Genomic Variation between PRSV Resistant Transgenic SunUp and Its Progenitor Cultivar Sunset Induced by Particle Bombardment Transformation

Jingping Fang
Fujian Normal University
Andrew Wood
University of Illinois at Urbana-Champaign
Youqiang Chen
Fujian Normal University
Jingjing Yue
Fujian Agriculture and Forestry University
Ray Ming (rming@life.uiuc.edu)
Fujian Agriculture and Forestry University https://orcid.org/0000-0002-9417-5789

Research article

Keywords: Carica papaya L., Whole-genome resequencing, Genomic variation, Nuclear plastid DNA (NUPT), Nuclear mitochondria DNA (NUMT)

Posted Date: November 12th, 2019

DOI: https://doi.org/10.21203/rs.2.17159/v1

Version of Record: A version of this preprint was published at BMC Genomics on June 12th, 2020. See the published version at https://doi.org/10.1186/s12864-020-06804-7.

Abstract

The safety of genetically transformed plants remains a subject of scrutiny. Genomic variants in PRSV resistant transgenic papaya will provide evidence to rationally address such concerns. In this study, a total of more than 74 million Illumina reads for progenitor ‘Sunset’ were mapped onto transgenic papaya ‘SunUp’ reference genome. 310,364 single nucleotide polymorphisms (SNPs), 34,071 Small Inserts/deletions (InDels) and 1,200 large structural variations (SVs) were detected between ‘Sunset’ and ‘SunUp’. Those variations have an uneven distribution across nine chromosomes in papaya. Only 0.27% of mutations were predicted to be high-impact mutations. ATP-related categories were highly enriched among these high-impact genes. The SNP mutation rate was about 8.4×10 -4 per site, comparable with the rate induced by spontaneous mutation over numerous generations. The transition-to-transversion ratio was 1.439 and the predominant mutations were C/G to T/A transitions. Spontaneous mutations were the leading cause of SNPs in transgenic papaya ‘SunUp’. A total of 3,430 nuclear plastid DNA (NUPT) and 2,764 nuclear mitochondrial DNA (NUMT) junction sites have been found in ‘SunUp’, which is proportionally higher than the predicted total NUPT and NUMT junction sites in ‘Sunset’ (3,346 and 2,745, respectively). Among all nuclear organelle DNA (norgDNA) junction sites, 96% of junction sites were shared by ‘SunUp’ and ‘Sunset’. The average identity between ‘SunUp’ specific norgDNA and corresponding organelle genomes was higher than that of norgDNA shared by ‘SunUp’ and ‘Sunset’. Six ‘SunUp’ organelle-like borders of transgenic insertions were nearly identical to corresponding sequences in organelle genomes (98.18~100%). None of the paired-end spans of mapped ‘Sunset’ reads were elongated by any ‘SunUp’ transformation plasmid derived inserts. Significant amounts of DNA were transferred from organelles to the nuclear genome during bombardment, including the six flanking sequences of the three transgenic insertions. Comparative whole-genome analyses between ‘SunUp’ and ‘Sunset’ provide a reliable estimate of genome-wide variations and evidence of organelle-to-nucleus transfer of DNA associated with biolistic transformation.

Background

Papaya (*Carica papaya* L.) is a diploid plant with a relatively small genome (2*n*=18, 372Mb) in the family *Caricaceae* [1]. It is one of the most popular tropical fruits owing to its exceptional nutritional and medicinal properties. However, *Papaya Ringspot Virus* (PRSV) has been recognized as the most destructive disease threatening worldwide papaya production. The papaya industry in Hawaii was severely damaged and its marketable papaya production drastically declined since the onset of PRSV’s spread in 1992 [2]. The development of PRSV-resistant transgenic papaya ‘SunUp’ and ‘Rainbow’ revived the industry.

‘SunUp’ papaya is a genetically modified (GM) version of its non-GM progenitor ‘Sunset’, and the hybrid cultivar ‘Rainbow’ derived from crosses between ‘SunUp’ and ‘Kapoho’ became the first transgenic virus-resistant fruit tree cultivar to be commercialized in the United States [3]. Over 25 generations of inbreeding led to an extremely low genetic heterozygosity level of 0.06% in the red-fleshed cultivar ‘Sunset’ [4]. PRSV-resistant cultivar ‘SunUp’ was developed based on the concept of pathogen-derived resistance (PDR) through biolistic transformation of a plasmid vector containing the PRSV HA 5-1 coat protein (*cp*) gene expression cassette [5, 6]. ‘SunUp’ was obtained by selecting transgenic progenies that were homozygous for the *cp* functional transgene, which confer PRSV resistance [7]. ‘SunUp’ has grown apart from ‘Sunset’ for more than 20 years, i.e. more than 20 rounds of meiosis.

Genomic variants comprise small changes in nucleotides including single nucleotide polymorphisms (SNPs) and small insertion/deletions (InDels), and large changes in chromosome structure (>50 bp), i.e. structural variants (SVs). SVs are considered to have a direct effect on behavior of the chromosome and cause variation in gene dosage [8]. Detection of genomic variants including unintended vector-derived fragments and other foreign fragments at the whole-genome level is characterized as an important criterion in the context of evaluation of GM organisms. The vector-derived inserts and transgene numbers in 'SunUp' were preliminarily determined by Southern analysis in previous research [7]. This study revealed that three plasmid vector elements inserted in the host nuclear genome during bombardment were stably inherited afterwards. One was a 9,789 bp functional insert, coding for intact functional transgenes PRSV *cp*, *nptII* and *uidA*; two were unintended and nonfunctional inserts, including a 290 bp partial *nptII* gene segment and a 1,533 bp plasmid-derived fragment consisting of a 222 bp truncated *tetA* gene, respectively. Nevertheless, at the genome-wide structural level, it remains unclear what unintended alterations were induced during bombardment and how many spontaneous mutations accumulated in more than two decades of independent cultivation. Conventional Southern blot, PCR and comparative genome hybridization (array-CGH) techniques are most commonly used to detect integration of vector sequences (>20 bp), whereas other small unintended incorporations of exogenous DNA fragments are below the detection limit of these techniques.

In many eukaryotes, the host nuclear genomes are prevalently faced with the modification of themselves by integrations of their symbiotic organellar genomes [9-13]. Such transfers occur from both plastid and mitochondrial genomes to the nucleus and are termed nuclear plastid sequences (NUPTs) and nuclear mitochondrial sequences (NUMTs), respectively. The organelle-derived fragments in the nucleus are collectively referred to as nuclear organelle DNA (norgDNA). The gene content and genome complexity of nuclear genomes differs among angiosperm taxa typically associated with these continuing intercompartmental DNA transfer events [12]. In contrast to those beneficial or nonfunctional long-existing nuclear organelle integrations, substantial numbers of newly formed norgDNA are more deleterious and are rapidly eliminated [14, 15]. The pattern and mechanism of organelle-to-nucleus DNA transfer has been analyzed in detail in a number of species [16, 17]. Continuous, mosaic structured, and inter/intra-chromosomal rearranged patterns are formed by NUPTs in the nuclear genome [18]. Non-homologous end joining of double-strand break repair (NHEJ-DSB repair) are suggested to be the integration mechanism as any other foreign sequences [18]. Recent evidence reveals that DNA methylation plays a pivotal role in regulating norgDNA, which may contribute to maintaining the genome stability and evolutionary dynamics of organellar and nuclear genomes [19]. NUPTs were shown to have integration preferences, simultaneous integration [20] and strong bias for nucleotide substitutions from C/G to T/A correlating with the time of integration [19]. It is intriguing that in Suzuki's study [7] all six genomic DNA segments flanking three inserts in 'SunUp' were nuclear organelle sequences. Five out of six were NUPTs, and one was NUMT. At present, no investigations have been conducted to determine whether bombardment affects the transfer frequency from cytoplasmic-to-nuclear genome or whether it was a consequence of insertion preference.

The last decade has witnessed revolutionary breakthroughs in next-generation sequencing (NGS) techniques, which enables fast and accurate re-sequencing of complete genomes at rather low costs. Whole-genome resequencing is a promising method for delivering information not only regarding inserts and their flanking sequences, but also about additional genome-wide assessments between genomes of transgenic lines versus

their progenitors. The integration of norgDNAs and subsequent nucleotide changes can be detected by conducting sequence similarity analysis between nuclear organelle sequences and the organelle genomes, likewise their changes in distribution according to the time of integration can be easily estimated. The available papaya nuclear and organelle genome provide a unique opportunity to study the genome-wide SVs and organelle-to-nucleus DNA shifts between GM papaya and its non-GM progenitor.

In the work presented here, we describe genome-wide comparative analysis of transgenic papaya 'SunUp' versus its progenitor 'Sunset', focusing on analysis of genomic variations such as small SNPs/InDels and large SVs, and the turnover and shuffling of nuclear organelle-derived sequences between the two varieties. These results will enable us to visualize the dynamic changes in 'SunUp' genome architecture after the integration of foreign sequences, provide evidence on where these norgDNA-like flanking sequences came from, and unravel the global impact of particle bombardment-mediated transformation on whole genome structure and organelle-to-nucleus DNA transfer.

Results

Whole-genome resequencing of 'Sunset'

The 'Sunset' genome was sequenced and assembled using a reference guided assembly approach using Illumina sequencing technology. The sequencing quality of these raw reads was generally high (90% with Phred quality score >27). After filtering, a total of 74 million high quality, 124 bp paired-end (PE) reads were generated. The total read length was 9.197 Gb, representing around 24.72× genome equivalents (**Table 1**). The sequencing depths were evenly dispersed along the papaya chromosomes. We first mapped the PE reads back to the 'SunUp' reference genome by BWA's short read aligner [21]. After removing multiple mapping reads and PCR duplicates, 48 million clean reads were retained for the following study. Of these 'Sunset' reads, as high as 99.97% matched unique 'SunUp' genomic locations, showing substantial consistency over most genome regions between 'SunUp' and 'Sunset'. The remaining 15,822 reads (0.03%) were unmapped, and likely correspond to the organelle genomes, 'Sunset'-specific region or highly repetitive regions that were unassembled in the reference 'SunUp' genome. Approximately 46 million (95.78%) clean reads mapped to reference genome in a properly paired orientation.

Detection and characterization of SNPs and small InDels in 'Sunset'

Polymorphisms between 'Sunset' and 'SunUp' were identified using SAMtools software suite [22] with strict parameters. Polymorphisms with coverage <10 or >100 and quality <50 were discarded to eliminate false positives in low coverage and highly repetitive regions respectively. Polymorphism sites with only one ALT were retained given the diploid nature of papaya. In total, 310,364 SNPs and 34,071 small InDels were found between 'Sunset' and the 'SunUp' reference genome (**Table 2**), with an average mutation rate of 0.084% for SNPs vs. 0.009% for InDels. The number of heterozygous SNPs was nearly 7 times higher than that of homozygous SNPs (269,493 vs. 40,871). A more even distribution was observed in the numbers of

homozygous and heterozygous InDels, with 19,135 and 14,936, respectively. The genome wide average for polymorphisms across the 'Sunset' genome was 84 SNPs per 100 kb and 9 InDels per 100 kb (**Table 3 and Fig. S1**). SNPs were substantially more prevalent at the genome-wide level than InDels. SNPs had an uneven distribution across the 9 chromosomes of papaya ranging from 24 SNPs per 100 kb in chromosome 2 to 165 SNPs per 100 kb in chromosome 6. InDels were more evenly dispersed across the 'Sunset' genome ranging from an average of 7 InDels per 100 kb in chromosome 2/9 to 13 InDels per 100 kb in chromosome 6.

All types of base changes were obtained and subdivided into transitions (Ts) and transversions (Tv) (**Table 4**). The total amount of Ts and Tv detected in all SNPs was 205,333 and 105,031 respectively, with a Ts/Tv ratio of 1.95. The average ratios of Ts to Tv for homozygous and heterozygous SNPs were 1.03 and 2.18, respectively. The amount of all four types of Ts were observed to have between 3.4- to 5.8-fold more than that of any types of Tv. The SNPs consisted of 104,312 G/C to A/T transitions (33.6%), 101,021 A/T to G/C transitions (32.6%), followed by 29,222 G/C to T/A transversions (9.4%), 28,910 A/T to C/G transversions (9.3%), 28,835 A/T to T/A (9.3%) and 18,064 G/C to C/G transversions (5.8%). Changes from G/C to A/T (Ts) were observed with the highest frequency whereas G/C to C/G (Tv) were the least frequent changes.

The length of small InDels ranged in size from 1 to 6 bp throughout the entire genome (**Fig. 1**), of which 1 bp-sized InDels were the most abundant, followed by 2 bp-sized InDels. In general, the amount of InDels decreased sharply as their size increased, especially for the shortest ones (1- to 2-bp) which showed the most dramatic drop in number. An exception was that the number of 3 bp-sized and 5 bp-sized InDels were slightly less than that of 4 bp-sized and 6 bp-sized InDels respectively.

Classification of SNPs and small InDels by potential impact on protein function

We predicted the variant effects of SNPs and small InDels according to their potential impact on protein function using SNPEff program [23] and self-built papaya data sets (**Fig. 2 and Table 5**). All variants that may have an effect on protein function could be categorized into 35 effect types, which were further grouped into the following four larger predefined impact categories on the basis of the assumed severity: HIGH, MODERATE, LOW, and MODIFIER (**Table 5**). The vast majority of variants (571,039, 97.4%) belonged to the MODIFIER category, which is usually comprised of intronic and intergenic variants and assumed to have only a weak or no impact on the protein. The LOW category is thought to be mostly harmless or unlikely to change protein behavior, such as synonymous mutations. A non-disruptive variant that might change protein effectiveness is defined as MODERATE, including in-frame deletions and missense mutations. In all 7,533 (1.28%) and 6,114 (1.04%) variants had possible MODERATE and LOW impacts on gene function. Only 1,591 variants with HIGH impacts were found, representing 0.27% of the total variants, which are assumed to have disruptive impacts on the protein, probably causing protein truncations, loss of function or triggering nonsense mediated decay. The most common types of mutations were frameshift variants in the HIGH category.

In terms of genomic distribution, approximately 48.5% of SNPs were identified in intergenic regions, while merely 8.4% were present in genic regions. Upstream promoter regions and downstream regulatory regions contained high proportions of SNPs, accounting for about 21% (**Fig. 2A**). Within the genic region, 5.9% and

2.5% of SNPs were present in the introns and coding sequence (CDS) regions, respectively (**Fig. 2B**). Overall, a similar distribution pattern of SNPs and InDels was observed across the entire genome. Likewise, a substantial number of InDels (~39%) were identified in intergenic regions (**Fig. 2A**), whereas only 9.9%were located in genic regions, consisting of 8.1% of intronic InDels and 1.8% of exonic InDels (**Fig. 2C**). The presence of InDels in the upstream and downstream regulatory regions of genes was also shown with a relatively high percentage (~25%) (**Fig. 2A**). In order to investigate the effect of SNPs on the amino acid alteration of a protein, the likelihood of non-synonymous and synonymous coding SNPs was estimated. Among all SNPs, 7,589 non-synonymous and 5,272 synonymous type modifications were detected in 'Sunset' (**Fig. 2B**). The ratio of non-synonymous to synonymous SNPs (NS/Syn ratio) was about 1.439. The predominant InDels within the coding regions were frameshift mutations (1,137, 95.7%), i.e. an indel size of which is not multiple of 3 (the length of a codon), whereas a significantly lower amount of codon insertions (31, 2.6%) and deletions (20, 1.7%) was observed (**Fig. 2C**).

With respect to gene function, all high-impact SNPs were predicted to affect 1,454 genes. For the global functional analysis of HIGH category genes, Gene Ontology (GO) terms were assigned to corresponding genes using BLAST2GO software [24]. Of 1,454 high-impact genes, 751 genes were associated with at least one GO term. We further applied GO category enrichment analysis to elucidate the functional enrichment of potentially high-impact genes, using Fisher's exact test with an FDR cutoff ≤ 0.05. A total of 31 GO terms were significantly enriched in biological processes and molecular functions (**See Table S1 and Fig. S3**). The biological process GO term "ATP catabolic process" was the most significantly and specifically overrepresented term, followed by "ribonucleotide catabolic process", and "purine nucleotide catabolic process". Enrichment in the biological process category was also reflected by high numbers of molecular function GO terms "nucleoside-triphosphatase activity", "hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides" and "ATPase activity", etc.

Detection of large structural variants

The BLAST result indicated that no additional plasmid derived inserts were found in the available 'SunUp' genome with the exception of three previously detected plasmid-derived inserts. In addition to SNPs and small InDels, the prevalence of some other types of larger structural variations (>50 bp) such as larger insertions (INS) and deletions (DEL), inversions (INV), intra-chromosomal translocations (ITX) and inter-chromosomal translocations (CTX) were also assessed using BreakDancer under stringent criteria. A total of 1,200 structural variants were identified in 'Sunset' (**Table 6**). Besides those ITXs detected by BreakDancer, which only consider intra-chromosomal translocation events that occurred between two non-adjacent portions of one scaffold, we detected more ITXs manually when an intra-chromosomal translocation happened between a non-adjacent scaffold pair from the same chromosome. In total, out of all 1,200 structural variants, 309 were identified as certain types including 128 DELs and 76 INSs, followed by 58 CTXs, 44 ITXs and 3 INVs, the remaining 891 were considered as uncertain CTXs due to unanchored scaffolds (**Table 6**). 220 out of 309 certain structural variations could be located on nine papaya chromosomes. ITX were in the range of 19 bp to 787.76 Kbp spread over Chr2 (7), Chr3 (7), Chr9 (4), Chr6 (3), Chr1 (2), Chr8 (2), Chr4 (1) and Chr7 (1). INV ranged from 37.76 Kbp to 270.08 Kbp. Chr1, Chr6 and Chr9 were found to each have one INV. CTX took place between 58

pair of chromosomes (**Table 6**). The size of the CTX was not taken into account as it is not relevant for inter-chromosomal translocations in BreakDancer output (see README of BreakDancer software). With respect to larger InDels, INS were in the range of 136 bp to 167 bp predominantly on Chr2 (13) and Chr6 (11), followed by Chr3 (6), Chr1 (5), Chr7 (5), Chr5 (4), Chr4 (2), Chr8 (1), and Chr9 (1). Compared with INS, DEL occurred more frequently with a wider range of sizes: 113 bp to 788.397 Kbp. They were also mainly distributed on chr2 (14) and chr6 (14), followed by chr5 (13), chr3 (10), chr1 (9), chr4 (8), chr9 (8), chr8 (5) and chr7 (3). The distribution of 1,200 SVs in different components of the genome were determined. As a result, the majority of variants (984, 82%) occurred within intergenic regions, which were roughly four times more likely to be present than SVs (216, 18%) found in genic regions. The SVs in genic regions may potentially affect 168 genes, among which 98 SVs were located in CDS regions of 87 genes.

We further validated these SVs by manual inspection of read alignments and found that all of SVs were unreliably predicted or false positives. Although each detected SV was supported by several reads, these regions were also covered by paired-end reads that supported the papaya reference genome arrangement. False positives were found to be located in the gap regions or regions with high levels of coverage (>100).

Shared and specific nuclear organelle integration sites between ‘SunUp’ and ‘Sunset’

With the aim of conducting genome-wide comparative analysis of the integration of nuclear organelle fragments between ‘SunUp’ and ‘Sunset’, two in-house software pipelines written in a mixture of python scripts (available upon request) were developed for automatic processing and identification of shared and variety-specific norgDNA integration sites between these two varieties. Schematic diagrams of pipelines are shown in **Fig. 3** and **Fig. 4**.

A total of 3,430 NUPT and 2,764 NUMT junction sites were obtained by searching against organelle genomes with the ‘SunUp’ reference genome as the query (**Table 7**). Out of all 3,430 NUPT junction sites, a large fraction of junction sites (3,327, 97%) were shared by ‘SunUp’ and ‘Sunset’. With BLASTN we identified that shared NUPTs matched the papaya chloroplast (pt) genome with an average identity of 91.92%. The remaining 3% (103) were specific in ‘SunUp’, with a higher average identity of 94.03% to the pt genome (further details of the 103 junction sites are provided in **Table S2**). Similar to the trend observed for the distribution of NUPTs, out of 2,764 NUMT junction sites, junction sites shared between ‘SunUp’ and ‘Sunset’ numbered 2,642 and account for the major share 95.6% whereas ‘SunUp’-specific junction sites only accounted for 4.4% (122) (further details of the122 junction sites are provided in **Table S3**). The average similarity in identity between ‘SunUp’-specific NUMTs and papaya mitochondria (mt) genome was 93.77%, which is slightly less than the identity between ‘SunUp’-specific NUPTs and the pt genome (94.03%) but a bit higher than the identity between shared NUMTs and the mt genome (92.97%). In general, higher similarities in identities were apparent between ‘SunUp’-specific norgDNAs and corresponding organelle genomes than between shared norgDNAs and corresponding organelle genomes. We next evaluated the performance of our pipeline through manual inspection of read alignments surrounding those identified as ‘SunUp’-specific norgDNA junction sites in the Integrative Genomics Viewer (IGV) software [25]. The visual display exhibited that no ‘Sunset’ reads aligned to or spanned any ‘SunUp’-specific junction site in the ‘SunUp’ reference genome as we had expected, thus those

'SunUp'-specific integration events predicted by our pipeline were bona fide. In the 'SunUp'-specific norgDNA regions, no reads mapped or having a read depth greater than 100× were observed, suggesting that those reads likely correspond to the organellar DNA. The results demonstrate the superior sensitivity and accuracy of our pipeline.

Overall, 'SunUp'-specific norgDNA integration junction sites were distributed non-randomly across nine chromosomes of papaya, with distinct regions of high and low variation (**Table 8**). The most distinct region was in Chr2 which had the highest frequency of NUPT junction sites with 11.65% compared to other chromosomes of the genome, followed by Chr6 and Chr8, with 8.74% each. Only a low proportion of NUPT junction sites were found in Chr3 (1.94%) and Chr2 (2.91%). Compared with NUPT junction sites, a smaller range of variation across chromosomes was found at NUMT junction sites. Similarly, NUMT junction sites were highly enriched in Chr6 (10.66%), Chr2 (9.84%) and Chr8 (9.02%), while less prevalent in Chr5 (4.92%) and Chr1 (5.74%).

Using a strict pipeline (**Fig. 4**), the 'Sunset' genome was also scanned for norgDNA integrations by searching the papaya chloroplast and mitochondria genomes. The total amount of either NUPT or NUMT integration junction sites in the 'Sunset' genome were slightly fewer than in the 'SunUp' genome, with 3,430 NUPT and 2,764 NUMT junction sites, respectively (**Table 7**). In contrast to 'SunUp'-specific NUPT integrations (103), the amount of 'Sunset'-specific NUPT integration junction sites sharply reduced to only 19, with an average sequence identity of 95.64% matching to the papaya pt genome; 'Sunset'-specific NUMT integration junction sites decreased to 103, having an average identity of 96.95% to the mt genome.

The origin of organelle-like borders of transgenic inserts in 'SunUp'

BLASTN search analysis of transgenic inserts' flanking sequences was conducted to investigate the possible identity of sequences around the insertion sites. All six genomic DNA segments flanking the three previously identified transgenic insertions were surprisingly found to share near sequence identity to the papaya organelle sequences (**Fig. 5A**). Both sides of the single, contiguous 9,789 bp functional transgene insertion encoding intact PRSV *cp*, *uidA* and *nptII* genes were identified to be NUPT sequences, consisting of a 4,000 bp and a 1, 790 bp plasmid-derived segments, which were highly homologous with *trn*, *rps* genes of the plasmid genome and part of the *ycf3* gene. The genomic DNA flanking both borders of the nonfunctional *nptII* fragment insert (290 bp) also exhibited homology with papaya plastid genome genes *ndhG* and *atpB, E*, with size of 363 bp and 827 bp, respectively. The contiguous 1,533 bp nonfunctional *tetA* fragment insert, in particular, had one border of NUPT sequence homologous to the plastid gene *ycf2*, reaching up to 6,199 bp. The other border of the *tetA* fragment was comprised of non-plastid DNA-like sequence and showed identity to a papaya mitochondria genome segment, totaling 1,708 bp. Sequences of three flanking pairs of transgenic inserts showed significant homology to papaya organelle genome segments, with a range of 98.18~100% identity. Especially two flanking pairs of the functional transgene insert and the nonfunctional *nptII* fragment insert, having identities approaching 100%. By contrast, organelle-like sequences at both borders of the *tetA* fragment insert experienced further rearrangements and showed lower similarities of 98.6% and 98.18% with

the pt genome and the mt genome, respectively. We estimated the homology between our previously assembled ‘Sunset’ norgDNAs and six flanking organelle-like sequences of inserts in ‘SunUp’. Through a rigid BLAST screening, there were respectively 12, 6, 1, 5, 43 and 2 best BLAST hits detected between ‘Sunset’ norgDNAs and six flanking norgDNAs, with combined lengths ranging from 49 to 4,180 bp (**Table 9**). Only one (best) hit was found between ‘Sunset’ norgDNAs and border A of the nonfunctional *nptII* fragment insert, with a size of 49 bp; whereas there was at least 1,231 bp of combined length for the remaining five borders. The sequence identity between ‘Sunset’ norgDNAs and six flanking norgDNAs of insertions in ‘SunUp’ varied from 93.68% to 99.09%. Of which all NUPT borders matched ‘Sunset’ norgDNAs with relatively lower identities in comparison to their matching with the papaya chloroplast. Meanwhile, the sole NUMT border had 99.09% similarity to ‘Sunset’ norgDNAs, which is higher than its similarity to the mt genome (98.18%).

We developed a strategy based on massive paired-end mapping (**Fig. 5B**) to further investigate whether these organelle-like sequences at both borders of three insertions were present in the genome prior to bombardment or not. If a deletion in ‘Sunset’ relative to the ‘SunUp’ insertion-with-the border region was found (**Fig. 5B**), we were able to deduce that those organelle-derived fragments flanking transgenic insertions were originally present in ‘Sunset’ prior to bombardment. Deletions were identified using paired ends spanning the specified genomic region in ‘SunUp’ that were longer than the transgenic insert size (*cutoff*).

As a result, a total of 217,890 ‘Sunset’ short reads could be aligned to the region of the functional transgene insertion with organelle-like borders, of which 183,488 reads were mapped to the reference region in properly paired orientations. According to statistical calculations, the inner distances between PE reads were far less than the *cp* functional transgenic insert size (9,789 bp), which ranged in size from 0 to 246 bp. Of these, the inner distances of 0 bp were significantly enriched, with 1,320 pairs. Meanwhile, there were 42,273 ‘Sunset’ short reads aligned to the region of the nonfunctional *nptII* transgene insert with organelle-like borders, among which 22,518 were mapped as a pair, with pairwise distances ranging from 0 to 169 bp in length. All pairwise distances were smaller than the size of the nonfunctional *nptII* fragment insert at 290 bp. A major fraction of PE reads (223 pairs) were found to have no distance between each other. Regarding the nonfunctional *tetA* transgene insert with both flanks, the total number of mapped reads were 418,697, including 150,110 optimally mapped PE reads. Of the latter, the sizes of inner distances were in the range of 0 to 969 bp, which is under the *cutoff* 1,533 bp. There was a significant enrichment for the 0 bp inner distance as well, containing 2,661 pairs. The distribution of mapped paired-end spans in regions of three inserts with flanks is shown in histograms (**Fig. 5C**). Except for the 0 bp distance, three histograms of paired-end inner spans were normally distributed and showed primary peaks at 59 bp (1,296 pairs), 57 bp (187 pairs) and 56 bp (1,993 pairs). In summary, in all cases of three transgenic insertion events, the inner distance of any pair of mapped PE reads was shorter than a transgenic insert size (*cutoff*), indicating that the distance between any pair of ‘Sunset’-derived PE reads was not elongated by an insertion. This serves as a strong hint that these flanking norgDNAs were not present in the genome prior to bombardment.

Discussion

Conformation of the presence or absence of unintended alterations in addition to target gene integration is a key issue in the evaluation of GM plants. Plant tissue culture processes required during post-transformation can introduce somaclonal variations, which could cause unintended genetic and epigenetic changes leading to heritable phenotypic alterations, as bombardment-mediated transformations can [26].

Revolutionary breakthroughs in NGS in conjunction with developments in bioinformatic software that are tailored to solve biological problems have assisted in the molecular characterization of GM crops and detecting their genome-wide genomic variants induced by somaclonal variations and transformations. For instance, a deep sequencing coverage of 75× in transgenic soybean has uncovered the insertion site of T-DNA [27]. In the case of transgenic rice OSCR11 expressing a seed-based edible vaccine against Japanese cedar pollinosis, 11.3-33.2× whole-genome sequencing was adopted to reveal that the genomic discrepancy between OSCR11 and its host a123 was small, and that nucleotide substitution profiles were analogous to somaclonal variation [28]. Whole-genome sequencing (7×) and CGH arrays were performed to evaluate the molecular composition of herbicide-tolerant mutant rice generated by *Agrobacterium*-mediated gene targeting (GT). In inspected GT rice plants, more than 1,000 SNPs and InDels were identified and over 300 somaclonal mutations were predicted to be induced between generations, although no integration of *Agrobacterium*-derived DNA fragments had been detected [29]. The availability of the ‘SunUp’ draft genome, the rapid evolution of deep-sequencing technology together with increasingly robust bioinformatics tools make it possible to decipher genome-wide structural perturbations at the single-base resolution level in the transgenic papaya genome after subjection to bombardment, tissue culture, and other spontaneous mutations during 20-year's separation.

Here, we carried out paired-end sequencing of DNA-Seq libraries prepared from genomic DNA isolated from young healthy leaves of non-transgenic host ‘Sunset’. High throughput sequencing generated more than 74 million filtered reads, which translates to an average depth of coverage of 24.72×. The sequencing depths were found to be evenly distributed amongst the nine papaya chromosomes, indicating a high randomicity performance of Illumina sequencing. After removing multiple mapping reads and PCR duplicates, nearly 100% reads could be uniquely mapped on the ‘SunUp’ reference genome, suggesting a well-assembled reference genome and high levels of similarity between ‘SunUp’ and ‘Sunset’ genome.

The Leading Factor Responsible for Polymorphisms in ‘SunUp’

In total there were 310,364 SNPs, 34,071 small InDels and 1,200 large SVs detected between ‘Sunset’ and the ‘SunUp’ reference genome. Detailed estimates of genetic relationships among papaya accessions and related species were revealed by AFLP makers, suggesting the smallest genetic variation among papaya cultivars derived from the same or similar gene pools [30]. A similar trend has also been observed between distantly related papaya varieties of ‘SunUp’ and nontransgenic ‘AU9’ [31]. Comparative genomic analysis between two homologous BACs from ‘AU9’ and ‘SunUp’ revealed 99% gapless sequence identity, further confirming the limited diversity among papaya varieties by virtue of self-pollination in hermaphrodite papaya and its coexistence and cross-breeding with dioecious varieties. In this study, transgenic papaya ‘SunUp’ was

transformed from its nontransgenic progenitor ‘Sunset’, therefore they share genetic similarity with each other, with an average SNP mutation rate of 0.084% in ‘SunUp’, i.e. around 8.4×10^{-4} bases per papaya genome, almost matching the genetic heterozygosity at 0.06% in the ‘SunUp’ genome [32]. This SNP mutation rate is about an order of magnitude greater than the 0.0077% SNP polymorphism rate between the X chromosome and its homologous X^h counterpart, but conversely one order lower than the 0.261% SNP rate between recently diverged (<7 MYA) Y and Y^h chromosomes from the same papaya varieties [33].

Compared with other species, this observed SNP rate between ‘SunUp’ and ‘Sunset’ is far lower than those reported from other wild-type plant species [34, 35]. SNP frequency varies from 0.53 to 0.78% between two cultivated rice subspecies *japonica* and *indica* [35]. The whole-genome resequencing of soybean MYMIV (*Mungbean yellow mosaic India virus*) resistant cultivar ‘UPSM-534’ and susceptible Indian cultivar ‘JS-335’ was performed to identify SNPs by their individual comparison with the reference genome *Glycine max* var. Williams 82 and an overall SNP rate of 0.17% was found [34]. In contrast, this SNP rate between ‘SunUp’ and ‘Sunset’ is much higher than the transformation-specific mutation rate and somaclonal mutation rate observed in other species [28, 36, 37]. The pattern of nucleotide base substitution in transgenic rice OSCR11 relative to its nontransgenic host was consistent with somaclonal variation, with a transformation-induced SNP rate of 0.68×10^{-7} per cell culture week [28]. This was highly comparable with the rate induced by somaclonal variation in *Arabidopsis* (0.86×10^{-7}) [36] and rice (0.85×10^{-7}) [37] per cell culture week.

A previous report indicated that gene mutation rate of transgenic plants was two orders of magnitude less than that observed between soybean cultivars [38], genetic variants which occurred spontaneously. Given that ‘SunUp’ and ‘Sunset’ have separated for more than 20 years, a theoretical mutagenesis rate in ‘SunUp’ compared with ‘Sunset’ was calculated by dividing the detected mutation rate (8.4×10^{-4}) by 20, resulting in a mutation rate of 4.2×10^{-5} per generation, within the range of spontaneous mutation rates (10^{-11}~10^{-4}). Ossowski et al. [39] reported a spontaneous mutation rate of 6.0×10^{-9} mutations/effective site calculated for *Arabidopsis*. The detected spontaneous mutation rate in papaya off-type SSR markers was rather high at 3% frequency after one meiosis [40]. Considering the transformation-induced SNPs in ‘SunUp’ cannot readily be distinguished from somaclonal and spontaneous variants, and the calculated rate corresponds well with the rate induced by spontaneous variation, we speculate that ongoing spontaneous mutations induced through propagation and regeneration during 20 years of separation is a primary mutation type in particle bombardment-mediated transformed ‘SunUp’. The genetic variants accumulated through ongoing spontaneous mutation over numerous generations were not found to pose any new risk to consumers, as they likely already evolved through natural selection [38].

Both SNPs and small InDels were randomly produced amongst papaya chromosomes (**Table 3**), indicating that biolistic based transformation could have a genome-wide effect on the papaya genome, not just specifically affecting the flanking sequences of insertion sites. Interestingly, an uneven distribution was observed for those mutations with the highest density in chromosome 6 and the lowest in chromosome 2. Although sequences of three ‘SunUp’ transformation plasmid vector derived inserts with genomic borders had been well characterized [7], their exact genomic positions in ‘SunUp’ remain enigmatic owing to technical limitation [41]. The increased levels of nucleotide variation in chromosome 6 imply this chromosome might experience strong disturbances in the genomic stability accompanied by transgene integration. As reported by

Doerfler et al. [42] exogenous DNA insertion can have genome-wide perturbations that are not limited to the insertion site, and possibly transmitted to neighboring DNA sequences, to chromatin structures and even to adjacent chromosomes that are in contact with the insertion site of the chromosome targeted by foreign DNA insertion. Multiple insertions separated by genomic DNA in one single chromosome were reported to be a common occurrence in biolistic based transformation [43, 44]. We surmise that three transgenic inserts were likely inserted in one chromosome. This conjecture remains to be further studied. Additionally, sources of bias and error, such as technical variability during library preparation and sequencing, sequencing bias and the inevitable error rates during short read alignment in highly repetitive regions especially repeat-rich gene-poor heterochromatin, may account for the relative high frequency of mutations in some regions.

In terms of base substitution type, bias towards G/C to A/T transitions was observed in this study (**Table 4**). This result support previous reports on the pattern of nucleotide substitutions, regardless of whether SNPs were caused by spontaneous and somaclonal mutations [28, 39], chemical and physical mutagens [45], or by *Agrobacterium*-mediated gene targeting and transformation [29, 46]. Overabundance of G/C to A/T fit the earlier theory that G:C sites in CpG contexts are more likely to be methylated [47], and spontaneous deamination of methylated cytosine would lead to thymine substitution [48, 49].

However, non-methylated G:C sites also had a higher rate of transition than A:T sites in *A. thaliana*, suggesting that other factors in addition to methylation are responsible for the high rate of transitions at G:C sites[39]. Other studies have shown that G/C to A/T transitions frequently happen at dipyrimidine sites where the C is adjacent to another C or to a T under ultraviolet (UV) radiation, which exists in natural conditions [50]. Another supporting theory was proposed that alkylated guanines are easily paired with thymines, leading to misplaced adenines at sites of guanines [51].

The combined effect of the deamination of methylated cytosines, alkylated guanines and UV-induced mutagenesis could explain the increased rate of transitions at G:C sites in our study. The determined Ts/Tv ratio was 1.95, which is comparable to ratios of 1.9781 and 1.9609 found in soybean MYMIV susceptible and resistant cultivars [34], lower than the 2.4~2.7 ratio reported for spontaneous mutations in *Arabidopsis* mutation accumulation lines [39], but obviously higher than the Ts/Tv ratios in transformation-induced SNPs or somaclonal variation induced SNPs of nearly 1.0 [36, 37]. Transitions are interchanges of two-ring purines (A/G) or of one-ring pyrimidines (C/T), and can be generated at higher frequency than transversions under natural conditions. Transversions, on the contrary, are reported to become more prevalent when there is considerably more genetic instability [52]. Oxidized guanines (8-hydroxy-G) are prone to pairing with adenines instead of cytosines and leading to misplaced thymines in the positions where guanines should be. Therefore, G/C to T/A transversions arise when DNA is oxidized resulting in a modified base lesion [53]. We conclude that the increased Ts/Tv ratio in our study, relative to transformation-induced and somaclonal variation, can be explained by the coupled effects of genetic transformation, somaclonal and spontaneous variants, but largely caused by spontaneous variants.

We observed an excess of 1 to 2 bp-sized InDels and a significant deficit of 5 bp-sized InDels (**Fig. 1**). Small InDels preferentially occurred in repetitive regions such as microsatellites and homopolymers [36], the mutational nature of which was mainly attributed to the DNA replication slipped-strand mispairing (SSM) mechanism [54, 55]. In agreement with previous studies, nearly all InDels directly engendered by

transformation and somaclonal mutation were 1 or 2 bp in size except for one 5-bp transformation-induced deletion [28, 36]. All of the 1- or 2-bp InDels occurred in a mono- or di-nucleotide context as a result of slippage during DNA replication. Since the exceptional 5-bp deletion was in a non-polymeric context, it may be attributable to the improper repair of a DNA double strand break (DSB) caused by transformation or it could have happened spontaneously.

Based on SNPEFF results, most SNPs and InDels were identified in intergenic regions. As compared with genic regions, SNPs and InDels were much denser in the upstream and downstream regulatory regions of genes (**Fig. 2**). The abundance of variations in the upstream and downstream regulatory regions of genes is expected on account of low sequence conservation and reduced purifying selection pressure in non-coding regulatory regions relative to coding regions [56]. The ratio of non-synonymous to synonymous SNPs was 1.439, and a collection of 1,454 high-impact genes affected by SNPs were predicted (**Table 5**). Functional annotation of these genes revealed putative roles of respective proteins in ATP catabolic process and ribonucleotide catabolic process. Information on those high-impact mutations would be useful for the development of DNA markers associated with disease-resistance related genes, which could accelerate genomics-assisted disease resistance breeding in papaya.

Following manual inspection of read alignments by IGV software, all SVs were identified as false positives largely owing to the incompleteness of the papaya genome. A more completely assembled and gapless papaya reference genome is needed in the future for dissecting large structural variations. Previous reports analyzing somaclonal variations in *Arabidopsis* and rice showed that no SVs were detected [36, 37], we surmise that large SVs were likely caused by integration position effects of particle bombardment transformation.

NorgDNAs Flanking the Inserts as a Result of the Transformation

Organelle-to-nucleus DNA transfers are continually ongoing in plant genomes [18, 57]. We developed two pipelines for the automatic identification of norgDNA junction sites in ‘SunUp’ and ‘Sunset’ in this study, and results showed that altogether 3,327 NUPT and 2,642 NUMT junction sites were shared by ‘SunUp’ and ‘Sunset’, covering at least 95% of total norgDNA junction sites. Our data provide direct evidence that norgDNAs are widely spread throughout the papaya genome and are highly conserved between the transgenic papaya ‘SunUp’ and its nontransgenic precedent cultivar ‘Sunset’. It can also be inferred from the high conservation that the vast majority of norgDNAs were older transfers predating the transgenic event and sparse organelle-to-nucleus integrations were triggered by transgenes. Those ancient norgDNAs might play a critical role in papaya genome evolution. This result agrees with earlier findings, which shows that newly formed norgDNAs tend to be fragmented, shuffled and rapidly eliminated [14, 58]. The transfer amount and rates of pt and mtDNA in the nucleus differs among species. The accumulation of NorgDNAs is driven by selective pressure or recombination suppression and NorgDNAs would accumulate to a varying extent even in different regions of the same genome. As previously reported [13], papaya HSY and MSY in the absence of recombination accumulated 4 times the amount of NUPTs than the papaya genome-wide average and nearly 12 times the

average in the corresponding region of the X chromosome. By contrast, NUMTs are less prevalent in the X and HSY chromosomes compared to the whole genome. Furthermore, shared norgDNA is sparse between X and HSY, with only 11% of pt and 12% of mt fragments conserved, respectively, indicating that the accelerated accumulation of norgDNAs occurred after the recombination suppression was seen in the HSY.

Those 'Sunset' or 'SunUp' regions where specific norgDNAs are detected could be newly formed via shuffling and the rearrangement of extant genomic norgDNA fragments when bombardment-induced exogenous DNA was integrated into the genome causing instability, or new transfer from organelle genomes which was accompanied by bombardment-mediated transformation. Older inserts from organelles are predicted to exhibit lower pt/mt DNA identities due to fragmentation and mutation that occurs over time [18, 59]. As well characterized in *Oryza* and *Arabidopsis* [60], clusters of NUPTs and NUMTs contained in the angiosperm nuclear genomes can be very fragmented and rearranged with respect to the extant organelle genomes. Hence, the evolutionary change of individual norgDNA fragments since integration into the nuclear genome can be estimated by comparison with organelle genomes in this current analysis. The variable matches to papaya organelle genomes indicate that the fragments have a range of insertion times, with some predating the bombardment and others taking place within the last 20 years. We also found that the average identity between 'SunUp'-specific norgDNAs and the extant organelle genomes was higher than that of conserved norgDNAs between 'SunUp' and 'Sunset' (**Table 7**). It can be inferred that biolistic based gene transformation could accelerate the DNA transfer frequency and amount from organelles into the papaya nuclear genome, and that new organelle-to-nucleus DNA integration probably occurred during bombardment.

Three transgenic inserts in 'SunUp' are surprisingly flanked by norgDNA segments, with five NUPTs and one NUMT. The higher ratio of NUPT:NUMT (5:1) is expected because it is proportionally close to the ratio of pt:mt genome (5.5:1) in the cell. The average read depths from the whole genome shotgun reads for the pt and mt genomes are 1044 and 189 respectively (data not shown). This predicts a pt:mt genome ratio of 5.5. NUPTs were observed to be more abundant than NUMTs on the genome-wide scale as well, according to our findings (**Table 7**). The distribution of norgDNAs showed a similar trend to SNPs, in that they are overrepresented in Chr6 compared to other chromosomes. This finding further implies that Chr6 may experience strong perturbations in genome structure in the event of foreign DNA being inserted. To estimate whether those organelle-like border fragments were present in the genome prior to bombardment or not, we initially examined the identities between six 'SunUp' organelle-like borders and papaya organelle genomes. All six border sequences, especially plastid-like borders, were nearly identical to the corresponding sequences in the extant papaya organelle genomes (98.18~100%), this being significantly higher than the identity of conserved norgDNAs compared with organelle genomes (91.92~92.97%). Six organelle-like borders with high nucleotide identity relative to organelle genomes likely represent newer transfers of DNA. Homology searches between norgDNAs in 'Sunset' and six organelle-like borders in 'SunUp' in the follow-up step showed that all five NUPT borders had relatively lower similarities to 'Sunset' norgDNAs (93.68~99.09%) than to the papaya pt genome (nearly 100%) (**Table 9**), demonstrating that new transfers of DNA from chloroplast to the nuclear genome occurred including five plastid-like borders following bombardment-induced foreign gene insertion. We did not expect that the NUMT border would match 'Sunset' norgDNAs with a slightly higher similarity (99.09%) than to the mt genome (98.18%). Based on the massive paired-end mapping strategy, we did not find the inner distance of mapped 'Sunset' PE reads was elongated by a transgenic insert. This further confirmed that six

organelle-like border sequences were not present in the recipient genome antecedent to particle bombardment-mediated transformation, and it is likely that they were newly added to the papaya nuclear genome from organelles in the wake of gene transfer although the integration mechanism underlying bombardment-induced norgDNA remains to be elucidated. Two hypotheses were put forward in this study. One hypothesis is that the acquisition of many bases of inserted DNA increased the instability of the papaya genome and likely altered the chromatin topology, enabling organelle DNA fragments to be readily integrated into the nuclear genome. When encountered DNA lesion such as DNA double-strand breaks triggered by exogenous sequences, cells respond by activating a DSB repair mechanism [61]. Accumulating evidence suggests that most norgDNA integrates into the nuclear genome via a non-homologous recombination or non-homologous end joining of double-stranded breaks repair mechanism as any other exogenous sequences [18]. Another possibility is that foreign genes initially insert into the chloroplast genome are spontaneously shifted into the nucleus with sections of adjacent chloroplast DNA. Several different studies have shown that a plastid transgene *nptII* was successfully transferred into the nuclear genome from the plastid, which was found to happen at a surprisingly high frequency of approximately one in five million cells [59, 62]. The foreign DNA tends to integrate randomly into the host genome via biolistic based transformation, so it is not possible to determine where the transgene initially inserted and evidence in support of this assumption is largely lacking. We hope that further studies based on our results will lead to remarkable breakthroughs in the filed of plant genetic engineering.

Conclusions

The primary objective of this study was to thoroughly inspect genome-wide discrepancies between the PRSV resistant transgenic papaya ‘SunUp’ and its progenitor cultivar ‘Sunset’, including small SNPs/InDels, large SVs, and nuclear organelle DNA integrations. Detected variations were randomly distributed amongst papaya chromosomes, whereas only 0.27% were predicted to have a disruptive impact on the protein function. Development of SNP/InDel markers that occurred in high-impact genes could facilitate marker-assisted PRSV disease resistance breeding in papaya. Genome-wide analysis of organelle-to-nucleus integration events confirmed that norgDNAs are ubiquitous in papaya genome and highly conserved before and after genetic transformation. Those conserved norgDNAs might play a pivotal role in papaya nuclear genome. We reasoned that biolistic transformation could speed up the organelle-to-nucleus transfer frequency and amount, and six organelle-like borders of transgenic inserts likely newly transferred to the nucleus in the wake of bombardment-induced foreign gene insertion. The newly integrated norgDNA induced by particle bombardment revealed the mechanisms underlying the process of foreign gene transformation. The major cause of polymorphisms in ‘SunUp’ is likely to be spontaneous mutation. Therefore, any speculated risk due to the unintended consequences of biolistic transformation in ‘SunUp’ should only merit the same consideration given to variations arising spontaneously from traditional breeding practices, which attests to the safety of transformation technology. A completely assembled papaya genome in the near future will complement the present study.

Methods

Plant Material and Next-generation Sequencing

The non-transgenic progenitor papaya cultivar ‘Sunset’ was grown under natural conditions at Kunia substation in Oahu, Hawaii by Hawaii Agriculture Research Center. Young and healthy leaf tissues from plants with at least two visible leaves were collected for DNA extraction. High quality genomic DNA (>100 ng/ul, OD 260/280 close to 1.8) was extracted from leaves using a modified approach for reduced organelle contamination [63]. The amounts and quality of DNA was estimated by NanoDrop 2000 spectrophotometer (Nano-Drop Technologies, USA). The frozen samples including leaf tissues and genomic DNA were preserved in Ming’s laboratory in University of Illinois at Urbana-Champaign (UIUC) and can be acquired with the voucher number Sunset-A07.

Sequencing of papaya ‘Sunset’ genome was carried out at the W.M. Keck Center for Comparative and Functional Genomics, UIUC. A paired-end library with a 300bp insert size was constructed and sequenced in a single lane of a sequencing flow cell on an Illumina HiSeq2000 platform (Illumina Inc., San Diego, CA, USA). Over 74 million 124 bp paired-end reads were generated from one lane of sequencing. Prior to any downstream processing, the empty reads, poor quality reads and adaptor sequences in the raw sequenced data (>30% of the bases with a Phred quality score of <Q20) were filtered out using the program IlluQC.pl in NGSQCToolkit [64] to obtain clean reads (**Fig. 3A**). Graphs showing QC statistics were generated. After filtering, the NGSQCToolkit was used to check the data quality again.

Genome-wide Detection of SNPs, Small InDels and Large Structural Variants

Sequences of three ‘SunUp’ transformation plasmid derived inserts with genomic borders could not be assembled into the ‘SunUp’ genome (http://www.plantgdb.org/CpGDB/) which was ascribed to technical limitations, therefore three inserts and their flanking sequences were not taken into account in the genome-wide detection of SNPs, InDels and SVs. In order to further detect other plasmid vector derived inserts in ‘SunUp’ reference genome in addition to the three aforementioned well-known plasmid-derived inserts, the BLASTN [65] program was conducted to search with the entire transformation plasmid (19,567 bp) as a nucleotide query against the whole ‘SunUp’ reference genome as a database.

The resulting paired-end reads of ‘Sunset’ were aligned to the most updated ‘SunUp’ genome using BWA’s short read aligner with default settings [21]. BWA can be used for mapping low-divergent sequences against a large reference genome and provide Sequence Alignments/Map (SAM) format outputs. Only uniquely mapped reads were retained by choosing the “@SQ|@PG|@RG|XT:A:U” tag in raw output SAM format file to ensure that a read only had a single mapped location. The unimap SAM file was then converted to Binary Alignment/Map (BAM) format, sorted according to chromosomal coordinates, treated for potential PCR duplicates removal and indexed using the SAMtools software suite [22]. SNP variants were called by performing the SAMtools ‘mpileup’ command with -ugDV parameters followed by ‘bcftools’ from the SAMtools package. Polymorphism results were saved in a variant call format (VCF) file. The raw variant calls were filtered with the SAMtools vcfutils.pl varFilter script and a custom script vcf_filter.py for read depth ≥10 and ≤100 and polymorphism site quality ≥50. An SNP site at which two or more alternate alleles (ALT) were called was removed for diploid organisms. In an output VCF file for diploids, the genotype “0/0” represents homozygotes of the reference

allele, with “0/1” for heterozygotes, “1/1” for homozygotes of the alternate allele and “./.” for unknown genotypes.

Variant effect analysis of SNPs and InDels were predicted on the basis of information on gene structure and function in papaya using SNPEff (ver. 4.1) [23]. Since papaya genomic annotation database is not available in the pre-built databases of SNPEff, we built a database for papaya using the ‘SunUp’ reference genome in FASTA format and its gene annotation file in GFF format. The potential effect of each variant on gene expression and protein structure or function was examined by SNPEff. GO terms describing the biological processes, molecular functions and cellular components were assigned to the high-impact genes using the Blast2GO program [24]. Further, GO enrichment analysis for high-impact genes was performed in the agriGO program [66] using the gene models of papaya reference genome as a background. The Fisher statistical test was applied to test for enrichment of functional categories with Bonferroni’s correction (FDR $\leq$0.05). BreakDancer [67] was used to detect genomic SVs using ‘Sunset’ read pairs that are mapped to ‘SunUp’ reference genome with unexpected separation distances or orientations. BreakDancer predicts five types of structural variants: insertions (INS), deletions (DEL), inversions (INV), inter- and intra-chromosomal translocations (ITX and CTX). The SVs were filtered by scores equal to 99 and number of reads $\geq$10 thereby selecting a highly confident set of SVs.

‘SunUp’-specific Nuclear Organelle DNA Junction Sites

The BLASTN [65] algorithm was used to search the ‘SunUp’ genome for nuclear plastid (NUPT) and nuclear mitochondria (NUMT) integrations with papaya (*Carica papaya*) organelle genomes as databases (**Fig. 3B**). The organelle genomes are available at Genbank, with accession number EU431223 for the chloroplast genome and accession number EU431224 for the mitochondria genome. An *E*-value cut-off of 1e-20 with >80 % homology is included in the analyses.

A set of clean single-end reads of ‘Sunset’ were aligned to the papaya ‘SunUp’ reference genome using BWA v0.7.12 default settings (**Fig. 3C**). After the alignment, the mixture of reads that aligned back to the reference genome were predicted to originate from different sources of DNA in ‘Sunset’ genome, including nuclear DNA (nuDNA), nuclear organelle DNA (norgDNA) and organelle DNA (orgDNA) (**Fig. 3E**). We labeled the joint position that lies at the junction of norgDNA and nuDNA as a junction site. Only a junction site in the ‘SunUp’ reference genome that was both mapped and spanned by ‘Sunset’ reads can be termed a shared junction site, and it is considered to be shared by both ‘SunUp’ and ‘Sunset’ genomes (**Fig. 3D**). In order to discriminate between these three categories of reads and obtain the reliable junction sites shared by ‘SunUp’ and ‘Sunset’, the flanking regions (5 bp upstream and downstream) of the junction sites are used as an indicator. Reliable norgDNA reads were selected if those reads were not only spanning the junction sites but also mapped at least to 5 bp of norgDNA or nuDNA (**Fig. 3E**). Otherwise, if there were no reads mapped to or no reliable norgDNA reads spanning the junction site, we considered this junction site a ‘SunUp’-specific norgDNA junction site (**Fig. 3F**). An in-house software pipeline written in a mixture of python scripts (available upon request) was developed for automatically processing and identifying norgDNA junction sites in ‘SunUp’. This pipeline is well documented and widely applicable to other diploid plants. The Integrative Genomics Viewer

(IGV) software [25] was used to visualize and ensure the validity and reliability of those identified as 'SunUp'-specific norgDNA junction sites by this pipeline.

'Sunset'-specific Nuclear Organelle DNA Junction Sites

We aligned a set of clean single-end reads of 'Sunset' to chloroplast and mitochondria as reference genome independently using Bowtie2 version 2.2.5 [68] (**Fig. 4A**). The CIGAR (Compact Idiosyncratic Gapped Alignment Report) strings of reads which represent the sequence alignment in SAM/BAM file were used to identify and extract soft-clipped reads with at least 5 bp mismatches at the extremity (**Fig. 4B**). Those reads were *de novo* assembled by SOAPdenovo 63 mer-V2.04 [69] with an optimized *K*-mer length of 63 to generate potential norgContigs that were as long as possible (**Fig. 4C**). The norgContigs were screened for sequence similarity by BLAST against corresponding organelle genome at an *E*-value cut-off of 1e-20 (**Fig. 4D**). Only hits of norgContigs with ≥30 bp mapped to organelle genomes and ≥5 bp unmatched on the edge were considered as reliable norgContigs and used in further study. The 'Sunset'-specific norgDNAs were obtained when no hits were determined by BLAST against 'SunUp' reference genome (*E*-value ≤1e-5) (**Fig. 4E**). An in-house software pipeline written in a mixture of python scripts (available upon request) was developed for automatically processing and identifying of norgDNA junction sites in 'Sunset'.

Identity between 'SunUp' Organelle-like Borders of Transgenic Inserts and 'Sunset' NorgDNA

Sequences of three 'SunUp' transformation plasmid derived inserts with borders are available at Genbank (accession numbers: FJ467933, FJ467932 and FJ467934) (**Fig. 5A**). Six organelle-like sequences flanking three 'SunUp' transgenic insertions could be extracted from them. Those organelle-like borders were screened for organelle genome similarity using BLAST (*E*-value ≤1e-5).

In order to see whether these six organelle-like border sequences were present in the genome prior to bombardment or not, we set out to examine the identity between norgDNAs in 'Sunset' and the flanking norgDNAs of inserts in 'SunUp'. To see the identity between them, searches between reliable 'Sunset' norgContigs as query against two organelle genomes and six organelle-like borders as databases were separately performed with BLASTN using an *E*-value cut-off of 1e-5 (**Fig. 4F**). Blast hits between 'Sunset' norgDNA and 'SunUp' organelle-like borders were considered the best hits if the query-start and query-end of one hit in blast-output2 matches the hit with the same query ID in blast-output1. The other option to be considered a best hit would be if the longest hit (which cannot be any longer and is shorter than norgDNA in corresponding hit in blast-output1) of one query ID in blast-output2 totally matches the corresponding part of hit in blast-output1.

Identification of the Origin of 'SunUp' Organelle-like Borders of Transgenic Inserts

In order to see whether these six organelle-like border sequences were present in the genome prior to bombardment or not, we developed a strategy which utilizes high-throughput and massive paired-end mapping to identify deletions in 'Sunset' relative to the reference genome (**Fig. 5**). Clean paired-end reads of

'Sunset' were aligned against three 'SunUp' transformation plasmid derived inserts with borders as a whole by using BWA's short read aligner with default parameters. After removing multiple mapping reads, the unimap alignments were converted from SAM format into BAM format. Aligned reads were then sorted, treated for potential PCR duplicates removal and indexed using SAMtools. Three BAM files of read alignments in regions of three inserts with borders could be separated according to reference names using SAMtools 'view' command. Name-sorted BAM files were converted to BED with bamToBed script from the BEDTools package [70]. A deletion in 'Sunset' relative to the reference genome was identified using paired-end reads spanning the transgenic insert region. If the inner distance of paired-end reads in the reference genome was longer than a transgenic insert size then a deletion had taken place. If this was found to be the case, the flanking norgDNA of transgenic inserts in transgenic cultivar 'SunUp' were identified as native to its progenitor 'Sunset'. Histogram plots of the inner distance of mapped paired-end reads in regions of three inserts with borders could be generated by the R version 3.2.1 statistical package (www.CRAN.R-project.org).

List Of Abbreviations

ALT: alternate alleles; array-CGH: comparative genome hybridization; BAM: Binary Alignment/Map; CDS: coding sequence; *cp*: coat protein; CTX: inter-chromosomal translocations; DEL: deletions; DSB: DNA double strand break; GM: genetically modified; GO: Gene Ontology; IGV: Integrative Genomics Viewer; InDels: Small Inserts/deletions; INS: insertions; INV: inversions; ITX: intra-chromosomal translocations; mt: mitochondria; MYMIV: *Mungbean yellow mosaic India virus*; NGS: next-generation sequencing; NHEJ-DSB repair: Non-homologous end joining of double-strand break repair; norgDNA: nuclear organelle DNA; NS/Syn: non-synonymous to synonymous; nuDNA: nuclear DNA; NUMT: nuclear mitochondrial DNA; NUPT: nuclear plastid DNA; orgDNA: organelle DNA; PDR: pathogen-derived resistance; PE: paired-end; PRSV: *Papaya Ringspot Virus*; pt: chloroplast; SAM: Sequence Alignments/Map; SNPs: single nucleotide polymorphisms; SSM: slipped-strand mispairing; SVs: structural variations; Ts: transitions; Tv: transversions; UV: ultraviolet; VCF: variant call format

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The Illumina DNA-sequencing raw reads of nontransgenic cultivar 'Sunset' are available from the NCBI Sequence Read Archive database (SRA; https://www.ncbi.nlm.nih.gov/sra/) under project accession number

of PRJNA578028. The pipelines used during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no conflict of interest.

Funding

This work was supported by a startup fund from Fujian Agriculture and Forestry University, the US National Science Foundation (NSF) Plant Genome Research Program Award DBI-1546890 to R.M, the Natural Science Foundation of Fujian Province, China (Grant Number 2019J0102 and 2018J01601) and Xiyuan River Scholar Project Fund from College of Life Science of Fujian Normal University (Grant Number FZSKG2018004).

Authors' contributions

RM and JF conceived the study and designed the experiments. JF, AW, YC, and JY carried out the experiments and analyzed the data. JF and RM wrote the manuscript. All authors read and approved the final paper.

Acknowledgements

We would like to thank the reviewers for their helpful comments on the manuscript.

Authors' information

[1] The Public Service Platform for Industrialization Development Technology of Marine Biological Medicine and Product of State Oceanic Administration, Key Laboratory of Developmental and Neural Biology, College of Life Sciences, Fujian Normal University, Fuzhou 350117, P. R. China

[2] Center of Engineering Technology Research for Microalgae Germplasm Improvement of Fujian, Southern Institute of Oceanography, Fujian Normal University, Fuzhou 350117, P. R. China

[3] FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou 350002, Fujian, P.R. China;

[4] Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

References

1. Fisher JB: **The Vegetative and Reproductive Structure of Papaya (*Carica papaya*)**. *Harold L Lyon Arboretum* 1980, **1**(4):191-208.
2. Ming R, Moore PH: **Genetics and Genomics of Papaya**. In: *Plant Genetics & Genomics Crops & Models.* Springer; 2014.
3. Gonsalves D: **Control of papaya ringspot virus in papaya: a case study**. *Annual Review of Phytopathology* 1998, **36**(1):415-437.
4. Storey W, Ferwerda F, Wit F: **Outlines of Perennial Crop Breeding in the Tropics**. Landbouwhogeschool Wageningen, Netherlands. Veenman & Zonen, Wageningen; 1969. pp389-408.
5. Fitch MM, Manshardt RM, Gonsalves D, Slightom JL, Sanford JC: **Stable transformation of papaya via microprojectile bombardment**. *Plant Cell Reports* 1990, **9**(4):189-194.
6. Fitch MM, Manshardt RM, Gonsalves D, Slightom JL, Sanford JC: **Virus resistant papaya plants derived from tissues bombarded with the coat protein gene of papaya ringspot virus**. *Nature Biotechnology* 1992, **10**(11):1466-1472.
7. Suzuki JY, Tripathi S, Fermín GA, Jan F-J, Hou S, Saw JH, Ackerman CM, Yu Q, Schatz MC, Pitz KY: **Characterization of insertion sites in Rainbow papaya, the first commercialized transgenic fruit crop**. *Tropical Plant Biology* 2008, **1**(3-4):293-309.
8. Zhang S, Chen W, Xin L, Gao Z, Hou Y, Yu X, Zhang Z, Qu SJHr: **Genomic variants of genes associated with three horticultural traits in apple revealed by genome re-sequencing**. *Horticulture Research* 2014, **1:14045. doi:10.1038/hortres.2014.45**.
9. Timmis JN, Scott NS: **Sequence homology between spinach nuclear and chloroplast genomes**. *Nature* 1983, **305**(5929):65-67.
10. Baldauf SL, Palmer J: **Evolutionary transfer of the chloroplast *tuf*A gene to the nucleus**. *Nature* 1990, **344**(6263):262-265.
11. Martin W, Herrmann RG: **Gene transfer from organelles to the nucleus: how much, what happens, and why?** *Plant Physiology* 1998, **118**(1):9-17.
12. Kleine T, Maier UG, Leister D: **DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis**. *Annual Review of Plant Biology* 2009, **60**:115-138.
13. VanBuren R, Ming R: **Organelle DNA accumulation in the recently evolved papaya sex chromosomes**. *Molecular Genetics and Genomics* 2013, **288**(5-6):277-284.
14. Sheppard AE, Timmis JN: **Instability of plastid DNA in the nuclear genome**. *PLoS Genetics* 2009, **5**(1):e1000323.
15. Yoshida T, Furihata HY, Kawabe A: **Analysis of nuclear mitochondrial DNAs and factors affecting patterns of integration in plant species**. *Genes and Genetic Systems* 2017, **92**(1):27-33.

16. Michalovova M, Vyskot B, Kejnovsky E: **Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization**. *Heredity* 2013, **111**(4):314-320.
17. Yoshida T, Furihata HY, Kawabe A: **Patterns of genomic integration of nuclear chloroplast DNA fragments in plant species**. *DNA Research* 2014, **21**(2):127-140.
18. Leister D: **Origin, evolution and genetic effects of nuclear insertions of organelle DNA**. *Trends in Genetics Tig* 2005, **21**(12):655-663.
19. Yoshida T, Furihata HY, To TK, Kakutani T, Kawabe A: **Genome defense against integrated organellar DNA fragments from plastids into plant nuclear genomes through DNA methylation**. *Scientific reports* 2019, **9**:2060. doi: 2010.1038/s41598-41019-38607-41596.
20. Noutsos, C.: **Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants**. *Genome Research* 2005, **15**(5):616-628.
21. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
23. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso*-2; *iso*-3**. *Fly* 2012, **6**(2):80-92.
24. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.
25. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nature Biotechnology* 2011, **29**(1):24-26.
26. Neelakandan AK, Wang K: **Recent progress in the understanding of tissue culture-induced genome level changes in plants and potential applications**. *Plant Cell Reports* 2012, **31**(4):597-620.
27. Kovalic D, Garnaat C, Guo L, Yan Y, Groat J, Silvanovich A, Ralston L, Huang M, Tian Q, Christian A: **The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterization of crops improved through modern biotechnology**. *The Plant Genome* 2012, **5**(3):149-163.
28. Kawakatsu T, Kawahara Y, Itoh T, Takaiwa F: **A whole-genome analysis of a transgenic rice seed-based edible vaccine against cedar pollen allergy**. *DNA Research* 2013, **20**(6):623-631.
29. Endo M, Kumagai M, Motoyama R, Sasaki-Yamagata H, Mori-Hosokawa S, Hamada M, Kanamori H, Nagamura Y, Katayose Y, Itoh T: **Whole-Genome Analysis of Herbicide-Tolerant Mutant Rice Generated by *Agrobacterium*-Mediated Gene Targeting**. *Plant and Cell Physiology* 2015, **56**(1):116-125.
30. Kim M, Moore P, Zee F, Fitch MM, Steiger D, Manshardt R, Paull R, Drew RA, Sekioka T, Ming R: **Genetic diversity of *Carica papaya* as revealed by AFLP markers**. *Genome* 2002, **45**(3):503-512.
31. Blas AL, Ming R, Liu Z, Veatch OJ, Paull RE, Moore PH, Yu Q: **Cloning of the papaya chromoplast-specific lycopene *β*-Cyclase, *CpCYC-b*, controlling fruit flesh color reveals conserved microsynteny and a**

recombination hot spot. *Plant Physiology* 2010, **152**(4):2013-2022.

32. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)**. *Nature* 2008, **452**(7190):991-996.
33. Yu Q, Navajas-Pérez R, Tong E, Robertson J, Moore PH, Paterson AH, Ming R: **Recent origin of dioecious and gynodioecious Y chromosomes in papaya**. *Tropical Plant Biology* 2008, **1**(1):49-57.
34. Yadav CB, Bhareti P, Muthamilarasan M, Mukherjee M, Khan Y, Rathi P, Prasad M: **Genome-Wide SNP Identification and Characterization in Two Soybean Cultivars with Contrasting *Mungbean Yellow Mosaic India Virus* Disease Resistance Traits**. *PLoS One* 2015, **10**(4):e0123897.
35. Project IRGS: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.
36. Jiang C, Mithani A, Gan X, Belfield EJ, Klingler JP, Zhu J-K, Ragoussis J, Mott R, Harberd NP: **Regenerant *Arabidopsis* lineages display a distinct genome-wide spectrum of mutations conferring variant phenotypes**. *Current Biology* 2011, **21**(16):1385-1390.
37. Miyao A, Nakagome M, Ohnuma T, Yamagata H, Kanamori H, Katayose Y, Takahashi A, Matsumoto T, Hirochika H: **Molecular spectrum of somaclonal variation in regenerated rice revealed by whole-genome sequencing**. *Plant and Cell Physiology* 2012, **53**(1):256-264.
38. Anderson JE, Michno JM, Kono TJY, Stec AO, Campbell BW, Curtin SJ, Stupar RM: **Genomic variation and DNA repair associated with soybean transgenesis: a comparison to cultivars and mutagenized plants**. *BMC Biotechnology* 2016, **16:41. doi: 10.1186/s12896-016-0271-z**.
39. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M: **The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana***. *Science* 2010, **327**(5961):92-94.
40. Fang J, Wood A, Chen R, Ming R: **Molecular basis of off-type microsatellite markers in papaya**. *Euphytica* 2016, **209**(2):323-339.
41. Fang J, Lin A, Qiu W, Cai H, Umar M, Chen R, Ming R: **Transcriptome profiling revealed stress-induced and disease resistance genes up-regulated in PRSV resistant transgenic papaya**. *Frontiers in Plant Science* 2016, **7:855. doi: 10.3389/fpls.2016.00855.**
42. Doerfler W, Hohlweg U, Müller K, Remus R, Heller H, Hertz JJAotNYAoS: **Foreign DNA integration--perturbations of the genome--oncogenesis**. *Annals of the New York Academy of Sciences* 2010, **945**(1):276-288.
43. Dai S, Zheng P, Marmey P, Zhang S, Tian W, Chen S, Beachy RN, Fauquet CJMB: **Comparative analysis of transgenic rice plants obtained by *Agrobacterium*-mediated transformation and particle bombardment**. *Molecular Breeding* 2001, **7**(1):25-33.
44. Kohli A, Leech M, Vain P, Laurie DA, Christou PJPotNAoSotUSoA: **Transgene organization in rice engineered through direct DNA transfer supports a two-phase integration mechanism mediated by the establishment of integration hot spots**. *Proceedings of the National Academy of Sciences* 1998, **95**(12):7203-7208.
45. Shirasawa K, Hirakawa H, Nunome T, Tabata S, Isobe S: **Genome-wide survey of artificial mutations induced by ethyl methanesulfonate and gamma rays in tomato**. *Plant Biotechnology Journal* 2016,

14(1):51-60.

46. Kashima K, Mejima M, Kurokawa S, Kuroda M, Kiyono H, Yuki Y: **Comparative whole-genome analyses of selection marker–free rice-based cholera toxin B-subunit vaccine lines and wild-type lines**. *BMC Genomics* 2015, **16:48. doi: 10.1186/s12864-015-1285-y.**
47. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: **Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning**. *Nature* 2008, **452**(7184):215-219.
48. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W: **Molecular basis of base substitution hotspots in *Escherichia coli***. *Nature* 1978, **274**(5673):775-780.
49. Duncan BK, Miller JH: **Mutagenic deamination of cytosine residues in DNA**. *Nature* 1980, **287**(5782):560-561.
50. Friedberg EC, Walker GC, Siede W, Wood RD: **DNA repair and mutagenesis**: American Society for Microbiology Press; 2005.
51. Cooper JL, Greene EA, Till BJ, Codomo CA, Wakimoto BT, Henikoff SJG: **Retention of induced mutations in a Drosophila reverse-genetic resource**. *Genetics* 2008, **180**(1):661-667.
52. Liu S, Liu W, Jakubczak JL, Erexson GL, Tindall KR, Chan R, Muller WJ, Adhya S, Garges S, Merlino G: **Genetic instability favoring transversions associated with ErbB2-induced mammary tumorigenesis**. *Proceedings of the National Academy of Sciences* 2002, **99**(6):3770-3775.
53. Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA: **8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G→T and A→C substitutions**. *Journal of Biological Chemistry* 1992, **267**(1):166-172.
54. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution**. *Molecular Biology and Evolution* 1987, **4**(3):203-221.
55. Tachida H, Iizuka M: **Persistence of repeated sequences that evolve by replication slippage**. *Genetics* 1992, **131**(2):471-478.
56. Jain M, Moharana KC, Shankar R, Kumari R, Garg R: **Genomewide discovery of DNA polymorphisms in rice cultivars with contrasting drought and salinity stress response and their functional relevance**. *Plant Biotechnology Journal* 2014, **12**(2):253-264.
57. Timmis JN, Ayliffe MA, Huang CY, Martin W: **Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes**. *Nature Reviews Genetics* 2004, **5**(2):123-135.
58. Chen H, Yu Y, Chen X, Zhang Z, Gong C, Li J, Wang AJF, Genomics I: **Plastid DNA insertions in plant nuclear genomes: the sites, abundance and ages, and a predicted promoter analysis**. *Functional and Integrative Genomics* 2015, **15**(2):131-139.
59. Stegemann S, Bock R: **Experimental reconstruction of functional gene transfer from the tobacco plastid genome to the nucleus**. *The Plant Cell* 2006, **18**(11):2869-2878.
60. Noutsos C, Richly E, Leister D: **Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants**. *Genome Research* 2005, **15**(5):616-628.
61. Khanna KK, Jackson SP, %J Nature Genetics: **DNA double-strand breaks: signaling, repair and the cancer connection**. *Nature Genetics* 2001, **27**(3):247-254.

62. Huang CY, Ayliffe MA, Timmis JN: **Direct measurement of the transfer rate of chloroplast DNA into the nucleus**. *Nature* 2003, **422**(6927):72-76.
63. Lutz KA, Wang W, Zdepski A, Michael TP: **Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing**. *BMC Biotechnology* 2011, **11:54. doi: 10.1186/1472-6750-11-54**.
64. Patel RK, Jain M: **NGS QC Toolkit: a toolkit for quality control of next generation sequencing data**. *PloS One* 2012, **7**(2):e30619.
65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of Molecular Biology* 1990, **215**(3):403-410.
66. Du Z, Zhou X, Ling Y, Zhang Z, Su Z: **agriGO: a GO analysis toolkit for the agricultural community**. *Nucleic Acids Research* 2010, **38**(S2):W64–W70.
67. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation**. *Nature Methods* 2009, **6**(9):677-681.
68. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-359.
69. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K: **De novo assembly of human genomes with massively parallel short read sequencing**. *Genome Research* 2010, **20**(2):265-272.
70. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.

Tables

Table 1. Papaya Sunset genome-wide assembly statistics

		Sunset genome wide
	Total read count	74,169,662
	Read length (bp)	124
	Total read length (Gb)	9.197
	Average coverage (×)	24.72
Remove multiple mapping and duplicates	Total read count	48,170,821
	Mapped read count	48,154,999
	Mapped read rate (%)	99.97
	Unmapped read count	15,822
	Properly paired read count	46,139,627
	Properly paired read rate (%)	95.78

Table 2. Number of homo/hetero SNPs and InDels detected before and after data filtering

	Raw	DP10-100Q50*
Homo SNPs	83,926	40,871
Hetero SNPs	603,970	269,493
Total SNPs	687,896	**310,364**
Homo InDels	41,218	19,135
Hetero InDels	29,504	14,936
Total InDels	70,722	**34,071**
Total	758,618	**344,435**

Notes: (*): Validated depth and quality. DP10-100Q50: The variant calls with read depths of <10 or >100 and polymorphism sites of quality <50 were filtered out.

Table 3. Summary of polymorphisms between SunUp and Sunset

Chrom.	Total size(bp)	No.of SNPs	No.of InDels	SNP per 1kb	In/Del per 1kb
CHROM_1	22,976,894	16,246	2,214	0.71	0.10
CHROM_2	28,675,255	6,842	1,893	0.24	0.07
CHROM_3	29,397,938	18,294	2,630	0.62	0.09
CHROM_4	27,056,416	12,813	2,426	0.47	0.09
CHROM_5	24,352,217	13,952	2,150	0.57	0.09
CHROM_6	30,516,430	50,463	3,821	1.65	0.13
CHROM_7	22,375,162	17,294	2,361	0.77	0.11
CHROM_8	21,952,264	12,610	2,001	0.57	0.09
CHROM_9	27,303,179	12,021	1,986	0.44	0.07
Unanchored scaffolds	135,176,073	149,829	12,589	1.11	0.09
Genome-wide	369,781,828	**310,364**	**34,071**	0.84	0.09

Table 4. Pattern of homozygous and heterozygous SNPs

SNP pattern		Homo SNPs	Hetero SNPs	Total SNPs
Transition	A/G	5,315	45,067	50,382
	T/C	5,768	44,871	50,639
	G/A	4,701	47,543	52,244
	C/T	4,908	47,160	52,068
	total(Ts)	**20,692**	**184,641**	**205,333**
Transversion	A/C	2,329	12,114	14,443
	A/T	2,327	11,999	14,326
	T/A	2,310	12,199	14,509
	T/G	2,274	12,193	14,467
	G/C	2,509	6,589	9,098
	G/T	3,020	11,576	14,596
	C/A	3,104	11,522	14,626
	C/G	2,306	6,660	8,966
	total(Tv)	**20,179**	**84,852**	**105,031**
	Ts/Tv	**1.03**	**2.18**	**1.95**

Table 5. Prediction of the effects of SNPs and InDels

Impact (count, percentage in Sunset)	Effect type	Count	Percentage (%)
HIGH (1591, 0.2714%)	frameshift_variant	1,033	0.1762
	frameshift_variant+splice_region_variant	66	0.0113
	frameshift_variant+start_lost	12	0.0020
	frameshift_variant+stop_gained	9	0.0015
	frameshift_variant+stop_gained+splice_region_variant	1	0.0002
	frameshift_variant+stop_lost	1	0.0002
	frameshift_variant+stop_lost+splice_region_variant	15	0.0026
	splice_acceptor_variant+intron_variant	75	0.0128
	splice_acceptor_variant+splice_region_variant+intron_variant	2	0.0003
	splice_donor_variant+intron_variant	87	0.0148
	splice_donor_variant+splice_region_variant+intron_variant	1	0.0002
	start_lost	24	0.0041
	start_lost+splice_region_variant	1	0.0002
	stop_gained	185	0.0316
	stop_gained+disruptive_inframe_insertion	1	0.0002
	stop_gained+splice_region_variant	6	0.0010
	stop_lost	23	0.0039
	stop_lost+inframe_insertion+splice_region_variant	1	0.0002
	stop_lost+splice_region_variant	48	0.0082
MODERATE (7533, 1.2849%)	missense_variant+splice_region_variant	130	0.0222
	disruptive_inframe_deletion	3	0.0005
	disruptive_inframe_insertion	7	0.0012
	inframe_deletion	17	0.0029
	inframe_insertion	22	0.0038
	missense_variant	7,354	1.2544
LOW (6114, 1.0429%)	initiator_codon_variant	9	0.0015
	splice_region_variant+intron_variant	833	0.1421
	splice_region_variant+stop_retained_variant	13	0.0022
	splice_region_variant+synonymous_variant	100	0.0171
	stop_retained_variant	4	0.0007
	synonymous_variant	5,155	0.8793
MODIFIER (571039, 97.4009%)	downstream_gene_variant	128,197	21.8663
	intergenic_region	278,076	47.4308
	intron_variant	36,054	6.1497
	upstream_gene_variant	128,712	21.9541

Notes: Variants (SNPs and InDels) that may affect protein function were categorized into 35 types. These types were further grouped into HIGH, MODERATE, LOW, and MODIFIER according to potential severity. The assignment criteria were pre-defined in the annotation program (SNPEff).

Table 6. Distribution of structural variations on nine pseudomolecules of papaya

Type of SVs	Chromosomes									Subtotal			Total
	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9		unanchored*	uncertain**	
CTX	5	8	0	3	7	11	7	6	11	58		891	949
ITX	2	7	7	1	0	3	1	2	4	27	17		44
INV	1	0	0	0	0	1	0	0	1	3	0		3
INS	5	13	6	2	4	11	5	1	1	48	28		76
DEL	9	14	10	8	13	14	3	5	8	84	44		128
Total	22	42	23	14	24	40	16	14	25	220	89	891	1200

Notes: *Scaffolds which have not been anchored to chromosomes to date. **inter-chromosomal translocation occur between a scaffold pair. To date, one or two of the scaffolds have not been anchored to chromosomes.

Table 7. Junction site numbers and identities of NUPT and NUMT

Junction site type	NUPT			NUMT		
	Count	Percentage	Identity (nupt/pt)*	Count	Percentage	Identity (numt/mt)*
SunUp	3430	100.00%		2764	100.00%	
Shared	3327	97.00%	91.92%	2642	95.59%	92.97%
Specific in SunUp	103	3.00%	94.03%	122	4.41%	93.77%
Sunset	3346	100.00%		2745	100.00%	
Shared	3327	99.43%	91.92%	2642	95.50%	92.97%
Specific in Sunset	19	0.57%	95.64%	103	4.50%	96.95%

Notes: (*): the identity between nupt/numt and corresponding organelle genome. chloroplast (pt); mitochondria (mt).

Table 8. The chromosome information for organelle DNA integration sites

Chromosome	Specific junction sites in SunUp			
	NUPT		NUMT	
	Count	Percentage	Count	Percentage
CHROM_1	3	2.91%	7	5.74%
CHROM_2	12	11.65%	12	9.84%
CHROM_3	2	1.94%	8	6.56%
CHROM_4	9	8.74%	10	8.20%
CHROM_5	6	5.83%	6	4.92%
CHROM_6	9	8.74%	13	10.66%
CHROM_7	6	5.83%	10	8.20%
CHROM_8	9	8.74%	11	9.02%
CHROM_9	8	7.77%	8	6.56%
Unanchored scaffolds	39	48.75%	37	30.33%
Total	103	100.00%	122	100.00%

Table 9. Comparative analysis of 6 organelle-like borders of 3 transgenic insertions

Insertion	Border	Sequence type	Length (bp)	Identity with orgDNA (%)	Sunset matches		
					Identity with inserts (%)	Count	Combined length (bp)
Functional insert: *cp*	A	pt	4000	100.00	97.01	12	4180
	B	pt	1790	99.94	99.09	6	1944
Nonfunctional insert: pseudo-*nptII*	A	pt	363	100.00	97.96	1	49
	B	pt	827	100.00	93.68	5	1231
Nonfunctional insert: pseudo-*tetA*	A	pt	6299	98.60	95.09	43	4242
	B	mt	1708	98.18	99.09	2	1738

Notes: chloroplast (pt); mitochondria (mt).

Figures

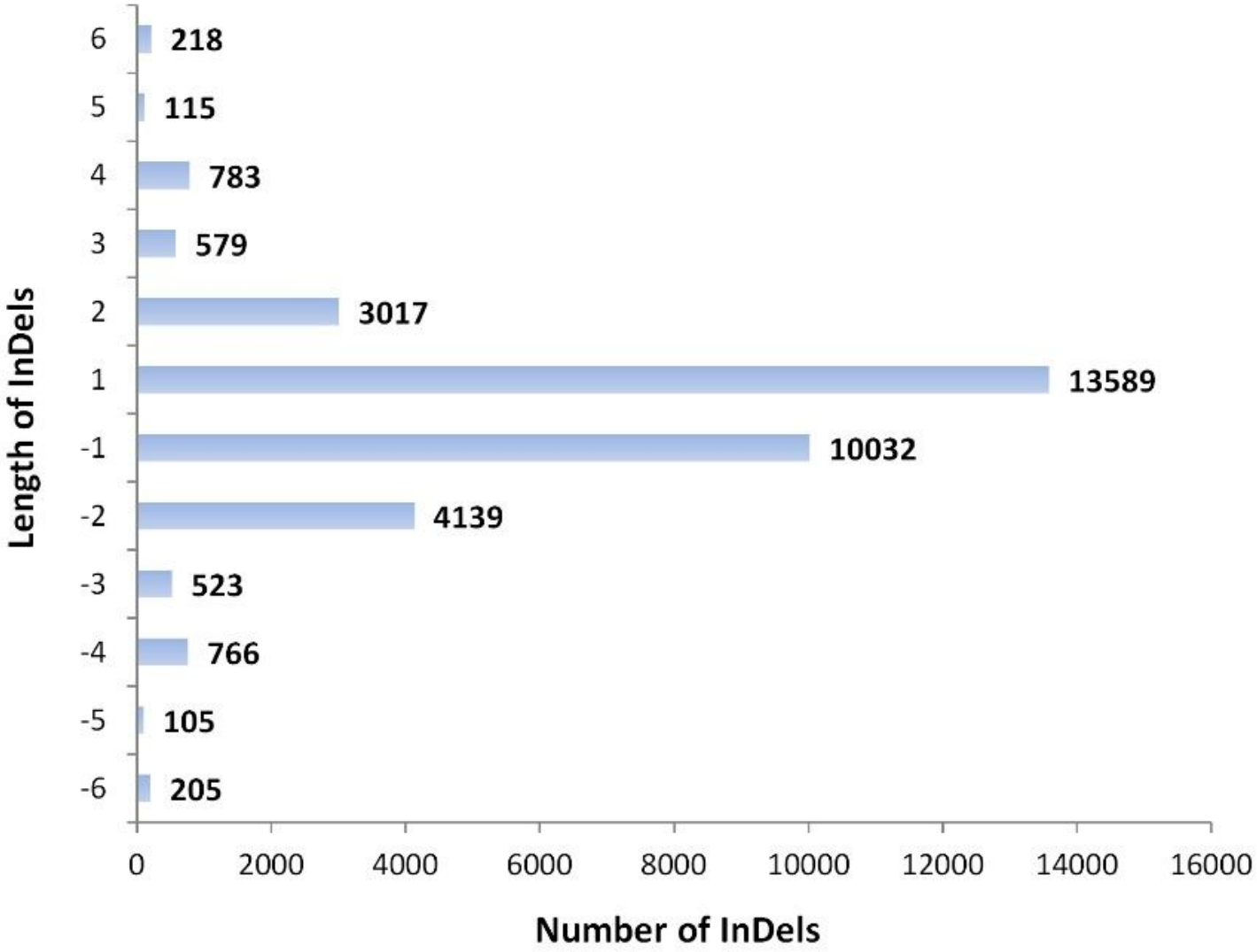

Figure 1

Histogram of InDels number and length in Sunset genome compared to SunUp reference genome.

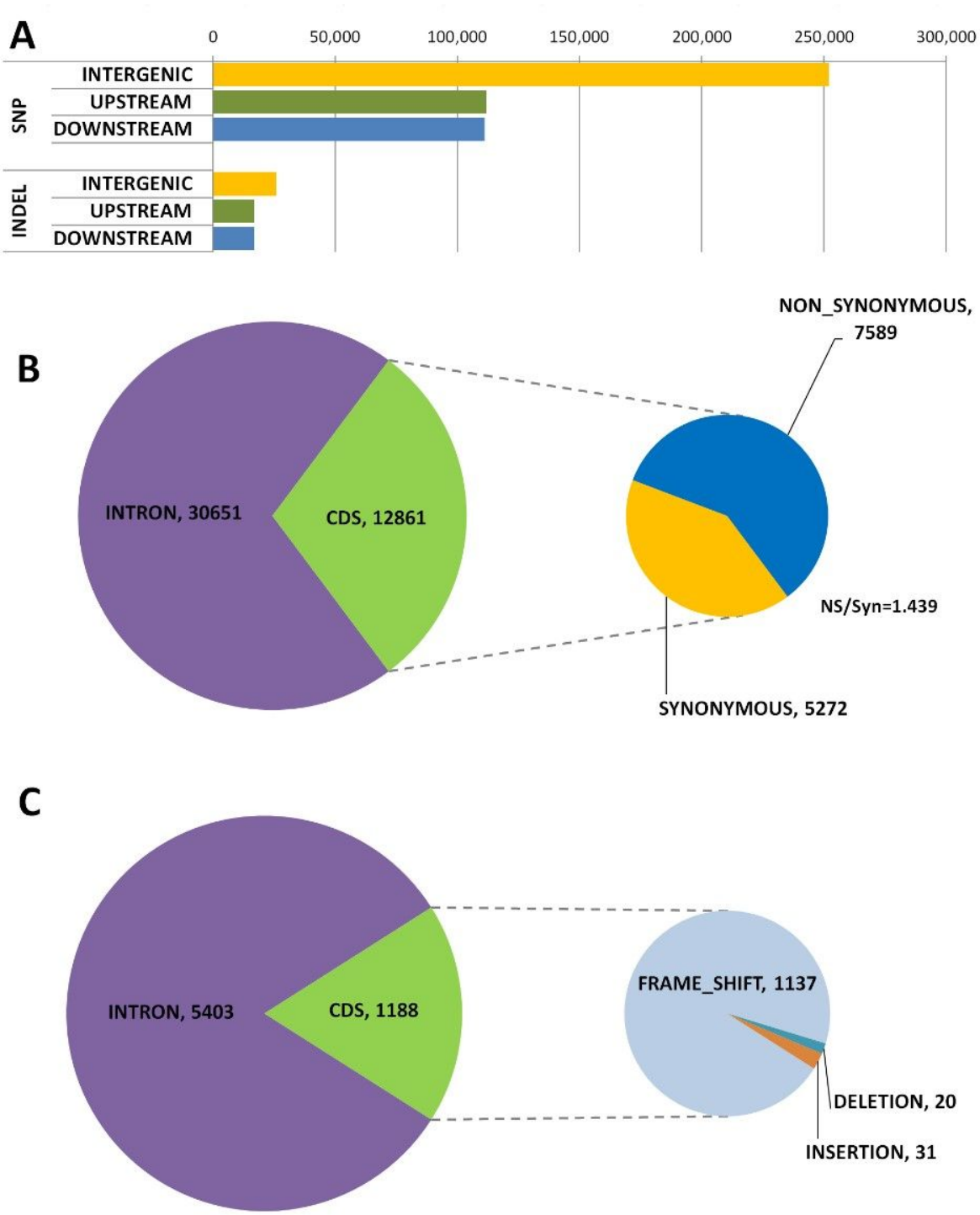

Figure 2

Annotation of single-nucleotide polymorphisms (SNPs) and InDels in Sunset genome compared to SunUp reference genome. A. Distribution of SNPs and InDels in intergenic, upstream and downstream regions. B. Distribution of SNPs in different genic regions. C. Distribution of InDels in genic regions. The number of synonymous and non-synonymous SNPs detected within the CDS region has also been shown.

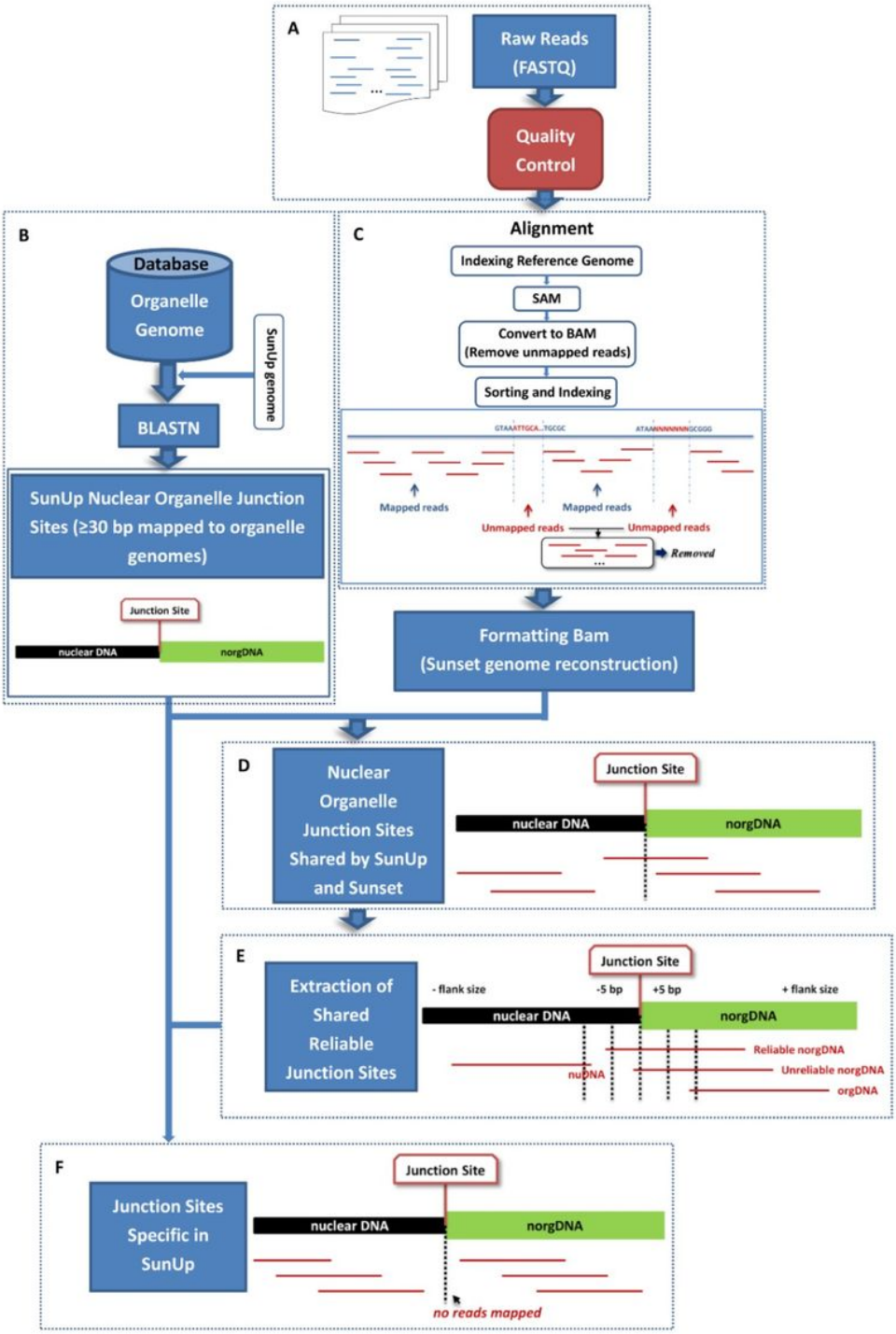

Figure 3

Pipeline of SunUp-specific genomic integration of nuclear organelle DNA fragments. A. Quality control of raw sequenced data. B. Searches for SunUp nuclear organelle junction sites by BLASTN [65]. The BLASTN algorithm was used to search SunUp genome for nuclear plastid DNA (NUPT) and nuclear mitochondria DNA (NUMT) integrations with papaya organelle genomes as databases. Only hits with ≥30 bp mapped to organelle genomes were considered. C. Alignment between Sunset reads and SunUp reference genome. Unmapped reads were removed after subsequent analysis. D. Nuclear organelle junction sites shared by SunUp and Sunset. A junction site was supposed to be shared by SunUp and Sunset genomes when there

were reads mapped to and spanning its position in the SunUp reference genome. E. Extraction of reliable shared junction sites. The mixture of reads that aligned back to the reference genome may originate from different sources of DNA in the Sunset genome, including nuclear DNA (nuDNA), nuclear organelle DNA (norgDNA) and organelle DNA (orgDNA). In order to discriminate these three categories of reads and extract the reliable junction sites shared by SunUp and Sunset, the flanking regions (5 bp upstream and downstream) of the junction sites are used as an indicator. Reliable norgDNA reads were selected if those reads were spanning the junction sites and mapped to at least 5 bp of norgDNA or nuDNA. F. Junction sites specific in SunUp. If there were no reads mapped to or no reliable norgDNA reads spanning the junction site, we considered this junction site as a SunUp-specific norgDNA junction site.

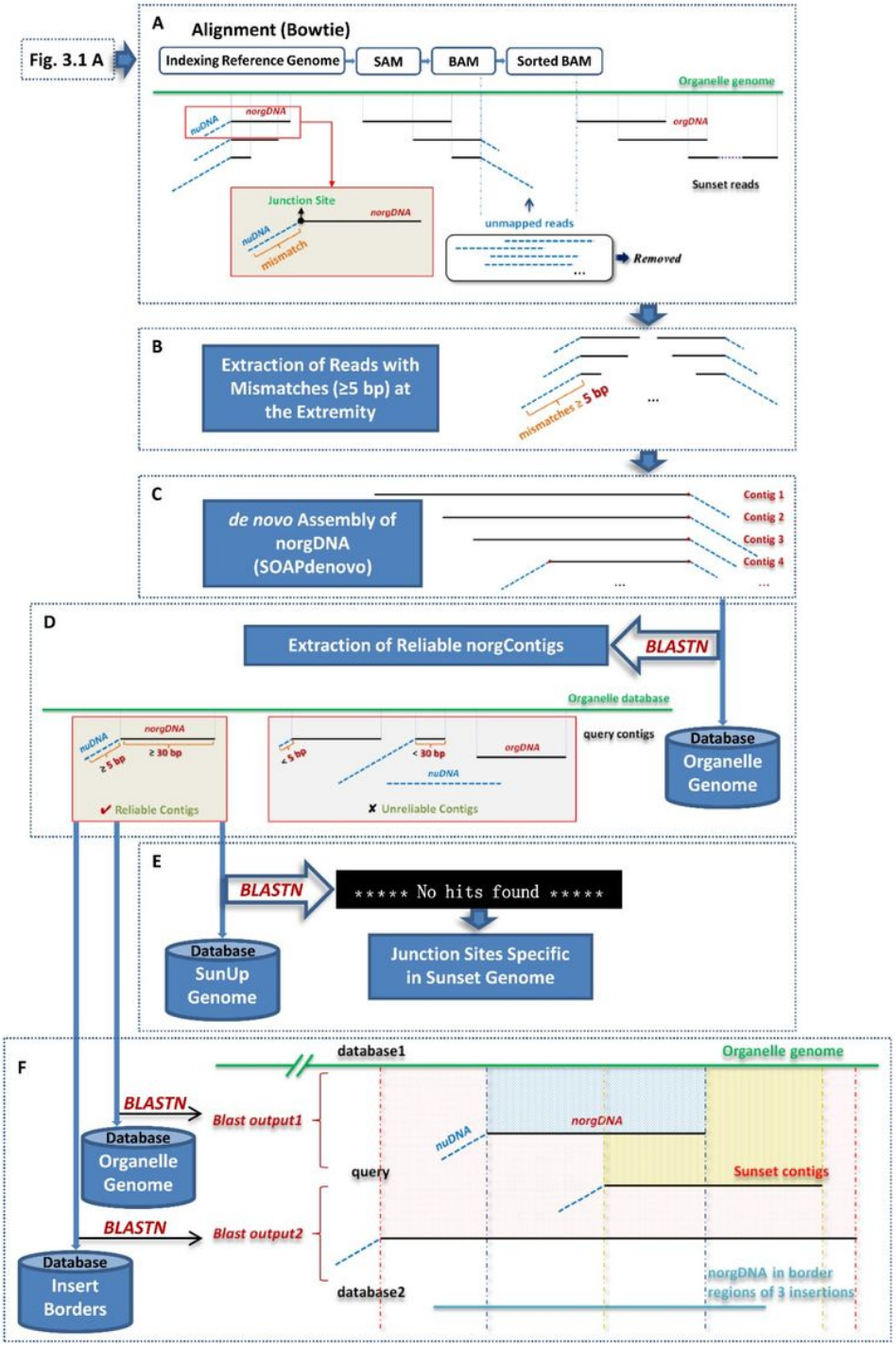

Figure 4

Pipeline of Sunset-specific genomic integration of nuclear organelle DNA fragments. A. Alignment between Sunset reads and organelle reference genome. Unmapped reads were removed after subsequent analysis. Soft-clipped reads were shown in the red box, which refers to reads with mismatches at the extremities. B. Extraction of reads with at least 5 bp mismatches (≥5 bp) at the extremities. C. de novo assembly of norgDNA by SOAPdenovo. D. Extraction of reliable Sunset norgContigs. Only blast hits of norg contigs with ≥30 bp mapped to organelle genomes and ≥5 bp unmatched on the edges were considered as reliable norgContigs. E. Junction sites specific in Sunset. The Sunset-specific norg sequences were obtained when no hits were determined using BLAST against the SunUp reference genome. F. Identity between the six organelle-like borders of transgenic insertions in SunUp and Sunset norgDNA.

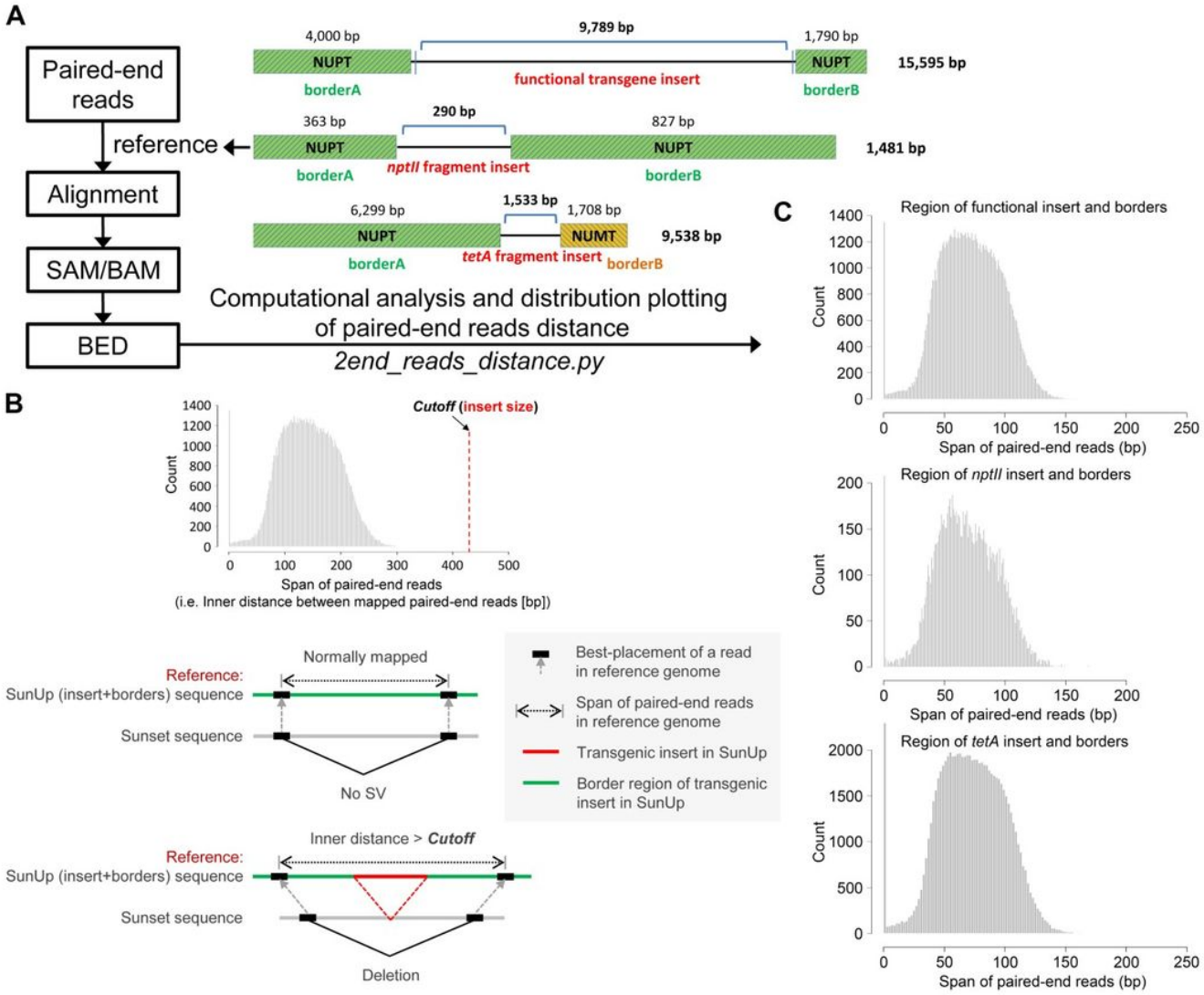

Figure 5

Workflow for the identification of the origin of the flanking norgDNA of transgenic inserts. A. Sequences of three SunUp transformation plasmid derived inserts with borders and the bwa alignment process. B. A strategy using high-throughput and massive paired-end mapping to identify deletions in Sunset relative to the reference genome. Insertions in SunUp were predicted from paired-end spans larger than a specified cutoff

(size of a transgenic insert). C. Histogram plots exhibiting the inner distance of mapped paired-ends in regions of three inserts with borders.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 20191026AdditionalfilesRMforBMCgenomics.docx