Genetic diversity and population structure in *Physalis peruviana* and related taxa based on InDels and SNPs derived from COSII and IRG markers

Gina A. Garzón-Martínez, Jaime A. Osorio-Guarín, Paola Delgadillo-Durán, Franklin Mayorga, Felix E. Enciso-Rodríguez, David Landsman, Leonardo Mariño-Ramírez, and Luz Stella Barrero

Gina A. Garzón-Martínez: ggarzon@corpoica.org.co; Jaime A. Osorio-Guarín: josoriog@corpoica.org.co; Paola Delgadillo-Durán: ldelgadillo@corpoica.org.co; Franklin Mayorga: fmayorgac@corpoica.org.co; Felix E. Enciso-Rodríguez: fenciso@corpoica.org.co; David Landsman: landsman@ncbi.nlm.nih.gov; Leonardo Mariño-Ramírez: marino@ncbi.nlm.nih.gov; Luz Stella Barrero: lbarredo@corpoica.org.co

**Abstract**

The genus *Physalis* is common in the Americas and includes several economically important species, among them *Physalis peruviana* that produces appetizing edible fruits. We studied the genetic diversity and population structure of *P. peruviana* and characterized 47 accessions of this species along with 13 accessions of related taxa consisting of 222 individuals from the Colombian Corporation of Agricultural Research (CORPOICA) germplasm collection, using Conserved Orthologous Sequences (COSII) and Immunity Related Genes (IRGs). In addition, 642 Single Nucleotide Polymorphism (SNPs) markers were identified and used for the genetic diversity analysis. A total of 121 alleles were detected in 24 InDels loci ranging from 2 to 9 alleles per locus, with an average of 5.04 alleles per locus. The average number of alleles in the SNP markers was two. The observed heterozygosity for *P. peruviana* with InDel and SNP markers was higher (0.48 and 0.59) than the expected heterozygosity (0.30 and 0.41). Interestingly, the observed heterozygosity in related taxa (0.4 and 0.12) was lower than the expected heterozygosity (0.59 and 0.25). The coefficient of population differentiation $F_{ST}$ was 0.143 (InDels) and 0.038 (SNPs), showing a relatively low level of genetic differentiation among *P. peruviana* and related taxa.

**Correspondence to:** Leonardo Mariño-Ramírez, marino@ncbi.nlm.nih.gov; Luz Stella Barrero, lbarredo@corpoica.org.co.

**Publisher’s Disclaimer:** This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

**Authors’ contributions**

Performed the experiments: GAG-M JAO-G FEE-R PD-D FM. Analyzed the data: GAG-M JAO-G FEE-R. Jointly prepared the manuscript: GAG-M JAO-G FEE-R LMR LSB. Conceived and oversaw the project and its components: FEE-R LSB. Contributed bioinformatics expertise for data analysis LMR DL. All authors read and approved the final manuscript.
Higher levels of genetic variation were instead observed within populations based on the AMOVA analysis. Population structure analysis supported the presence of two main groups and PCA analysis based on SNP markers revealed two distinct clusters in the *P. peruviana* accessions corresponding to their state of cultivation. In this study, we identified molecular markers useful to detect genetic variation in *Physalis* germplasm for assisting conservation and crossbreeding strategies.

**Keywords**
- Cape gooseberry; germplasm; COSII; IRGs; SNPs; Genetic variation

1. Introduction

The genus *Physalis* consists in more than 90 species, native of the Americas, being Mexico the center of diversity of the husk tomato (Vargas-Ponce et al., 2010). The genus includes different species with nutritional, nutraceutical and commercial interests. Among them, *P. peruviana* and other related taxa as *P. philadelphica*, *P. pruinosa* and *P. longifolia* has been characterized for different health related compounds with anti-inflammatory and antioxidant properties (Martínez et al., 2010; Yen et al., 2010; Maldonado et al., 2011; Ramadan, 2011; Jin et al., 2012; Kindscher et al., 2014; Takimoto et al., 2014), as well as others with valuable nutritional properties including vitamins A, B and C, polyunsaturated fatty acids, proteins and minerals (Puente et al., 2011; Ramadan, 2011). Therefore the commercial interest of *P. peruviana*, also known as Cape gooseberry, has increased to the point of being currently the main exported fruit after banana in Colombia (Bonilla et al., 2009).

Accurate knowledge of genetic diversity and relationships among preserved germplasm collections of any crop is essential and important for establishing, managing and ensuring long-term success of appropriate crop improvement programs through breeding (Gwag et al., 2010). Thus, the study on genetic diversity and population structure of germplasm collections has been useful in supporting conservation and genetic improvement strategies (Rao and Hodgkin, 2002; Grandillo, 2014). Furthermore, natural biodiversity found in non-cultivated relatives of crop species are reservoirs representing an important source of genetic variation essential for any crop-breeding program (Grandillo, 2014). Although some *Physalis* species have been widely recognized by their nutraceutic and economic importance, little is known about their genetic diversity at the molecular level, mainly because of the lack of available markers in accordance with their current status as orphan species. Dominant markers RAMs (Random Amplified Microsatellites) were the first type used to study the genetic diversity of a Colombian *P. peruviana* collection where high expected heterozigosity (He=0.2559) was found (Bonilla et al., 2008; Morillo Paz et al., 2011). Later, next-generation sequencing technologies for the rapid identification of SSR loci derived from ESTs were used (Simbaqueba et al., 2011; Garzón-Martínez et al., 2012). Nevertheless, since all SSRs loci were located in the UTR regions of the transcriptome, a low polymorphic rate (22%) was found in a panel of 8 accessions from *P. peruviana* and the related species *P. floridana*, therefore we decided to develop alternative markers. More recently, 97 tomato markers (COS, SSRs and InDel markers) and 25 *P. peruviana* SSRs
markers were used for genetic diversity analysis in 38 accessions of *Physalis* (Wei et al., 2012). This study suggested the efficient use of tomato markers in genetic studies of *Physalis* as species from the family Solanaceae, and a high level of polymorphism (92.7% of markers were polymorphic) in accordance with a broad genetic at DNA level in wild and cultivated species. Furthermore, Berdugo et al. (2015), used 328 COSII and 154 Immunity Related Genes (IRGs) to evaluate an F1 population generated between contrasting pathogen response parents. This population showed a total of 127 alleles with an average of 3.18 per locus, a PIC of 0.358 and high values of heterozygosity (Ho: 0.737 and He: 0.449).

Recent advances in sequencing technologies allowed the identification of large sets of Single Nucleotide Polymorphisms (SNPs) (Patel et al., 2015; Yang et al., 2015). SNP genotyping has become the most useful technique in model and non-model species for genetic diversity and population structure analysis, marker-assisted selection and association studies (Frascaroli et al., 2013). Recently, Enciso et al. (2013) identified 74 IRGs in *P. peruviana*, from which 17 markers were selected and sequenced in a small subset of *P. peruviana* and related taxa allowing the identification of one candidate SNP associated to the resistance response against the fungal pathogen *Fusarium oxysporum*, one of the main constraints to *P. peruviana* production in Colombia.

Since the development and discovery of new markers for non-model species as well as the use of Conserved Ortholog Set of markers (COSII) (Fulton et al., 2002; Wu et al., 2006) are valuable resources for genetic studies of *Physalis*, the aim of the present study was to investigate the genetic diversity and population structure of 60 accessions with a large representation of *P. peruviana*, by the use and comparison of InDels and SNPs derived from COSII and IRG markers, to contribute knowledge on the germplasm genetic base for conservation and use for breeding strategies.

### 2. Material and methods

#### 2.1. Plant material and DNA isolation

Young leaves of 222 plants belonging to 47 *Physalis peruviana* accessions (an average of 3 plants per accession, each derived from a single seed) and 13 related taxa (an average of 4 plants per accession) were collected from an *in vitro* germplasm collection maintained at the Colombian Corporation for Agricultural Research (CORPOICA) (Table 1, Supplementary Table 1). The leaves were stored at −70°C and utilized for genomic DNA isolation using the modified Dellaporta et al. (1983) method, as described by Enciso et al. (2013).

#### 2.2. Molecular marker selection

A total of 454 molecular markers were tested for genetic diversity analyses on *P. peruviana* and related taxa. We selected 327 COSII markers based on their distribution across the 12 linkage groups of the tomato genome, considering their orthologous nature in a broad range of Solanaceae family species (Fulton et al., 2002; Wu et al., 2006; Bedoya-Reina and Barrero, 2010). Furthermore, 33 COSII genes were selected, based on resistance to biotic factors related ontologies (i.e. against phytopathogenic microorganisms), as a strategy to look for polymorphisms within homologous defense/resistance genes. Each primer pair was
obtained from the Solanaceae network database (http://www.solgenomics.net). The remaining 94 markers belonged to IRGs, previously developed from the *P. peruviana* leaf transcriptome (Enciso-Rodríguez et al., 2013).

The 360 COSII markers were screened on five accessions with contrasting responses to *Fusarium oxysporum* which includes *P. peruviana* and related species (Enciso-Rodríguez et al., 2013). In addition, 94 IRGs were tested on six contrasting accessions as described by Enciso et al. (2013). Polymorphic COSII and IRGs markers were selected and used to genotype 222 samples from 60 accessions. From the original set of markers, we selected monomorphic COSII and IRGs molecular markers for SNP identification on a 64-plant panel (Table 1). Complete marker information is shown in Supplementary Table 2.

2.3. PCR amplification and marker visualization

Polymerase chain reactions (PCR) were performed each in 15 µl final reaction volume using 1X PCR Buffer, 2.0 mM MgCl$_2$, 0.2 µM dNTPs, 0.2 µM of each primer, 0.05 U/µl Taq DNA polymerase and 5 ng of genomic DNA. Thermal cycler conditions included a denaturation step at 94°C for 5 min followed by 35 cycles at 94°C for 30 s, 56°C to 60°C for 1 min (depending on the marker), 72°C for 2 min, ending with and extension at 72°C for 10 min. Amplifications were performed in a i-Cycler thermal Cycler (Bio-Rad, Hercules, CA, USA). PCR products were separated on 2% agarose gels stained with ethidium bromide (0.5 µg/mL). Molecular size of the COSII and IRGs bands was estimated using a 1 kb plus ladder as reference.

2.4. Marker sequencing

Each forward and reverse COSII and IRG primer selected for sequencing was modified with a 10-base Multiplex Identifier (MID) adaptor sequence and a 454-sequencing primer (A-primer and B-primer). The MID adapters were used for post-sequencing sample identification. Following the PCR amplification as described above, each of four-sample reactions was pooled. The pooled products were sequenced in two sequencing plates divided in eight regions, using the 454 GS-FLX Titanium platform (Roche) (Engencore Sequencing Facility, Columbia, SC, USA). The SFF files were submitted to the NCBI Sequence Read Archive (SRA, accession number SRX216233).

2.5. Data processing and analyses

COSII and IRG InDels were scored as co-dominants markers using the GeneTools software (Syngene, Frederick, MD). Sequence data were processed using the Mothur software package (Schloss et al., 2009), which involves trimming of adapter sequences and sample identification. Reference sequences for the set of markers were generated from Sanger and 454 reads (Enciso-Rodríguez et al., 2013) by aligning the 454 reads of each sample using the Burrows-Wheeler Alignment tool (BWA) (Li and Homer, 2010). SNP calling was performed with BWA read mapping files (BAM formats) using the SAMtools package (Li et al., 2009), retaining SNPs with a minimum read depth of 20 and a minimum base quality of 30. The final consensus sequence of each genotype was aligned using Muscle (v3.7) (Edgar, 2004), and then manually edited using Bioedit (http://www.mbio.ncsu.edu/bioedit/).
Finally, SNPs with a minor allele frequency (MAF) ≤ 0.05 were selected by Tassel software v4.1 (Bradbury et al., 2007).

Standard measures of diversity including the average number of alleles per locus (a), expected heterozygosity (He), observed heterozygosity (Ho), polymorphic index content (PIC) were calculated by PowerMarker (version 3.25) (Liu and Muse, 2005) using InDels (COSII and IRG) and SNP markers independently. GenAlex (Peakall and Smouse, 2012) was used to perform Wright’s F statistics and an analysis of molecular variance (AMOVA).

Population structure analysis was carried out by Structure 2.3.4 (Pritchard et al., 2000). Number of populations (K) was set from 1 to 10, repeated 10 times, with a burn-in period of 50,000 iterations and 100,000 MCMC (Markov Chain Monte Carlo) repeats and a model allowing for admixture and correlated frequencies. The K optimum was evaluated by Evanno et al. (2005), using Structure Harvester (Earl and VonHoldt, 2011). The Neighbor-Joining algorithm was used for cluster analyses based on the Nei’s genetic distance using Phylip 3.695 (Felsenstein, 1989) with 1,000 bootstraping replicates. Principal component analysis (PCA) was performed using the R Stats package (R Development Core Team, 2009).

3. Results

3.1. DNA amplification and marker polymorphism

From the 360 COSII and the 94 IRGs molecular markers selected (Table 2), 24 polymorphic InDels (18 COSII and 6 IRGs) were identified and used to screen the entire population (222 plants) (Table 3). The amplicon average size of the 24 InDels markers ranged from 210 to 1700 bp (Supplementary Table 3). Molecular markers with low level of polymorphism (<10% of amplification in the whole population) were discarded.

Monomorphic markers on agarose gels (15 COSII and 18 IRGs) based on their size (less than 400 pb) and random distribution across the tomato chromosomes were selected for sequencing in order to uncover SNPs on a 64-plant panel (Table 2, 3). A multi-fasta alignment showed a total length of 13,000 nt per plant and 832,000 nt over all sequences. From these, we identify a total of 642 SNPs for the whole collection, 463 SNPs for P. peruviana population and 1,600 SNPs for related taxa germplasm.

3.2. Genetic diversity and population structure

The 24 InDel markers detected a total of 121 alleles in the whole collection, with an average number of alleles per locus of 5.04, ranging from 2 to 9. For SNP markers, the allele number was 2. Based on InDel markers the average observed heterozygosity was 0.47 on average for the whole collection, ranging from 0 (C2_At5g51040 marker) to 1.0 (C2_At1g18165 marker). Null values in observed heterozygosity were found in individuals for some markers suggesting homozygosity. Furthermore, the mean expected heterozygosity value (0.42) was lower than the observed heterozygocity, which ranged from 0.08 for PpIRG64 to 0.718 for C2_At2g3000. Similarly, SNPs markers showed a lower expected heterozygosity value (0.33) than the observed heterozygocity (0.38) for all taxa. Polymorphism information content values (PIC) ranged from 0.094 to 0.663, with a mean value of 0.37 for InDels.
markers, while SNP markers showed a mean PIC value of 0.26. These parameters are summarized in Table 4 and Supplementary Tables 4 and 5.

When *P. peruviana* and related taxa were analyzed as two different populations, the average number of alleles per locus for InDels were higher for related taxa (4.75) than *P. peruviana* (2.75) while for SNPs markers the average number of alleles was equal in both populations (2). The observed heterozygosity for *P. peruviana* with both markers was higher (InDels = 0.48 and SNPs = 0.59) than the expected heterozygosity (0.30 and 0.41). Nonetheless, *Ho* values in related taxa (InDels = 0.4 and SNPs = 0.12) were lower than gene diversity values (He) (InDels = 0.59 and SNPs = 0.25) for both kinds of markers (Table 4).

Genetic structure of the whole population was analyzed using neighbor joining (NJ), PCA and Bayesian cluster approaches. The cluster analysis based on Nei’s genetic distance derived from InDel and SNP markers showed two major groups (Figure 1) comprised of related taxa and *P. peruviana* accessions, with high bootstrap values (>50%), as expected. Two exceptions were observed; the related species *P. philadelphica* 09U056 was grouped with the *P. peruviana* population while the *P. peruviana* 09U289 accession was clustered with related taxa with bootstrap values above 70%, using either type of markers. Moreover, NJ tree for InDels (Figure 1a) showed some plants from *P. peruviana* accessions (09U132-1, 09U138-6, 09U118-1, 09U134-4, 09U108-3, 09U116-1, 2, 3, 4) clustered with related taxa.

PCA analyses from InDel and SNP marker data showed a similar result to the NJ clustering approach, supporting the presence of two populations (Related taxa and *P. peruviana*), and highlighting the misclassification of 09U289 and 09U056 accessions (Figure 2a–b). Besides, PCA analysis based on SNP markers revealed two clusters separated by state of cultivation on *P. peruviana* population: wild and cultivated accessions (Figure 2b). InDel based analyses showed that the first three components accounted for 61% of the total variation, where each component explained 46%, 8% and 7% of the overall variation, respectively; while SNP based analyses showed that the first to the third component explained 22%, 10% and 7%, respectively, for a total variation of 39%. Non-geographical distribution pattern was found with any type of markers by PCA analysis (data not shown).

The identification of genetically homogeneous groups of individuals was performed using the method based on ΔK suggested by Evanno *et al.* (2005), which showed a peak with a maximum value around K=2 (Figure 3), indicating the presence of two main clusters in the whole collection (*P. peruviana* and related taxa), consistent with the results displayed by NJ and PCA analysis. Population differentiation analysis for the whole collection using InDel markers showed a higher value (0.143±0.022) among populations than the SNPs F<sub>ST</sub> value (0.038±0.002), while FIS values for both kinds of markers were 0.044±0.103 for InDels and −0.015±0.029 for SNPs markers.

The analysis of molecular variance (AMOVA) for InDel markers showed that genetic variance within accessions contributed with 73% to genetic diversity (Table 5) while the variance among populations (*P. peruviana* and related taxa) and among accessions accounted for 23% and 4% of the total variance, respectively. Similarly, for SNPs markers,
most of the variation was found within accessions with 95%, while 5% of the variance was observed among populations.

4. Discussion

4.1. Identification of marker polymorphisms by InDels and SNPs

The reduced availability of molecular markers in non-model commercially important crop species such as *P. peruviana* has been a bottleneck for the assessment of genetic variability that could support breeding strategies. Nonetheless, dominant RAM markers as well as codominant markers such as SSRs and COSII, conserved in related species as tomato have been used in genetic studies in *Physalis* (Bonilla et al., 2008; Morillo Paz et al., 2011; Wei et al., 2012). COSII markers represent a strong strategy to integrate diversity estimations over extended genetic distances and to map highly variable gene spots (Bedoya-Reina and Barrero, 2010). Besides, due to the fact they are derived from universal primers transferred among species, these are an economic alternative to other types of co-dominant markers such as SSRs (Enciso-Rodríguez et al., 2010). Although, identification of SSRs markers in *P. peruviana* has been reported (Simbaqueba et al., 2011), one of the main drawbacks of these markers are the presence of null alleles related to mutations in the primer annealing sites which may lead to errors in genotyping scoring (Kumar et al., 2009), and may also be the case for other type of markers. SSRs markers from *P. peruviana* were designed from expressed sequence tags and UTR regions (Simbaqueba et al., 2011), and usually, markers found in untranslated regions are less polymorphic than genomic SSRs (Ellis and Burke, 2007) and less transferable to other species from the same genera as shown by Lara-Cabrera and Spooner (2005) in potato and its wild relatives. Thus, in the present study we increased the number of co-dominant markers available for the genus *Physalis* and evaluated these new markers to contribute knowledge of the genetic variation within the genus especially on *P. peruviana* species. The study showed a low polymorphism level of COSII and IRG markers with 5% of the 360 COSII markers and 8% of the 94 IRGs behaving as polymorphic on agarose gels.

To increase the number of markers for further analyses, this study took advantage of next generation sequencing technologies to uncover SNPs on selected COSII and IRG monomorphic markers, which resulted in the identification of 642 SNPs in 15 COSII and 18 IRGs for the whole collection. According to several reports a larger number of SNPs (7 to 11 or more times as compared to SSRs) can have the same level of information as SSRs (Yu et al., 2009; Van Inghelandt et al., 2010a; Filippi et al., 2015), suggesting that the number used in the present study is within a suitable range. The average number of alleles per locus in these SNP markers was two, suggesting that despite the variable nature in ploidy level of *Physalis* spp. (Liberato et al., 2014), the species in the genus might behave as allotetraploids, hypothesis that will need further investigation through segregation analyses. The biallelic and multiallelic markers found in this study were used for diversity and population structure analyses as well as for establishing comparisons between InDels and SNPs derived from previously reported markers.

*Plant Gene*. Author manuscript; available in PMC 2015 December 01.
4.2. Genetic diversity and population structure

A broad genetic base is important for genetic resources and plant breeding applications which include but are not limited to generation of core collections, parental lines selection, quantitative trait loci identification, association mapping or genomic selection, all of which require previous information of genetic diversity (Rao and Hodgkin, 2002; Zhu et al., 2008; Rauf et al., 2010). In this study, the application of InDels and SNP markers on the whole collection (all taxa) and the two subpopulations identified (*P. peruviana* and related taxa) revealed a moderately to high genetic diversity.

InDel markers showed an average PIC value of 0.37 for the entire population, which placed them in the range of loci with intermediate polymorphism (0.25 to 0.5) according to Ge et al. (2013). Highly polymorphic markers such as C2_At1g53000 and C2_At2g42750 derived from COSII (0.66 and 0.60 PIC values, respectively) could be used in future studies to discriminate species in the genus *Physalis*. Likewise, markers C2_At1g53000 and C2_At3g18165 were identified as the most informative to evaluate the *P. peruviana* collection with PIC values of 0.592 and 0.545, respectively. PIC values of SNP markers reached until 0.5 due to SNPs biallelic nature and were in general lower than InDels PIC values.

InDel markers have been used to characterize diversity of *Physalis* and other related species from the same family. A recent study by Wei et al. (2012) with 96 InDel markers, revealed an average number of 3.5 alleles per marker, lower in comparison with the average number for the whole population in this study (5.04 alleles). Nevertheless, this value is difficult to compare, because the sample size used in Wei et al. (2012) study was smaller (38 accessions) and the present study included 60 accessions along with the differences in the populations and marker loci used. Compared to SNPs markers, diversity values are lower that those estimated by InDel markers, probably explained by the biallelic nature of SNP markers. Usually a maximum gene diversity with biallelic markers is 0.5, whereas for multi-allelic markers such as SSRs or InDels can approach 1 (Van Inghelandt et al., 2010b).

Additionally, the importance of the number and selection of SNPs, as well as the use of a large panel of accessions to prevent ascertainment bias and enhance the accuracy of diversity studies (Moragues et al., 2010; Emanuelli et al., 2013) should be considered as well.

Some differences are found in the level of diversity when compared to a previous study in a *P. peruviana* collection (Bonilla et al., 2008). Values in the present study (He = 0.30 and 0.41 for InDels and SNPs, respectively) were slightly higher than those observed by Bonilla et al. (2008) (He = 0.2559). Nevertheless, it should be considered that Bonilla et al. (2008) used a different population and a set of dominant (RAM markers), which may underestimate allelic diversity as compared to the co-dominant nature of markers used in the present investigation. On the other hand, Berdugo et al. (2015) showed slightly higher diversity values (He = 0.44) for InDels markers than this study. However, this difference is due to the fact that, Berdugo et al. (2015) used an F1 population which is expected to have high heterozygosity values.
Diversity estimations from COSII and IRG as well SNP derived markers used in this study should be interpreted with caution due to the fact that these markers were developed and discovered from conserved domains in the Solanaceae and thus probably had been under selective pressures (Wu et al., 2006). Nonetheless, relatively high diversity values were found related to the fact that Cape gooseberry is 53% outcrossing (Lagos et al., 2008) and it became cultivated from the wild without being subjected to a long domestication process as compared to other fruit bearing species of the same family as tomato (Labate et al., 2009; Sim et al., 2012). It is probable then, that its domestication is still in process and natural selection is still playing an important role retaining heterogeneous populations with a broad genetic variability and adaptability (Rauf et al., 2010). These results could have important implications in plant breeding purposes, since the presence of a high genetic diversity is a prerequisite for the efficiency of selection processes (Rao and Hodgkin, 2002; Rauf et al., 2010).

The clustering analysis based on Nei’s genetic distance did not show any geographical distribution pattern. This result might be associated to gene flow via seed exchanges among different regions occurred frequently or possibly by human transfer trying to mitigate the effects of some diseases like F. oxysporum. In other analysis, NJ trees from the P. peruviana population were run separately for each marker type and no clustering was observed in relation to geographical distribution or state of cultivation (Data not shown). Additionally, the position in the dendrogram of accession 09U289 suggests a misclassification; supported by the flow cytometry studies where this accession presented a different ploidy level (24/2n=4x; of DNA= 2.33 pg) from other P. peruviana accessions (ranging from 24/2n=4x; of DNA = 2.33 pg to 48/2n=4x; of DNA = 8.12 pg) (Liberato et al., 2014). On the other hand, the unexpected clustering of some plants from P. peruviana accessions with the related taxa group using InDels markers, was possibly due to the fact that InDels had a considerable number of missing data (9.3%) in the population as compared to SNP data (4.3% missing).

All three-population analyses supported that the whole population represented by all taxa has two different genetic populations, related taxa and P. peruviana, as expected. Based on the PCA approach, it was possible to identify two different populations within the P. peruviana group, differentiated by state of cultivation, commercial and wild accessions with some mixtures between the two subgroups. Wild samples found in a cluster with commercial materials probably came from commercial fields usually characterized by non-homogeneous materials due to the lack of breeding cultivars grown for this species in Colombia. Most of the accessions used in the present study were cultivars that followed a primary artificial selection for certain desirable characteristics without a breeding program. Besides, the low percentage of variation explained by the first three components in the PCA analysis and the different nature of markers indicates that differentiation of some individuals may not be well captured. Nonetheless, it is important to consider the outcrossing nature of the species, where a greater number of SNPs markers across all genome may be required to detect population structure (Zhu et al., 2008).

Based on the $F_{ST}$ values (InDels = 0.143 and SNPs = 0.038) for the whole collection we did not find a high level of differentiation between the two subpopulations (P. peruviana and...
related taxa). The $F_{ST}$ values were closer to zero and the AMOVA values among populations represented 5% and 23% for SNPs and InDels markers respectively, while within populations was much higher for both markers. This indicates more variation within than between populations, aspect that should be considered for conservation and breeding strategies where more variation wants to be captured. According with Hamrick and Godt (1989) the reproductive biology is an important factor in determining the genetic structure, this is because the inbreeding species maintain more of their genetic diversity among populations rather than within populations than do outcrossers. In outcrossers a higher proportion of their variation occurs within individual populations. Furthermore, Silvertown (2009) affirms that the outcrossing populations such as Physalis show lower $F_{ST}$ values than self-fertilizing populations; consistent with the high genetic variation that we found in the entire population and the lower $F_{IS}$ values typical of populations with heterozygote excess. 

The relatively high diversity values of the whole collection support that the $P.$ peruviana population and the related taxa germplasm, is a valuable source to assist crop improvement programs for more efficient and strategic use of genetic resources. This information should be used to identify duplicates or wrong classifications in germplasm collections towards establishment of core collections, conservation and breeding purposes.

Understanding genetic diversity and population structure is essential for conservation efforts and the basis for selection is the first step to establish a crossing program of different genotypes or populations of Physalis peruviana or its related taxa. This is a strategy for capturing novel desirable alleles for valuable traits such as fruit quality, yield or reduced susceptibility to pests.

5. Conclusions

This study found a moderate to high genetic diversity on a 64 plant-panel of $P.$ peruviana and related taxa species. A total of 121 alleles were detected on 24 InDels loci and 642 SNPs were identified in a 222 individuals sampled. Diversity and population structure analysis showed two clusters related to Physalis peruviana species and related taxa. The evaluation and characterization of germplasm through INDels and SNPs markers was useful to infer diversity and population structure analyses that may have important implications for managing conserved germplasm, for avoiding duplicates and wrong classification of the collection, for developing core collections and for selecting material for breeding strategies in this orphan crop.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Edgar Jimenez for supplying in vitro plant material for analyses. This study was partially funded by the Colombian Ministry of Agriculture Contract No. 054/08072-2008 to LSB. We thank Aureliano Bombarely for data analysis advice. This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and National Center for Biotechnology Information. We thank Roxana Yockteng for her critical review of this manuscript.
References

Bedoya-Reina OC, Barrero LS. Preliminary assessment of COSII gene diversity in lulo and a relative species: initial identification of genes potentially associated with domestication. Gene. 2010; 458:27–36. [PubMed: 20302924]

Berdugo Cely JA, Enciso Rodríguez FE, González Almario C, Barrero Meneses LS. Genetic variability of parentals and inter and interspecific F1 populations of Physalis peruviana L. and P. floridana Rydb. Revista Brasileira de Fruticultura. 2015; 37:179–192.

Bonilla, M.; Arias, P.; Landínez, L.; Moreno, J.; Cardozo, F.; Suárez, M. Agenda prospectiva de investigación y desarrollo tecnológico para la cadena productiva de la uchuva en fresco para exportación en Colombia. Bogotá: Ministerio de Agricultura y Desarrollo Rural, MADR; 2009. p. 152

Bonilla M, Piedrahíta KE, Mauricio A, Terranova P, Amariles DV, Eduardo J, Flórez M. Caracterización molecular de 43 accesiones de uchuva de seis departamentos de Colombia. Acta. Agron. 2008; 57:109–115.

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. [PubMed: 17586829]

Dellaporta S, Wood J, Hicks J. A plant DNA minipreparation: version II. Plant Mol. Biol. Rep. 1983; 1:19–21.

Earl DA, VonHoldt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 2011; 4:359–361.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]

Ellis J, Burke J. EST-SSRs as a resource for population genetic analyses. Heredity (Edinb.). 2007; 99:125–132. [PubMed: 17519965]

Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefannini M, Troggio M, Myles S, Martinez-Zapater JM, Zyprian E, Moreira FM, Grando MS. Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biol. 2013; 13:39. [PubMed: 23497049]

Enciso-Rodríguez F, Martínez R, Lobo M, Barrero LS. Genetic variation in the Solanaceae fruit bearing species lulo and tree tomato revealed by Conserved Ortholog (COSII) markers. Genet. Mol. Biol. 2010; 33:271–278. [PubMed: 21637482]

Enciso-Rodríguez FE, González C, Rodríguez EA, López CE, Landsman D, Barrero LS, Mariño-Ramírez L. Identification of Immunity related genes to study the Physalis peruviana Fusarium oxysporum pathosystem. PLoS One. 2013; 8:e68500. [PubMed: 23844210]

Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005; 14:2611–2620. [PubMed: 15969739]

Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 1989; 5:164–166.

Filippi CV, Aguirre N, Rivas JG, Zubrzycki J, Puebla A, Cordes D, Moreno MV, Fusari CM, Alvarez D, Heinz RA. Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. BMC Plant Biol. 2015; 15:52. [PubMed: 25848813]

Frascaroli E, Schrag TA, Melchinger AE. Genetic diversity analysis of elite European maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. Theor. Appl. Genet. 2013; 126:133–141. [PubMed: 22945268]

Fulton TM, Hoveen RVD, Eannetta NT, Tanksley SD. Identification, analysis, and utilization of Conserved Ortholog Set Markers for comparative genomics in higher plants. Plant Cell. 2002; 14:1457–1467. [PubMed: 12119367]

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. [PubMed: 22553342]

Garzón-Martínez et al. Page 11

Plant Gene. Author manuscript; available in PMC 2015 December 01.
Ge H, Liu Y, Jiang M, Zhang J, Han H, Chen H. Analysis of genetic diversity and structure of eggplant populations (Solanum melongena L.) in China using simple sequence repeat markers. Sci. Hortic. 2013; 162:71–75.

Grandillo, S. Introgression libraries with wild relatives of crops, Genomics of Plant Genetic Resources. Springer; 2014. p. 87-122.

Gwag JG, Dixit A, Park YJ, Ma KH, Kwon SJ, Cho GT, Lee GA, Lee SY, Kang HK, Lee SH. Assessment of genetic diversity and population structure in mungbean. Genes Genomics. 2010; 32:299–308.

Hamrick, JL.; Godt, MJW. Allozyme diversity in plant species. In: Brown, AHD.; Clegg, MT.; Kahler, AL.; Weir, BS., editors. Plant population genetics, breeding and genetic resources. Sunderland, MA: Sinauer; 1989. p. 43-63.

Jin Z, Mashuta MS, Stolowich NJ, Vaisberg AJ, Stivers NS, Bates PJ, Lewis WH, Hammond GB. Physangulidines A, B, and C: three new antiproliferative withanolides from Physalis angulata L. Organic Letters. 2012; 14:1230–1233. [PubMed: 22329497]

Kindscher K, Cao C-M, Gallagher RJ, Zhang H, Long Q, Timmermann BN. Comparison of bioactive secondary metabolites in experimental and natural populations of wild tomatillos, Physalis longifolia Nutt. Ethnobotany Research and Applications. 2014; 12:175–182.

Kumar P, Gupta V, Misra A, Modi D, Pandey B. Potential of molecular markers in plant biotechnology. Plant Omics J. 2009; 2:141–162.

Labate, Ja; Robertson, LD.; Wu, F.; Tankslsy, SD.; Baldo, AM. EST, COSII, and arbitrary gene markers give similar estimates of nucleotide diversity in cultivated tomato (Solanum lycopersicum L.), Theor. Appl. Genet. 2009; 118:1005–1014. [PubMed: 19153710]

Lagos T, Vallejo F, Criollo H, Muñoz J. Sexual reproduction of the cape gooseberry. Acta. Agron. 2008; 57:81–87.

Lara-Cabrera, SI.; Spooner, DM. Taxonomy of Mexican diploid wild potatoes (Solanum sect. Petota): morphological and microsatellite data, A Festschrift for William G. D’arcy: The legacy of a Taxonomist. St. Louis, MO, USA: Missouri Botanical Garden; 2005. p. 199-225.

Li H, Handsaker B, Wysocker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. 2010; 11:473–483. [PubMed: 20460430]

Liberato SA, Sánchez-Betancourt E, Argüelles JH, González CA, Núñez VM, Barrero LS. Cytogenetics of Physalis peruviana L. and Physalis floridana Rydb. genotypes with differential response to Fusarium oxysporum. Revista Corpóica Ciencia y Tecnología Agropecuaria. 2014; 15:49–59.

Maldonado E, Pérez-Castorena AL, Garcés C, Martínez M. Philadelphicalactones C and D and other cytotoxic compounds from Physalis philadelphica. Steroids. 2011; 76:724–728. [PubMed: 21497618]

Martínez W, Ospina LF, Granados D, Delgado G. In vitro studies on the relationship between the anti-inflammatory activity of Physalis peruviana extracts and the phagocytic process. Immunopharmacol. Immunotoxicol. 2010; 32:63–73. [PubMed: 19678736]

Moragues M, Comadran J, Waugh R, Milne I, Flavell a.J, Russell JR. Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theor. Appl. Genet. 2010; 120:1525–1534. [PubMed: 20157694]

Morillo Paz AT, Villota Cerón DE, Lagos Burbano TC, Ordóñez Jurado HR. Morphological and Molecular Characterization of 18 Introductions of Cape Gooseberry Physalis peruviana L. collection of the University of Nariño. Revista Facultad Nacional de Agronomía, Medellín. 2011; 64:6043–6053.

Patel, DA.; Zander, M.; Dalton-Morgan, J.; Batley, J. Advances in plant genotyping: where the future will take us, Plant Genotyping. Springer, 2015. p. 1-11.

Peakall R, Smouse P. GenAlEx 6.5 : Genetic analysis in Excel. Population genetic soft- ware for teaching and research - an update. Bioinformatics. 2012; 1:6–8.
Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. [PubMed: 10835412]

Puente LA, Pinto-muñoz CA, Castro ES, Cortés M. Physalis peruviana Linnaeus, the multiple properties of a highly functional fruit: A review. Food Res. Int. 2011; 44:1733–1740.

Ramadan MF. Bioactive phytochemicals, nutritional value, and functional properties of cape gooseberry (Physalis peruviana): An overview. Food Res. Int. 2011; 44:1830–1836.

Rao VR, Hodgkin T. Genetic diversity and conservation and utilization of plant genetic resources. Plant Cell Tissue Organ Cult. 2002; 68:1–19.

Rauf S, Teixeira da Silva J, Khan AA, Naveed A. Consequences of plant breeding on genetic diversity. International Journal of Plant Breeding. 2010; 4:1–21.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 2009; 75:7537–7541. [PubMed: 19801464]

Silvertown, J.; Charlesworth, D. Introduction to Plant Population Biology. John Wiley & Sons; 2009.

Sim SC, Van Deynze A, Stoffel K, Douches DS, Zarka D, Ganal MW, Chetelat RT, Hutton SF, Scott JW, Gardner RG, Panthee DR, Mutschler M, Myers JR, Francis DM. High-density SNP genotyping of tomato (Solanum lycopersicum L.) reveals patterns of genetic variation due to breeding. PLoS One. 2012; 7:e45520. [PubMed: 23029069]

Simbaqueba J, Sanchez P, Sanchez E, Nunez Zarantes VM, Chacon MI, Barrero LS, Marino-Ramirez L. Development and characterization of microsatellite markers for the Cape gooseberry Physalis peruviana. PLoS One. 2011; 6:e26719. [PubMed: 22039540]

Takimoto T, Kanbayashi Y, Toyoda T, Adachi Y, Furuta C, Suzuki K, Miwa T, Bannai M. 4β-Hydroxywithanolide E isolated from Physalis pruinosa calyx decreases inflammatory responses by inhibiting the NF-κB signaling in diabetic mouse adipose tissue. Int. J. Obes. 2014; 38:1432–1439.

Van Inghelandt D, Melchinger AE, Lebreton C, Stich B. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor. Appl. Genet. 2010a; 120:1289–1299. [PubMed: 20063144]

Van Inghelandt D, Melchinger AE, Lebreton C, Stich B. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor. Appl. Genet. 2010b; 120:1289–1299. [PubMed: 20063144]

Vargas-Ponce O, Pérez-Álvarez LF, Zamora-Tavares P, Rodríguez A. Assessing Genetic Diversity in Mexican Husk Tomato Species. Plant Mol. Biol. Rep. 2010; 29:733–738.

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. [PubMed: 23166835]

Wu F, Mueller La, Crouzillat D, Pétiard V, Tanksley SD. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics. 2006; 174:1407–1420. [PubMed: 16951058]

Yang H, Li C, Lam H-M, Clements J, Yan G, Zhao S. Sequencing consolidates molecular markers with plant breeding practice. Theor. Appl. Genet. 2015;1:1–17. [PubMed: 25239219]

Yen CY, Chiu CC, Chang FR, Chen JY, Hwang CC, Hsieh JC, Yang HL, Hsu YJ, Lee AW, Tsai MT, Guo ZL, Cheng YS, Li CY, Lan YH, Chang YC, Ko YC, Chang HW, Wu YC. 4beta-Hydroxywithanolide E from Physalis peruviana (golden berry) inhibits growth of human lung cancer cells through DNA damage, apoptosis and G2/M arrest. BMC Cancer. 2010; 10:46. [PubMed: 20167063]

Yu J, Zhang Z, Zhu C, Tabanai DA, Pressoir G, Tuinstra MR, Kresovich S, Todhunter RJ, Buckler ES. Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. Plant Genome. 2009; 2:63–77.

Zhu C, Gore M, Buckler ES, Yu J. Status and prospects of association mapping in plants. Plant Genome. 2008; 1:5–20.
Figure 1.
Neighbor-joining tree based on Nei’s genetic distance from the dataset of (a) 24 molecular markers across 222 plants and (b) 642 SNP markers across 64 plants. Bootstrap values (in percentage) are indicated at the main branches. Related taxa: red, *P. peruviana*: blue.
a
Figure 2.
Principal component analysis of the Cape gooseberry germplasm collection on (a) 24 molecular markers and (b) 642 SNP markers.
Figure 3.
Inferred population structure of 60 accessions based on 24 InDel markers. SNPs markers showed a similar pattern. RT=Related taxa, Pp=Physalis peruviana
| Species               | Number of accessions | Number of plants used for InDels* | Number of plants sequenced |
|-----------------------|----------------------|-----------------------------------|---------------------------|
| Physalis peruviana    | 47                   | 171                               | 51                        |
| Physalis philadelphica| 4                    | 18                                | 4                         |
| Physalis floridana    | 2                    | 9                                 | 2                         |
| Physalis angulata     | 2                    | 5                                 | 2                         |
| Physalis viscosa      | 1                    | 3                                 | 1                         |
| Physalis pruinosa     | 1                    | 3                                 | 1                         |
| Physalis isocarpa     | 1                    | 2                                 | 1                         |
| Nicandra physalodes   | 1                    | 3                                 | 1                         |
| Solanum auriculatum  | 1                    | 8                                 | 1                         |
| **Total**             | **60**               | **222**                           | **64**                    |

* InDel markers = Polymorphic COSII and IRGs on agarose gel
### Table 2
Summary of markers used for InDels studied and SNPs discovery

| Type of study | Marker type | No. of markers | No. of polymorphic markers | No. of SNPs |
|---------------|-------------|----------------|-----------------------------|-------------|
| InDels analysis | COSII       | 360            | 18                          | --          |
|               | IRGs        | 94             | 6                           | --          |
| Sub-Total     |             | 454            | 24                          | --          |
| SNPs discovery | COSII*      | 15             | --                          | 642         |
|               | IRGs*       | 18             | --                          | 642         |
| Sub-Total     |             | 33             | --                          | 642         |
| Total         |             | 487            | 24                          | 642         |

*Monomorphic markers selected for SNPs discovery
Table 3
Summary of *Polymorphic markers used for InDels study and \( \xi \) Monomorphic markers used for SNPs discovery

| Marker Type | Marker ID | Tomato Chromosome | Forward | Reverse | Amplicon Size (pb) | Annealing temperature (°C) |
|-------------|-----------|-------------------|---------|---------|--------------------|---------------------------|
| COSII       | C2_A4g38240* | 1 | aAGATTAGCcaGCAgCAAGGAAG | ATTGcGcGcAcTCTGTTGCT | 280–530 | 56 |
| COSII       | C2_A1g02140* | 3 | TcCggTATCTCAcAATTCGACC | TGGTTGCTTGCGTCACAAAC | 280–1100 | 56 |
| COSII       | C2_A5g49970* | 3 | AaTGGCGaGCaTTGATGCTGC | tCCACCATTTGCACCAGGACC | 500–1350 | 56 |
| COSII       | C2_A5g51040* | 3 | tCgCGaGATATcGTATCTTGCC | tCCAAAGACACCTGACCAGG | 210–550 | 50 |
| COSII       | C2_A5g10920* | 4 | TcTATGAgGATTCAGAAGGATC | TTGGGTATCTCAGCTCGACC | 280–1450 | 60 |
| COSII       | C2_A1g46480* | 4 | TAAGTACCTTCAAGCTGGAAG | TATcTGTCCTGCGTCACAAAC | 280–700 | 60 |
| COSII       | C2_A4g29490* | 7 | AaGAGCaAACtCAGTGGTTC | AcAAGTACCTTGCACTGACC | 500–600 | 60 |
| COSII       | C2_A2g42750* | 7 | TCaCGTGAAAAgATAGTTGATG | aCTCTcGCCTTCCAAGTCTC | 300–750 | 60 |
| COSII       | C2_A3g09920* | 9 | AaGCAATCTATcAGGACAGG | AgACCATTGCGaATATCTTGCC | 800–1650 | 56 |
| COSII       | C2_A1g53000* | 10 | AgATcTCCCGCAGCCATGATCC | ATGcTCTGCGCTTTCCAGTGTC | 750–1700 | 56 |
| COSII       | C2_A5g03905* | 10 | tGcGAGGGCGGCAAGTCTG | TCgGcGTTCTCATCAGCACTATT | 410–1400 | 56 |
| COSII       | C2_A2g27450* | 11 | tCAGATGAGaGCaGCCTGTTTC GC | TCTcAATTATCTCAGCT | 370–750 | 60 |
| COSII       | C2_A1g14790* | 5 | TGG AGA GAA CAC TAA TAT TCT CAA GG | ATG TCT TGA CCC CAG GAA AAG | 630–650 | 60 |
| COSII       | C2_A1g71260* | NA | TGT TTT ATC AGC AAC AGA GAC TGG | ACG AAA TAG CCA TCC CCA TCA GC | 280–330 | 58 |
| COSII       | C2_A2g10920* | 6 | TGG CCT GGT GTG GAC AAA GAC | TGC AAG TAG TAT GCG TGG TTCC C | 650–1500 | 58 |
| COSII       | C2_A3g18165* | NA | ATG GGC ACA AAA GAC GGC GAC | TGA TGC TGC AGT AAT CGC TGA GC | 250–450 | 58 |
| COSII       | C2_A4g23100* | NA | TGC AAG TCC TCC AAC AGA GGA TGC | AGC CCA ATA ATC TGG TCA CCC TCC | 840–900 | 60 |
| COSII       | C2_A5g51700* | 11 | AGA TGC CAC CAG GGA TTA TTT TG | AGC AGT GTG GTC ATT TTC CTT TTC | 425–500 | 56 |
| COSII       | C2_A5g58220* | 1 | AATTCAACAGCTCTGCTACCCACTGACTAG | TTcTcCtGAATCAGCATTGCC | 290 | 56 |
| COSII       | C2_A4g62010* | 1 | TACCCaTcATAcATGCGCAAC | ATGcTcATGCTTTAAGCAGCCAGTCC | 325 | 56 |
| COSII       | C2_A2g35920* | 1 | TCGTGCAACcAACCACAGCTGAG | aatgcTcTATGCTTGATGAGG | 400 | 56 |
| COSII       | C2_A4g01880* | 1 | ATGGCCCAATTCcAACACAGCCT | TcCcAGAATATGCGCTGCGTTC | 425 | 56 |
| COSII       | C2_A5g27620* | 1 | ATcTaCTATGCTGATTGGAAC | TCGTCTGCTTGCCAGCCAGTCC | 350 | 56 |
| COSII       | C2_A5g06580* | 1 | TGCTTCAACTACAGTCTGAAAG | aATGCcTATGCTTGCGAGG | 310 | 56 |
| COSII       | C2_A3g07100* | 2 | AtTCAGGATGCGCCTGATATAC | AaGACCCTAGGcTAATGCTCGTTC | 275 | 50 |
| COSII       | C2_A4g34090* | 2 | TGdATCGATGCGCCaACTGCCC | TCTTCCaCACAaCACCTGAG | 425 | 56 |
| COSII       | C2_A5g45950* | 2 | TGAGATTCAATGTTGATAGCTGCC | TGTGCGCTTAGAAAGCCTTGTGCTCAT | 490 | 56 |
| Marker Type | Marker ID | Tomato Chromosome | Forward | Reverse | Amplicon Size (pb) | Annealing temperature (°C) |
|-------------|-----------|-------------------|---------|---------|-------------------|--------------------------|
| COSII       | C2_A2g40760f | 3 | AgGAAGAgGCTtTTCATCCTGGAC | TCTtTCAATTTTCGAAAGAACAGAAGG | 290 | 56 |
| COSII       | C2_A5g41480f | 6 | TATTCGTgCTGTTGCAAGATGGC | ATGATCCTTACCGTTCACCCATAGC | 490 | 56 |
| COSII       | C2_A4g24690f | 6 | TgtCCtCCGAGGAAAGTTTGG | TCAAAATTCCTGAGCGAGTGTTT | 320 | 56 |
| COSII       | C2_A4g30580f | 7 | TGGCCgGTGCTTACGTCAC | TgaAgAcTgAaCTTACCCAAGAAG | 425 | 56 |
| COSII       | C2_A2g63190f | 9 | TTTGTCGAgCTGGAATCCTGACC | TCCACATTTCGTTGATCTACCC | 490 | 56 |
| COSII       | C2_A1g67740f | 10 | ATGTAAGAcGCGATTTGCAATC | ATCTCATCCGTAaCTCGTAAAGA | 350 | 56 |
| IRG         | PpIRG-7*   | NA | AGTCCAAGGATCGAAGAGCA | ACCTCCATGCAATTCGTTG | 240–500 | 59.95 |
| IRG         | PpIRG-24*  | NA | TCTTCTCGTCTCGTGTTG | GCAATCCTTTCGCTCAAAG | 800–850 | 59.84 |
| IRG         | PpIRG-42*  | NA | GCTCCTCGATTTGGATTA | TGCTGGAAGTAAGCTAAAC | 510–540 | 59.79 |
| IRG         | PpIRG-64*  | NA | GATTCCTCAATCGGACCAAT | TGATGTTGAGGTAAGGGA | 470–650 | 59.86 |
| IRG         | PpIRG-84*  | NA | GCCAAATACAAAGCAGCAATTT | TCCTAGCATTGCGAATTGG | 450–1000 | 60.10 |
| IRG         | PpIRG-86*  | NA | GTTTTCTCCTCCCTCCCTTCC | GCTTCAACATGCTTGGTTCC | 400–450 | 59.91 |
| IRG         | PpIRG-45*  | NA | TCCTGGGGTGGCTCAGCTTA | CCATAGAAGGAGCTGGTTTC | 480 | 59.97 |
| IRG         | PpIRG-55*  | NA | AGCAGCAAGAGAGCATGT | CACCAGGAAATTTCATGCTT | 500 | 60.02 |
| IRG         | PpIRG-30*  | NA | AATCTGGAGGCTGCAATTGAA | CATTGGAGGGTGTAAGCATGG | 310 | 59.72 |
| IRG         | PpIRG-31*  | NA | CAAGCTGGCGGCATCTACATA | AGGCATGGAAAAGCACTAC | 510 | 60 |
| IRG         | PpIRG-32*  | NA | CTGATCCTGGAGGATGGTTTCC | AGCCATGGAATGAGGTTTC | 490 | 59.65 |
| IRG         | PpIRG-33*  | NA | GATGTCGAACGCTGCAATTA | CTTCAACATCATTGACCTCA | 410 | 60.59 |
| IRG         | PpIRG-43*  | NA | AACATTGGAGGAGGTCATC | TGAAAGGCCCAGCAACTTTCA | 320 | 59.85 |
| IRG         | PpIRG-44*  | NA | TGTATGAGCTGGAGATCCTGGA | GCTCAACATTCAAGCAGCTGTA | 390 | 60.05 |
| IRG         | PpIRG-46*  | NA | TCTGCCTGCGGCAATGATGTT | GAGACCCAGACATTTCA | 520 | 59.99 |
| IRG         | PpIRG-51*  | NA | GGTGTTGCGGTATGAGGA | GCCACGGTCATGAACAGAT | 480 | 59.32 |
| IRG         | PpIRG-53*  | NA | TCCATGAGCAAAAGCAGATA | GCAATGGCGCCAGCTTAA | 520 | 59.38 |
| IRG         | PpIRG-56*  | NA | CATGGAAGGTTGGAGCAAA | AGGCTCAGAATGAGGAAAGA | 490 | 59.40 |
| IRG         | PpIRG-60*  | NA | TGGAAACATCCTCCAGCATCA | GAGGAGATGGCAAAAGCAGAA | 490 | 60.24 |
| IRG         | PpIRG-61*  | NA | GAAAGTTGCGATGTTTGGCTT | TGTCACCAGAAAGACTGTTT | 490 | 60.12 |
| IRG         | PpIRG-63*  | NA | CTTCTGATACCTCGTGGCTATT | GCCAATTCCTTTCACCAAGA | 500 | 59.85 |
| IRG         | PpIRG-79*  | NA | GGTGIACTGGAATAGGAGGA | TCCAGAAAGATGGGGATGG | 500 | 59.89 |
| IRG         | PpIRG-90*  | NA | GCCATCCAAACCGTGATTA | AGGAGAGTTGCGCTGAAAGA | 430 | 60.59 |
| Marker Type | Marker ID | Tomato Chromosome | Forward | Reverse | Amplicon Size (pb) | Annealing temperature (°C) |
|-------------|-----------|-------------------|---------|---------|-------------------|--------------------------|
| IRG         | PpIRG-93  | NA                | ATTTGCTGCATCCTGAAAAC | GTTTTCGTTCTCCGTCCA | 600                | 60.23                    |
Table 4

Summary statistics of genetic diversity calculated with 24 InDel markers and 642 SNP markers for all taxa, 463 SNPs for *P. peruviana* population and 1,600 SNPs for related taxa germplasm

| Sample       | N   | A    | He  | Ho  | PIC |
|--------------|-----|------|-----|-----|-----|
| All Taxa     | 222 | 5.04 | 0.42| 0.47| 0.37|
| InDel P. peruviana | 171 | 2.75 | 0.30| 0.48| 0.24|
| Related Taxa | 51  | 4.75 | 0.59| 0.4 | 0.54|
| All Taxa     | 64  | 2    | 0.33| 0.38| 0.26|
| SNPs P. peruviana | 51  | 2    | 0.41| 0.59| 0.32|
| Related Taxa | 13  | 2    | 0.25| 0.12| 0.21|

N = sample size; a = mean number of alleles per locus; He = Unbiased expected heterozygosity; Ho = observed heterozygosity; PIC = Polymorphic information content.
### Table 5

AMOVA analysis for the 222 and 64 plant population based on InDel and SNP markers for the whole collection

| Source of variation | d.f. | SS   | MS   | Variance component | Total variance percentage | P-value |
|---------------------|------|------|------|--------------------|--------------------------|---------|
| **Indel markers**   |      |      |      |                    |                          |         |
| Among Populations   | 1    | 258.2| 258.2| 1.607              | 23%                      | 0.01    |
| Among Accessions    | 220  | 1,255.9| 5.7 | 0.314              | 4%                       |         |
| Within Accessions   | 222  | 1,128.0| 5.0 | 5.081              | 73%                      |         |
| Total               | 437  | 2,642.0| 7.002|                    | 100%                     |         |
| **SNPs markers**    |      |      |      |                    |                          |         |
| Among Populations   | 1    | 555.3| 555.3| 13.7               | 5%                       | 0.01    |
| Within Accessions   | 62   | 16,719.8| 269.6| 269.6              | 95%                      |         |
| Total               | 63   | 17,275.1| 283.4|                    | 100%                     |         |