

## RESEARCH ARTICLE

# Optimization of the genotyping-by-sequencing SNP calling for diversity analysis in cape gooseberry (*Physalis peruviana* L.) and related taxa

Felix E. Enciso-Rodríguez, Jaime A. Osorio-Guarín , Gina A. Garzón-Martínez, Paola Delgadillo-Duran, Luz Stella Barrero \*

Centro de Investigación Tibaitatá, Corporación Colombiana de Investigación Agropecuaria–Agrosavia, Mosquera, Cundinamarca, Colombia

\* [lbarrero@agrosavia.co](mailto:lbarrero@agrosavia.co)



## OPEN ACCESS

**Citation:** Enciso-Rodríguez FE, Osorio-Guarín JA, Garzón-Martínez GA, Delgadillo-Duran P, Barrero LS (2020) Optimization of the genotyping-by-sequencing SNP calling for diversity analysis in cape gooseberry (*Physalis peruviana* L.) and related taxa. PLoS ONE 15(8): e0238383. <https://doi.org/10.1371/journal.pone.0238383>

**Editor:** Shailendra Goel, University of Delhi, INDIA

**Received:** April 23, 2020

**Accepted:** August 14, 2020

**Published:** August 26, 2020

**Copyright:** © 2020 Enciso-Rodríguez et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All the supporting data are included as additional files (Additional file 1: [S1 Table](#), [S2 Table](#) and [S3 Table](#)). Raw data of the cape gooseberry transcriptome is available at NCBI Bioproject ID No. PRJNA67621.

**Funding:** This research was funded by the Colombian Ministry of Agriculture Contract No. 054/08072-2008 and there was no additional external funding received for this study.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

A robust Genotyping-By-Sequencing (GBS) pipeline platform was examined to provide accurate discovery of Single Nucleotide Polymorphisms (SNPs) in a cape gooseberry (*Physalis peruviana* L.) and related taxa germplasm collection. A total of 176 accessions representing, wild, weedy, and commercial cultivars as well as related taxa from the Colombian germplasm bank and other world repositories were screened using GBS. The pipeline parameters mnLCov of 0.5 and a mnScov of 0.7, tomato and potato genomes, and cape gooseberry transcriptome for read alignments, were selected to better assess diversity and population structure in cape gooseberry and related taxa. A total of 7,425 SNPs, derived from *P. peruviana* common tags (unique 64 bp sequences shared between selected species), were used. Within *P. peruviana*, five subpopulations with a high genetic diversity and allele fixation ( $H_E$ : 0.35 to 0.36 and  $F_{IS}$ : -0.11 to -0.01, respectively) were detected. Conversely, low genetic differentiation ( $F_{ST}$ : 0.01 to 0.05) was also observed, indicating a high gene flow among subpopulations. These results contribute to the establishment of adequate conservation and breeding strategies for Cape gooseberry and closely related *Physalis* species.

## Introduction

Cape gooseberry (*Physalis peruviana* L.) is an herbaceous Solanaceae species, native to the Andean region, with enormous potential for biomedical research and commercial purposes [1]. The mating system of cape gooseberry is mostly outcrossed that, along with the occurrence of mixed-ploidy, indicates a transitioning state from wild to cultivated [2, 3], possibly due to a lack of a domestication process as has occurred in major crops of the Solanaceae family such as tomato or potato [4, 5]. Globally, this berry-bearing species is known for its nutritional value, possessing high contents of vitamins and minerals, as well as anti-inflammatory, antioxidant and disease control (diabetes and hypertension) properties [6–10]. These characteristics

have enabled its distribution as a crop to other parts of the world, including Africa, Asia, and Oceania [11, 12]. Colombia is the world's leading producer of this exotic fruit, with roughly 16,109 metric tons produced in 2018 [13], followed by Ecuador, Zimbabwe, Malaysia, China, Kenya, and South Africa [14].

Cape gooseberry production in the Andean region has been diminished due to important phytosanitary problems caused mainly by bacterial (*Xanthomonas* sp., and *Ralstonia solanacearum*), Oomycete (*Phytophthora* sp.), and fungal pathogens (*Alternaria* sp., *Cladosporium* sp., *Cercospora* sp., *Phoma* sp., *Sclerotinia sclerotiorum*, *Fusarium oxysporum*, among others) [14, 15]. Among them, *F. oxysporum* sp. *physali* (*Foph*) [16], represents one of the most damaging diseases. Therefore, leveraging favorable allelic combinations from wild, weedy, landraces and related taxa into cultivated cape gooseberry populations will contribute to reducing production losses and environmental impacts; since, to date, the only way to manage these pathogens is through chemical control.

The available genetic diversity of cultivated and wild related taxa has helped to establish appropriate conservation, management, and sustainable utilization strategies of different crops [17]. In particular, for cape gooseberry and related taxa, the proper characterization of available germplasm collections will contribute to the identification of resistant or tolerant sources to biotic and abiotic stresses. Not only will this help with battling pathogens, but will increasingly assist with future challenges due to climate change [18].

Molecular markers have become valuable tools to assess genetic variation of worldwide plant repositories. Population studies, marker assisted selection, mapping and association studies, among other applications, have been used for this purpose [19]. To date, some studies have been conducted to characterize the genetic diversity of cape gooseberry. Random amplified microsatellites (RAM) markers have been used by two independent studies on panels of 43 and 18 accessions respectively. They detected a low differentiation and high heterozygosity levels among the majority of the accessions [20, 21]. Other studies used more informative markers, such as simple sequence repeats (SSRs), conserved ortholog sequences II (COSII), immunity related genes (IRGs) and single nucleotide polymorphisms (SNPs) to analyze the diversity levels and population structure in natural and breeding *P. peruviana* populations [22–26]. However, these low-throughput platforms provide a limited ability to estimate the extent of cape gooseberry genetic variability, since they focus on limited genomic regions.

Recently, next generation sequencing (NGS) technologies have accelerated the screening of germplasm collections, identifying thousands of SNP markers in a cost effective and timely way [27, 28]. Notably, genotyping-by-sequencing (GBS), a highly multiplexed method based on the reduction of genome complexity through methylation-sensitive restriction enzymes [29], has become a popular approach for detecting genome-wide variation in plants [30]. Previous studies in cape gooseberry have identified about 50,000 SNPs using GBS in 100 accessions from the Colombian germplasm collection, for association studies related to fruit quality and *Foph* resistance [31, 32].

Although these previous advances in genetic diversity and association mapping have contributed to the exploration of cape gooseberry genetic resources, the species still lacks a more comprehensive GBS-SNP pipeline platform for SNP discovery. This study mainly seeks to provide a robust GBS-SNP calling pipeline for this species and related taxa by obtaining common genomic regions between tomato, potato, and cape gooseberry using the previously developed TasselPipelineGBS [33]. Moreover, it aims to provide an extensive study of a larger germplasm collection using GBS, which comprises 158 accessions of *P. peruviana* and 18 wild related species from the *Physalis* genus, including 95 technical replicates, to leverage the genetic diversity and population structure of this Andean crop towards conservation and sustainable utilization strategies.

## Materials and methods

### Plant material and DNA isolation

One hundred and fifty-eight cape gooseberry (*Physalis peruviana* L.) and 18 related taxa accessions, containing one to seven individuals each, for a total of 644 individuals, were used in this study (Table 1). This germplasm collection is maintained by the Colombian Corporation for Agricultural Research (AGROSAVIA). These accessions, collected mainly across the Colombian Andean mountains, were selected based largely on geographic distribution and state of cultivation (S1 Table).

The genomic DNA was isolated from young leaves using the DNeasy Plant Mini Kit (QIAGEN, Germany) following the manufacturer's procedure. DNA quantity was determined using NanoDrop (Thermo Scientific® ND 2000) and  $\lambda$  DNA/HindIII Ladder (Promega, Madison, USA). DNA quality was assessed using EcoRI restriction enzyme digestions (New England Biolabs, Beverly, MA) and visualized on 1% (w/v) agarose gels stained with ethidium bromide (0.5  $\mu$ g/mL).

### Genotyping and read alignment

GBS library generation and Illumina sequencing were conducted at the Institute for Genomic Diversity (IGD) from Cornell University (Ithaca, New York, USA). An additional 95 individuals previously sequenced [31] were included as technical replicates in this study (S1 Table). FASTQ files containing 739 individuals (644 plus 95 technical replicates) were processed using the GBS pipeline implemented on Tassel standalone V4.3.5 [34]. Since cape gooseberry does not have a reference genome, we used the closely related sequenced species tomato (*Solanum lycopersicum*) and potato (*Solanum tuberosum*) [35, 36] as well as the cape gooseberry leaf transcriptome (SRA: SRP005904) [37] for reads alignment and later SNP discovery.

### Parameter selection for SNP calling

Different GBS parameters in the DiscoverySNPCaller and GBSHapMapFilters plugins were used for SNP calling and hapmap filtering, respectively (Table 2), using a minimum read depth of 5. Additionally, homologue genomic regions among tomato and potato genomes, and cape gooseberry transcriptome were used. This was done to avoid possible bias caused by copy number variations and ploidy complexity during SNP calling. This approach was performed by selecting common tags (a unique sequence of 64 bp in length, excluding the barcode and shared among the selected species) using the TagsOnPhysicalMap (TOPM) file. Finally, to further reduce marker redundancy, the high linkage disequilibrium filter was implemented.

**Table 1. Cape gooseberry and related taxa accessions used in this study.**

Species	Number of accessions	Number of individuals
<i>Physalis peruviana</i>	158	587
<i>Physalis angulata</i>	2	9
<i>Physalis floridana</i>	2	8
<i>Physalis ixocarpa</i>	1	2
<i>Physalis philadelphica</i>	10	34
<i>Physalis pruinosa</i>	1	1
<i>Physalis viscosa</i>	1	1
<i>Physalis</i> sp.	1	2
<b>Total</b>	<b>176</b>	<b>644</b>

<https://doi.org/10.1371/journal.pone.0238383.t001>

Table 2. GBS parameters used for SNP calling and Hapmap filtering in this study.

Plugin*	Parameter	Parameter Abbreviation	Threshold Values	Description
DiscoverySNPCaller	Minimum locus coverage	mnLCov	0.1, 0.5, and 0.7	Uses the proportion of individuals with at least one tag present from the TagLocus covering a SNP
	Minimum minor allele count	mnMAC	20	Selects SNPs that pass the specified mnMAC
	Average sequencing error per base	errRate	0.05	Decides between heterozygous and homozygous calls
	Minimum minor allele frequency	mnMAF	0.05	Selects SNPs that pass the specified mnMAF
GBSHapMapFilters	Minimum site coverage	mnScov	0.1, 0.5, and 0.7	Uses the minimum taxon call rate for a SNP to be included in the output where taxon call rate is the proportion of the taxa with individuals that are not missing for that SNP
	Minimum taxon coverage	mnTCov	0.5	Uses the minimum taxon call rate for a SNP to be included in the output where call rate is the proportion of the SNP individuals for a taxon that are not missing.

\* Information gathered from: <https://bytebucket.org/tasseladmin/tassel-5-source/wiki/docs/TasselPipelineGBS.pdf>

<https://doi.org/10.1371/journal.pone.0238383.t002>

## Cluster and principal component analyses

The clustering patterns of the 739 individuals were compared through genetic distances generated by Tassel V4.3.5 in order to verify which parameters were correctly assessed. The parameter selection was based on genetic trees in which most of the technical replicates clustered together with their corresponding counterpart and those which better grouped the accessions according to their passport data (S1 Table). A total of 54 different genetic trees were generated based on the neighbor-joining (NJ) method [38], using filtered SNPs derived from datasets with or without common tags. From the selected SNP data set, a principal component analysis (PCA) was performed, in which each SNP marker was scored with 0 for the homozygous allele aa, 1 for the heterozygous allele Aa and 2 for the homozygous allele AA, based on the allele counting for each marker. The newly generated matrices were subsequently analyzed using the gdsfint and SNPRelate packages implemented on the statistical software R [39].

## Population structure and genetic diversity analyses

The population structure and genetic diversity analyses were carried out only for *P. peruviana* accessions using the best GBS parameter combinations selected from the NJ analysis. The software ADMIXTURE V.1.23 [40], was used to estimate the individual's genetic ancestry, calculating the optimum subpopulation (K) number, by the use of cross-validation tests ranging from K = 1 to K = 10.

Genetic diversity, referred to as expected heterozygosity ( $H_E$ ), was calculated as  $1 - \sum_{i=1}^k p_i^2$ , where  $p_i$  is the frequency of the  $i^{th}$  allele for  $k$  alleles. Observed heterozygosity ( $H_O$ ) and  $H_E$  for each locus were estimated using Genepop version 4.7.5 [41]. Finally, to estimate the overall genetic divergence among subpopulations within the cape gooseberry germplasm collection, the genetic differentiation ( $F_{ST}$ ) and fixation ( $F_{IS}$ ) indices were calculated using the software mentioned above.

## Results

### SNP calling and dataset selection

Two major categories were used for SNP calling, using independent assemblies with the tomato and potato genomes, and the cape gooseberry transcriptome, based on whether or not

**Table 3. SNPs identified for cape gooseberry and related taxa using different parameters in the standalone script of Tassel.** Two reference genomes (tomato, potato) and a transcriptome (cape gooseberry) were used for SNP calling.

Parameters			Without common tags			With common tags		
			<i>S. lycopersicum</i>	<i>S. tuberosum</i>	<i>P. peruviana</i>	<i>S. lycopersicum</i>	<i>S. tuberosum</i>	<i>P. peruviana</i>
mnLCov 0.1	mnScov 0.1	SNPs	83,792	91,692	52,179	16,986	16,552	15,772
		Miss <sup>†</sup>	48.17	48.27	44.31	42.81	43.20	41.65
		Het <sup>§</sup>	16.13	16.20	17.03	17.70	18.01	18.51
	mnScov 0.5	SNPs	40,101	43,411	29,339	9,637	9,370	9,393
		Miss	18.35	18.11	18.29	16.90	16.88	16.86
		Het	32.06	32.42	29.26	29.84	30.94	30.20
	mnScov 0.7	SNPs	28,969	31,765	21,994	7,632	7,324	7,459
		Miss	9.91	9.96	11.14	10.67	10.49	10.78
		Het	40.29	40.31	35.53	34.89	36.55	35.20
mnLCov0.5	mnScov 0.1	SNPs	40,567	44,304	30,649	9,666	9,649	9,591
		Miss	21.05	21.11	21.20	19.60	19.63	19.56
		Het	26.14	26.21	26.36	28.27	27.92	28.57
	mnScov 0.5	SNPs	39,581	43,117	29,666	9,372	9,369	9,312
		Miss	19.25	19.23	19.01	17.21	17.40	17.28
		Het	26.85	26.92	27.28	29.28	28.81	29.48
	mnScov 0.7	SNPs	29,257	32,068	22,302	7,573	7,431	7,425 <sup>§</sup>
		Miss	12.46	12.63	12.63	11.97	11.98	12.08
		Het	32.45	32.40	32.61	33.38	33.39	33.91
mnLCov0.7	mnScov 0.1	SNPs	8,952	9,213	6,549	2,397	2,475	2,443
		Miss	7.22	7.62	8.47	8.55	8.7	8.20
		Het	44.05	43.42	42.51	41.90	41.71	42.41
	mnScov 0.5	SNPs	8,695	8,921	6,238	2,318	2,365	2,361
		Miss	4.7	4.83	4.73	5.03	5.17	4.75
		Het	45.48	44.92	44.65	43.88	43.73	44.23
	mnScov 0.7	SNPs	8,602	8,809	6,189	2,285	2,328	2,331
		Miss	4.36	4.39	4.47	4.63	4.74	4.45
		Het	45.57	45.08	44.72	44	43.90	44.27

<sup>†</sup> Percentage of missing data.

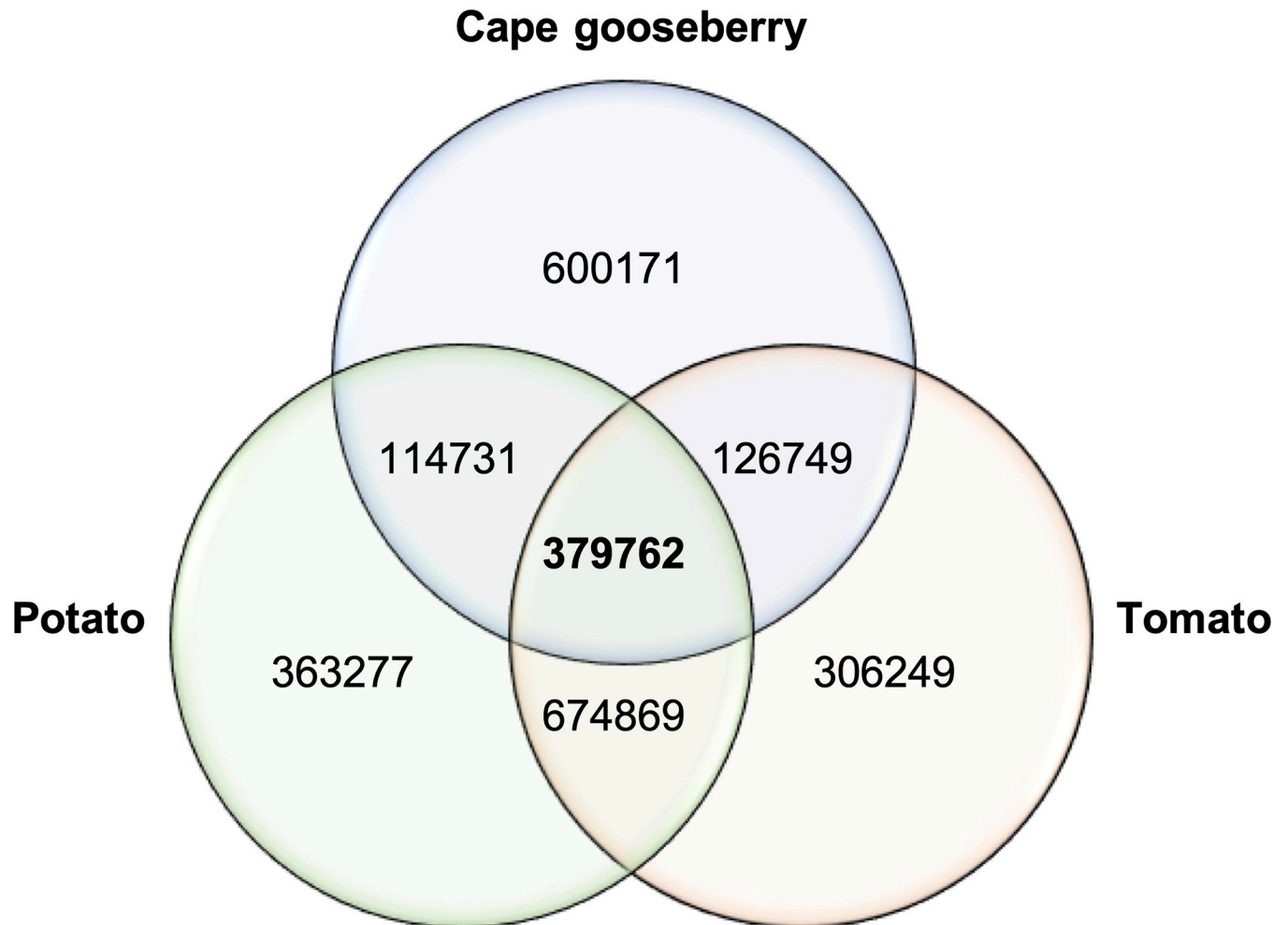
<sup>§</sup> Percentage of heterozygosity.

<sup>§</sup> SNP number selected for subsequent analysis, according to the best parameter combination.

<https://doi.org/10.1371/journal.pone.0238383.t003>

common tags were used. For those SNP calls derived without common tags, around 9 million tags were generated. Between 91,692 and 6,189 SNPs within each reference genome/transcriptome were identified (Table 3) with an average depth of 6.3. As expected, with less stringent parameters (mnLCov = 0.1, and mnScov = 0.1), the percentage of missing data was higher, ranging from 44.31 to 48.27%. With more stringent parameters (mnLCov = 0.7, and mnScov = 0.7), the percentage of missing data was reduced by a factor of 10 (4.36 to 4.47%, Table 3). Conversely, the number of heterozygous SNPs was lower, ranging from 16.13% (low parameter values) to 45.57% (high parameter values, Table 3).

With common tags, low tag numbers were obtained between tomato, potato and cape gooseberry (379,762), after the Illumina reads alignment (Fig 1). These common tags were later used for SNP calling using the GBS pipeline with the parameters described herein. The SNPs identified with common tags, ranged from 2,285 to 16,986, for high and low parameter values respectively. The percentage of missing data ranged between 4.45 and 43.2%, and an



**Fig 1. Venn diagram showing tags in common between the tomato, potato and cape gooseberry obtained after read alignment using the GBS pipeline. A common tag refers to a 64 bp read-length that is shared among the three selected species.**

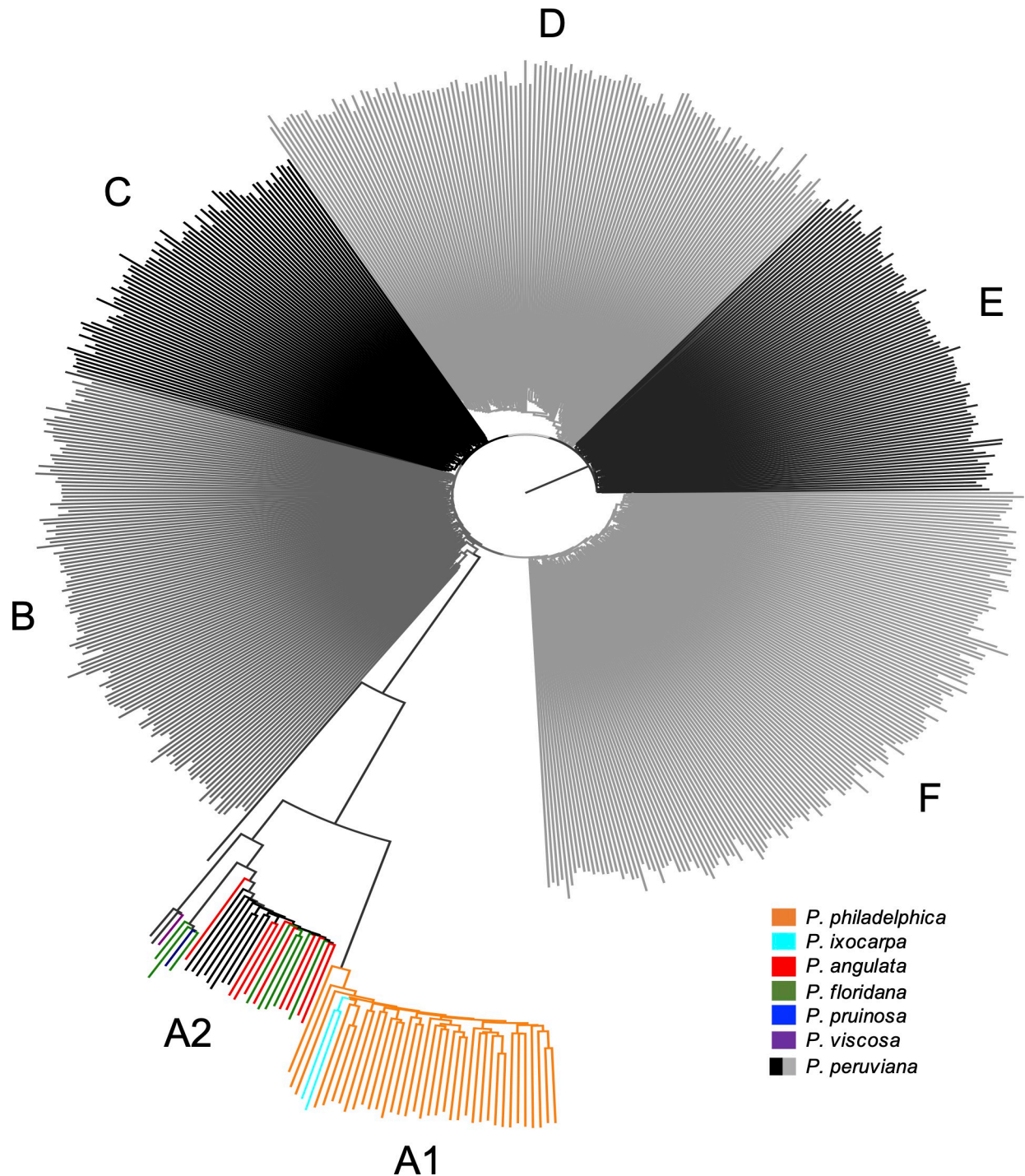
<https://doi.org/10.1371/journal.pone.0238383.g001>

observed heterozygosity between 17.7 and 44.27% (Table 3), for high and low parameter values, respectively.

### Cluster and PCA analyses for cape gooseberry and related taxa

A total of 54 genetic trees were generated, corresponding to each parameter and whether or not common tags were evaluated (Table 3), using the 95 technical replicates and their counterparts within the initial 644 selected individuals. The parameter arrangement with a mnLCov of 0.5 and a mnScov of 0.7 showed the closest genetic distance of all the trees generated, with approximately 70% of the technical replicates grouping in the same cluster as their counterparts. From this parameter combination, a total of 7,425 high quality SNPs (resulting from common tags for *P. peruviana* transcriptome) were selected to evaluate the clustering pattern in cape gooseberry and related taxa.

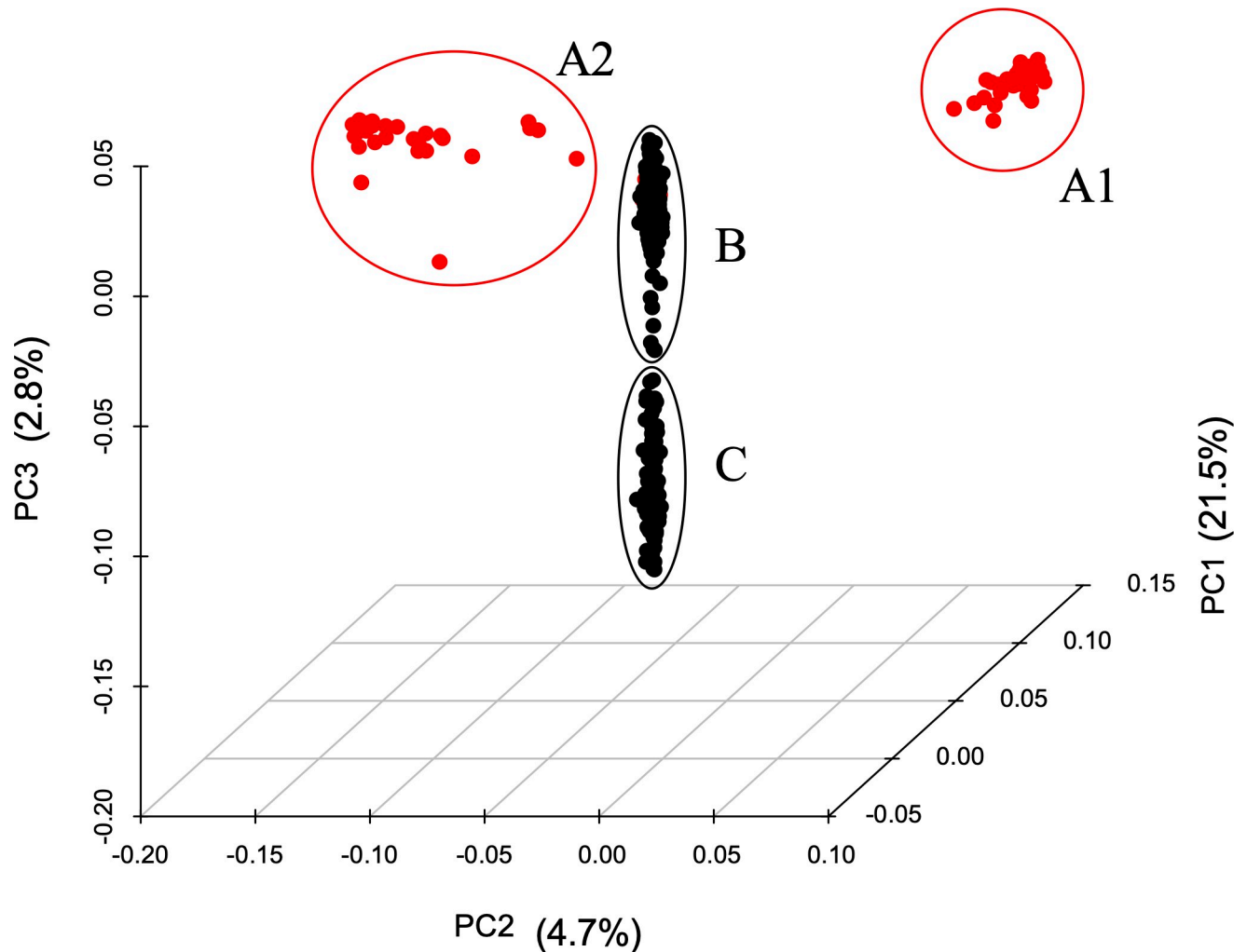
The best NJ tree includes six clusters in which clear patterns emerged for the population under study (Fig 2). Cluster A includes two sub-groups comprised of all related taxa. Specifically, group A1 harbors all *P. philadelphica* accessions and one accession from *P. ixocarpa*, and A2 harbors two accessions from both *P. angulata*, and *P. floridana*, and one accession from



**Fig 2. Neighbor-joining (NJ) tree for cape gooseberry and related taxa generated using 7,425 SNPs derived from cape gooseberry common tags.** Groups A1 and A2 represent related taxa clusters, and B-F represent five sub-populations within the *P. peruviana* population. Each branch represents an individual plant.

<https://doi.org/10.1371/journal.pone.0238383.g002>

both *P. pruinosa* and *P. viscosa*. We also observed three *P. peruviana* accessions (09U207, 09U289 and 09U291) in group A2. Furthermore, most cape gooseberry accessions (including cultivated, weedy, and wild), clustered in five different groups. Clusters B, C, D and E



**Fig 3. Population structure of cape gooseberry and related taxa as revealed by a Principal Component Analysis (PCA).** The first three components uncovered four clusters. Most *P. peruviana* accessions grouped in two clusters at the main centroid (B and C, black dots), and the related taxa were located in two separate clusters (A1 and A2, red dots).

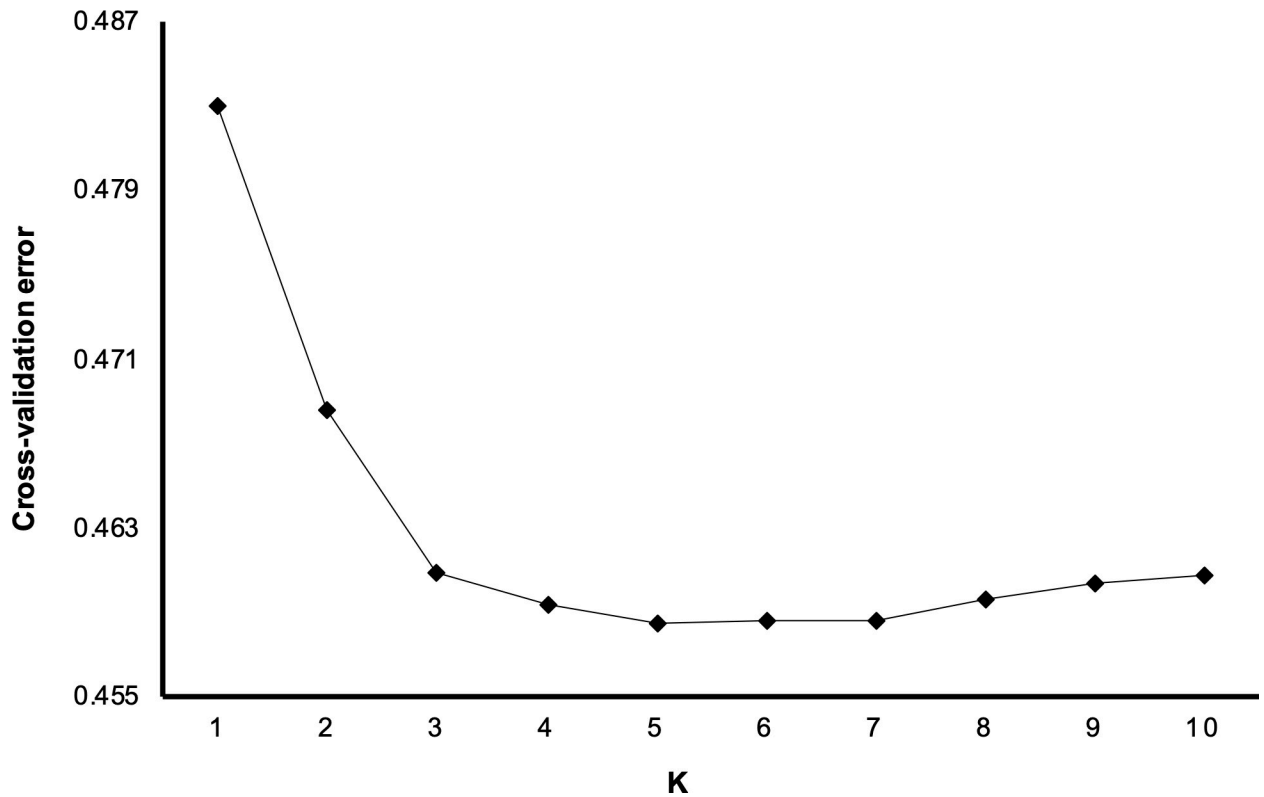
<https://doi.org/10.1371/journal.pone.0238383.g003>

concentrated most accessions from the cape gooseberry producing regions, while cluster F harbors a high number of wild accessions from the Antioquia and Nariño departments of Colombia.

Moreover, the PCA revealed that the first three principal components explained 29% of the total variance, in which the first component contributed for almost all the variance observed, with 21.5%, followed by the second and third component with 4.7% and 2.8%, respectively (Fig 3).

In Fig 3, each cluster represents a distinctive cape gooseberry and related taxa grouping pattern. Clusters B and C include most of the cultivated accessions coming from the main cape gooseberry producing regions in Colombia (Boyacá, Cundinamarca, Antioquia and Nariño departments). In contrast, clusters A1 and A2 contain mostly related taxa. Specifically, group A1 harbors all *P. philadelphica* accessions and one accession from *P. ixocarpa*, while Group A2 includes *P. angulata*, *P. floridana*, *P. pruinosa* and *P. viscosa* as well as three *P. peruviana* accessions (09U207, 09U289 and 09U291, S1 Table) as observed in the NJ tree.



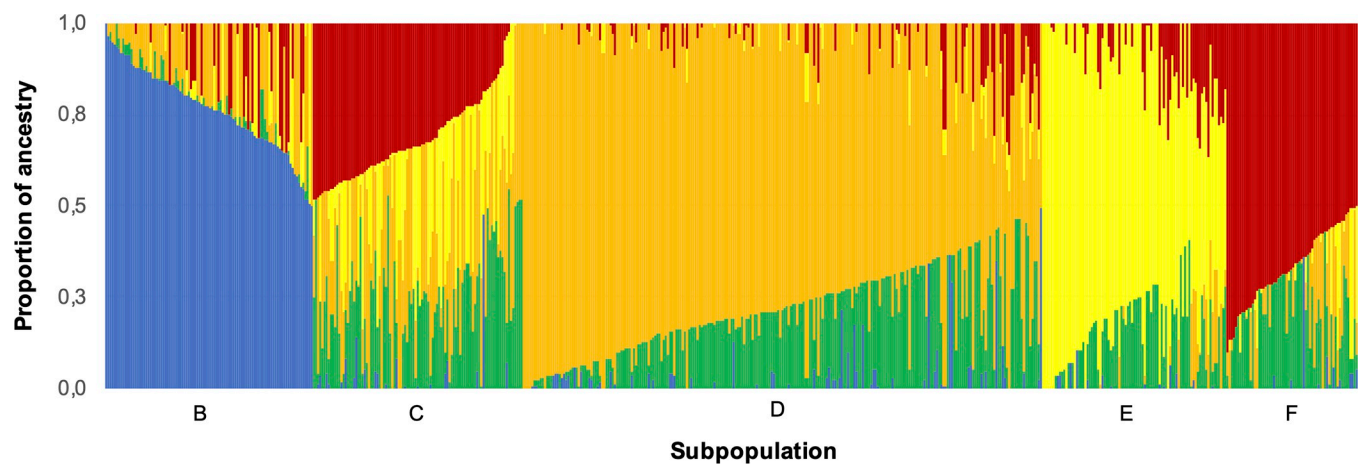


**Fig 4. Cross validation plot for subpopulation (K) estimation in cape gooseberry.** The smallest cross validation error value was observed when  $K = 5$ .

<https://doi.org/10.1371/journal.pone.0238383.g004>

### Population structure and genetic diversity in cape gooseberry

To further analyze the population structure and genetic diversity within *P. peruviana*, the related taxa and *P. peruviana* accessions clustering in sub-groups A1 and A2 in the NJ and PCA analyses were removed. The cross-validation procedure implemented in ADMIXTURE enabled the identification of five subpopulations within *P. peruviana* ( $K = 5$ , Fig 4), in accordance to the number of clusters shown in the NJ tree for this species (Fig 2). In Fig 5,



**Fig 5. Population stratification of *P. peruviana* as revealed by ADMIXTURE.** Five subpopulations were classified based on ancestry values.

<https://doi.org/10.1371/journal.pone.0238383.g005>

**Table 4. Average observed heterozygosity ( $H_O$ ), expected heterozygosity ( $H_E$ ) and inbreeding coefficient ( $F_{IS}$ ) per five *P. peruviana* subpopulations previously identified by ADMIXTURE.**

Subpopulation	$H_O$	$H_E$	$F_{IS}$
B	0.37	0.36	-0.01
	(0.25)*	(0.18)	(0.37)
C	0.38	0.35	-0.02
	(0.24)	(0.18)	(0.29)
D	0.41	0.35	-0.11
	(0.26)	(0.19)	(0.29)
E	0.39	0.35	-0.07
	(0.25)	(0.19)	(0.30)
F	0.38	0.35	-0.04
	(0.24)	(0.19)	(0.30)

\*In parenthesis, standard deviation.

<https://doi.org/10.1371/journal.pone.0238383.t004>

subpopulations B and D include accessions mostly from the Colombian departments with the largest cape gooseberry production (Cundinamarca and Boyacá), while group E includes most accessions from southern Colombia (Nariño and Valle) and international repositories (Ecuador, New Zealand, India and Nepal). Groups C and F combine accessions from all over Colombia, with slightly high accession numbers coming from Boyacá and Nariño in group C. Regarding the state of cultivation, groups B and C gather mostly cultivated accessions; group C contains a moderate number of weedy accessions. However, there were a high number of accessions with neither geographical origin (subpopulation C and F) nor state of cultivation (all subpopulations) information, preventing the full categorization of each subpopulation according to their passport data.

The genetic diversity estimated here as the average  $H_E$  per locus was relatively high within *P. peruviana* subpopulations, ranging from 0.35 to 0.36 (Table 4). In contrast, the  $F_{IS}$  values were low, ranging from -0.11 to -0.01 (Table 4). These results indicate an excess of heterozygosity within *P. peruviana* in agreement with the mixed mating system characteristic of the species. The  $H_O$  values ranged from 0.37 to 0.41 (Table 4). These values did not differ substantially from the  $H_E$  values reported, suggesting that *P. peruviana* could be close to a Hardy-Weinberg equilibrium due to the lack of distinct groups and high allele sharing.

These results were further supported by the genetic differentiation between the five subpopulations detected by ADMIXTURE. The  $F_{ST}$  correlation values among subpopulations uncovered a low genetic differentiation within *P. peruviana*, with  $F_{ST}$  values ranging from 0.01 to 0.05 (Table 5), revealing a high gene flow within the species. Similar diversity and population

**Table 5. Pairwise  $F_{ST}$  estimates among the five cape gooseberry subpopulations previously identified by admixture.**

Subpopulation*	B	C	D	E	F
B	-	0.04	0.04	0.05	0.04
C	0.11	-	0.01	0.01	0.01
D	0.13	0.05	-	0.02	0.02
E	0.13	0.04	0.09	-	0.01
F	0.12	0.02	0.08	0.06	-

\*Above diagonal: Average  $F_{ST}$  values per locus. Below diagonal: Standard deviation.

<https://doi.org/10.1371/journal.pone.0238383.t005>

differentiation results were also observed when the *P. peruviana* was grouped according to geographic origin (S2 and S3 Tables).

## Discussion

Molecular markers have allowed the detailed study of germplasm collections, revealing the history of crop domestication, discovering novel genetic diversity [42], and improving the efficiency of conventional plant breeding schemes through marker-assisted and genomic selection [43–45]. Markers such as COSIs, IRGs, and SSRs have been used for genetic diversity studies in Colombian cape gooseberry populations, due to their highly polymorphic nature [22, 46]. However, the use of SNP markers is now more common to characterize the genetic diversity of germplasm collections, considering its abundance, reproducibility, discriminative power, and cost-effectiveness [47]. In this study, a standardized GBS pipeline was chosen to identify a new set of SNP markers, in order to evaluate a large Colombian germplasm collection of cape gooseberry.

GBS represents an innovative method for large scale SNP detection and genotyping of genetic resources [29]. In an orphan species such as cape gooseberry, molecular markers derived from high throughput sequencing technologies like GBS have allowed, in a cost-effective and time-efficient way, access to its genetic diversity and population structure [31, 32]. It has also permitted the identification of genes related to the resistance response to *Foph* [31] and quality-related traits [32] through genome-wide association studies (GWAS) in a small *P. peruviana* germplasm collection.

### Intermediate call SNP rates, technical replicates and passport data allowed for the selection of SNP datasets

In NGS and GBS studies, one of the most important applications is the ability to accurately and comprehensively identify genetic variation. Obtaining unbiased results usually requires a complex multi-step processing pipeline that includes pre-processing, read alignment, and variant calling. Each of these steps uses its own set of modifiable parameters, creating a significant amount of possible distinct pipelines, which vary significantly in the resulting called variants [48]. Therefore, evaluating the performance of calling methods is not straightforward and particular metrics and data sets can introduce bias into the performance test. Although previous studies in cape gooseberry obtained a high number of markers [31, 32], there is a need to optimize the SNP calling pipeline by increasing population size, as well as ensuring technical replicates to test for appropriate parameter algorithms. Here, we proposed the use of the available *P. peruviana* transcriptome assembly [49], for reads alignment, due to the absence of a high-quality genome for this species. Furthermore, we leveraged the common tags (64 bp reads) among well annotated genomes from the *Solanaceae* genus (*S. lycopersicum* and *S. tuberosum*) and *P. peruviana* to reduce the bias caused by the large amount of repetitive sequences, structural variations, and complex polyploid genomes present in plants [50]. In particular, cape gooseberry represents a mixed-ploidy genome as has been shown by classic cytogenetic analyses [2, 51]. In addition, the use of intermediate stringent parameters (mnLCov: 0.5 and mnScov: 0.7) was useful to decrease the missing data and the underestimation of diversity because of the presence-absence variations [52, 53]. As a consequence, the selected intermediate strict parameters had a reliable effect on population clustering, as was verified by the grouping of the 95 technical replicates with high genetic distance congruence (70%).

These parameters were selected based upon their ability to call SNPs in terms of read and taxon coverage, using additional values below and above to those reported in previous studies for cape gooseberry [31, 32]. Similarly, given the nature of the germplasm collection, in which

related taxa were also included, the use of the aforementioned parameters and stringent mnMAF value (5%), could contribute to reduced false SNP calling and error rates, as well as homozygosity overestimation.

Moreover, the use of reference genomes to identify additional sources of diversity that went undetected when using a single reference, as has been previously used in similar species like tomato [54], was an advantage in this study. Considering the lack of reference genome in cape gooseberry, the strategy of combining related reference genomes and the use of common tags, based on the well-known synteny between the Solanaceae species (Table 1) [55], has been valuable.

### The standardized GBS pipeline allowed to differentiate the *Physalis* species in the germplasm under study

The NJ tree and PCA analyses enabled the identification of two main groups, containing *P. peruviana* and related taxa, as reported in previous studies [22, 31]. However, unlike those studies, the NJ tree and PCA allowed the differentiation of the related taxa into two subgroups, separating *P. philadelphica* and *P. ixocarpa* accessions from the others (Figs 2 and 3).

The *Physalis* genus, a member of the plant family Solanaceae, includes more than 90 commercial and ornamental species with high morphological diversity [56]. For this reason, different morphological and molecular studies have been carried out in order to resolve the relationships between the *Physalis* species. In concordance with the present study, Hu *et al.* [57] reported a closer phylogenetic relationship between *P. floridana* and *P. peruviana* than between *P. peruviana* with *P. philadelphica* using chloroplast markers. Similarly, a recent study that used complete chloroplast genomes conducted by Fen *et al.* [58] identified a close similarity between *P. angulata* and *P. peruviana* genomes. Both results were found in the NJ and PCA analyses reported in this study (Figs 2 and 3). Finally, the results of this study are further supported by the research conducted by Beest *et al.* [59]. Based on 22 morphological traits, it was found that *P. philadelphica* was close to *P. ixocarpa*, clustering apart from *P. pruinose* and *P. viscosa*, which were grouped with *P. peruviana*.

In particular, *P. peruviana* and *P. philadelphica* are considered the species with the most significant advances in their cultivation within the *Physalis* genus [26, 60]. However, the former has a center of diversification in the Andean Mountains of South America [61, 62], while the latter has a center of diversification located in North America and Central America [56, 60, 63], which is supported by this study where the two species were separated into different clusters. Overall, the results of this study demonstrated the usefulness of the common tags-derived SNPs to assess the genetic relationships in the *Physalis* genus.

The PCA revealed a grouping pattern in which most cape gooseberry cultivated accessions clustered together. However, the NJ tree uncovered additional subgroups (subpopulations B-F) within *P. peruviana*. Three *P. peruviana* accessions grouped with the related taxa subgroup A2, in both PCA and NJ analyses. These accessions are from the Nariño (09U207) and Tolima (09U291) departments of Colombia, and Poland (09U289). In another study based on orthologous genes Wei *et al.* [23], found that three *P. peruviana* accessions cluster together with related taxa (such as *P. philadelphica* and *P. angulata*). One of these accessions corresponds to the 09U289 (PI28570597GI) from Poland, supporting the grouping pattern obtained in this study. Additionally, one possible explanation of finding *P. peruviana* accessions grouping together with related taxa could be associated to misclassifications within the cape gooseberry germplasm collection as previously reported [22, 31]. For this reason, the above-mentioned accessions were removed for the *P. peruviana* population structure and diversity analyses.

## Cape gooseberry exhibits high genetic diversity with low population differentiation

The population structure analyses using ADMIXTURE did not show a clear separation among *P. peruviana* accessions neither by geographical origin nor by state of cultivation, as observed in other studies [22, 31, 32]. Likewise, the  $H_E$  values indicates a high diversity within this species, possibly as the sum of multiple factors including mixed ploidy, heterozygosity, mating system, and marker informativeness. The  $H_E$  values in this study agree with those found for the same species using SNP, SSR and COSII markers [23, 26]. However, depending on the molecular marker nature, marker number, and population size, different genetic diversity values have been reported for *P. peruviana* germplasm collections. For instance, low  $H_E$  values (0.22–0.25) have been uncovered in a small cape gooseberry population using random amplified polymorphic DNA and SSRs markers [20, 26], in which the dominant nature and population size, respectively, could underestimate the real genetic diversity. Furthermore, the seed movement of cultivated material between the different production regions may contribute to obtaining low  $F_{ST}$  values (0.01 to 0.05, Table 5) in this study, leading consequently to high homogenous but heterozygous cape gooseberry materials across production regions as revealed by the  $F_{IS}$  and  $F_{ST}$  values (Tables 4 and 5) and the  $H_C/H_E$  values. Based on the diversity and genetic differentiation analyses using subpopulations derived from ADMIXTURE as well as geographical origin (S2 and S3 Tables), we obtained very similar findings, regardless of the clustering strategy. These results provide evidence of high allele sharing amongst subpopulations given the similar allele frequencies. Therefore, we inferred that the *P. peruviana* population could be behaving close to HWE, in accordance with its outcrossing nature and the fact that it has not undergone a domestication process [2, 3].

## Implications for cape gooseberry conservation and breeding

Characterizing plant genetic diversity is an important challenge, considering that diversity is a source of novel allele combinations that can be crucial for addressing climate and health challenges, ensuring food security, and improving nutrition of future generations. As a fruit crop, cape gooseberry possesses a remarkable gene pool that could be leveraged for germplasm conservation, human nutrition, and biotechnology applications. In this study, we characterize a large primary *P. peruviana* and related *Physalis* species germplasm collection that potentially represent secondary or tertiary gene pools that could increase the genetic variability needed to address future challenges for cape gooseberry production and value-chain.

As a member of the Solanaceae family, *P. peruviana* was proven in this study to be a representative example of the high diversity present in this family. The NJ and PCA analyses discriminate a germplasm collection with two main groups (related taxa and *P. peruviana*, Figs 2 and 3) and within *P. peruviana* five different subpopulations (Fig 4) with high genetic diversity (Table 4). Similarly, the  $F_{ST}$  pairwise revealed a relatively high gene flow between *P. peruviana* subpopulations regardless of the grouping method (i.e. ancestry or geographical origin, Table 5, S2 and S3 Tables) as has been found in previous studies [22].

Likewise, the results support a common origin in cultivated *P. peruviana*, given the distribution of the accessions within the different subpopulations as revealed by the  $F_{IS}$  values (Table 4). This statement is also supported by the inability to clearly classify these accessions according to their passport data as observed in ADMIXTURE given the similar allele frequencies within the *P. peruviana* subpopulations.

The study also reinforces previous inferences about the transition of cape gooseberry from wild/landrace to a cultivated state by farmer selection [2, 3]. For instance, the Nariño department (with accessions grouped in subpopulation E), considered to be the entry point of cape

gooseberry from its center of origin into Colombia [3], presented a mix of wild and cultivated ancestries. These unique combinations, along with the related species diversity, could contain genetic variability that can be used for breeding to increase production, quality, and tolerance against biotic or abiotic factors.

Moreover, this study has implications for proper conservation and classification of cape gooseberry germplasm banks. Thus, the SNP dataset used in the study allowed the identification of potentially misclassified accessions within the cape gooseberry (i.e., cluster A2 in Figs 2 and 3) [22, 31]. Despite the fact that some accessions have poor or missing passport data, the common tags-derived SNPs were sufficient to capture the genetic variability of the population under study. This assessment will contribute to the establishment of core collections based on the GBS-SNP pipeline approximation, for conserving the species variability and diversity for its safeguard and sustainable use.

## Conclusions

The results of this study provide a comprehensive insight into the genetic diversity and population structure of a relatively large cape gooseberry repository in Colombia. The population structure and genetic diversity of cape gooseberry was assessed employing a standardizing GBS pipeline, using sequenced genomes from closely related species, as well as transcriptomic information from the same species. Through different SNP-calling parameters, technical replicates and passport data, the selected SNP dataset enabled the separation of *P. peruviana* from related taxa accessions using the NJ and PCA grouping methods. High genetic diversity but low subpopulation differentiation was observed for *P. peruviana*. The selection of SNPs derived from common homologue regions between closely annotated related species and cape gooseberry will allow the accurate inference of gene function in future GWAS and genomic selection studies.

## Supporting information

**S1 Table. Summary of the accessions of the *P. peruviana* and related taxa collection used in this study.**

(XLS)

**S2 Table. Average observed heterozygosity ( $H_O$ ), expected heterozygosity ( $H_E$ ) and inbreeding coefficient ( $F_{IS}$ ) per five *P. peruviana* subpopulations grouped according to their geographical regions.**

(XLS)

**S3 Table. Pairwise  $F_{ST}$  estimates among the five *P. peruviana* subpopulations grouped according to their geographical regions.**

(XLS)

## Acknowledgments

The authors acknowledge Edgar Jimenez for his support in the *in vitro* propagation and maintenance of the germplasm collection. The authors thank Ivania Cerón for assistance in revising the final version of the manuscript.

## Author Contributions

**Conceptualization:** Felix E. Enciso-Rodríguez, Jaime A. Osorio-Guarín, Gina A. Garzón-Martínez, Luz Stella Barrero.

**Data curation:** Felix E. Enciso-Rodríguez, Jaime A. Osorio-Guarín.

**Formal analysis:** Felix E. Enciso-Rodríguez, Jaime A. Osorio-Guarín, Gina A. Garzón-Martínez, Luz Stella Barrero.

**Funding acquisition:** Luz Stella Barrero.

**Investigation:** Felix E. Enciso-Rodríguez, Luz Stella Barrero.

**Methodology:** Felix E. Enciso-Rodríguez, Jaime A. Osorio-Guarín, Paola Delgadillo-Duran, Luz Stella Barrero.

**Project administration:** Felix E. Enciso-Rodríguez, Gina A. Garzón-Martínez, Luz Stella Barrero.

**Writing – original draft:** Felix E. Enciso-Rodríguez, Jaime A. Osorio-Guarín, Gina A. Garzón-Martínez, Luz Stella Barrero.

**Writing – review & editing:** Felix E. Enciso-Rodríguez, Jaime A. Osorio-Guarín, Gina A. Garzón-Martínez, Paola Delgadillo-Duran, Luz Stella Barrero.

## References

1. Ramadan MF. Bioactive phytochemicals, nutritional value, and functional properties of cape gooseberry (*Physalis peruviana*): An overview. *Food Res Int*. 2011; 44: 1830–1836. <https://doi.org/10.1016/j.foodres.2010.12.042>
2. Liberato SA, Sánchez-Betancourt E, Argüelles JH, González C, Núñez VM, Barrero LS. Citogenética de genotipos de uchuva, *Physalis peruviana* L., y *Physalis floridana* Rydb., con respuesta diferencial a *Fusarium oxysporum*. *Cienc Tecnol Agropecuaria*. 2015; 15: 51–61. [https://doi.org/0.21930/rcta.vol15\\_num1\\_art:396](https://doi.org/0.21930/rcta.vol15_num1_art:396)
3. Lagos T. Biología reproductiva, citogenética, diversidad genética y heterosis en parentales de uvilla o uchuva *Physalis peruviana* L. Universidad Nacional de Colombia. Palmira, Colombia. 2006.
4. Hardigan MA, Laimbeer FPE, Newton L, Crisovan E, Hamilton JP, Vaillancourt B, et al. Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc Natl Acad Sci U S A*. 2017; 114: E9999 LP–E10008. <https://doi.org/10.1073/pnas.1714380114> PMID: 29087343
5. Giovannoni J. Tomato multiomics reveals consequences of crop domestication and improvement. *Cell*. 2018; 172: 6–8. <https://doi.org/10.1016/j.cell.2017.12.036> PMID: 29328921
6. Wu S-J, Ng L-T, Huang Y-M, Lin D-L, Wang S-S, Huang S-N, et al. Antioxidant activities of *Physalis peruviana*. *Biol Pharm Bull*. 2005; 28: 963–966. <https://doi.org/10.1248/bpb.28.963> PMID: 15930727
7. Deng K, Zang L, Lan X, Zhong Z, Xiong B, Zhang Y, et al. Antioxidant components from cape gooseberry. *J Food Process Pres*. 2016; 40: 893–898. <https://doi.org/10.1111/jffp.12667>
8. Ballesteros-Vivas D, Alvarez-Rivera G, León C, Morantes SJ, Ibáñez E, Parada-Alfonso F, et al. Antiproliferative bioactivity against HT-29 colon cancer cells of a withanolides-rich extract from golden berry (*Physalis peruviana* L.) calyx investigated by Foodomics. *J Funct*. 2019; 63: 103567. <https://doi.org/10.1016/j.jff.2019.103567>
9. Puente L, Nocetti D, Espinosa A. *Physalis peruviana* Linnaeus, an update on its functional properties and beneficial effects in human health. In: Mariod AA, editor. *Wild Fruits: Composition, Nutritional Value and Products*. Cham: Springer International Publishing; 2019. pp. 447–463. [https://doi.org/10.1007/978-3-030-31885-7\\_34](https://doi.org/10.1007/978-3-030-31885-7_34)
10. Rodrigues E, Rockenbach II, Cataneo C, Gonzaga LV, Chaves ES, Fett R. Minerals and essential fatty acids of the exotic fruit *Physalis peruviana* L. *Ciênc Tecnol Aliment*. 2009; 29: 642–645. <https://doi.org/10.1590/S0101-20612009000300029>
11. Menzel MY. The Cytotaxonomy and genetics of *Physalis*. *Proc Am Philos Soc*. 1951; 95: 132–183. Available: [www.jstor.org/stable/3143331](http://www.jstor.org/stable/3143331)
12. National Research Council. Goldenberry (Cape Gooseberry). Popenoe H, King S, León J, Kalinowski L, editors. *Lost crops of the Incas: little-known plants of the Andes with promise for worldwide cultivation*. Washington, DC: National Academy Press; 1989. <https://doi.org/10.17226/1398>
13. Agronet. Reportes estadísticos [Internet]. 2020.

14. Fischer G, Almanza-Merchan PJ, Miranda D. Importancia y cultivo de la uchuva (*Physalis peruviana* L.). Rev Bras Frutic. 2014; 36: 1–15. <https://doi.org/10.1590/0100-2945-441/13>
15. Smith A. Reconocimiento de las enfermedades y plagas en el cultivo de la uchuva. In: Diaz A, Smith A, Zapata J, Mesa P, editors. Avances en el manejo y control de *Fusarium oxysporum* en el cultivo de uchuva (*Physalis peruviana*). Bogota, Colombia: Produmedios; 2012. pp. 9–12.
16. Simbaqueba J, Catanzariti A-M, González C, Jones DA. Evidence for horizontal gene transfer and separation of effector recognition from effector function revealed by analysis of effector genes shared between cape gooseberry and tomato-infecting formae speciales of *Fusarium oxysporum*. Mol Plant Pathol. 2018; 19: 2302–2318. <https://doi.org/10.1111/mpp.12700> PMID: 29786161
17. Odong TL, van Heerwaarden J, Jansen J, van Hintum TJL, van Eeuwijk F a. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? Theor Appl Genet. 2011; 123: 195–205. <https://doi.org/10.1007/s00122-011-1576-x> PMID: 21472410
18. Henry RJ. Genomics strategies for germplasm characterization and the development of climate resilient crops. Front Plant Sci. 2014; 5: 68. <https://doi.org/10.3389/fpls.2014.00068> PMID: 24616732
19. Nadeem MA, Nawaz MA, Shahid MQ, Doğan Y, Comertpay G, Yıldız M, et al. DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip. 2018; 32: 261–285. <https://doi.org/10.1080/13102818.2017.1400401>
20. Bonilla M, Piedrahíta K, Posso A, Vásquez H, Muñoz J. Caracterización molecular de 43 accesiones de uchuva de seis departamentos de Colombia. Acta Agron. 2008; 57: 109–116.
21. Morillo AT, Villota DE, Lagos TC, Ordóñez HR. Morphological and molecular characterization of 18 introductions of cape gooseberry *Physalis peruviana* L. collection of the University of Nariño. Rev Fac Nac Agron Medellín. 2011; 64: 6043–6053.
22. Garzón-Martínez GA, Osorio-Guarín JA, Delgadillo-Durán P, Mayorga F, Enciso-Rodríguez FE, Landsman D, et al. Genetic diversity and population structure in *Physalis peruviana* and related taxa based on InDels and SNPs derived from COSII and IRG markers. Plant Gene. 2015; 4: 29–37. <https://doi.org/10.1016/j.plgene.2015.09.003> PMID: 26550601
23. Wei J, Hu X, Yang J, Yang W. Identification of single-Copy orthologous genes between *Physalis* and *Solanum lycopersicum* and analysis of genetic diversity in *Physalis* using molecular markers. PLoS One. 2012; 7: E50164. <https://doi.org/10.1371/journal.pone.0050164> PMID: 23166835
24. Spindel J, Wright M, Chen C, Cobb J, Gage J, Harrington S, et al. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theor Appl Genet. 2013; 126: 2699–2716. <https://doi.org/10.1007/s00122-013-2166-x> PMID: 23918062
25. Berdugo J, Enciso-Rodríguez F, González C, Barrero M. LS. Variabilidad genética de parentales y poblaciones F1 inter e intraespecíficas de *Physalis peruviana* L. y *P. floridana* Rydb. Rev Bras Frutic. 2015; 37: 179–192. <https://doi.org/10.1590/0100-2945-002/14>
26. Chacón MI, Sánchez Y del P, Barrero LS. Genetic structure of a Colombian cape gooseberry (*Physalis peruviana* L.) collection by means of microsatellite markers. Agron Colomb. 2016; 34: 5–16. <https://doi.org/10.15446/agron.colomb.v34n1.52960>
27. Kilian B, Graner A. NGS technologies for analyzing germplasm diversity in genebanks. Brief Funct Genomics. 2012; 11: 38–50. <https://doi.org/10.1093/bfgp/elr046> PMID: 22257472
28. Yadav P, Vaidya E, Rani R, Yadav NK, Singh BK, Rai PK, et al. Recent perspective of next generation sequencing: applications in molecular plant biology and crop improvement. Proc Natl Acad Sci, India, Sect B Biol Sci. 2018; 88: 435–449. <https://doi.org/10.1007/s40011-016-0770-7>
29. Elshire RJ, Glaubitz JC, Sun Q, Poland J a, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011; 6: e19379. <https://doi.org/10.1371/journal.pone.0019379> PMID: 21573248
30. Scheben A, Batley J, Edwards D. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnol J. 2017; 15: 149–161. <https://doi.org/10.1111/pbi.12645> PMID: 27696619
31. Osorio-Guarín JA, Enciso-Rodríguez FE, González C, Fernández-Pozo N, Mueller LA, Barrero LS. Association analysis for disease resistance to *Fusarium oxysporum* in cape gooseberry (*Physalis peruviana* L.). BMC Genomics. 2016; 17: 248. <https://doi.org/10.1186/s12864-016-2568-7> PMID: 26988219
32. García-Arias FL, Osorio-Guarín JA, Núñez Zarrantes VM. Association study reveals novel genes related to yield and quality of fruit in cape gooseberry (*Physalis peruviana* L.). Front Plant Sci. 2018. p. 362. <https://doi.org/10.3389/fpls.2018.00362> PMID: 29616069



33. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLoS One*. 2014; 9: e90346. Available: <https://doi.org/10.1371/journal.pone.0090346> PMID: 24587335
34. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*. 2007; 23: 2633–2635. <https://doi.org/10.1093/bioinformatics/btm308> PMID: 17586829
35. Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*. 2012; 485. <https://doi.org/10.1038/nature11119> PMID: 22660326
36. Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, et al. Genome sequence and analysis of the tuber crop potato. *Nature*. 2011; 475: 189–195. <https://doi.org/10.1038/nature10158> PMID: 21743474
37. Garzón-Martínez GA, Zhu ZI, Landsman D, Barrero LS, Mariño-Ramírez L. The *Physalis peruviana* leaf transcriptome: assembly, annotation and gene model prediction. *BMC genomics*. 2012; 13: 151. <https://doi.org/10.1186/1471-2164-13-151> PMID: 22533342
38. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987; 4: 406–425. <https://doi.org/10.1093/oxfordjournals.molbev.a040454> PMID: 3447015
39. R Core Team. R: A Language and environment for statistical computing [Internet]. Vienna, Austria; 2015. Available: <https://www.r-project.org>
40. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009; 19: 1655–64. <https://doi.org/10.1101/gr.094052.109> PMID: 19648217
41. Rousset F. Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol Ecol Resour*. 2008; 8: 103–106. <https://doi.org/10.1111/j.1471-8286.2007.01931.x> PMID: 21585727
42. Bevan MW, Uauy C, Wulff BBH, Zhou J, Krasileva K, Clark MD. Genomic innovation for crop improvement. *Nature*. 2017; 543: 346–354. <https://doi.org/10.1038/nature22011> PMID: 28300107
43. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*. 2013; 193: 327 LP–345. <https://doi.org/10.1534/genetics.112.143313> PMID: 22745228
44. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, et al. Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci*. 2017; 22: 961–975. <https://doi.org/10.1016/j.tplants.2017.08.011> PMID: 28965742
45. Kamaluddin Khan MA, Kiran U, Ali A, Abdin MZ, Zargar MY, et al. Molecular markers and marker-assisted selection in crop plants. In: Abdin MZ, Kiran U, Kamaluddin, Ali A, editors. *Plant Biotechnology: Principles and Applications*. Singapore: Springer Singapore; 2017. pp. 295–328. [https://doi.org/10.1007/978-981-10-2961-5\\_12](https://doi.org/10.1007/978-981-10-2961-5_12)
46. Simbaqueba J, Sanchez P, Sanchez E, Nunez Zarantes VM, Chacon MI, Barrero LS, et al. Development and characterization of microsatellite markers for the Cape gooseberry *Physalis peruviana*. *PLoS One*. 2011; 6: 26719. <https://doi.org/10.1371/journal.pone.0026719> PMID: 22039540
47. Chang L-Y, Toghiani S, Ling A, Aggrey SE, Rekaya R. High density marker panels, SNPs prioritizing and accuracy of genomic selection. *BMC Genetics*. 2018; 19: 4. <https://doi.org/10.1186/s12863-017-0595-2> PMID: 29304753
48. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*. 2013/01/21. Oxford University Press; 2014; 15: 256–278. <https://doi.org/10.1093/bib/bbs086> PMID: 23341494
49. Garzon G, Zhu I, Landsman D, Barrero L, Marino-Ramirez L. The *Physalis peruviana* leaf transcriptome: assembly, annotation and gene model prediction. *BMC genomics*. 2012; 13. <https://doi.org/10.1186/1471-2164-13-151> PMID: 22533342
50. Schatz MC, Witkowski J, McCombie WR. Current challenges in de novo plant genome sequencing and assembly. *Genome Biol*. 2012; 13: 243. <https://doi.org/10.1186/gb4015> PMID: 22546054
51. Trevisani N, Melo RC de, Pierre PMO, Colli MP, Coimbra JLM, Guidolin AF. Ploidy and DNA content of cape gooseberry populations grown in southern Brazil. *Caryologia*. 2018; 71: 414–419. <https://doi.org/10.1080/00087114.2018.1494440>
52. Gabur I, Chawla HS, Lopisso DT, von Tiedemann A, Snowdon RJ, Obermeier C. Gene presence-absence variation associates with quantitative *Verticillium longisporum* disease resistance in *Brassica napus*. *Sci Rep*. 2020; 10: 4131. <https://doi.org/10.1038/s41598-020-61228-3> PMID: 32139810
53. Zhang L, Luo H, Liu Z-Q, Yi Z, Luo J-C, Hao D-Y, et al. Genome-wide patterns of large-size presence/absence variants in sorghum. *J Integr Plant Biol*. 2014; 56: 24–37. <https://doi.org/10.1111/jipb.12121> PMID: 24428208

54. Wu X, Heffelfinger C, Zhao H, Dellaporta SL. Benchmarking variant identification tools for plant diversity discovery. *BMC Genomics*. 2019; 20: 701. <https://doi.org/10.1186/s12864-019-6057-7> PMID: 31500583
55. Frary A, Doganlar S, Frary A. Synteny Among Solanaceae Genomes BT—The Tomato Genome. In: Causse M, Giovannoni J, Bouzayen M, Zouine M, editors. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016. pp. 217–243. [https://doi.org/10.1007/978-3-662-53389-5\\_12](https://doi.org/10.1007/978-3-662-53389-5_12)
56. Vargas-Ponce O, Pérez-Álvarez LF, Zamora-Tavares P, Rodríguez A. Assessing genetic diversity in Mexican husk tomato species. *Plant Molecular Biology Reporter*. 2011; 29: 733–738. <https://doi.org/10.1007/s11105-010-0258-1>
57. Hu J, Saedler H. Evolution of the Inflated Calyx Syndrome in Solanaceae. *Molecular biology and evolution*. 2007; 24: 2443–2453. <https://doi.org/10.1093/molbev/msm177> PMID: 17827172
58. Feng S, Zheng K, Jiao K, Cai Y, Chen C, Mao Y, et al. Complete chloroplast genomes of four *Physalis* species (Solanaceae): lights into genome structure, comparative analysis, and phylogenetic relationships. *BMC Plant Biology*. 2020; 20: 242. <https://doi.org/10.1186/s12870-020-02429-w> PMID: 32466748
59. Beest M, Berg R, Brandenburg W. A taxonomic analysis of the species of *Physalis* L. (Solanaceae) based on morphological characters. In: Biodiversity, taxonomy and conservation of flowering plants: Calicut, Kerala, India, 1998 / Sivadasan, M and P Mathew—Calicut, Kerala, India: Mentor Books, 1999. 1999.
60. Zamora P, Vargas-Ponce O, Sánchez-Martínez J, Cabrera-Toledo D. Diversity and genetic structure of the husk tomato (*Physalis philadelphica* Lam.) in Western Mexico. *Genetic Resources and Crop Evolution*. 2014; 62: 141–153. <https://doi.org/10.1007/s10722-014-0163-9>
61. Arenas P, Kamienkowski N. Ethnobotany of the Genus *Physalis* L. (Solanaceae) in the South American Gran Chaco. *Candollea*. 2013; 68. <https://doi.org/10.15553/c2012v682a9>
62. Legge AP. Notes on the history, cultivation and uses of *Physalis peruviana* L. *Journal of the Royal Horticultural Society*. 1974; 99: 310–314.
63. Vargas-Ponce O, Sánchez Martínez J, Zamora Tavares M del P, Valdivia Mares LE. Traditional management of a small-scale crop of *Physalis angulata* in Western Mexico. *Genetic Resources and Crop Evolution*. 2016; 63: 1383–1395. <https://doi.org/10.1007/s10722-015-0326-3>