Supplementary Note 1

Assembly and annotation of the *Petunia* genomes.

Aureliano Bombarely [1], Michel Moser [2], Noe Fernandez-Pozo [3], Lukas A. Mueller [3], Cris Kuhlemeier [2] and Thomas L. Sims[4]

Affiliations

1- Department of Horticulture, Virginia Tech, Blacksburg, VA 24061 USA.
2- University of Bern, Altenbergrain 21, CH-3013 Bern, Switzerland.
3- Boyce Thompson Institute for Plant Research, 533 Tower Rd, Ithaca, NY 14853-1801, USA.
4- Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115-2861, USA.

EXTENDED MATERIAL AND METHODS FOR GENOME ASSEMBLY AND ANNOTATION

Plant Material and DNA preparation

Petunia inflata S6 was obtained from the University of Florida in 1974 and maintained by inbreeding (for >15 generations at Free University of Amsterdam and University of Amsterdam). *Petunia axillaris* N was provided by the Botanical Garden of Rostock to the Botanical Garden of Bern and deposited in the Amsterdam collection under the designation of *P. axillaris* S26. Plants used for DNA extraction were grown axenically in tissue culture containers. Mature plants (leaves and stems) were harvested, flash-frozen in liquid nitrogen and stored at -80C until used for DNA extractions. Plant material (approximately 15 g) was extracted using a modification of methods designed to isolate high molecular weight DNA from nuclei (Fischer and Goldberg, 1982; Carrier et al., 2011). Briefly, frozen plant material was homogenized in a blender with liquid nitrogen until a fine powder was obtained. Powdered material was thawed in 1X SEB plus mercaptoethanol (10 mM Tris pH 8.0, 100 mM KCl, 10 mM Na_2EDTA, 0.5 M sucrose, 4 mM spermidine, 1 mM spermine, 0.13% carbamic acid, 0.25% PVP-40, 0.2% β-mercaptoethanol), then filtered through Nitex mesh. Triton X-100 was added to 0.5% and nuclei isolated and washed by repeated low speed centrifugation and washing with SEB. Nuclei were lysed by adding an equal volume of NLB (2% Sodium N-lauryl sarcosine, 40 mM Na_2EDTA, 0.1 M Tris-HCl pH 8.0 and 1mg/ml proteinase K) followed by incubation at 55 C for 1 hour. Cesium chloride was added to 50% w/w along with ethidium bromide to a final concentration of 0.4%. DNA gradients were centrifuged in a 70.1 Ti rotor at 40,000 rpm for 36 hours followed by re-banding in a Vti65.2 rotor at 60,000 rpm for 6 hours. Ethidium bromide was removed by extraction with SSC-saturated isopropanol and the remaining solution dialyzed against TNE (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 0.1 mM EDTA) for 24 hours. DNA was precipitated, washed with 70% ethanol, dried and resuspended in EB (10 mM Tris pH 8.0) to a final concentration of 150 µg/ml.

For the PacBio sequencing the DNA was prepared in a different manner. Briefly, genomic DNA from *Petunia axillaris* subsp. *axillaris* N was extracted from sterile plant cultures. The plants were grown at 16 h photoperiod and constant 20 °C. Young folded leaves from 3 plants were collected and ground in liquid nitrogen. Genomic DNA was extracted using a CTAB protocol (Clarke, 2009) and DNA was treated with RNase digestion and an extra ethanol washing step before library preparation.

For the BAC PacBio sequencing, the BAC DNA from a *Petunia axillaris* subsp. *axillaris* N library (in house, unpublished) was isolated with columns from the QIAGEN Large-Construct kit (QIAGEN Redwood City, CA, USA) following the manual supplied by the QIAGEN Large-Construct kit.

DNA Sequencing

Illumina library preparation and sequencing

Illumina reads were produced in two phases. In the first one, BGI-Zhenzhen prepared four pair-ends libraries with insert sizes of 170, 350, 500 and 800 bp and two mate-pair libraries with insert sizes of 2 and 5 Kb with TruSeq library preparation kit (Illumina, San Diego, CA, USA) and sequenced in a HiSeq1000 system (2x100). In the second one, three more libraries were prepared with the Nextera DNA library prep. kit (Illumina, San Diego, CA, USA) and sequenced at University of Illinois Roy J. Carver Biotechnology Center, one pair-end library with an insert size of 1 Kb, sequenced as 2x150 in a HiSeq2000 Illumina system and two mate-pair libraries with insert sizes of 8 and 15 Kb sequenced as 2x100 in a HiSeq2500 Illumina system **(Supplementary table 1)**.

Petunia axillaris PacBio library preparation and sequencing

High molecular weight genomic DNA was sheared in a Covaris g-TUBE (Covaris, Woburn, MA, USA) to obtain 18 Kb fragments. After shearing the DNA size distribution was checked on a DNA Fragment Analyzer (Advanced Analytical Technologies, Ames, IA, USA). The sheared DNA was used to prepare a SMRTbell library with the PacBio DNA Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA,

USA) according to the manufacturer's recommendations. The recovered library was sequenced with P4/C2 chemistry and MagBeads on a PacBio RSII system (Pacific Biosciences, Menlo Park, CA, USA) at 120 min movie length. The library was sequenced with a total of 62 SMRT Cells. Sequencing produced 7,501,054 reads longer than 500 bp and a total of 21x coverage (29 .1 Gb). The average read length was 3,888 bp and maximal read length reached 24,597 bp.

Petunia axillaris BAC Library Preparation and Sequencing
BAC DNA from 7 BAC clones and a total of 40 microgram genomic DNA was sent on ice to the Lausanne Genomic Technologies Facility (LGTF) for library preparation and sequencing. Per BAC clone, a single library was prepared. High molecular weight DNA from BACs was sheared in a Covaris g-TUBE (Covaris, Woburn, MA, USA) to obtain 10-20 Kb fragments. After shearing the DNA size distribution was checked on a DNA Fragment Analyzer (Advanced Analytical Technologies, Ames, IA, USA). The sheared DNA was used to prepare a SMRTbell library with the PacBio DNA Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA) according to the manufacturer's recommendations. The recovered library was sequenced with XL/C2 chemistry and MagBeads on a PacBio RSII system (Pacific Biosciences, Menlo Park, CA, USA) at 120 min movie length using one SMRT Cell per BAC clone.

Read processing, sequence assembly and draft genome evaluation
Read processing
Illumina reads were processed in three different steps: 1) Quality filtering and adapter removal using Fastq-mcf (Aronesty, 2013) with a minimum quality score of 30 (-q 30) and a minimum length of 50 bp (-l 50). 2) Read pair duplications filtering using Prinseq (Schmieder and Edwards, 2011). 3) Read error correction using Musket with the default parameters (Liu et al., 2013). Pacbio reads were processed using the SMRT Analysis pipeline (v.2.0.1). PacBio reads were filtered for a minimum length of 500 bp and minimum quality score of 0.70.

Genome assembly
Illumina reads were assembled using SOAPdenovo2 (Luo et al., 2012) with different k-mer sizes (23, 31, 39, 47, 55, 63, 71, 79, 87). Different assembly results were compared using its corresponding total assembly size, N50/L50 and N90/L90, selecting the assembly with the best stats (kmer=79 for both species, *Petunia axillaris* N and *P. inflata* S6). Gaps between contigs were completed using GapCloser, from the SOAPdenovo2 package with the default parameters (Luo et al., 2012). For the *P. axillaris* assembly, where Pacbio reads were also available, the assembly was broken in its contigs using a Perl script, BreakScaffolds (https://github.com/aubombarely/GenoToolBox/blob/master/SeqTools/BreakScaffolds) and then a new assembly was performed using AHA (Bashir et al., 2012) to combine Pacbio reads and the contigs produced by the Illumina read assembly (4 iterations using only reads longer than 3000 bp). Gaps produced by the hybrid assembly were filled using PBJelly (English et al., 2012b), and then new scaffolding using the Illumina pair ends and mate pairs was performed using SSPACE (Boetzer et al., 2011). Finally a last round of gap filling was performed using PBJelly (English et al., 2012a).

Genome size estimation
Genome size estimation was performed through the k-mers abundance distribution (Li et al., 2010) with the Illumina 1000 paired end libraries. Jellyfish was used to count k-mers (Marçais and Kingsford, 2011) with a kmer of 31. K-mer distribution was analyzed with R.

Organelle genomes
Chloroplastic and mitochondrial PacBio sequences from *P. axillaris* were filtered out using the chloroplast genomes of S*olanum lycopersicum* LA3032 (GI:84371962) and *Nicotiana tabacum* (GI:76559634) and mitochondrial DNA of *Nicotiana tabacum* (GI:56806513) with the alignment tool BLASR (Chaisson and Tesler, 2012). Filtered reads were assembled separately using HGAP (Chin et al., 2013). The chloroplast genome of *P. axillaris* contains 158,794 bp and was assembled in a single

contig with a GC content of 37.7 %. The mitochondrial genome of *P. axillaris* has a total length of 440,114 bp, fragmented over 10 contigs.

Petunia axillaris BAC Clone Assembly
De novo assembly of the P. *axillaris* BAC inserts was performed using the SMRT Analysis (v.2.0.1) pipeline. The complete BAC clones were assembled with HGAP followed by consensus calling with Quiver (Chin et al., 2013). The resulting contig was blasted against the vector sequence (pCC1BAC) and the identified vector sequence was cut out to retain the BAC insert. PacBio reads were mapped back to the BAC insert using BLASR (Chaisson and Tesler, 2012). Alignments were reviewed for possible miss-assemblies by visualizing read depth in IGV (Robinson et al., 2011). The 7 BAC clones could be assembled into single contigs each of sizes between 90 Kb and 155 Kb. As 2 pairs of BAC clones were extracted from the BAC library using the same marker, their sequence could be overlaid and merged. The final 5 BAC clone sequences had an average length of 170,478 bp and a total length of 852,393 bp.

Genome quality assessment
Gene space completeness of the assembly was evaluated running CEGMA v2.5 (Parra et al., 2007) with the default parameters. We mapped the 248 Core Eukaryotic Genes (CEGs) to assess the completeness of both assemblies and found 239 (94%) and 243 (98 %) in the assembly of *P. axillaris* and *P. inflata*, respectively.

BAC sequences were used to evaluate genome accuracy of P*. axillaris* assembly. Similarity searches were conducted using BLASTN (McGinnis and Madden, 2004) in ungapped mode with an E-value cut-off of 1E-10. The percent nucleotide identity was calculated by matcher from EMBOSS tools version 2.0u4 (www.ebi.ac.uk/Tools/psa/emboss_matcher/nucleotide.html). The five BAC sequences mapped against the *P. axillaris* scaffolds showed an average similarity of 95 % (Supplementary figure 1).

Heterozygosity Evaluation
The heterozygosity of the sequenced genomes was estimated mapping the Illumina reads to the assembly using Bowtie2 (Langmead and Salzberg, 2012), calling SNPs using FreeBayes (Garrison and Marth, 2012) with a minimum mapping score of 20 and a minimum read depth of 5. SNPs were annotated using SnpEff (Cingolani et al., 2014). Heterozygosity was calculated following two methodologies: A- Number of heterozygous variants divided by the whole genome sequence with a minimum read depth of 5 (total heterozygosity). The total heterozygosity was 0.27 and 0.34 variants/Kb (0.03%) for *P. axillaris* N and *P. inflata* S6 genomes respectively; B- Using Vcftools (Danecek et al. 2011) to calculate the variants/Kb in 10 Kb genomic bins (119,652 *P. axillaris* and 124,348 *P. inflata* bins). In this case, regions with read mapping depth below 5 and over 2 times the read mapping depth mean (56.7 for *P. axillaris* and 52.7 for *P. inflata*) were removed from analysis to avoid possible paralog regions collapsing problems yielding 117,797 and 122,481 bins respectively. The estimated heterozygosity calculated by 10 Kb bins was 0.25 and 0.37 variants/Kb (0.03 and 0.04 %) respectively. The distribution was represented in the Supplementary figure 2.

Genome annotation
Structural annotation
The genome structural annotation was performed using Maker (Cantarel et al., 2007). This program creates the gene model annotation using *ab-initio* gene predictions programs such as Augustus (Stanke et al., 2006) and SNAPP (Korf, 2004) integrating this results with experimental data such as ESTs, RNAseq and protein alignments. Augustus was trained with 400 *P. axillaris* N gene models curated manually using Web Apollo (Lee et al., 2013), experimental data from 454 (Zenoni et al., 2011) and Illumina RNASeq reads from several tissues and developmental stages (Supplementary table 2), and compared with protein alignments from the tomato genome project, ITAG2.4 (Tomato Genome Consortium, 2012) and SwissProt Solanaceae proteins set (Magrane and Consortium, 2011).

RNAseq Illumina data was mapped using Tophat2 and assembled with Cufflinks (Trapnell et al., 2009). The same datasets were also used as experimental supporting data for Maker. Additionally Maker annotates the repeat content. A previous repeat analysis was done using RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html). tRNA were annotated using tRNAscan (Lowe and Eddy, 1997). After the automatic annotation, more than 500 gene models were structurally curated using Web Apollo (Lee et al., 2013) including the genes described in the different sections of this publication. The resulting annotation contained a total of 35,812 and 39,408 genes, thereof 32,928 and 36,697 being protein-coding genes for *P. axillaris* and *P. inflata*, respectively (Supplementary table 3). As non-coding RNAs, 546 rRNAs and 2884 tRNAs and 558 rRNAs and 2711 tRNAs were predicted for *P. axillaris* and *P. inflata*, respectively.

Functional annotation

The gene functional annotation was performed in two steps: 1) Functional description association by sequence homology search from different sources such as protein datasets from SwissProt (Magrane and Consortium, 2011), arabidopsis genome annotation version TAIR10 (Lamesch et al., 2012) and tomato genome annotation version ITAG2.3 (Tomato Genome Consortium, 2012) and GenBank (Benson et al., 2013) using BlastX (McGinnis and Madden, 2004) and protein domains using InterProScan (Mulder and Apweiler, 2007). 2) Functional annotation filtering, scoring and integration using AHRD (https://github.com/groupschoof/AHRD). Swissprot, TAIR10, ITAG2.3 and GenBank were scored 100, 50, 50 and 30 respectively. 58, 73 and 88% of the proteins showed at least one hit with a minimum e-value of 1e-20 with Swissprot, TAIR10 and GenBank respectively. 83% presented at least one protein domain from the InterPro database. Gene Ontology (GO) annotations were inferred from the protein domain hits.

Supplementary table 1

Sequencing machine	Insert-size [bp]	Read length / type	P. axillaris Processed reads coverage (X)*	P. inflata Processed reads coverage (X)*
Illumina HiSeq1000	170	2x100 bp / paired-end	17	22
Illumina HiSeq1000	350	2x100 bp / paired-end	12	6
Illumina HiSeq1000	500	2x100 bp / paired-end	10	25
Illumina HiSeq1000	800	2x100 bp / paired-end	12	15
Illumina HiSeq1000	2000	2x100 bp / mate-pair	11	15
Illumina HiSeq1000	5000	2x100 bp / mate-pair	8	20
Illumina HiSeq2500	1000	2x150 bp / paired-end	48	NA
Illumina HiSeq2500	8000	2x150 bp / mate-pair	10	17
Illumina HiSeq2500	15000	2x150 bp / mate-pair	9	15
Total short reads:	-	-	137	135
PacBio RS 2	-	-	21	NA

*see section Read processing for details about read processing

Supplementary table 2

Species	Tissue/ Develop. Stage	Treatment	Reads	SRA Accession	Publication
P. axillaris[1]	Mixed tissues	NA	578,107	SRX036998	Zenoni et al. 2011
P. axillaris	Floral buds	NA	34,150,939	SRX1402727	Sheehan et al. 2015
P. axillaris	Floral buds	NA	44,731,386	SRX1402609	Sheehan et al. 2015
P. axillaris	Floral buds	NA	52,405,543	SRX1402587	Sheehan et al. 2015
P. axillaris	Trichomes	NA	23,620,517	SRX710299	Guo et al. 2015
P. axillaris	Seedling	NA	32,715,526	SRX710174	Guo et al. 2015
P. axillaris	Apical shoots	NA	35,221,375	SRX709906	Guo et al. 2015
P. axillaris	Mature flower	NA	27,178,746	SRX709957	Guo et al. 2015
P. axillaris	Callus	NA	28,505,785	SRX710298	Guo et al. 2015
P. integrifolia[1]	Mixed tissues	NA	602,753	SRX036999	Zenoni et al. 2011
P. integrifolia	Trichomes	NA	23,957,225	SRX711430	Guo et al. 2015
P. integrifolia	Seedling	NA	25,099,736	SRX711429	Guo et al. 2015
P. integrifolia	Apical shoots	NA	30,675,565	SRX711427	Guo et al. 2015
P. integrifolia	Mature flower	NA	32,220,783	SRX711426	Guo et al. 2015
P. integrifolia	Callus	NA	28,961,576	SRX711385	Guo et al. 2015
P. inflata	Mixed tissues[5]	NA	83,739,290	In preparation	No published
P. inflata	Mixed tissues[5]	NA	41,077,858	In preparation	No published
P. inflata	Mixed tissues[5]	NA	71,909,258	In preparation	No published
P. inflata	Mixed tissues[5]	NA	71,334,072	In preparation	No published
P. inflata	Mixed tissues[5]	NA	92,583,881	In preparation	No published
P. hybrida[2]	Mature leaves	Control NaCl 0h	14,299,136	SRX1530795	Villarino et al. 2014
P. hybrida[2]	Mature leaves	Control NaCl 0h	23,084,764	SRX1530797	Villarino et al. 2014
P. hybrida[2]	Mature leaves	Control NaCl 0h	18,140,264	SRX1530799	Villarino et al. 2014
P. hybrida[2]	Mature leaves	Control NaCl 6h	21,904,108	SRX1530801	Villarino et al. 2014
P. hybrida[2]	Mature leaves	Control NaCl 6h	13,362,088	SRX1530803	Villarino et al. 2014
P. hybrida[2]	Mature leaves	Control NaCl 6h	14,459,178	SRX1530805	Villarino et al. 2014
P. hybrida[2]	Mature leaves	Control NaCl 24h	12,733,646	SRX1530807	Villarino et al. 2014
P. hybrida[2]	Mature leaves	Control NaCl 24h	14,364,874	SRX1530809	Villarino et al. 2014
P. hybrida[2]	Mature leaves	Control NaCl 24h	16,181,450	SRX1530811	Villarino et al. 2014
P. hybrida[2]	Mature leaves	150 mM NaCl 6h	16,771,012	SRX1530813	Villarino et al. 2014
P. hybrida[2]	Mature leaves	150 mM NaCl 6h	16,458,846	SRX1530814	Villarino et al. 2014
P. hybrida[2]	Mature leaves	150 mM NaCl 6h	18,340,684	SRX1530815	Villarino et al. 2014
P. hybrida[2]	Mature leaves	150 mM NaCl 24h	14,368,884	SRX1530822	Villarino et al. 2014
P. hybrida[2]	Mature leaves	150 mM NaCl 24h	16,404,210	SRX1530823	Villarino et al. 2014
P. hybrida[2]	Mature leaves	150 mM NaCl 24h	16,508,928	SRX1530825	Villarino et al. 2014
P. hybrida[3]	Mature petals	NA	15,259,555	In preparation	In preparation
P. hybrida[3]	Mature petals	NA	9,829,522	In preparation	In preparation
P. hybrida[3]	Mature petals	NA	11,170,745	In preparation	In preparation
P. hybrida[3]	Mature petals	NA	11,523,112	In preparation	In preparation
P. hybrida[3]	Mature petals	NA	13,949,735	In preparation	In preparation
P. hybrida[3]	Mature petals	NA	8,617,423	In preparation	In preparation
P. hybrida[3]	Mature petals	NA	11,569,546	In preparation	In preparation
P. hybrida[3]	Mature petals	NA	15,817,998	In preparation	In preparation
P. hybrida[4]	Mature petals	NA	13,060,169	In preparation	In preparation
P. hybrida[4]	Mature petals	NA	11,883,830	In preparation	In preparation

Notes: All the sequenced libraries were Illumina HiSeq2000 Pair Ends except for (1) that were 454 libraries publically available at NCBI SRA. The Petunia hybrida sequenced cultivars were Mitchell (2), R27 (3) and R143 (4). For more details about the plant growth, RNA extraction, library preparation and sequencing, consult the corresponding publication. For the libraries were no publication was specified, the plants were growth in standard conditions of light and temperature (long day, 22C). RNA was extracted with RNeasy Qiagen kit. (5) For library produced with mixed tissues, RNA for the most representative tissues (leaves, roots, flowers, stems…) were mixed to maximize the gene space sequenced.

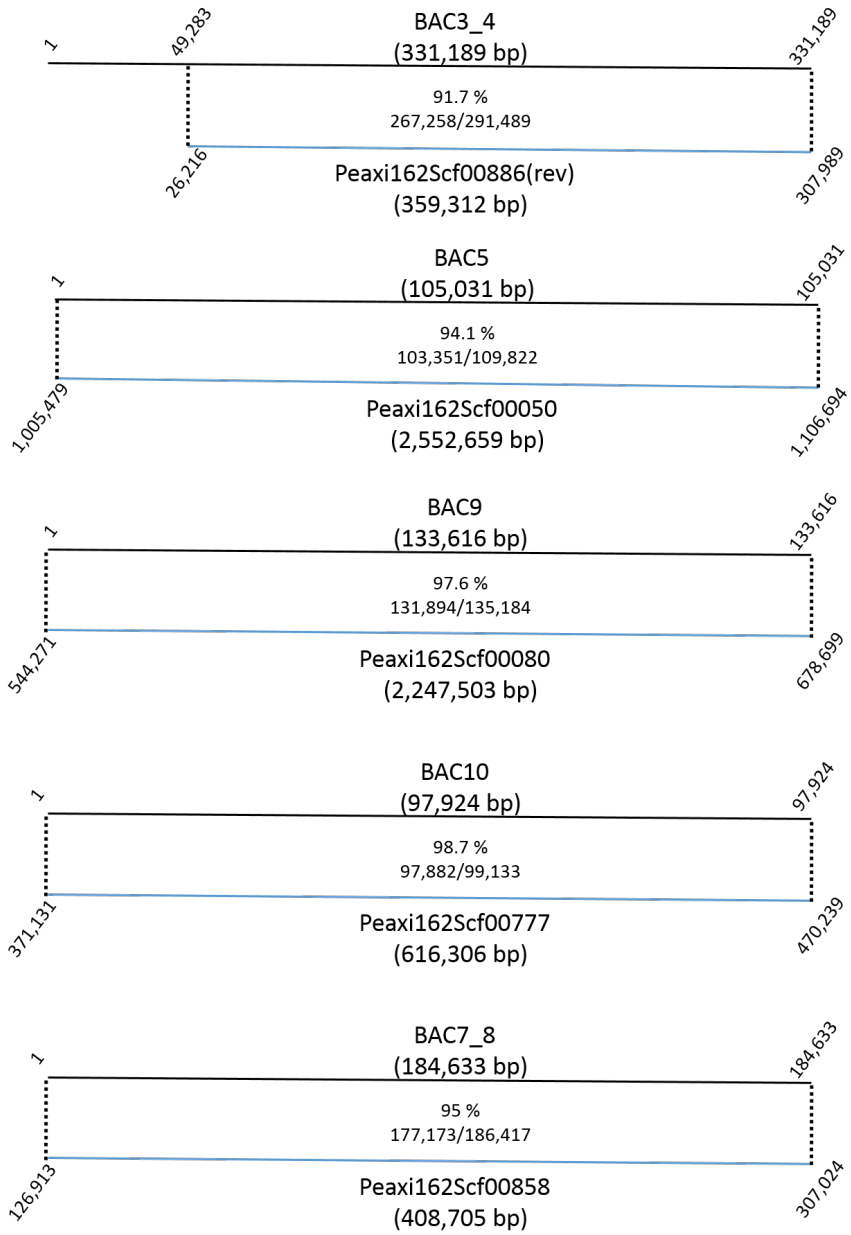
Supplementary table 3

Protein-coding gene statistics	*P. axillaris*	*P. inflate*	*S. lycopersicum*
Gene number	32,928	36,697	34,725
Average gene length [bp]	4252	4152	3162
Exon number	173,712	188,372	160,001
Average exon number per gene	5.27	5.13	4.61
Average exon length [bp]	238.2	238.0	261.4
Total exon length [bp]	41,387,658	44,837,367	41,821,567
Total CDS length [bp]	38,480,559	42,303,603	35,813,852
Average exonic length per gene [bp]	1256.9	1221.8	1204.4
Intron number	138,743	150,011	125,276
Total intron length [bp]	98,287,394	106,892,756	67,748,290
Average intron length [bp]	708.4	712.6	540.8

Protein coding gene statistics of the two Petunia assemblies in comparison with S. lycopersicum (ITAG 2.4)

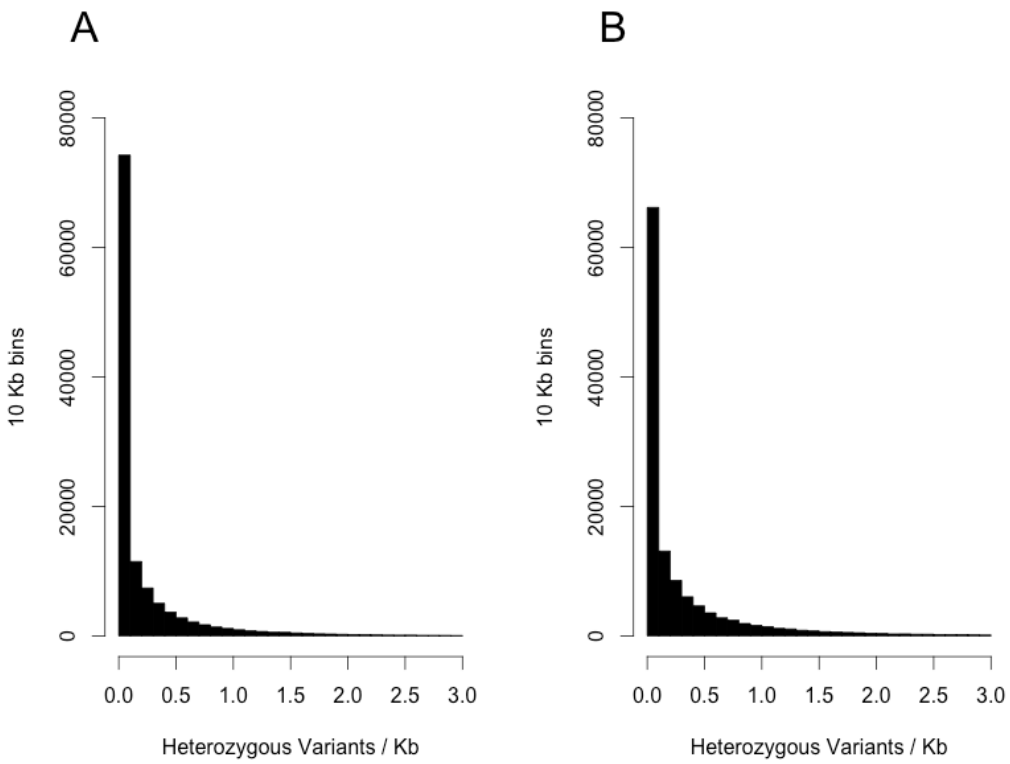
Supplementary Figure 1

BAC sequences were mapped against the scaffolds of the *P. axillaris* assembly using BLAST v. Black lines and blue lines indicate the BAC inserts and the *P. axillaris* scaffolds, respectively. Numbers along the lines indicate the start and end of sequences and their overlapping regions which are connected by dotted lines. Percent nucleotide identity for each homologous region is indicated along with the length of the matching basepairs out of the total alignment length.



Supplementary Figure 2

Heterozygosity distribution for 10 Kb genomic bins for: A- 117,797 bins for *P. axillaris N* genome assembly v1.6.2 (Variants/Kb mean = 0.25); B- 122,481 bins for *P. inflata* S6 genome assembly v1.0.1 (Variants/Kb mean = 0.37).



REFERENCES

Aronesty, E. (2013). Comparison of sequencing utility programs. Open Bioinform J. 7:1-8

Bashir, A. et al. (2012). A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol **30**: 701–707.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2013). GenBank. Nucleic Acids Res **41**: D36–42.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27**: 578–579.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2007). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res **18**: 188–196.

Carrier, G., Santoni, S., Rodier-Goud, M., Canaguier, A., Kochko, A. de, Dubreuil-Tranchant, C., This, P., Boursiquot, J.-M., and Le Cunff, L. (2011). An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods. Am J Bot **98**: e13–5.

Chaisson, M.J. and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. Bmc Bioinformatics **13**: 238.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., Turner, S.W., and Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Meth **10**: 563–569.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2014). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) **6**: 80–92.

Clarke, J.D. (2009). Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation. Cold Spring Harb Protoc **2009**: pdb.prot5177–pdb.prot5177.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R., Lunter G., Marth G., Sherry S.T., McVean G., Durbin R. and 1000 Genomes Project Analysis Group (2011). The Variant Call Format and VCFtools. Bioinformatics **27**:2156-2158.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., and Gibbs, R.A. (2012a). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE **7**: e47768.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., and Gibbs, R.A. (2012b). Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE **7**: e47768.

Fischer, R.L. and Goldberg, R.B. (1982). Structure and flanking regions of soybean seed protein genes. Cell **29**: 651–660.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv:**1207**.3907

Guo, Y., Wiegert-Rininger, K.E., Vallejo, V.A., Barry, C.S. and Warner, R.M., 2015. Transcriptome-enabled marker discovery and mapping of plastochron-related genes in Petunia spp. *BMC genomics*, *16*(1), p.726.

Korf, I. (2004). Gene finding in novel genomes. Bmc Bioinformatics **5**: 59.

Lamesch, P. et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res **40**: D1202–10.

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth **9**: 357–359.

Lee, E., Helt, G.A., Reese, J.T., and Munoz-Torres, M.C. (2013). Web Apollo: a web-based genomic annotation editing platform. Genome Biology. **14**:R93 doi:10.1186/gb-2013-14-8-r93

Li, R. et al. (2010). The sequence and de novo assembly of the giant panda genome. Nature **463**: 311–317.

Liu, Y., Schroder, J., and Schmidt, B. (2013). Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinformatics **29**: 308–315.

Lowe, T.M. and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25**: 955–964.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., and He, G. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience **1**: 18.

Magrane, M. and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) **2011**: bar009–bar009.

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

McGinnis, S. and Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. **32** (Web Server issue):W20-5

Mulder, N. and Apweiler, R. (2007). InterPro and InterProScan. In Comparative Genomics, Methods in Molecular Biology. (Humana Press: Totowa, NJ), pp. 59–70.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**: 1061–1067.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol **29**: 24–26.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics **27**: 863–864.

Sheehan, H., Moser, M., Klahre, U., Esfeld, K., Dell'Olivo, A., Mandel, T., Metzger, S., Vandenbussche, M., Freitas, L. and Kuhlemeier, C., (2015). MYB-FL controls gain and loss of floral UV absorbance, a key trait affecting pollinator preference and reproductive isolation. *Nature genetics*. doi:10.1038/ng.3462

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res **34**: W435–W439.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**: 635–641.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**: 1105–1111.

Villarino, G.H., Bombarely, A., Giovannoni, J.J., Scanlon, M.J. and Mattson, N.S., (2014). Transcriptomic analysis of Petunia hybrida in response to salt stress using high throughput RNA sequencing. *PloS one*, *9*(4), p.e94651.

Zenoni, S., D'Agostino, N., and Tornielli, G.B. (2011). Revealing impaired pathways in the an11 mutant by high-throughput characterization of Petunia axillaris and Petunia inflata transcriptomes. The Plant Journal 68(1):11-27