**The genome sequence of the octocoral *Paramuricea clavata* – a key resource to study the impact of climate change in the Mediterranean**

**SUPPLEMENTARY INFORMATION**

1. **Estimation of Ploidy Level**

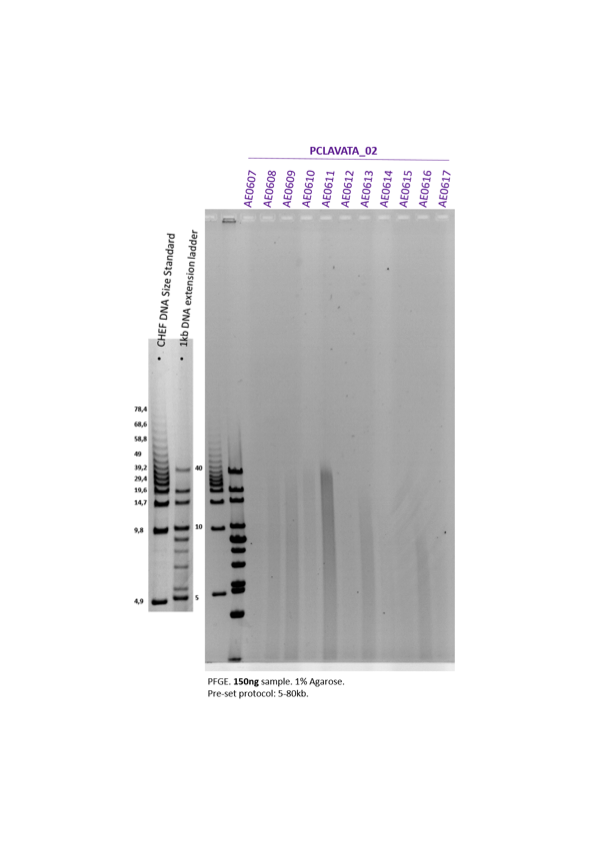
In order to determine the ploidy level of the red gorgonian (*Paramuricea clavata*) we have used *Smudgeplot* v0.2.1 (Ranallo-Benavidez et al. 2020), a k-mer based approach that allows profiling a genome by counting the frequency of different classes of heterozygous k-mers pairs. Following the author recommendations, we used the default k-mer length of 21. We first counted the 21-mers in the raw paired-end reads (PE400) with Jellyfish version 2.2.6 (Marçais and Kingsford 2011) and options: *jellyfish count -m 21 -C -s 1000000000*. Then we ran *smudgeplot.py* to determine the lower and upper coverage cut-offs. Initially, these were determined to be 54 and 2,700. The next step consisted in extracting the 21-mers with a coverage between 54 and 2,700. The extraction was performed using *jellyfish dump -c -L 54 -U 2700 21mer\_counts.jf*. Then, we computed the k-mer pairs from the extracted or *dumped* 21-mers by running *smudgeplot.py hetkmers*. Finally, we produced a first smudgeplot using the coverage of the identified k-mer pairs (.tsv file).

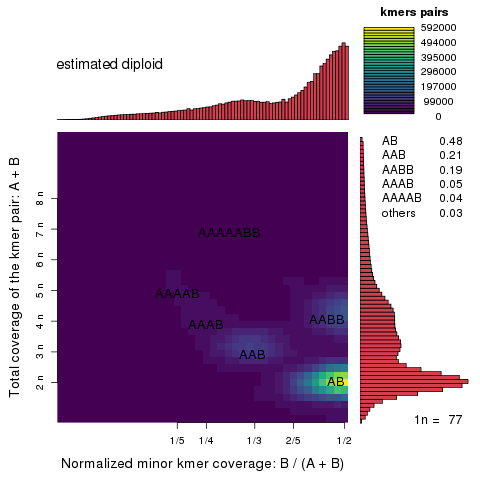
The analysis *Estimated Ploidy* was 2 and reported that the lower coverage threshold used (54) was higher than half of the 1n coverage estimate (1n / 2 = 39.07). Therefore, we used as lower coverage threshold (1n / 2) – 10 to ensure the inclusion of the required k-mer pairs. The value of 1n was estimated more precisely by running Genomescope2 with k=21 and p=2 to be 84.4. Thus, we repeated the *smudgeplot* analysis with lower coverage limit of 32. The final result supports a diploid genome with a proportion of heterozygosity carried by paralogs (i.e non-diploid k-mer pairs) of 0.52 (see Figure S2, Table S1 and Table S2).

1. **Purging Alternative Haplotypes**

Our first version of the genome free of contaminants and mitochondrion (pcla6) had a length of 711Mb and contig N50 of 15.85 Kb. It also presented a k-mer spectra that was symptomatic of a poor collapse of alternative haplotypes (Figure S3a) that could also be inflating the assembly size with respect to the real genome size. This is also reflected in the percentage of duplicated BUSCOs. In order to obtain a more haploid genome reference, we used *Purge Haplotigs* (Roach et al. 2018) as described in the main text. Figure S4 shows the genome coverage distribution obtained after mapping reads back to the pcla6 assembly. This was used to determine the coverage thresholds required to purge the alternative scaffolds from the assembly. As a result, the comparison of the KAT stacked histograms for the assembly before and after purging haplotigs suggests that we removed almost 50% of the alternative alleles from the heterozygous peak (Figure S3a and Figure S3b). This is also consistent with a reduction in the percentage of duplicated BUSCOs (Table S3).

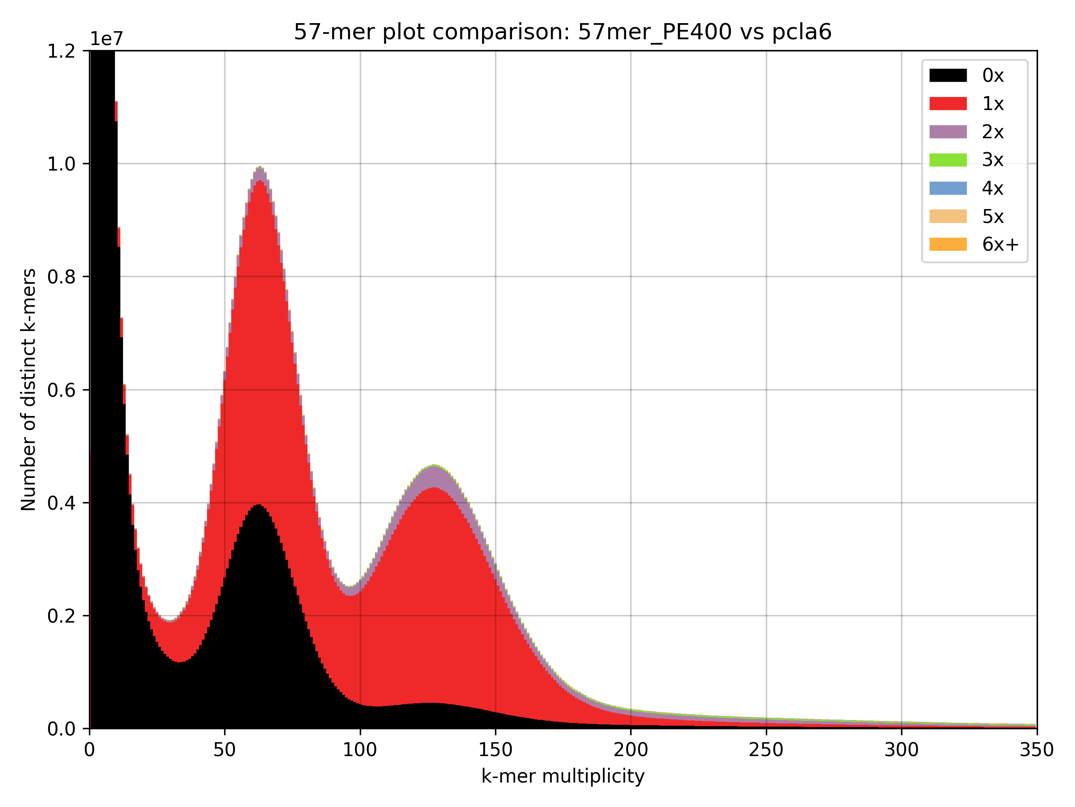
**Figure S1 – DNA extractions quality check.** The quality of DNA extractions was checked usinga Pippin Pulse electrophoresis instrument (Sage Science). This gel picture is representative of the quality of the DNA extractions for *Paramuricea clavata*. Here, the 1% agarose gel was loaded with 150ng of each DNA sample and run for 16h in 0.5X TBE running buffer. The markers used were: 1kb DNA extension ladder (Invitrogen; for sizing DNA fragments from 500 bp to 40 kb) and, CHEF DNA Size Standard (Bio-Rad Laboratories; for size estimation of DNA fragments of 4.9–98 kb). The sample AE0610 corresponds to the fragment section 5\_27-10 used for the long read whole genome sequencing. The smearing pattern is likely due to poor quality DNA (e.g., fragmented), which may explain the low N50 value obtained for the long read whole genome sequencing.



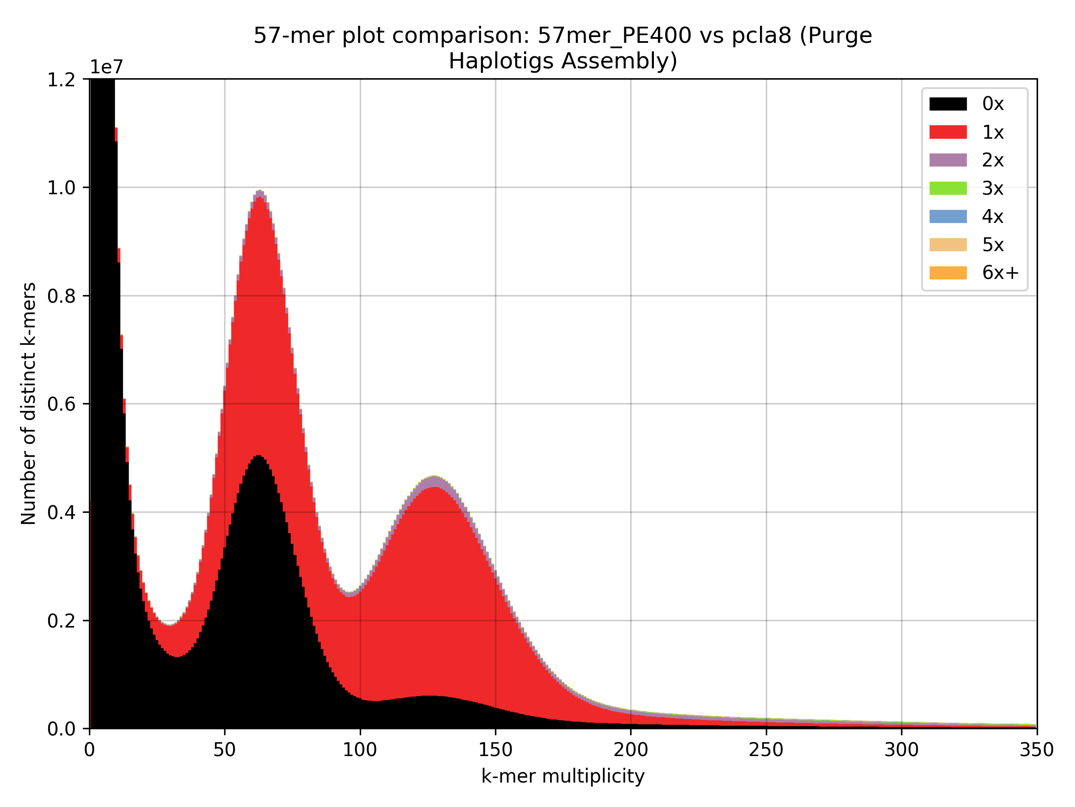


**Figure S2.** Smudgeplot of *P. clavata* genome extracting 21-mers from raw PE reads. Smudgeplot computed from raw PE reads showing the haplotype structure using heterozygous k-mer pairs. Here we used a k-mer length of 21. The color intensity reflects the approximate number of k-mers per bin. The most abundant is the diploid heterozygous k-mer pair AB.

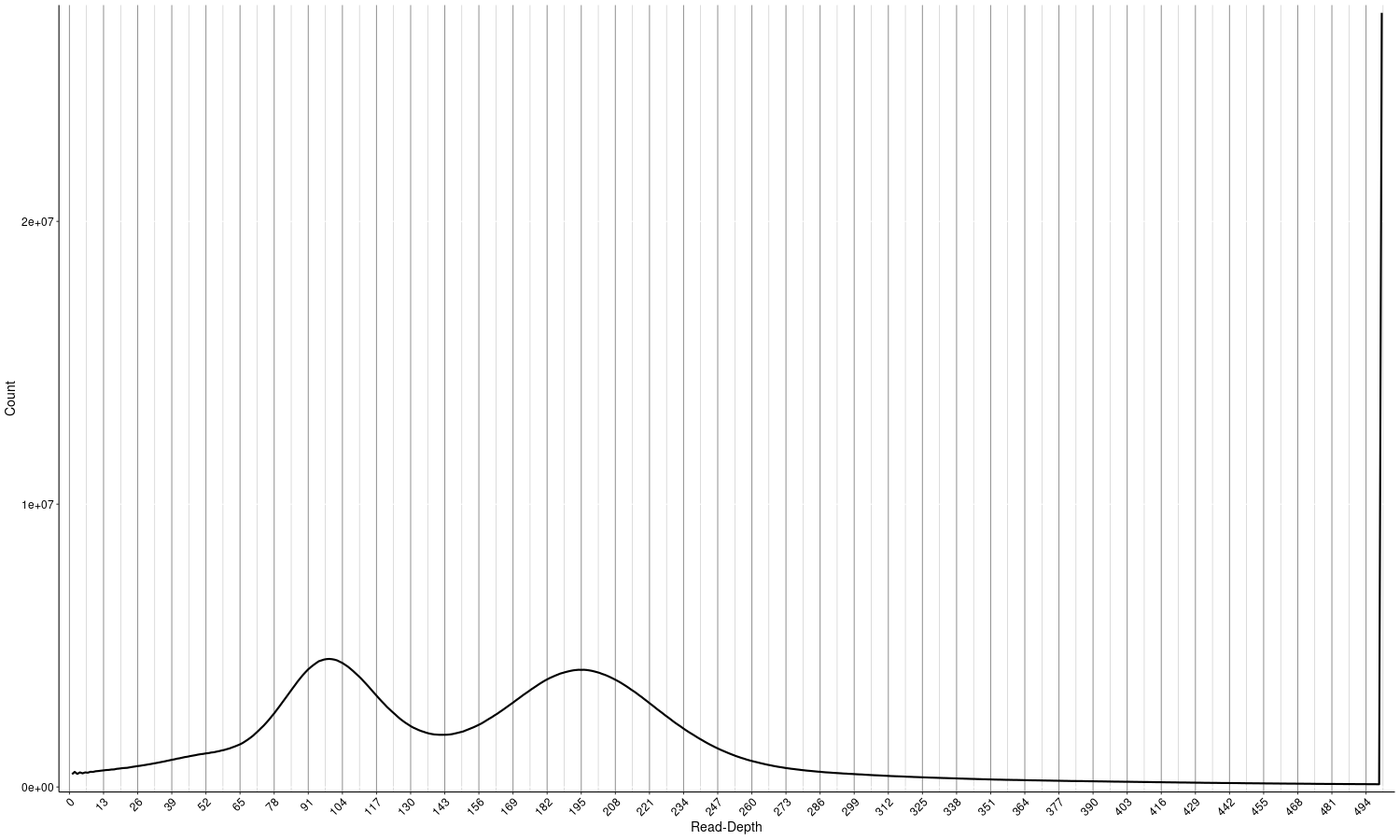
**A**

****

**B**

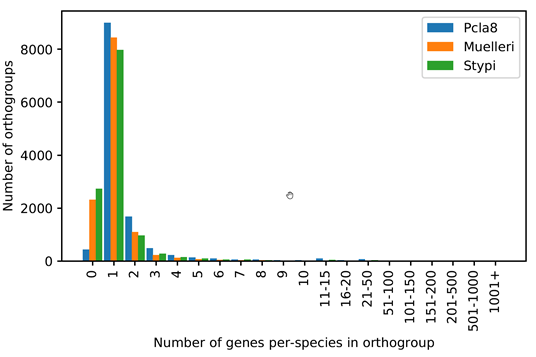
****

**Figure S3.** Stacked histograms for Pcla6 and Pcla8.Plots obtained with KAT v2.3.3 by comparing the filtered PE400 reads (no contaminants) with: a) the pcla6 assembly (before running Purge Haplotigs) and b) the pcla8 assembly (after running Purge Haplotigs).

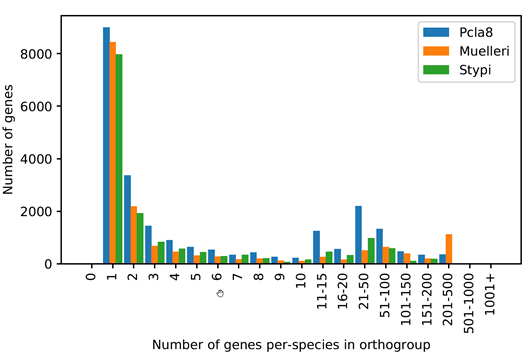


**Figure S4.** Coverage Histogram.Plot obtained with (Purge Haplotigs v1.1.0 by running the command *Purge hist -b PE400.merged.bam -g pcla6.fa -d 500*. This step determined the coverage cut-offs to guide the purging.

**A**



**B**



**Figure S5.** Number of orthogroups with a concrete number of genes per species (a) and

number of genes in orthogroups with a concrete number of genes per species (b).

**Table S1. Proportion of heterozygosity in k-mer pairs with different genome copies**

|  |  |
| --- | --- |
| **genome\_copies** | **propotion\_of\_heterozygosity** |
| 2 | 0.48 |
| 3 | 0.21 |
| 4 | 0.24 |
| 5 | 0.04 |
| 6 | 0 |
| 7 | 0.03 |

**Table S2. Summary of peaks detected by Smudgeplot**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **peak** | **kmers [#]** | **kmers [proportion]** | **summit B / (A + B)** | **summit A + B** |
| **AB** | **11981969** | **0.48** | **0.49** | **163.7** |
| **AAB** | **5376242** | **0.21** | **0.33** | **255.25** |
| **AABB** | **4741549** | **0.19** | **0.49** | **328.49** |
| **AAAB** | **1187013** | **0.05** | **0.26** | **328.49** |
| **AAAAB** | **1069760** | **0.04** | **0.21** | **420.04** |
| **AAAAABB** | **852228** | **0.03** | **0.33** | **511.59** |

AB are heterozygous k-mer pairs consistent with two haplotypes and diploid coverage

**Table S3. Statistics of all *Paramuricea clavata* assemblies**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Assembly** | **BUSCO statistics** | **Total Length** | **ctgN50** | **scfN50** | **NOTES** |
| pcla1 | C:76.2%[S:71.8%,D:4.4%],F:8.8%,M:15.0% | 760,259,452 | 14,323 | 15,331 | Hybrid assembly with MASURCA\_3.2.6. All PE400 2x250bp reads |
| pcla4 | C:75.8%[S:71.8%,D:4.0%],F:8.7%,M:15.5% | 724,624,315 | 15,858 | 19,717 | Hybrid assembly with MASURCA\_3.2.6. Hard-trimmed PE400 library 2x150bp |
| pcla5 | C:77.0%[S:72.1%,D:4.9%],F:8.9%,M:14.1% | 712,445,758 | 15,851 | 19,721 | Bacterial decontamination |
| pcla6 | C:75.5%[S:71.2%,D:4.3%],F:8.8%,M:15.7% | 712,409,171 | 15,851 | 19,718 | Removal of mitochondrion scaffold |
| pcla7 | C:72.6%[S:68.4%,D:4.2%],F:11.0%,M:16.4% | 707,632,445 | 15,965 | 15,965 | Collapsed pcla6 contigs with pseudohaploid (e.g., Chen et al. 2019) |
| **pcla8** | **C:75.8%[S:73.4%,D:2.4%],F:9.4%,M:14.8%** | **606,969,498** | **19,152** | **23,918** | **Collapsed pcla6 scaffolds with Purge Haplotigs** |

BUSCO database used for all assemblies was *metazoa\_odb9* (n:978 BUSCOs)

**Table S4. GO term enrichment of genes not assigned to orthogroups**

|  |  |  |
| --- | --- | --- |
| GO.ID | Term | classicFisher |
| GO:0051276 | chromosome organization | 4.10E-10 |
| GO:0060249 | anatomical structure homeostasis | 4.80E-10 |
| GO:0000723 | telomere maintenance | 8.40E-10 |
| GO:0032200 | telomere organization | 8.40E-10 |
| GO:0042592 | homeostatic process | 2.00E-09 |
| GO:0006259 | DNA metabolic process | 3.20E-09 |
| GO:0065008 | regulation of biological quality | 1.30E-08 |
| GO:1902589 | single-organism organelle organization | 2.00E-07 |
| GO:0006996 | organelle organization | 1.40E-05 |
| GO:0000712 | resolution of meiotic recombination intermediates | 4.10E-05 |
| GO:0007131 | reciprocal meiotic recombination | 4.10E-05 |
| GO:0035825 | reciprocal DNA recombination | 4.10E-05 |
| GO:0045132 | meiotic chromosome segregation | 4.10E-05 |
| GO:0051307 | meiotic chromosome separation | 4.10E-05 |
| GO:0007126 | meiotic nuclear division | 4.50E-05 |
| GO:0007127 | meiosis I | 4.50E-05 |
| GO:0051321 | meiotic cell cycle | 4.50E-05 |
| GO:1903046 | meiotic cell cycle process | 4.50E-05 |
| GO:0001539 | cilium or flagellum-dependent cell motility | 4.60E-05 |
| GO:0016043 | cellular component organization | 4.60E-05 |
| GO:0006928 | movement of cell or subcellular component | 5.00E-05 |
| GO:0040011 | locomotion | 9.20E-05 |
| GO:0065007 | biological regulation | 0.00012 |
| GO:0006508 | proteolysis | 0.00017 |
| GO:0048584 | positive regulation of response to stimulus | 0.00017 |
| GO:0051304 | chromosome separation | 0.00019 |
| GO:0048870 | cell motility | 0.00024 |
| GO:0051674 | localization of cell | 0.00024 |
| GO:0002376 | immune system process | 0.00041 |
| GO:0048583 | regulation of response to stimulus | 0.00041 |
| GO:0016579 | protein deubiquitination | 0.00056 |
| GO:0070646 | protein modification by small protein removal | 0.00056 |
| GO:0098813 | nuclear chromosome segregation | 0.00096 |
| GO:0051259 | protein oligomerization | 0.00098 |
| GO:0051260 | protein homooligomerization | 0.00098 |
| GO:0007154 | cell communication | 0.00099 |
| GO:0006955 | immune response | 0.0011 |
| GO:0006260 | DNA replication | 0.0013 |
| GO:0002253 | activation of immune response | 0.00156 |
| GO:0050778 | positive regulation of immune response | 0.00156 |
| GO:0050776 | regulation of immune response | 0.00197 |
| GO:0007059 | chromosome segregation | 0.00215 |
| GO:0071840 | cellular component organization or biogenesis | 0.00272 |
| GO:0002682 | regulation of immune system process | 0.0031 |
| GO:0048518 | positive regulation of biological process | 0.00317 |
| GO:0031347 | regulation of defense response | 0.00319 |
| GO:0080134 | regulation of response to stress | 0.00319 |
| GO:0002684 | positive regulation of immune system process | 0.0032 |
| GO:0045087 | innate immune response | 0.00407 |
| GO:0002218 | activation of innate immune response | 0.00424 |

**Table S5. Microsatellite mapping**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Microsatellite | F Primer | R Primer | Annotated Repeat Unit | Resolved Repeat Unit | Scafolds name | Position within protein | Protein name |
| Pcla-09 | Yes | Yes | CA | CA | pcla8\_s015043 |  |  |
| Pcla-10 | Yes | Yes | GT | GT | pcla8\_s000816 | Intron | ATP-sensitive inward rectifier potassium channel 12-like |
| Pcla-12 | Yes | Yes | AC | AC | pcla8\_s000450 | Intron | A disintegrin and metallo ase with thrombospondin motifs 16-like |
| Pcla-14 | Yes | Yes | CA | CA | pcla8\_s000419 |  |  |
| Pcla-17 | Yes | Yes | AC | AC | pcla8\_s004585 | Intron | AT-rich interactive domain-containing 3C-like isoform X2 |
| Par-d | Yes | Yes | (GTT)xGTGTC(GTT)y | (GTT)xGTGTC(GTT)y | pcla8\_s000531 | Repeat in intron; 2 primers in exon | Collagen alpha-4(VI) chain |
| Par-a | Yes | Yes | GTT | GTT | pcla8\_s004379 | Intron | methylmalonyl-CoA epimerase |
| PC3-81 | Yes | Yes | ACA | ACA | pcla8\_s000481 | Repeat in intron; 1 primer in exon | hypothetical protein pdam\_00015477 |
| Pcla-20 | Yes | Yes | TTAT | TA | pcla8\_s001445 | Intron | Protein\_Alignment08754 |
| Pcla-21 | Yes | Yes | ATA | ATA | pcla8\_s006213 | 1 primer in exon | protein LOC113686782, partial |
| Pcla-22 | Yes | Yes | (ATCA)xATCC(ATCA)y | (ATCA)xATCC(ATCA) | pcla8\_s003139 | Intron | protein AWC38\_SpisGene675 |
| Pcla-23 | Yes | Yes | TGC | TGC | pcla8\_s011764 |  | adhesion G-coupled receptor D1-like |
| Pcla-24 | Yes | Yes | AATA | AATA | pcla8\_s000905 |  |  |
| Pcla-25 | Yes | Yes | TAA | TAA | pcla8\_s010731 |  |  |
| Pcla-26 | No | No | TAT | Not in the assembly | | | |
| Pcla-27 | Yes | Yes | TTGA | TTGA | pcla8\_s000419 | Repeat in intron; 1 primer in exon | fibrinogen C domain-containing protein 1-like isoform X1 |
| Pcla-28 | Yes | Yes | ATGT | ATGT | pcla8\_s003787 | Intron | PREDICTED: uncharacterized protein LOC107344584 |
| Pcla-29 | Yes | Yes | TGCG | TGCG | pcla8\_s004991 | Intron | tyramine receptor tyra-2 |

**References**

Chen, L.-Y., R. VanBuren, M. Paris, H. Zhou, X. Zhang *et al.*, 2019 The bracteatus pineapple genome and domestication of clonally propagated crops. *Nature Genetics* 51 (10):1549-1558.

Marçais, G., and C. Kingsford, 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27 (6):764-770.

Ranallo-Benavidez, T.R., K.S. Jaron, and M.C. Schatz, 2020 GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11 (1):1432.

Roach, M.J., S.A. Schmidt, and A.R. Borneman, 2018 Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19 (1):460.