetzer et al., 2019; Mokoboki et al., 2009) that can serve as a good model trait for association analysis using SNP data. The association of SNPs with fruit quality has earlier been done in peach (*Prunus persica* L.; Font i Forcada et al., 2019) and Chinese jujube (*Ziziphus jujuba* Mill.; Hou et al., 2020). Therefore, the identified SNPs provide a valuable molecular resource for exploring trait-marker associations within the South African *Opuntia* germplasm.

The genomic characterization of SSRs has been done in many non-model plants and recent examples include sago palm (*Metroxylon sagu* Rottb.; Purwoko et al. 2019), olive (*Olea europaea* L.; Li et al. 2020), Siberian wild rye (*Elymus sibiricus* L.; Xiong et al. 2021), the Australian silver oak (*Grevillea robusta* A. Cunn. ex R. Br.; Dabral et al. 2021) and black pepper (*Piper nigrum* L.; Negi et al. 2022). Unlike the 59,963 loci detected in the partial *Opuntia* pan-genome, SSRs reported in the above studies ranged from 13,335 loci in the partial genome of silver oak (Dabral et al., 2021) to as many as 276,230 loci across the entire black pepper genome (Negi et al., 2022). Almost all studies including the current observed 2-mers to be the most frequent motif (however, 2-mers were left out of the analysis on olive accessions; Li et al. 2020) and this finding correlates with reports from other plant genomes (Kalia et al., 2011; Ranade et al., 2014). A random selection of 10 SSR loci in the *Opuntia* pan-genome showed that not only were they amplifiable, but eight of the loci also appeared to be polymorphic. Indeed, the eight loci have since been used to distinguish and assess levels of diversity in 44 cultivars that makeup the South African *Opuntia* germplasm (Modise et al., 2024). Therefore, the generated library of SSRs expands the collection of molecular markers that can be used to examine genetic relationships among specimens of *Opuntia* (Caruso et al., 2010; Reis et al., 2018; Samah et al., 2016).

5. Conclusions

Opuntia species are succulent plants known for occupying and surviving in extreme temperature and water-scarce environments. They are also good sources for many nutrients and bioactive compounds. Notwithstanding, very little has been done in terms of deciphering the very large genomes of *Opuntia* species – a feat that has already been achieved in a few members of the Cactaceae family. The current study pioneered the sequencing and assembly of an *Opuntia* pan-genome using DNA originating from 10 different cultivars. The generated sequence data corresponded to a low coverage of about 12-fold, relative to the predicted 2.4 Gbp genome of *O. ficus-indica*. Thus, the assembled pan-genome remains mostly unfinished, though it contained many genes that correlated with actively growing tissues and survival in water-stressed conditions. In the future, further sequencing using different NGS platforms (thus, short- and long-read sequencers) should permit the assembly of a complete *Opuntia* reference genome and validate the annotation of genes identified herein. Screening the partial pan-genome for sequence variants returned >118,000 biallelic SNPs and ~60,000 SSR loci. This collection of SNPs presents a valuable genetic resource for developing cultivar-specific and trait-linked markers that are missing in *Opuntia*, while a few SSR loci have already been found to be useful in discriminating and measuring the levels of genetic diversity in *Opuntia* cultivars that form the South African germplasm (Modise et al., 2024). Overall, the current study has set the genomics foundation in *Opuntia* and future research should aim to improve the *Opuntia* pan-genome as well as associate the identified genetic diversity with variation in phenotypic characters such as fruit colour and quality (Felker et al., 2008).

Funding

The authors wish to declare that they have no affiliations or interest in any organizations/entities that have financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Declaration of competing interest

The authors declare that no external funding was received to help with the execution of this study or preparation of this manuscript.

CRediT authorship contribution statement

Mathabatha F. Maleka: Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Tsholofelo J. Modise:** Writing – review & editing, Writing – original draft, Methodology. **Morné G. Du Plessis:** Writing – review & editing, Visualization, Validation, Software. **Gesine M. Coetzer:** Writing – review & editing, Writing – original draft, Supervision, Resources, Funding acquisition, Conceptualization.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Amaral, D.T., Bombonato, J.R., da Silva Andrade, S.C., Moraes, E.M., Franco, F.F., 2021. The genome of a thorny species: comparative genomic analysis among South and North American Cactaceae. Planta 254, 44. https://doi.org/10.1007/s00425-021-03690-5.

Aranzana, M.J., Carbó, J., Arús, P., 2003. Using amplified fragment-length polymorphisms (AFLPs) to identify peach cultivars. J. Am. Soc. Hortic. Sci. 128, 672–677. https://doi.org/10.21273/jashs.128.5.0672.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29.

Avraham, S., Tung, C.-W., Ilic, K., Jaiswal, P., Kellogg, E.A., McCouch, S., Pujar, A., Reiser, L., Rhee, S.Y., Sachs, M.M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Zapata, F., Ware, D., 2008. The plant ontology database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Res. 36, D449–D454. https://doi.org/10.1093/nar/gkm908.

Bandelj, D., Jakse, J., Javornik, B., 2002. DNA fingerprinting of olive varieties by microsatellite markers. Food Technol. Biotechnol. 40, 185–190.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. https://doi.org/10.1089/cmb.2012.0021.

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176.

Caruso, M., Currò, S., Casas, G.Las, la Malfa, S., Gentile, A., 2010. Microsatellite markers help to assess genetic diversity among *Opuntia ficus-indica* cultivated genotypes and their relation with related species. Plant Syst. Evol. 290, 85–97. https://doi.org/10.1007/s00606-010-0351-9.

Chapman, B., Mondragon Jacobo, C., Bunch, R.A., Paterson, A.H., 2002. Breeding and biotechnology. In: Nobel, P.S. (Ed.), Cacti: Biology and Uses. University of California Press, Los Angeles, pp. 255–271.

Chawla, V., Kumar, R., Shankar, R., 2016. Identifying wrong assemblies in *de novo* short read primary sequence assembly contigs. J. Biosci. 41, 455–474. https://doi.org/10.1007/s12038-016-9630-0.

Chen, H., Morrell, P.L., Ashworth, V.E.T.M., De La Cruz, M., Clegg, M.T., 2009. Tracing the geographic origins of major avocado cultivars. J. Hered. 100, 56–65. https://doi.org/10.1093/jhered/esn068.

Chen, J., Zhang, S., Tang, W., Du, X., Yuan, Y., Wu, S., 2022. The complete chloroplast genome sequence of *Opuntia sulphurea* (Cactaceae). Mitochondrial DNA Part B, Resour. 7, 361–362. https://doi.org/10.1080/23802359.2022.2035837.

Choudhary, P., Khanna, S.M., Jain, P.K., Bharadwaj, C., Kumar, J., Lakhera, P.C., Srinivasan, R., 2012. Genetic structure and diversity analysis of the primary gene pool of chickpea using SSR markers. Genet. Mol. Res. 11, 891–905. https://doi.org/10.4238/2012.April.10.5.

Coetzer, G.M., De Wit, M., Fouché, H.J., Venter, S.L., 2019. Climatic influences on fruit quality and sensory traits of cactus pear (*Opuntia ficus-indica*): a 5-year evaluation. Acta Hortic. 1247, 23–30. https://doi.org/10.17660/ActaHortic.2019.1247.4.

Copetti, D., Búrquez, A., Bustamante, E., Charboneau, J.L.M., Childs, K.L., Eguiarte, L.E., Lee, S., Liu, T.L., McMahon, M.M., Whiteman, N.K., Wing, R.A., Wojciechowski, M.F., Sanderson, M.J., 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proc. Natl. Acad. Sci. U.S.A. 114, 12003–12008. https://doi.org/10.1073/pnas.1706367114.

Dabral, A., Shamoon, A., Meena, R.K., Kant, R., Pandey, S., Ginwal, H.S., Bhandari, M.S., 2021. Genome skimming-based simple sequence repeat (SSR) marker discovery and characterization in *Grevillea robusta*. Physiol. Mol. Biol. Plants 27, 1623–1638. https://doi.org/10.1007/s12298-021-01035-w.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., Group, 1000 Genomes Project Analysis, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

de Araújo, F.F., de Paulo Farias, D., Neri-Numa, I.A., Pastore, G.M., 2021. Underutilized plants of the Cactaceae family: nutritional aspects and technological applications. Food Chem. 362, 130196. https://doi.org/10.1016/j.foodchem.2021.130196.

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M.B., Hirsch, C.N., 2021. How the pan-genome is changing crop genomics and improvement. Genome Biol. 22, 3. https://doi.org/10.1186/s13059-020-02224-8.

Diederichsen, A., Fu, Y.-B., 2006. Phenotypic and molecular (RAPD) differentiation of four infraspecific groups of cultivated flax (*Linum usitatissimum* L. subsp. *usitatissimum*). Genet. Resour. Crop Evol. 53, 77–90. https://doi.org/10.1007/s10722-004-0579-8.

Dolezel, J., Bartos, J., Voglmayr, H., Greilhuber, J., Thomas, R.A., 2003. Nuclear DNA content and genome size of trout and human. Cytom. Part A 51, 127–129.

Felker, P., Stintzing, F.C., Müssig, E., Leitenberger, M., Carle, R., Vogt, T., Bunch, R., 2008. Colour inheritance in cactus pear (*Opuntia ficus-indica*) fruits. Ann. Appl. Biol. 152, 307–318. https://doi.org/10.1111/j.1744-7348.2008.00222.x.

Font i Forcada, C., Guajardo, V., Chin-Wo, S.R., Moreno, M.Á., 2019. Association mapping analysis for fruit quality traits in *Prunus persica* using SNP markers. Front. Plant Sci. 9. https://doi.org/10.3389/fpls.2018.02005.

Forapani, S., Carboni, A., Paoletti, C., Moliterni, V.M.C., Ranalli, P., Mandolino, G., 2001. Comparison of hemp varieties using random amplified polymorphic DNA markers. Crop Sci. 41, 1682–1689. https://doi.org/10.2135/cropsci2001.1682.

Franco, F.F., Amaral, D.T., Bonatelli, I.A.S., Romeiro-Brito, M., Telhe, M.C., Moraes, E.M., 2022. Evolutionary genetics of cacti: research biases, advances and prospects. Genes 13, 452. https://doi.org/10.3390/genes13030452.

Gaillochet, C., Lohmann, J.U., 2015. The never-ending story: from pluripotency to plant developmental plasticity. Development 142, 2237–2249. https://doi.org/10.1242/dev.117614.

Ge, S.X., Jung, D., Yao, R., 2020. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics 36, 2628–2629. https://doi.org/10.1093/bioinformatics/btz931.

Gottlieb, A., Müller, H.-G., Massa, A.N., Wanjugi, H., Deal, K.R., You, F.M., Xu, X., Gu, Y.Q., Luo, M.-C., Anderson, O.D., Chan, A.P., Rabinowicz, P., Devos, K.M., Dvorak, J., 2013. Insular organization of gene space in grass genomes. PLoS One 8, e54101.

Granados-Aguilar, X., Palomino, G., Martínez-Ramón, J., Arias, S., 2022. Genome evolution and phylogenetic relationships in *Opuntia tehuacana* (Cactaceae, Opuntioideae). Brazilian J. Bot. 45, 957–969. https://doi.org/10.1007/s40415-022-00821-4.

Guerrero-Garibay, S., Olvera-Martı́nez, F., Aceves-Monreal, D., López de Alba, P.L., Cruz-Hernández, A., 2023. Molecular strategies for prickly pear (*Opuntia* sp.) studies and its improvement. Acta Hortic. 495–498. https://doi.org/10.17660/ActaHortic.2023.1362.66.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

Henry, R.J., 2022. Progress in plant genome sequencing. Appl. Biosci. 1, 113–128. https://doi.org/10.3390/applbiosci1020008.

Heyduk, K., 2021. The genetic control of succulent leaf development. Curr. Opin. Plant Biol. 59, 101978. https://doi.org/10.1016/j.pbi.2020.11.003.

Hou, L., Chen, W., Zhang, Z., Pang, X., Li, Y., 2020. Genome-wide association studies of fruit quality traits in jujube germplasm collections using genotyping-by-sequencing. Plant Genome 13, e20036. https://doi.org/10.1002/tpg2.20036.

Hou, S., Zhu, G., Li, Y., Li, W., Fu, J., Niu, E., Li, L., Zhang, D., Guo, W., 2018. Genome-wide association studies reveal genetic variation and candidate genes of drought stress related traits in cotton (*Gossypium hirsutum* L.). Front. Plant Sci. 9, 1276. https://doi.org/10.3389/fpls.2018.01276.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R., Dhawan, A.K., 2011. Microsatellite markers: an overview of the recent progress in plants. Euphytica 177, 309–334. https://doi.org/10.1007/s10681-010-0286-9.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2015. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462. https://doi.org/10.1093/nar/gkv1070.

Kishor, D.S., Noh, Y., Song, W.-H., Lee, G.P., Park, Y., Jung, J.-K., Shim, E.-J., Sim, S.-C., Chung, S.-M., 2021. SNP marker assay and candidate gene identification for sex expression via genotyping-by-sequencing-based genome-wide associations (GWAS) analyses in Oriental melon (*Cucumis melo* L. var. makuwa). Sci. Hortic. 276, 109711. https://doi.org/10.1016/j.scienta.2020.109711.

Köhler, M., Reginato, M., Souza-Chies, T.T., Majure, L.C., 2020. Insights into chloroplast genome evolution across Opuntioideae (Cactaceae) reveals robust yet sometimes conflicting phylogenetic topologies. Front. Plant Sci. 11, 729. https://doi.org/10.3389/fpls.2020.00729.

Kumar, S., Palsaniya, D.R., Kumar, T.K., Misra, A.K., Ahmad, S., Rai, A.K., Sarker, A., Louhaichi, M., Hassan, S., Liguori, G., Ghosh, P.K., Govindasamy, P., Mahawer, S.K., Bhargavi, H.A., 2023. Survival, morphological variability, and performance of *Opuntia ficus-indica* in a semi-arid region of India. Arch. Agron. Soil Sci. 69, 708–725. https://doi.org/10.1080/03650340.2022.2031998.

Laurie, S., Halford, N.G., 2001. The role of protein kinases in the regulation of plant growth and development. Plant Growth Regul. 34, 253–265. https://doi.org/10.1023/A:1013311807626.

Li, C., Lin, F., An, D., Wang, W., Huang, R., 2017. Genome sequencing and assembly by long reads in plants. Genes 9, 6. https://doi.org/10.3390/genes9010006.

Li, D., Long, C., Pang, X., Ning, D., Wu, T., Dong, M., Han, X., Guo, H., 2020. The newly developed genomic-SSR markers uncover the genetic characteristics and relationships of olive accessions. PeerJ. 8, e8573. https://doi.org/10.7717/peerj.8573.

Li, F.-W., Harkess, A., 2018. A guide to sequence your favorite plant genomes. Appl. Plant Sci. 6, e1030. https://doi.org/10.1002/aps3.1030.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, W., Liu, J., Zhang, H., Liu, Z., Wang, Y., Xing, L., He, Q., Du, H., 2022. Plant pan-genomics: recent advances, new challenges, and roads ahead. J. Genet. Genomics 49, 833–846. https://doi.org/10.1016/j.jgg.2022.06.004.

Lischer, H.E.L., Shimizu, K.K., 2017. Reference-guided *de novo* assembly approach improves genome reconstruction for related species. BMC Bioinformatics 18, 474. https://doi.org/10.1186/s12859-017-1911-6.

Ma, X., Yan, H., Yang, J., Liu, Y., Li, Z., Sheng, M., Cao, Y., Yu, X., Yi, X., Xu, W., Su, Z., 2022. PlantGSAD: a comprehensive gene set annotation database for plant species. Nucleic Acids Res. 50, D1456–D1467. https://doi.org/10.1093/nar/gkab794.

Majure, L.C., Murphy, T.H., Köhler, M., Puente, R., Hodgson, W.C., 2023. Evolution of the xerocarpa clade (*Opuntia*; Opuntieae): evidence for the role of the Grand Canyon in the giogeographic history of the iconic beavertail cactus and relatives. Plants 12, 2677. https://doi.org/10.3390/plants12142677.

Marks, R.A., Hotaling, S., Frandsen, P.B., VanBuren, R., 2021. Representation and participation across 20 years of plant genome sequencing. Nat. Plants 7, 1571–1578. https://doi.org/10.1038/s41477-021-01031-8.

Martins, M., Ribeiro, M.H., Almeida, C.M.M., 2023. Physicochemical, nutritional, and medicinal properties of *Opuntia ficus-indica* (L.) Mill. and its main agro-industrial use: a review. Plants 12, 1512. https://doi.org/10.3390/plants12071512.

Mashope, B.K., Herselman, L., Labuschagne, M.T., 2011. Genetic diversity among South African cactus pear genebank accessions using AFLP markers. Bradleya 103–114. https://doi.org/10.25223/brad.n29.2011.a12.

Miller, J.R., Koren, S., Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. Genomics 95, 315–327.

Modise, T.J., Maleka, M.F., Fouché, H., Coetzer, G.M., 2024. Genetic diversity and differentiation of South African cactus pear cultivars (*Opuntia* spp.) based on simple sequence repeat (SSR) markers. Genet. Resour. Crop Evol. 71, 373–384. https://doi.org/10.1007/s10722-023-01629-1.

Mokoboki, K., Kgama, T., Mmbi, N., 2009. Evaluation of cactus pear fruit quality at Mara ADC, South Africa. African J. Agric. Res. 4, 028–032.

Naithani, S., Preece, J., D'Eustachio, P., Gupta, P., Amarasinghe, V., Dharmawardhana, P.D., Wu, G., Fabregat, A., Elser, J.L., Weiser, J., Keays, M., Fuentes, A.M.-P., Petryszak, R., Stein, L.D., Ware, D., Jaiswal, P., 2017. Plant reactome: a resource for plant pathways and comparative analysis. Nucleic Acids Res. 45, D1029–D1039. https://doi.org/10.1093/nar/gkw932.

Negi, A., Singh, K., Jaiswal, S., Kokkat, J.G., Angadi, U.B., Iquebal, M.A., Umadevi, P., Rai, A., Kumar, D., 2022. Rapid genome-wide location-specific polymorphic SSR marker discovery in black pepper by GBS approach. Front. Plant Sci. 13, 846937. https://doi.org/10.3389/fpls.2022.846937.

Neupane, D., Mayer, J.A., Niechayev, N.A., Bishop, C.D., Cushman, J.C., 2021. Five-year field trial of the biomass productivity and water input response of cactus pear (*Opuntia* spp.) as a bioenergy feedstock for arid lands. GCB Bioenergy 13, 719–741. https://doi.org/10.1111/gcbb.12805.

Nguyen, N.N., Kim, M., Jung, J.-K., Shim, E.-J., Chung, S.-M., Park, Y., Lee, G.P., Sim, S.-C., 2020. Genome-wide SNP discovery and core marker sets for assessment of genetic variations in cultivated pumpkin (*Cucurbita* spp.). Hortic. Res. 7, 121. https://doi.org/10.1038/s41438-020-00342-9.

Oliveira, H.R., Hagenblad, J., Leino, M.W., Leigh, F.J., Lister, D.L., Penã-Chocarro, L., Jones, M.K., 2014. Wheat in the Mediterranean revisited - tetraploid wheat landraces assessed with elite bread wheat Single Nucleotide Polymorphism markers. BMC Genet. 15. https://doi.org/10.1186/1471-2156-15-54.

Omar, A.A., ElSayed, A.I., Mohamed, A.H., 2021. Genetic diversity and ecotypes of *Opuntia* spp. BT - *Opuntia* spp.: chemistry, bioactivity and industrial applications, in: Ramadan, M.F., Ayoub, T.E.M., Rohn, S. (Eds.). Springer International Publishing, Cham, pp. 181–199. 10.1007/978-3-030-78444-7_8

Park, J.-S., Kang, M.-Y., Shim, E.-J., Oh, J., Seo, K.-I., Kim, K.S., Sim, S.-C., Chung, S.-M., Park, Y., Lee, G.P., Lee, W.-S., Kim, M., Jung, J.-K., 2022. Genome-wide core sets of SNP markers and Fluidigm assays for rapid and effective genotypic identification of Korean cultivars of lettuce (*Lactuca sativa* L.). Hortic. Res. 9, uhac119. https://doi.org/10.1093/hr/uhac119.

Parks, E.J., Moyer, J.W., 2004. Evaluation of AFLP in Poinsettia: polymorphism selection, analysis, and cultivar identification. J. Am. Soc. Hortic. Sci. 129, 863–869. https://doi.org/10.21273/jashs.129.6.0863.

Peña-Valdivia, C.B., Luna-Cavazos, M., Carranza-Sabas, J.A., Reyes-Agüero, J.A., Flores, A., 2008. Morphological characterization of *Opuntia* spp.: a multivariate analysis. J. Prof. Assoc. Cactus Dev. 10, 1–21.

Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L., 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428. https://doi.org/10.1093/bioinformatics/bts174.

Potgieter, J.P., Mashope, B.K., 2009. Cactus pear (*Opuntia* spp.) germplasm conservation in South Africa. Acta Hortic. 811, 47–54. https://doi.org/10.17660/Acta-Hortic.2009.811.2.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. https://doi.org/10.1086/519795.

Purwoko, D., Cartealy, I.C., Tajuddin, T., Dinarti, D., Sudarsono, S., 2019. SSR identification and marker development for sago palm based on NGS genome data. Breed. Sci. 69, 1–10. https://doi.org/10.1270/jsbbs.18061.

Ranade, S.S., Lin, Y.-C., Zuccolo, A., Van de Peer, Y., García-Gil, M.del R., 2014. Comparative *in silico* analysis of EST-SSRs in angiosperm and gymnosperm tree genera. BMC. Plant Biol. 14, 220. https://doi.org/10.1186/s12870-014-0220-8.

Reis, C.M.G., Raimundo, J., Ribeiro, M.M., 2018. Assessment of genetic diversity in *Opuntia* spp. Portuguese populations using SSR molecular markers. Agronomy 8, 55. https://doi.org/10.3390/agronomy8040055.

Russell, C.E., Felker, P., 1987. The prickly-pears (*Opuntia* spp., Cactaceae): a source of human and animal food in semiarid regions. Econ. Bot. 41, 433–445.

Samah, S., De Teodoro Pardo, C.V., Serrato Cruz, M.A., Valadez-Moctezuma, E., 2016. Genetic diversity, genotype discrimination, and population structure of Mexican *Opuntia* sp., determined by SSR markers. Plant Mol. Biol. Rep. 34, 146–159. https://doi.org/10.1007/s11105-015-0908-4.

Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O., Thompson, J.D., 2020. A benchmark study of *ab initio* gene prediction methods in diverse eukaryotic organisms. BMC Genomics 21, 293. https://doi.org/10.1186/s12864-020-6707-9.

Segura, S., Scheinvar, L., Olalde, G., Leblanc, O., Filardo, S., Muratalla, A., Gallegos, C., Flores, C., 2007. Genome sizes and ploidy levels in Mexican cactus pear species *Opuntia* (Tourn.) Mill. series Streptacanthae Britton et Rose, Leucotrichae DC., Heliabravoanae Scheinvar and Robustae Britton et Rose. Genet. Resour. Crop Evol. 54, 1033–1041.

Shalev, T.J., Gamal El-Dien, O., Yuen, M.M.S., Shengqiang, S., Jackman, S.D., Warren, R.L., Coombe, L., van der Merwe, L., Stewart, A., Boston, L.B., Plott, C., Jenkins, J., He, G., Yan, J., Yan, M., Guo, J., Breinholt, J.W., Neves, L.G., Grimwood, J., Rieseberg, L.H., Schmutz, J., Birol, I., Kirst, M., Yanchuk, A.D., Ritland, C., Russell, J.H., Bohlmann, J., 2022. The western redcedar genome reveals low genetic diversity in a self-compatible conifer. Genome Res. 32, 1952–1964. https://doi.org/10.1101/gr.276358.121.

Simpson, M.G., 2010. Diversity and classification of flowering plants: Eudicots, in: Simpson, M.G.B.T.-P.S, 2nd Ed. Academic Press, San Diego, pp. 275–448. https://doi.org/10.1016/B978-0-12-374380-0.50008-7.

Snyman, H.A., Fouché, H.J., Avenant, P.L., Ratsele, C., 2007. Frost sensitivity of *Opuntia ficus-indica* and *O. robusta* in a semiarid climate of South Africa. J. Prof. Assoc. Cactus Dev. 9, 1–21.

Sohn, J., Nam, J.-W., 2018. The present and future of *de novo* whole-genome assembly. Brief. Bioinform. 19, 23–40. https://doi.org/10.1093/bib/bbw096.

Srivastava, L., 2002. Plant growth and development: hormones and environment. Elsevier Science, California.

Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B., 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32, W309–W312. https://doi.org/10.1093/nar/gkh379.

Stintzing, F.C., Carle, R., 2005. Cactus stems (*Opuntia* spp.): a review on their chemistry, technology, and uses. Mol. Nutr. Food Res. 49, 175–194. https://doi.org/10.1002/mnfr.200400071.

Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., Guo, L., 2022. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. 27, 391–401. https://doi.org/10.1016/j.tplants.2021.10.006.

Talavera, A., Soorni, A., Bombarely, A., Matas, A.J., Hormaza, J.I., 2019. Genome-wide SNP discovery and genomic characterization in avocado (*Persea americana* Mill.). Sci. Rep. 9, 20137. https://doi.org/10.1038/s41598-019-56526-4.

The UniProt Consortium, 2023. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 51, D523–D531. https://doi.org/10.1093/nar/gkac1052.

Thiel, T., Michalek, W., Varshney, R., Graner, A., 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor. Appl. Genet. 106, 411–422.

Tian, H., Lv, B., Ding, T., Bai, M., Ding, Z., 2017. Auxin-Br interaction regulates plant growth and development. Front. Plant Sci. 8, 2256. https://doi.org/10.3389/fpls.2017.02256.

Valadez-Moctezuma, E., Samah, S., Mascorro-Gallardo, J.O., Marbán-Mendoza, N., Aranda-Osorio, G., Flores-Girón, E., Brito-Nájera, G., Rodríguez de la O, J.L., 2023. The first transcriptomic analyses of fruits and cladodes for comparison between three species of *Opuntia*. Genet. Resour. Crop Evol. 70, 951–970. https://doi.org/10.1007/s10722-022-01480-w.

Wallace, R.S., Gibson, A.C., 2002. Evolution and systematics. In: Nobel, P., Portes, A. (Eds.), Cacti: Biology and Uses. University of California Press, pp. 1–21. https://doi.org/10.1525/california/9780520231573.003.0001.

Wang, X., Chen, S., Ma, X., Yssel, A.E.J., Chaluvadi, S.R., Johnson, M.S., Gangashetty, P., Hamidou, F., Sanogo, M.D., Zwaenepoel, A., Wallace, J., Van de Peer, Y., Bennetzen, J.L., Van Deynze, A., 2021. Genome sequence and genetic diversity analysis of an under-domesticated orphan crop, white fonio (*Digitaria exilis*). Gigascience 10, giab013. https://doi.org/10.1093/gigascience/giab013.

Wessels, A.B., Van Der Merwe, L.L., Du Plessis, H., 1997. Yield variation in clonally propagated *Opuntia ficus-indica* (L.) Mill. Plants when terminal cladodes are used. Acta Horticulturae. International Society for Horticultural Science (ISHS), Leuven, Belgium, pp. 73–76. https://doi.org/10.17660/ActaHortic.1997.438.8.

Wickham, H., 2016. ggplot2: Elegant Graphics For Data Analysis. Springer-Verlag, New York.

Xia, W., Luo, T., Zhang, W., Mason, A.S., Huang, D., Huang, X., Tang, W., Dou, Y., Zhang, C., Xiao, Y., 2019. Development of high-density SNP markers and their application in evaluating genetic diversity and population structure in *Elaeis guineensis*. Front. Plant Sci. 10, 130. https://doi.org/10.3389/fpls.2019.00130.

Xiong, Yi, Lei, X., Bai, S., Xiong, Yanli, Liu, W., Wu, W., Yu, Q., Dong, Z., Yang, J., Ma, X., 2021. Genomic survey sequencing, development and characterization of single- and multi-locus genomic SSR markers of *Elymus sibiricus* L. BMC Plant Biol. 21, 3. https://doi.org/10.1186/s12870-020-02770-0.

Xu, C., Jiao, C., Sun, H., Cai, X., Wang, X., Ge, C., Zheng, Y., Liu, W., Sun, X., Xu, Y., Deng, J., Zhang, Z., Huang, S., Dai, S., Mou, B., Wang, Quanxi, Fei, Z., Wang, Quanhua, 2017. Draft genome of spinach and transcriptome diversity of 120 *Spinacia* accessions. Nat. Commun. 8, 15275. https://doi.org/10.1038/ncomms15275.

Xu, H., Li, Z., Tong, Z., He, F., Li, X., 2020. Metabolomic analyses reveal substances that contribute to the increased freezing tolerance of alfalfa (*Medicago sativa* L.) after continuous water deficit. BMC Plant Biol. 20, 15. https://doi.org/10.1186/s12870-019-2233-9.

Yuan, X., Li, Z., Xiong, L., Song, S., Zheng, X., Tang, Z., Yuan, Z., Li, L., 2022. Effective identification of varieties by nucleotide polymorphisms and its application for essentially derived variety identification in rice. BMC Bioinformatics 23, 30. https://doi.org/10.1186/s12859-022-04562-9.

Zheng, J., Meinhardt, L.W., Goenaga, R., Zhang, D., Yin, Y., 2021. The chromosome-level genome of dragon fruit reveals whole-genome duplication and chromosomal co-localization of betacyanin biosynthetic genes. Hortic. Res. 8, 63. https://doi.org/10.1038/s41438-021-00501-6.

Zhu, X., Leung, H.C.M., Wang, R., Chin, F.Y.L., Yiu, S.M., Quan, G., Li, Y., Zhang, R., Jiang, Q., Liu, B., Dong, Y., Zhou, G., Wang, Y., 2015. misFinder: identify mis-assemblies in an unbiased manner using reference and paired-end reads. BMC Bioinformatics 16, 386. https://doi.org/10.1186/s12859-015-0818-3.