# SUPPLEMENTAL INFORMATION

## The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives

Riccardo Aversano[a], Felice Contaldi[a], Maria Raffaella Ercolano[a], Valentina Grosso[a], Massimo Iorizzo[a], Filippo Tatino[a], Luciano Xumerle[b], Alessandra Dal Molin[b], Carla Avanzato[b], Alberto Ferrarini[b], Massimo Delledonne[b], Walter Sanseverino[c], Riccardo Aiese Cigliano[c], Salvador Capella-Gutierrez[d,e], Toni Gabaldón[d,e,f], Luigi Frusciante[a], James M. Bradeen[g], Domenico Carputo[a,1]

[a] Department of Agricultural Sciences, University of Naples Federico II, Via Università 100, 80055 Portici, Italy;
[b] Center of Functional Genomics, Department of Biotechnologies, University of Verona, Strada le Grazie 15, 37134 Cà Vignal, Italy;
[c] Sequentia Biotech srl, Campus UAB (CRAG building) Bellaterra, Cerdanyola del Vallès, 08193 Barcelona, Spain;
[d] Center for Genomic Regulation, Dr. Aiguader, 88. 08003, Barcelona, Spain;
[e] Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain;
[f] Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.
[g] Department of Plant Pathology and Stakman-Borlaug Center for Sustainable Plant Health, University of Minnesota, 495 Borlaug Hall, 1991 Upper Buford Circle, Saint Paul, MN, USA

**Table of content**

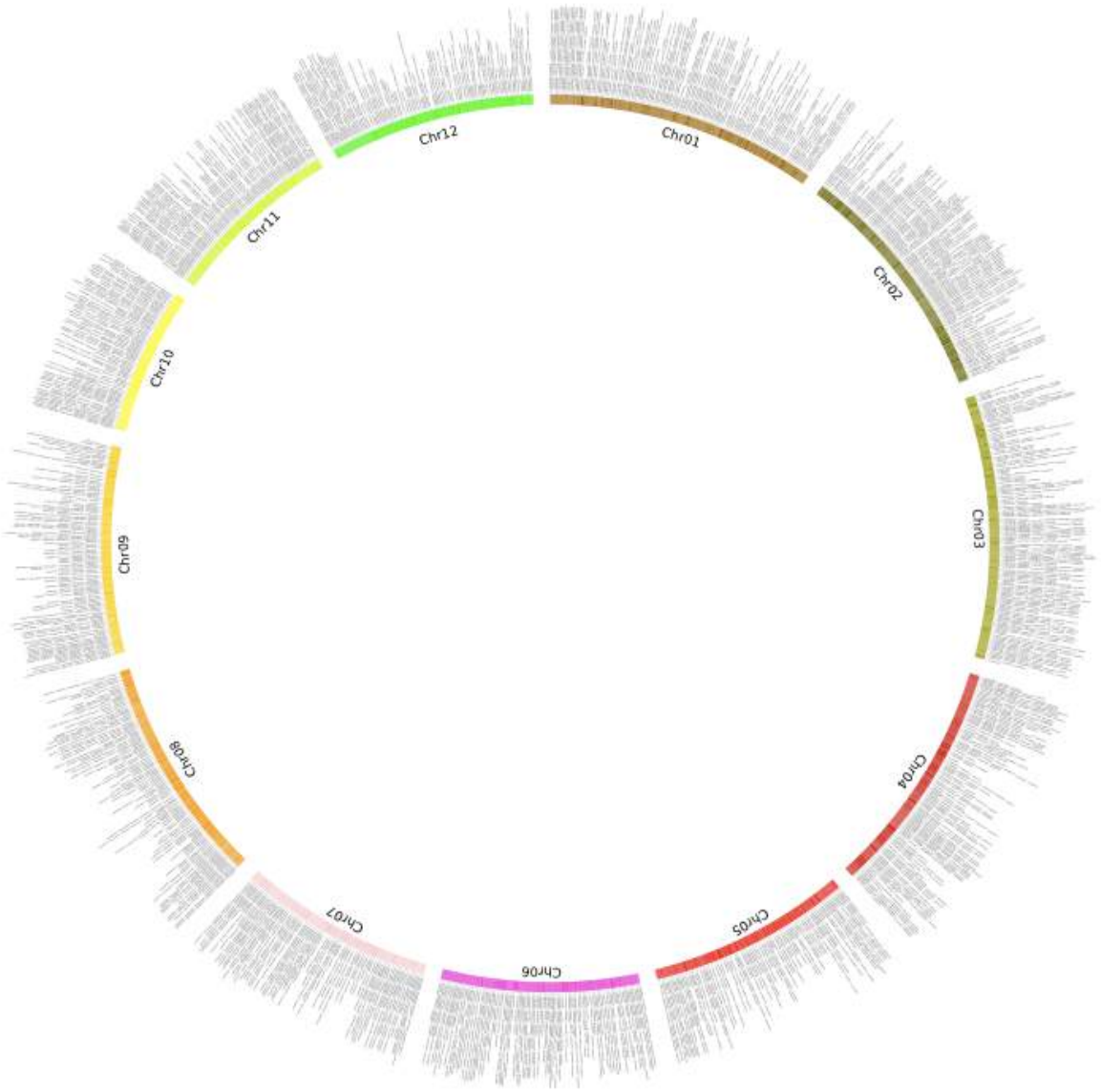
**Supplemental Figures** ..... **3**
**Supplemental Figure 1**. Ideograms of the 12 pseudochromosomes of *S. commersonii* ..... 3
**Supplemental Figure 2**. Phenotype and cytogenetic analysis ..... 4
**Supplemental Figure 3**. Distribution of Illumina 23-kmer frequency ..... 5
**Supplemental Figure 4**. Distribution of gap length within the scaffold assembly ..... 6
**Supplemental Figure 5**. Percentage of Core Eukaryotic Genes (CEGs) mapping on *S. commersonii* draft genome ..... 7
**Supplemental Figure 6**. SNP spacing in the *S. commersonii* genome ..... 8
**Supplemental Figure 7**. Proportion of transcriptome mapping to genome assembly ..... 9
**Supplemental Figure 8**. Functional annotation of *S. commersonii* transcriptome ..... 10
**Supplemental Figure 9**. R1 cluster in *S. commersonii* and *S. tuberosum* ..... 11
**Supplemental Figure 10**. Cold responsive genes annotation analysis ..... 12
**Supplemental Figure 11**. Common and differentially expressed genes between AC and NAC conditions ..... 13
**Supplemental Figure 12**. TFs with known DNA binding domains ..... 14
**Supplemental Figure 13**. CBF1 (A) and CBF2 (B) protein alignments ..... 15
**Supplemental Figure 14**. CBF1 (A) and CBF2 (B) protein alignments ..... 16

**Supplemental tables** ..... **17**
**Supplemental Table 1**. Summary of sequence read statistics of the mate pair and paired-end libraries used in WGS sequencing ..... 17
**Supplemental Table 2**. Summary of the *S. commersonii* genome assembly ..... 18
**Supplemental Table 3**. CG content in *S. commersonii* genome ..... 19
**Supplemental Table 4**. Heterozygosity in *S. commersonii* genome ..... 20
**Supplemental Table 5**. Annotation of SNPs detected in *S. commersonii* ..... 21
**Supplemental Table 6**. SINE families in *S. commersonii*. ..... 22
**Supplemental Table 7.** *De novo* assembled transcripts ..... 23
**Supplemental Table 8**. Micro RNA statistics ..... 24
**Supplemental Table 9**. Putative miRNA precursors showing miRNA/miRNA* duplexes and similarity to know miRNAs ..... 25
**Supplemental Table 10**. Transcripts annotated as responsive to cold stress and of their potential miRNA regulators ..... 27
**Supplemental Table 11**. Overview of the species used for the comparative genomics analyses ..... 29
**Supplemental Table 12.** Detected one-to-one orthologs between a given species and *S. commersonii* ..... 30
**Supplemental Table 13**. Number of duplication events detected in single gene trees according to their relative ages ..... 31
**Supplemental Table 14**. Functional enrichment analysis results after removing redundancy for the 10 biggest clusters of specifically expanded clusters of proteins in *S. commersonii* with statistically significant enriched functional terms ..... 32
**Supplemental Table 15**. Enrichment of functional categories among differentially expressed genes in nonacclimated (NAC, *) and acclimated (AC, **) conditions ..... 33
**Supplemental Table 16**. Number of non-redundant protein families annotated with the Gene Ontology term cold acclimation (CA), cellular response to cold (CRTC), and proteins as response to cold (RTC) and related number of proteins in *A. thaliana* and *S. tuberosum* ..... 35

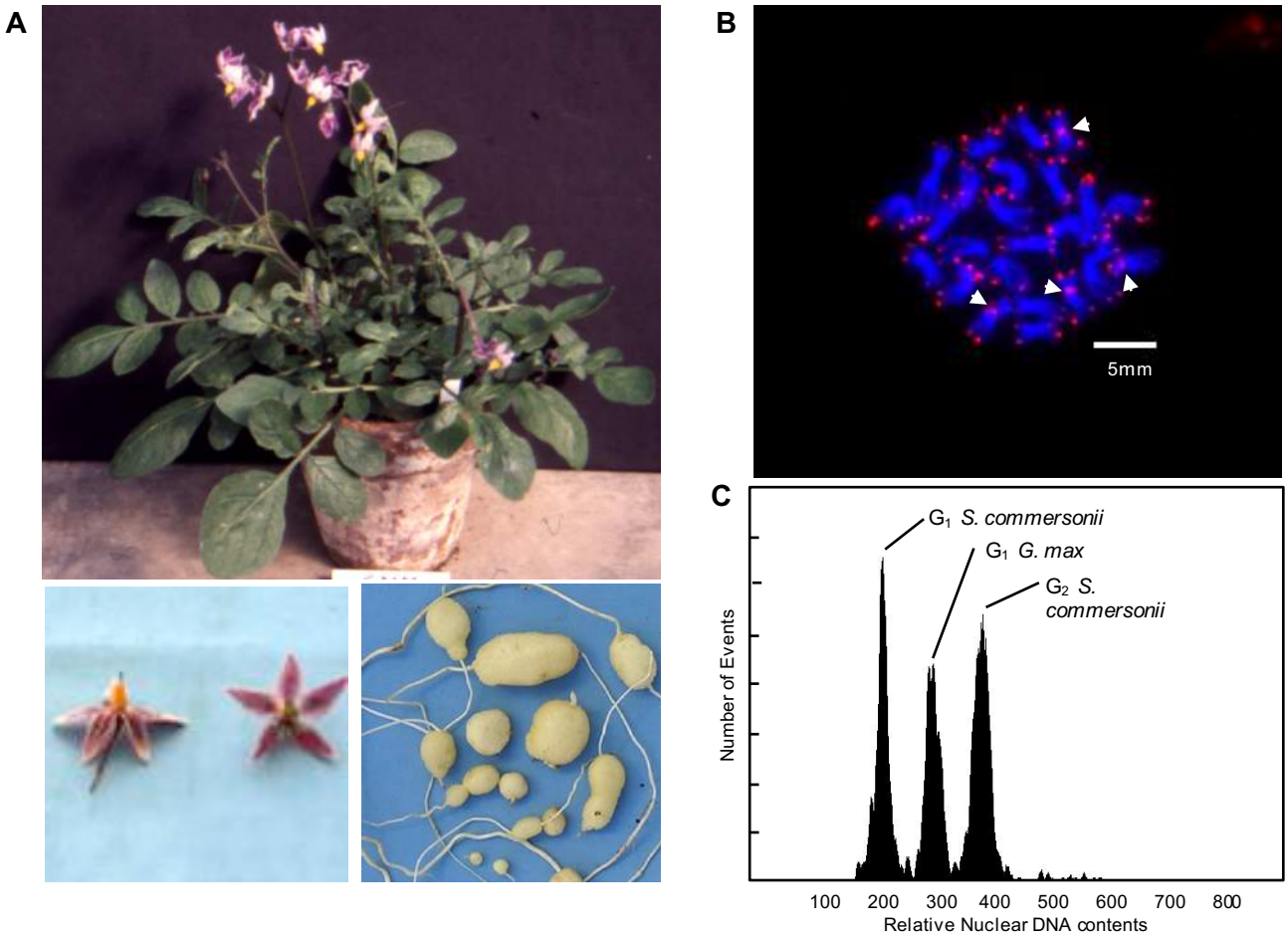
**Supplemental methods.** ..... **36**

**Additional references** ..... **41**

## Supplemental Figures




**Supplemental Figure 1.** Ideograms of the 12 pseudochromosomes of *S. commersonii* (in Mb scales).

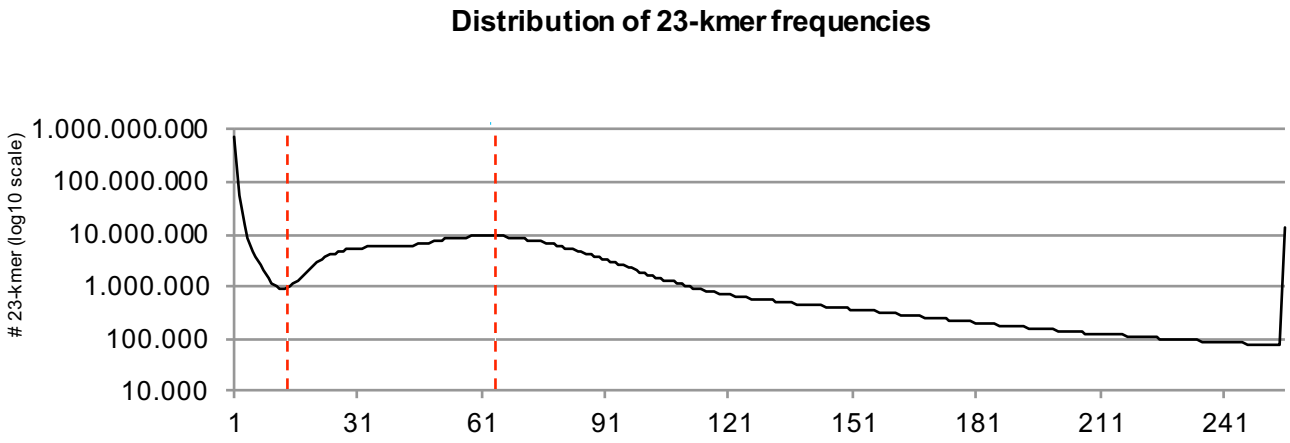



**Supplemental Figure 2.** Phenotype and cytogenetic analysis of *S. commersonii*.

*A. S. commersonii*, clone cmm1t (PI243504) whole plant, flowers and tubers.

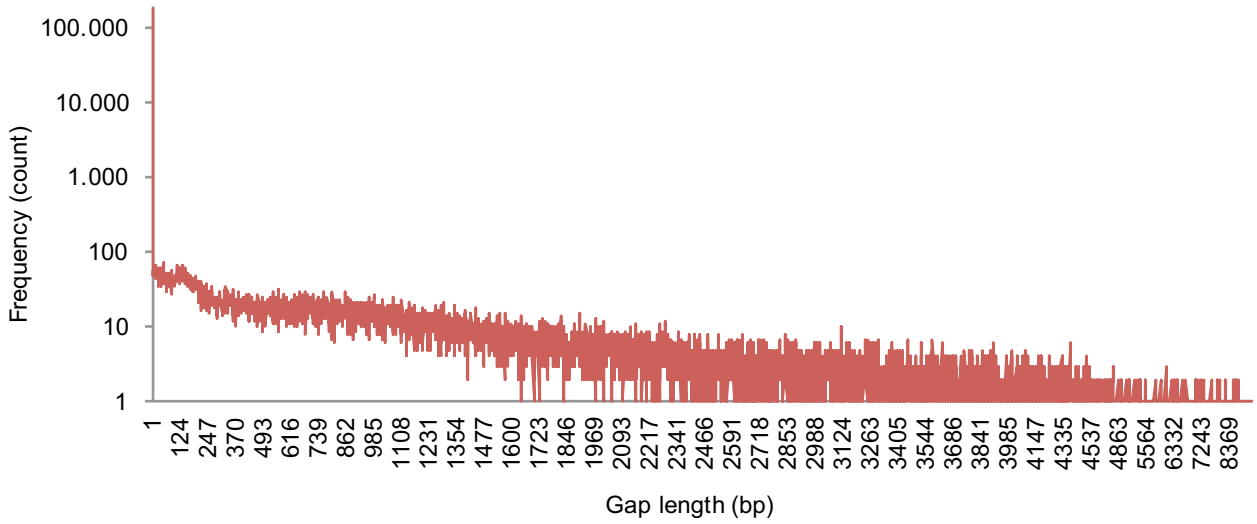
B. Fluorescence in situ hybridization in *S. commersonii* using a telomeric DNA probe. The mitotic metaphase chromosomes were stained in blue by DAPI (4',6-Diamidino-2-phenylindole). The telomeric probe, a $(TTTAGGG)_4$ oligonucleotide labeled at the 5'-end with carboxytetramethylrhodamine (TAMRA), generated signals at the ends of each chromosome (in red). In addition, interstitial telomeric repeats were detected in the pericentromeric regions of at least four chromosomes (white arrows). Photo kindly provided by Dr. Marina Iovene.

C. Estimation of absolute nuclear DNA amount (genome size) in *S. commersonii*. The histogram of relative DNA content was obtained after flow cytometric analysis of propidium iodide-stained nuclei of *S. commersonii* and *Glycine max*, which were isolated, stained and analysed simultaneously. Soybean (*Glycine max* 'Polanka', 2C= 2.50 pg DNA) served as internal reference standard. The absolute DNA amount of *Solanum commersonii* was calculated based on the values of $G_1$ peak means as follow: ($G_1$ peak means *S. commersonii*/ $G_1$ peak means of *G. max*) × *G. max* DNA content. Genome size of the *S. commersonii* was estimated to be 830 Mb.

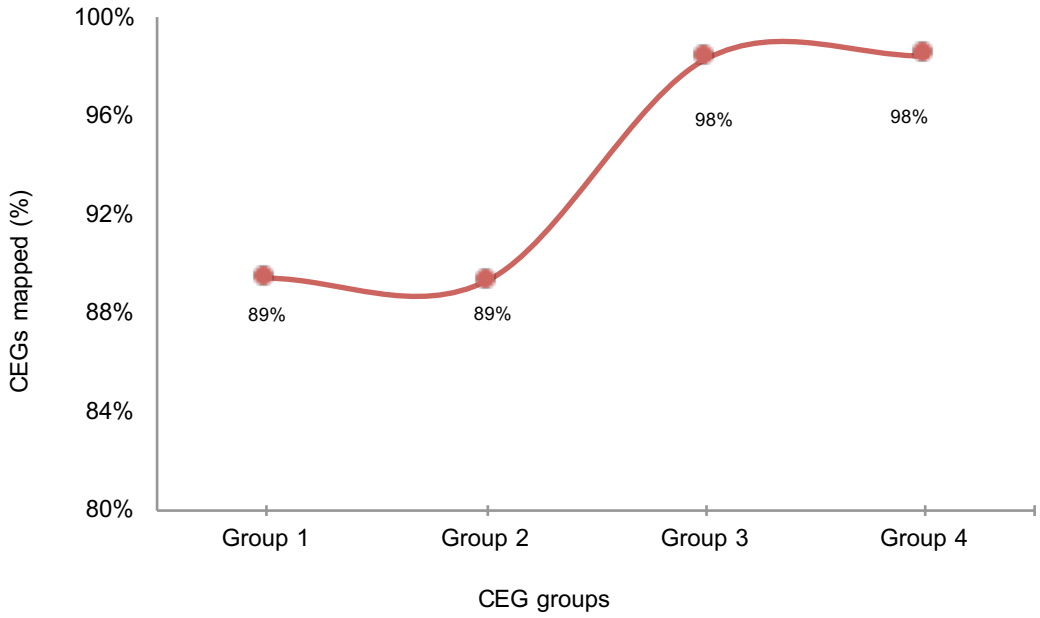



**Supplemental Figure 3.** Distribution of Illumina 23 k-mer frequency for *S. commersonii*.

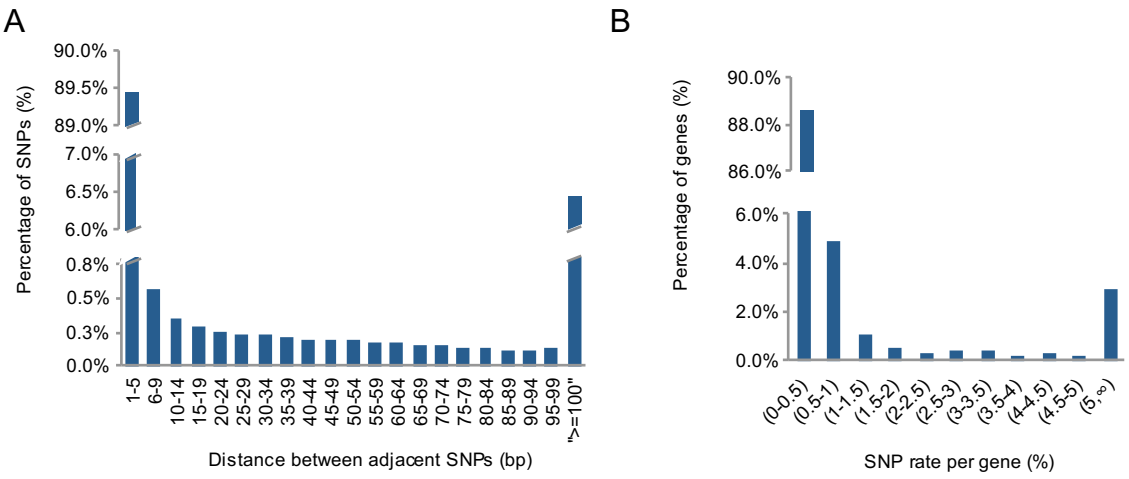
The volume of K-mers is plotted against the frequency at which they occur. The left-hand, truncated, peak at low frequency and high volume represents K-mers containing essentially random sequencing errors, while the right-hand distribution represents proper (putatively error-free) data. The total K- mer number is 54.703.986.536, and the volume peak is 64. The genome size can be estimated as (total K-mer number)/(the volume peak), which is 838 Mb.




**Supplemental Figure 4**. Distribution of gap length within the scaffold assembly of *S. commersonii*.



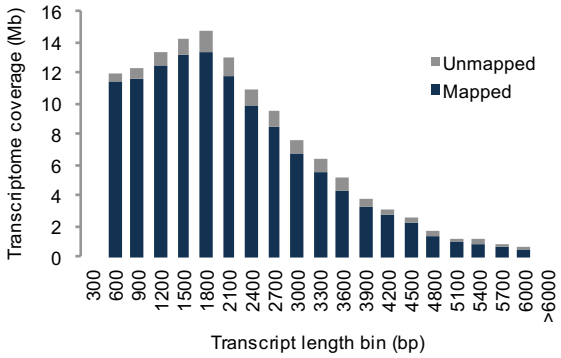

**Supplemental Figure 5.** Percentage of Core Eukaryotic Genes (CEGs) mapping on the *S. commersonii* draft genome. Group 1 represents the least conserved genes while Group4 the most conserved. Overall, 233 out of the 248 CEGs were detected (94%).



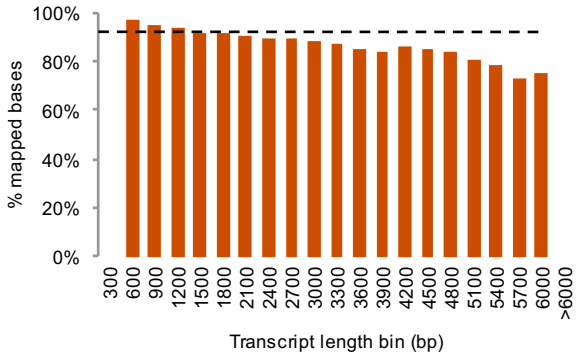

**Supplemetal Figure 6.** SNP spacing in the *S. commersonii* genome.

A. The distribution of distance between SNPs
B. SNP frequency per gene

A

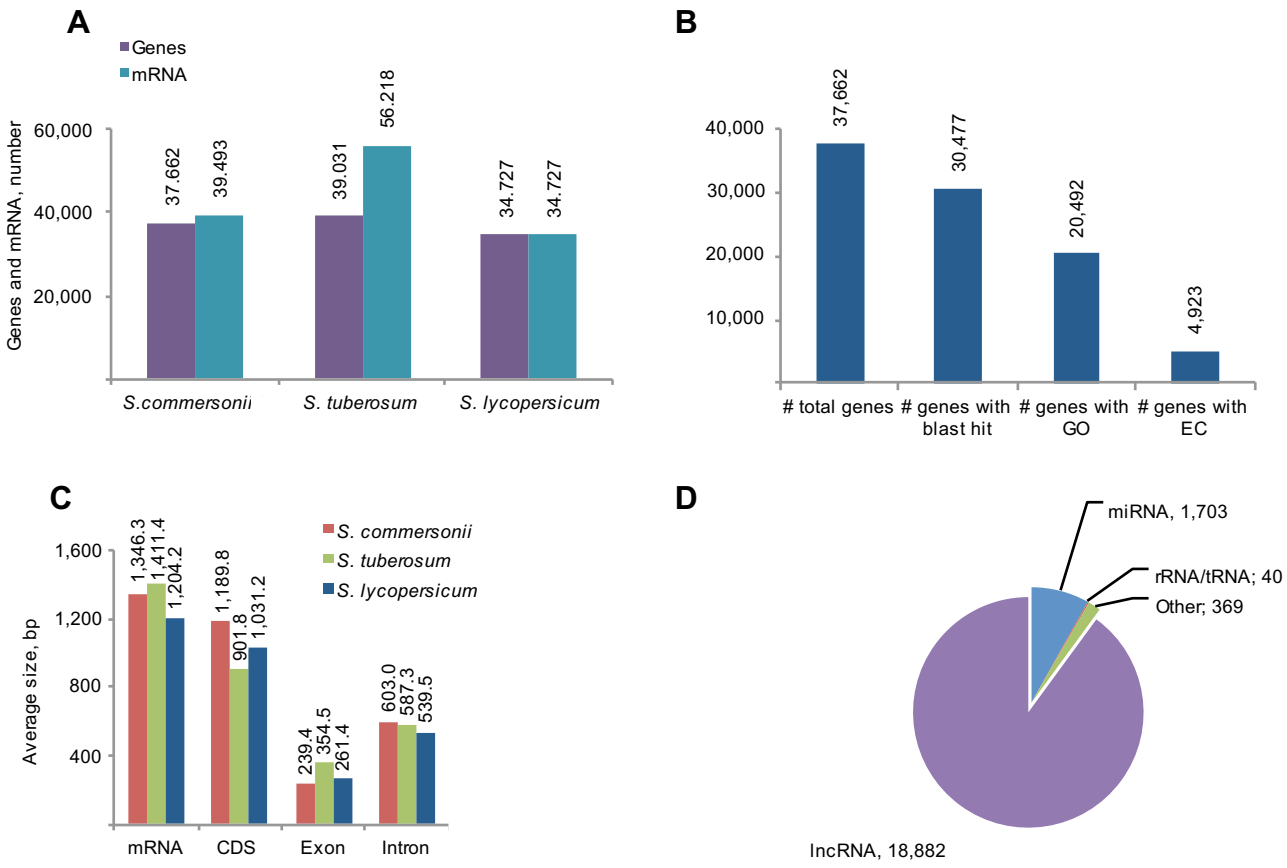



B



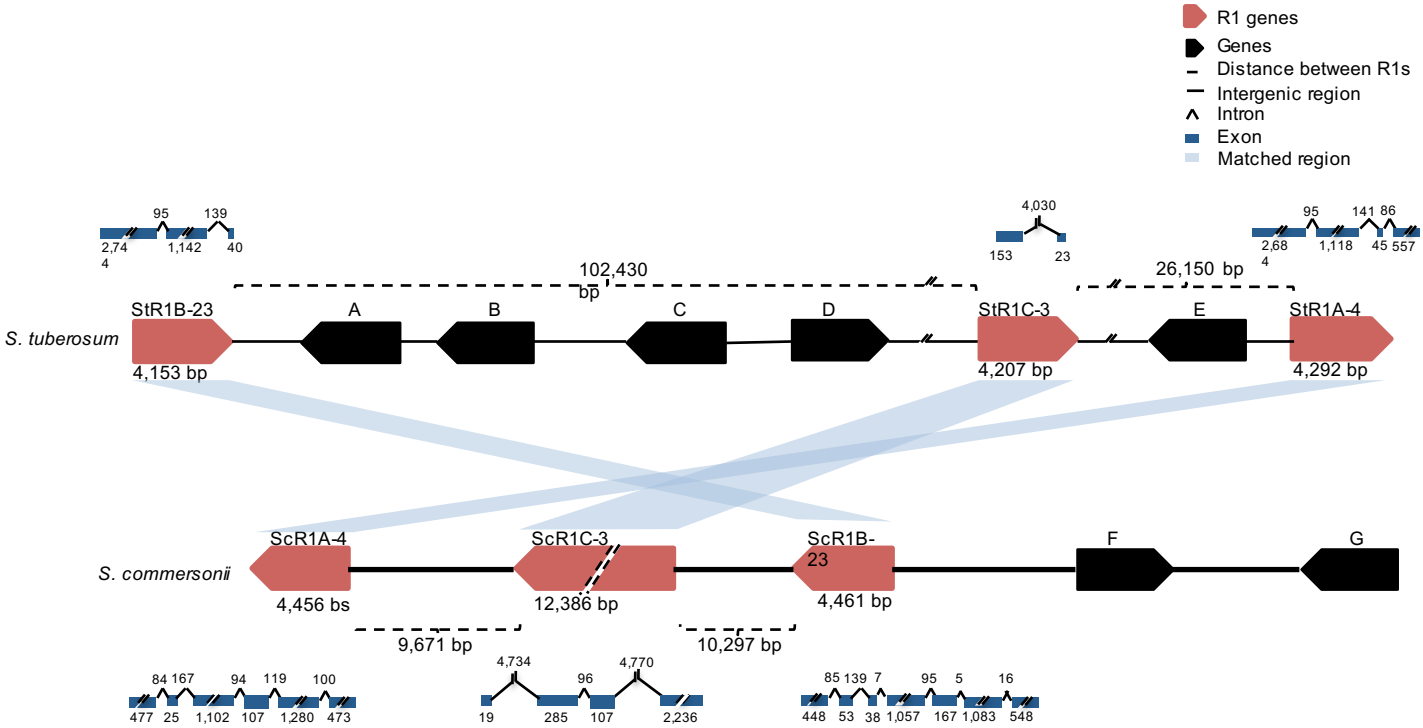

**Supplemental Figure 7.** Proportion of transcriptome mapping to *S. commersonii* genome assembly.

A. A histogram showing the number of bases in the transcript assembly that could be mapped to the genome at 98% sequence identity, as a function of transcript length in 300 bp bins.
B. The proportion of transcriptome bases that could be mapped to the genome for the same bins listed in (A). The black dashed line indicates the proportion of the transcriptome that is accounted for in the genome assembly.

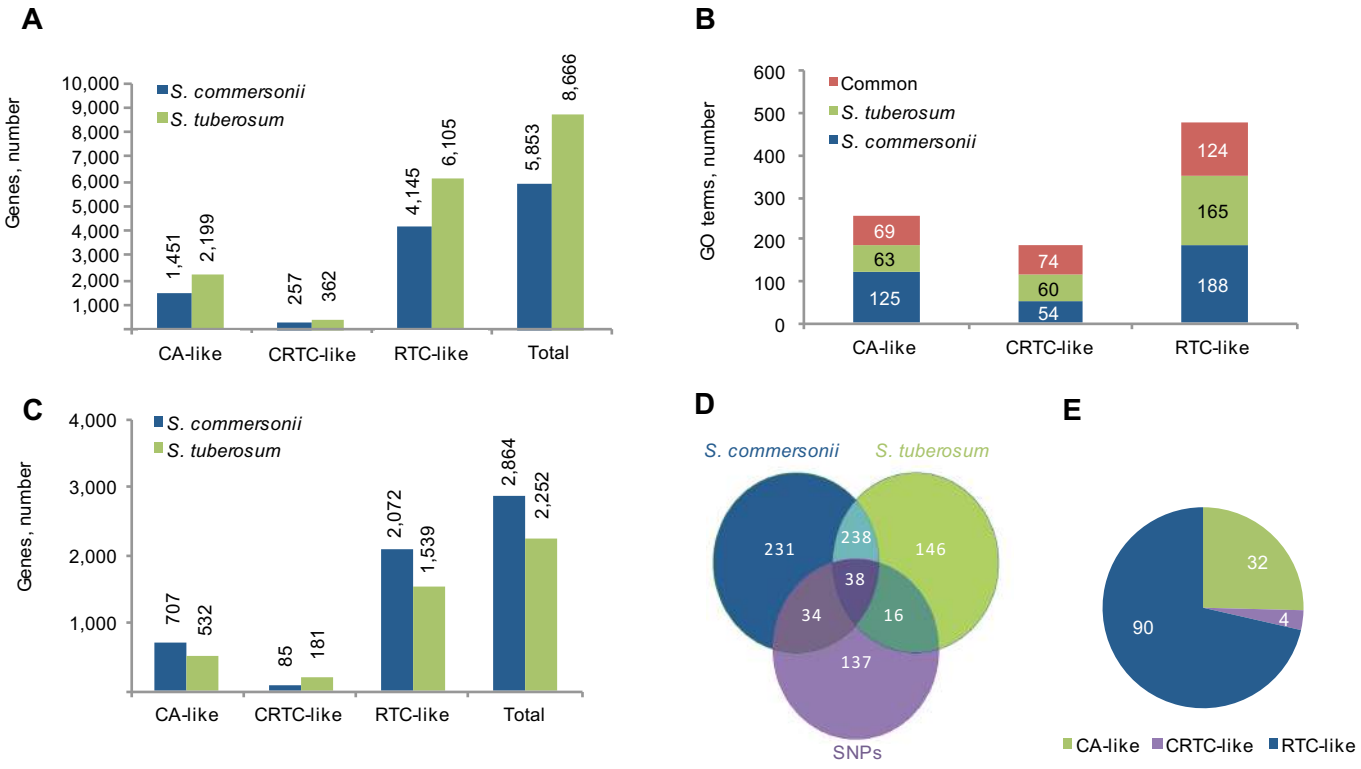



**Supplemental Figure 8**. Functional annotation of *S. commersonii* transcriptome.

A. Comparison of gene (AED≤0.5) and mRNA numbers in *S. commersonii*, *S. tuberosum* and *S. lycopersium*.
B. Number of predicted protein-encoding genes with significant BLAST similarity, with GO annotation and with a 4-digit EC number.
C. mRNA, CDS, exon and intron average size in *S. commersonii*. The mean number of exons and intron per gene are reported as well.
D. Non-coding RNA gene classes in *S. commersonii*, including long non-coding RNA (lncRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA (miRNA). Small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) were included in "other" category.



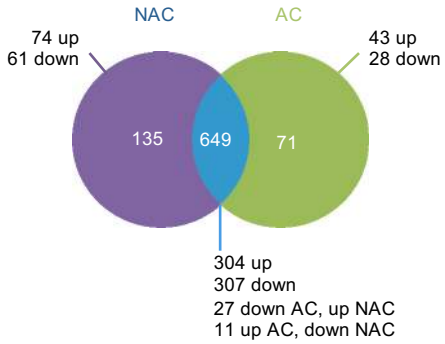

**Supplemental Figure 9**. R1 cluster in *S. commersonii* and *S. tuberosum.*
R1-gene homologues and genes are indicated in red and black filled oriented boxes, respectively. Numbers below the R1 homologue boxes indicate their length (bp). For each R1 homologue intron-exon structure is shown. Intergenic regions are drawn as thicker solid lines, whereas thick dashed lines indicate distance between R1-gene homologues. Blue-shaded areas between *S. tuberosum* and *S. commersonii* genotypes designate homology among R1 sequences. Figure not drawn to scale.



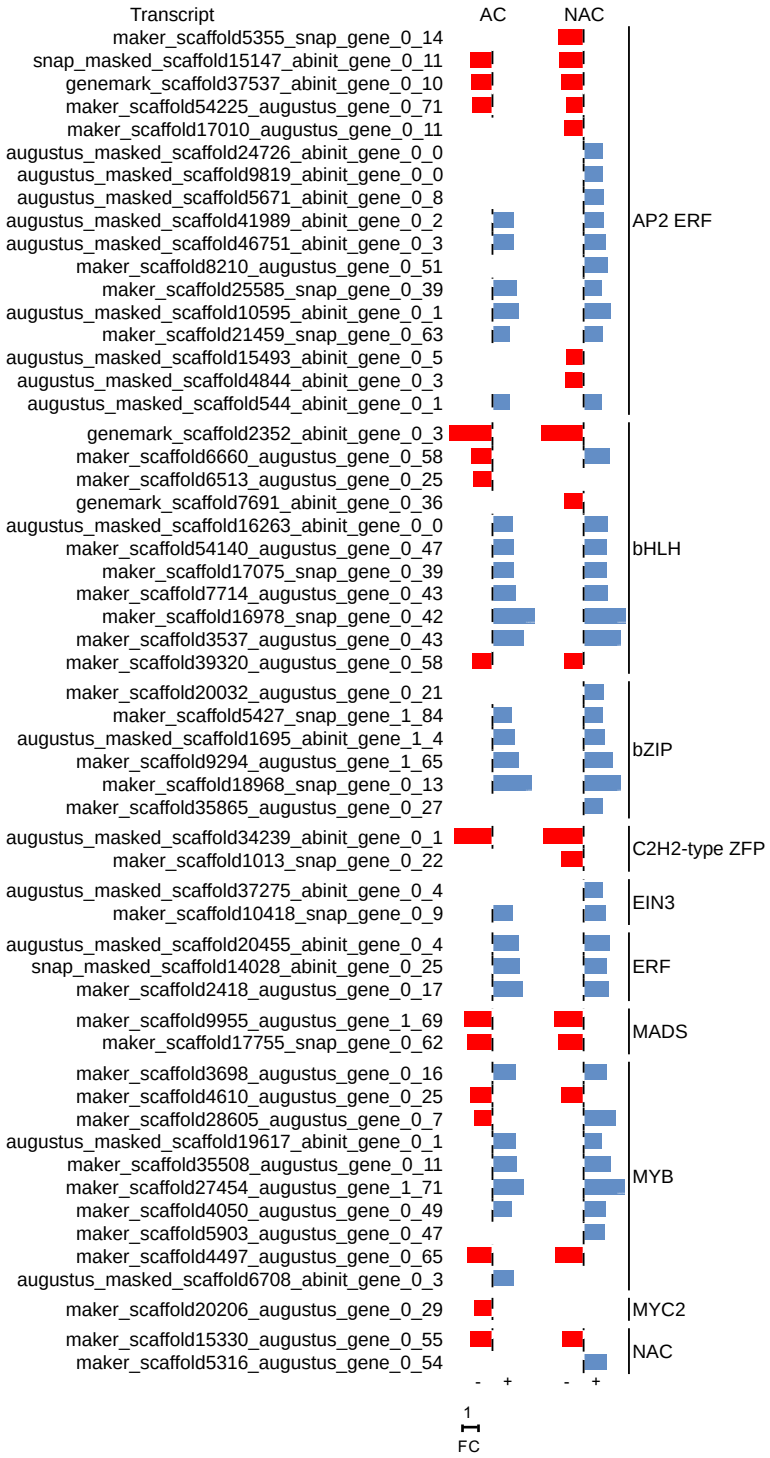

**Supplemental Figure 10.** Cold responsive genes annotation analysis.

To annotate putative cold resistance genes, a set of reference proteins was selected from *Arabidopsis thaliana*. CA: Cold Acclimation; CRTC: Cellular Response To Cold; RTC: Response To Cold.

A. Number of genes having putative binding sites for transcription factors related to responsive to cold.
B. Results of enrichment GO analysis.
C. Number of genes with unique GO term in *S. commersonii* and *S. tuberosum.*
D. Cold-responsive GO Terms significantly enriched (FDR < 0.05) in genes containing SNPs both in *S. commersonii* and *S. tuberosum.*
E. Number of unique genes involved in tolerance to cold in *S. commersonii*.



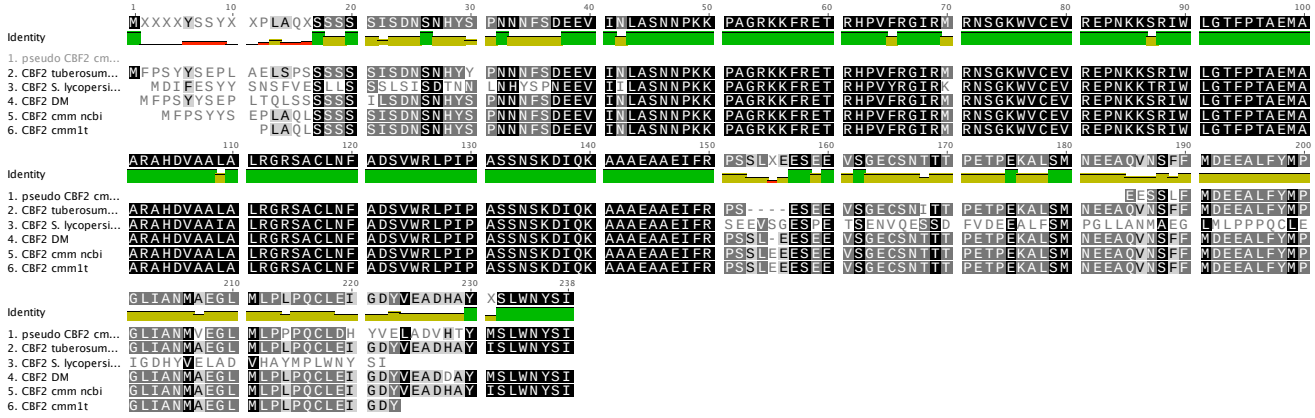

**Supplemental Figure 11.** Common and differentially expressed genes between AC and NAC conditions.



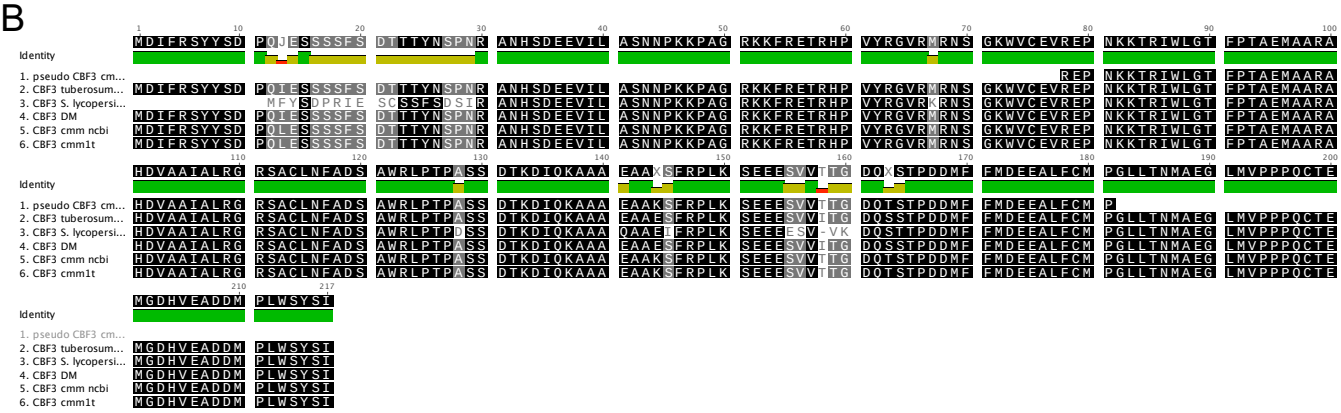

**Supplemental Figure 12.** Transcription Factors with known DNA binding domains

For each transcript the down- or up-regulation under NAC and AC are reported as red or blue bar, respectively. AP2, APETALA2; ERF, ethylene-responsive element binding factor; EIN: ethylene-insensitive; ERF: ethylene responsive factor; MYC: v-myc avian myelocytomatosis viral oncogene homolog; NAC: no apical meristem; C2H2: Cys2His2 (C2H2)-type zinc fingers; ZPF: zinc finger protein; bZIP: basic Leucine Zipper; bHLH: basic helix-loop-helix leucine zipper; MYB: myeloblastosis: MADS: Mcm1-Agamous-Deficiens-SRF domains.

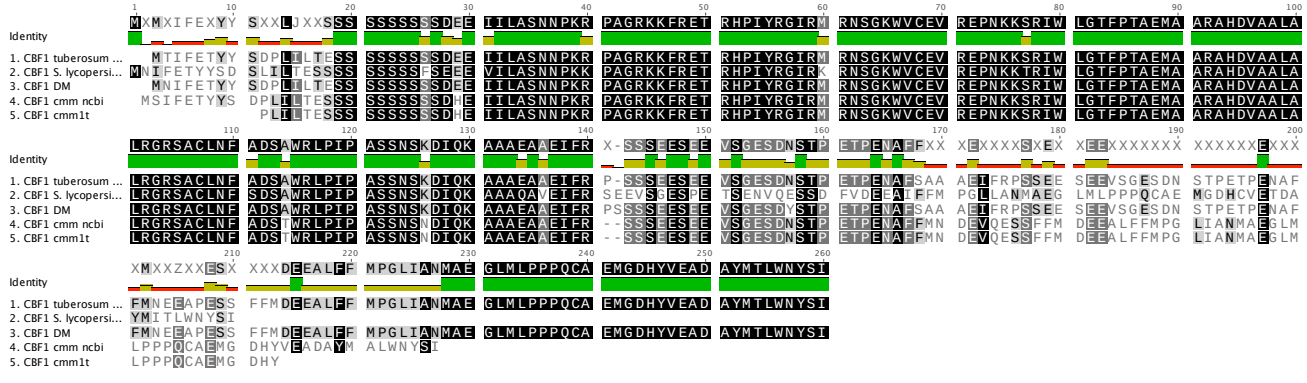
A



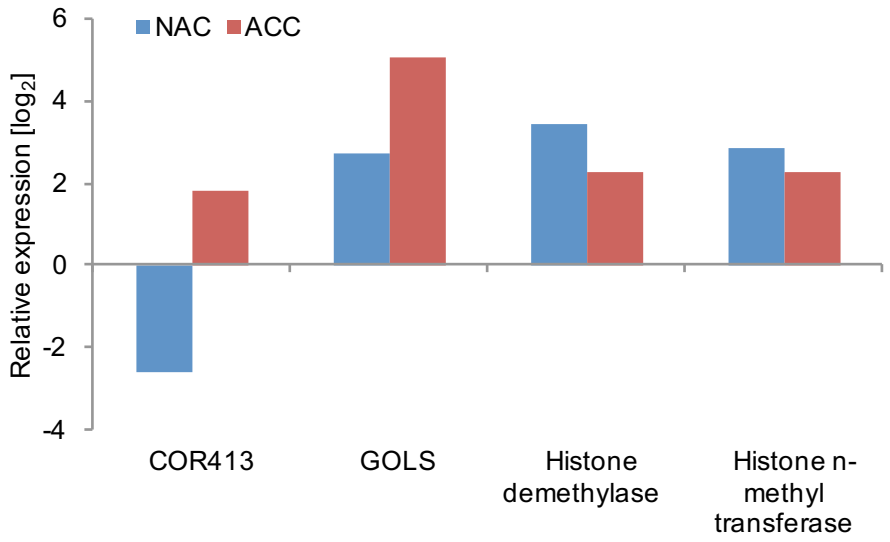
B



C



**Supplemental Figure 13**. Comparison between CBF2 (A), CBF3 (B) and CBF1 (C) protein sequences of *S. commersonii* (clone cmm1t) and the orthologous sequences of *S. commersonii* (NCBI, Pennycook et al. 2009), *S. tuberosum* DM1-3 516 R44, *S. tuberosum* cv. Umatilla and *S. lycopersicum*. For *ScCBF3 and ScCBF2* the corresponding pseudogenes were reported in A and B, respectively.




**Supplemental Figure 14**. Real Time qPCR on four target genes in NAC and AC conditions.

**Supplemental Tables**

**Supplemental Table 1.** Summary of sequence read statistics of the mate pair and paired-end libraries used in WGS sequencing

| Insert Size (bp) | Raw reads, number | Raw reads, Gb | Filtered reads, no | Filtered reads, Gb | Percentage of reads passing filter | Maximum length (bp) | Average length (bp) | Minimum length (bp) | Total Bases Length (Gb) | Effective Depth* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Pair End* | | | | | | | | | | |
| ~400 | 383,470,362 | 37.58 | 275,152,186 | 26.96 | 71.75 | 101 | 98 | 20 | 26.96 | 44.84 |
| ~550 | 344,071,682 | 33.72 | 305,005,284 | 29.89 | 88.65 | 101 | 98 | 20 | 30.30 | 40.23 |
| ~700 | 204,640,608 | 20.05 | 133,474,966 | 13.08 | 65.22 | 101 | 98 | 20 | 13.14 | 23.92 |
| *Mate Pair* | | | | | | | | | | |
| ~3,000 | 291,114,562 | 26.78 | 111,845,726 | 10.29 | 38.42 | 101 | 92 | 20 | 10.29 | 31.96 |
| ~5,000 | 227,189,440 | 20.67 | 68,843,984 | 6.26 | 30.30 | 101 | 91 | 20 | 6.30 | 24.66 |
| ~10,000 | 80,011,094 | 7.12 | 7,058,308 | 0.63 | 8.82 | 101 | 89 | 20 | 0.63 | 8.49 |

* estimated genome size: 838Mb

**Supplemental Table 2**. Summary of the *S. commersonii* genome assembly

| | **Conting** | | **Scaffold** | |
|---|---|---|---|---|
| | Size (bp) | Number | Size (bp) | Number |
| N90 | 1,178 | 146,855 | 5,763 | 26,615 |
| N80 | 2,108 | 94,918 | 12,735 | 15,653 |
| N70 | 3,258 | 63,880 | 21,439 | 10,432 |
| N60 | 4,628 | 42,804 | 31,743 | 7,132 |
| N50 | 6,506 | 27,829 | 44,298 | 4,833 |
| Longest | 170,543 | - | 458,668 | - |
| Total number (>100 bp) | - | 278,460 | | - |
| Total number (>500 bp) | - | 226,195 | | - |
| Total number (> 1kb) | - | - | | 64,655 |
| Total number (> 2kb) | - | - | | - |

**Supplemental Table 3**. CG content in *S. commersonii* genome

| Feature | # A | # C | # G | # T | # N | %GC content |
|---|---|---|---|---|---|---|
| Total | 267,803,084 | 141,392,099 | 140,663,862 | 266,680,757 | 45,924,484 | 34.54% |
| Intergenic | 212,916,516 | 109,799,013 | 109,139,327 | 211,976,646 | 40,869,266 | 34.01% |
| Genic | 54,886,568 | 31,593,086 | 31,524,535 | 54,704,111 | 5,055,218 | 36.55% |
| Intronic | 36,299,327 | 18,750,225 | 18,721,983 | 36,136,268 | 5,050,255 | 34.09% |
| Exonic | 18,587,241 | 12,842,861 | 12,802,552 | 18,567,843 | 4,963 | 40.84% |

**Supplemental Table 4**. Heterozygosity in *S. commersonii* genome

| Features | Bases affected | Length | Frequency |
|---|---|---|---|
| Genome* | 9,894,571 | 662,040,919 | 1.4946% |
| Gene | 261,398 | 149,307,299 | 0.1751% |
| Intron | 159,793 | 99,092,644 | 0.1613% |
| Exons | 141,821 | 50,152,571 | 0.2828% |
| 3' UTR | 14,216 | 4,660,982 | 0.3050% |
| 5' UTR | 10,594 | 4,070,026 | 0.2603% |
| UTR | 24,810 | 8,731,008 | 0.2842% |
| CDS | 117,011 | 41,421,563 | 0.2825% |

* only reliable positions are considered, not the whole genome

**Supplemental Table 5**. Annotation of SNPs detected in *S. commersonii*

| SNP Effect* | Count | Genes Affected, number | Percentage, % |
|---|---|---|---|
| Intergenic | 8,340,599 | - | 84.29 |
| Intragenic | 70,012 | - | 0.71 |
| Upstream | 294,797 | - | 2.98 |
| Downstream | 375,589 | - | 3.80 |
| Intron | 281,752 | 19,142 | 2.85 |
| UTR_5_prime | 18,865 | 2,199 | 0.19 |
| UTR_3_prime | 24,747 | 2,856 | 0.25 |
| Splice site acceptor | 3,017 | 1,710 | 0.03 |
| Splice site donor | 3,027 | 1,697 | 0.03 |
| Start lost | 1,687 | 1,037 | 0.02 |
| Non synonymous start | 459 | 462 | 0.00 |
| Stop lost | 1,546 | 914 | 0.02 |
| Stop gained | 25,404 | 4,127 | 0.26 |
| Non synonymous coding | 330,095 | 16,571 | 3.34 |
| Codon change | 1,017 | 269 | 0.01 |
| Synonymous start | 2 | 2 | 0.00 |
| Synonymous stop | 298 | 289 | 0.00 |
| Synonymous coding | 106,405 | 13,196 | 1.08 |
| Not processed** | 15,253 | - | 0.15 |
| Total SNP | 9,894,571 | | |

* only the most deleterious effect for each SNP is considered, thus every SNP is counted one time

** variants that software (SnpEff) cannot classify due to java errors

**Supplemental Table 6.** SINE families in *S. commersonii*

| Family | Number | Similarity (%) | Consensus (bp)[a] | Poly(A) (bp)[b] |
|---|---|---|---|---|
| SolS-Ia | 338 | 83,74 | 174 | 11 |
| SolS-Ib | 234 | 84,76 | 194 | 10 |
| SolS-II | 185 | 89,43 | 203 | 9 |
| SolS-IIIa | 503 | 92,84 | 231 | 11 |
| SolS-IV | 334 | 93,01 | 193 | 12 |
| SolS-V | 300 | 93,00 | 106 | 11 |
| SolS-VI | 2 | 96,73 | 226 | 14 |
| SolS-VII | 1 | 93,75 | | |
| TS | 5 | 76,98 | 164 | 7 |
| AU | 23 | 78,45 | 169 | 4 |

[a] Consensus sequence without poly(A).

[b] Averaged length

**Supplemental Table 7.** *De novo* assembled transcripts

| | |
|---|---|
| Assembled sequences. number | 117,816 |
| Maximum length. bp | 53,539 |
| Average length. bp | 1,369.13 |
| Minimum length. bp | 301 |
| Median | 1,026 |
| N50 | 1,887 |
| | |
| no mapping against assembly | 113,559 |
| % mapping against assembly | 96.39% |

**Supplemental Table 8**. Micro RNA statistics

| | |
|---|---|
| Predicted miRNA precursors | 1703 |
| Prediction of mature miRNAs | 1515 |
| Putative target transcripts | 4437 |
| | |
| Average Nr of targets per miRNA | 12 |
| Minimum Nr of targers per miRNA | 1 |
| Maximum Nr of targers per miRNA | 64 |
| | |
| Average Nr of miRNA per target | 2.2 |
| Minimum Nr of miRNA per target | 1 |
| Maximum Nr of miRNA per target | 70 |
| | |
| Nr of putative target loci in | |
| Cold Acclimation-Like | 277 |
| Cellular Response to cold-Like | 45 |
| Response to Cold-Like | 654 |

**Supplemental Table 9.** Putative miRNA precursors showing miRNA/miRNA* duplexes and similarity to know miRNAs. Similarity with known miRNAs was checked by blasting against miRBase and with RFAM

| Transcript_id | miRBase hit | RFAM hit |
|---|---|---|
| TCONS_00001190 | mtr-miR319a-5p | |
| TCONS_00002050 | ahy-miR3508 | |
| TCONS_00002051 | ahy-miR3508 | |
| TCONS_00005906 | stu-miR7997c | |
| TCONS_00012360 | stu-miR7985 | |
| TCONS_00019794 | stu-miR7998 | |
| TCONS_00020996 | stu-miR6023 | |
| TCONS_00022572 | gma-miR1520o | |
| TCONS_00024816 | pab-miR3698 | |
| TCONS_00025232 | bdi-miR5164 | |
| TCONS_00029235 | peu-miR2916 | |
| TCONS_00031426 | osa-miR5837.1 | |
| TCONS_00031603 | stu-miR8025-3p | |
| TCONS_00043860 | ppt-miR1033e | |
| TCONS_00045720 | stu-miR7998 | |
| TCONS_00047702 | stu-miR7998 | |
| TCONS_00049373 | stu-miR7998 | |
| TCONS_00053681 | gma-miR4995 | |
| TCONS_00055885 | ptc-miR169af | |
| TCONS_00058398 | mtr-miR2670g | |
| TCONS_00060297 | stu-miR7988 | |
| TCONS_00064460 | mtr-miR5298d | |
| TCONS_00067712 | stu-miR7988 | |
| TCONS_00068368 | sly-miR1918 | |
| TCONS_00075645 | osa-miR1863a | |
| TCONS_00020719 | stu-miR7986 | |
| TCONS_00031602 | stu-miR8025-5p | mir-399 |
| TCONS_00038446 | gma-miR4995 | |
| TCONS_00046799 | stu-miR7981-3p | |
| TCONS_00076957 | stu-miR8006-5p | mir-166 |
| TCONS_00058937 | | mir-598 |
| TCONS_00028908 | | mir-308 |
| TCONS_00033773 | | MIR1023 |
| TCONS_00034001 | | MIR396 |
| TCONS_00036819 | | mir-785 |
| TCONS_00050504 | | MIR821 |
| TCONS_00032245 | | lin-4 |
| TCONS_00017293 | | mir-198 |
| TCONS_00059748 | | mir-62 |
| TCONS_00006315 | | mir-156 |
| TCONS_00063573 | | MIR477 |
| TCONS_00055774 | | mir-48 |
| TCONS_00062719 | | MIR821 |

| | |
|---|---|
| TCONS_00005261 | MIR1122 |
| TCONS_00059744 | MIR807 |
| TCONS_00059483 | mir-598 |
| TCONS_00006798 | MIR820 |

**Supplemental Table 10.** Transcripts annotated as responsive to cold stress and of their potential miRNA regulators

| Transcript_target | Annotation | miRNA_precursor | miRBase hit | RFAM hit |
|---|---|---|---|---|
| augustus_masked_scaffold2559_abinit_gene_0_8 | avr9 cf-9 rapidly elicited protein 275 | TCONS_00020996 | stu-miR6023 | |
| augustus_masked_scaffold27265_abinit_gene_0_2 | cf-9 precursor | TCONS_00020996 | stu-miR6023 | |
| augustus_masked_scaffold31010_abinit_gene_0_1 | rna recognition motif-containing protein | TCONS_00067712 | stu-miR7988 | |
| augustus_masked_scaffold370_abinit_gene_0_0 | wd40 yvtn repeat and bromo-wdr9-i-like domain-containing protein | TCONS_00060297 | stu-miR7988 | |
| augustus_masked_scaffold40820_abinit_gene_0_3 | receptor-like protein 12-like | TCONS_00020996 | stu-miR6023 | |
| augustus_masked_scaffold4372_abinit_gene_0_0 | avr9 cf-9 rapidly elicited protein 275 | TCONS_00020996 | stu-miR6023 | |
| augustus_masked_scaffold5238_abinit_gene_0_0 | cf-9 precursor | TCONS_00020996 | stu-miR6023 | |
| augustus_masked_scaffold7053_abinit_gene_0_1 | phosphoglycerate mutase | TCONS_00029235 | peu-miR2916 | |
| augustus_masked_scaffold712_abinit_gene_0_0 | avr9 cf-9 rapidly elicited protein 275 | TCONS_00020996 | stu-miR6023 | |
| genemark_scaffold21357_abinit_gene_0_4 | peru 2 | TCONS_00020996 | stu-miR6023 | |
| genemark_scaffold363_abinit_gene_0_19 | g-type lectin s-receptor-like serine threonine-protein kinase rlk1-like | TCONS_00031602 | stu-miR8025-5p | mir-399 |
| genemark_scaffold363_abinit_gene_0_19 | g-type lectin s-receptor-like serine threonine-protein kinase rlk1-like | TCONS_00031603 | stu-miR8025-3p | |
| maker_scaffold10612_augustus_gene_0_22 | probable lrr receptor-like serine threonine-protein kinase at5g10290-like | TCONS_00020996 | stu-miR6023 | |
| maker_scaffold10960_snap_gene_0_61 | leucine-rich repeat protein kinase-like protein | TCONS_00060297 | stu-miR7988 | |
| maker_scaffold15760_snap_gene_0_35 | peru 1 | TCONS_00020996 | stu-miR6023 | |
| maker_scaffold1691_snap_gene_1_59 | protein kinase chloroplast | TCONS_00046799 | stu-miR7981-3p | |
| maker_scaffold1754_snap_gene_0_10 | peru 1 | TCONS_00020996 | stu-miR6023 | |
| maker_scaffold17583_augustus_gene_0_17 | g-type lectin s-receptor-like serine threonine-protein kinase rlk1-like | TCONS_00031602 | stu-miR8025-5p | mir-399 |
| maker_scaffold17583_augustus_gene_0_17 | g-type lectin s-receptor-like serine threonine-protein kinase rlk1-like | TCONS_00031603 | stu-miR8025-3p | |
| maker_scaffold20925_augustus_gene_0_30 | arginine serine-rich-splicing factor rsp40-like | TCONS_00046799 | stu-miR7981-3p | |
| maker_scaffold20968_augustus_gene_0_48 | vacuolar cation proton exchanger 5-like | TCONS_00005906 | stu-miR7997c | |
| maker_scaffold23900_augustus_gene_0_18 | peru 1 | TCONS_00020996 | stu-miR6023 | |
| maker_scaffold24560_snap_gene_0_68 | pentatricopeptide repeat-containing protein | TCONS_00060297 | stu-miR7988 | |
| maker_scaffold2531_augustus_gene_0_75 | receptor-like protein kinase | TCONS_00020996 | stu-miR6023 | |
| maker_scaffold27257_snap_gene_0_14 | peru 1 | TCONS_00020996 | stu-miR6023 | |

| | | | |
|---|---|---|---|
| maker_scaffold27265_augustus_gene_0_25 | peru 1 | TCONS_00020996 | stu-miR6023 |
| maker_scaffold32581_snap_gene_0_11 | peru 2 | TCONS_00020996 | stu-miR6023 |
| maker_scaffold4372_snap_gene_0_34 | lrr receptor-like serine threonine-protein kinase gso2-like | TCONS_00020996 | stu-miR6023 |
| maker_scaffold7225_snap_gene_0_60 | cationic peroxidase isozyme 40k precursor | TCONS_00031603 | stu-miR8025-3p |
| maker_scaffold8156_snap_gene_1_55 | receptor-like protein kinase | TCONS_00020996 | stu-miR6023 |
| maker_scaffold8450_snap_gene_0_34 | receptor-like protein kinase | TCONS_00020996 | stu-miR6023 |
| maker_scaffold8450_snap_gene_0_34 | receptor-like protein kinase | TCONS_00029235 | peu-miR2916 |
| snap_masked_scaffold15959_abinit_gene_0_11 | protein | TCONS_00068368 | sly-miR1918 |
| snap_masked_scaffold16580_abinit_gene_0_9 | transcriptional adapter ada2-like | TCONS_00049373 | stu-miR7998 |
| snap_masked_scaffold6229_abinit_gene_0_56 | catalase | TCONS_00029235 | peu-miR2916 |
| snap_masked_scaffold6491_abinit_gene_0_41 | udp-d-glucuronate 4-epimerase 2 | TCONS_00064460 | mtr-miR5298d |

**Supplemental Table 11**. Overview of the species used for the comparative genomics analyses

| Species Name | Genes | Unique, longest transcripts | Source | As in |
|---|---|---|---|---|
| *Solanum commersonii* | 37.662 | 37.477 | Genome Project | 04/2014 |
| *Solanum tuberosum* | 39.021 | 38.781 | Ensembl Plants - Release 22 | 04/2014 |
| *Solanum lycopersicum* | 34.727 | 34.635 | International Tomato Annotation Group | 02/2012 |
| *Mimulus guttatus* | 28.140 | 27.980 | Phytozome 10 by JGI | 04/2014 |
| *Beta vulgaris* | 27.421 | 27.363 | CRG | 11/2012 |
| *Cucumis melo* | 27.427 | 27.376 | melonomics,upv,es | 04/2011 |
| *Arabidopsis thaliana* | 27.416 | 27.233 | Ensembl Plants - Release 17 | 04/2013 |
| *Glycine max* | 54.174 | 53.821 | Ensembl Plants - Release 17 | 04/2013 |
| *Triticum aestivum* | 98.779 | 94.236 | Ensembl Plants - Release 22 | 04/2014 |
| *Zea mays* | 39.475 | 38.773 | Ensembl Plants - Release 22 | 04/2014 |
| *Brachypodium distachyon* | 26.552 | 26.470 | Ensembl Plants - Release 22 | 04/2014 |
| *Oryza sativa subsp, japonica* | 35.679 | 35.445 | Ensembl Plants - Release 22 | 04/2014 |

**Supplemental Table 12**. Detected one-to-one orthologs between a given species and *S. commersonii*

| **Species Name** | **one-to-one orthologs** |
|---|---|
| *S. tuberosum* | 17.297 |
| *S. lycopersicum* | 16.821 |
| *M. guttatus* | 7.058 |
| *B. vulgaris* | 6.799 |
| *C. melo* | 6.684 |
| *A. thaliana* | 5,.862 |
| *G. max* | 1.667 |
| *T. aestivum* | 1.160 |
| *Z. mays* | 3.913 |
| *B. distachyon* | 4.968 |
| *O. sativa* subsp. *japonica* | 4,492 |

**Supplemental Table 13**. Statistics about the number of duplication events detected in single gene trees according to their relative ages

| **Age** | **Events** | **Trees with events (all trees: 35,182)** | **Ratio (events / all trees)** |
|---|---|---|---|
| **1**: *S. commersonii* specific | 23,133 | 9,445 | 0.6575 |
| **2**: Potato Ancestor | 32,680 | 7,316 | 0.9289 |
| **3**: Solanum Ancestor | 33,185 | 14,61 | 0.9432 |
| **4**: Basal to Asterids | 2,331 | 1,814 | 0.0663 |

**Supplemental Table 14**. Functional enrichment analysis results after removing redundancy for the 10 biggest clusters of specifically expanded clusters of proteins in *S. commersonii* with statistically significant enriched functional terms

| Cluster | Size | Ontology | Go Term | Go Term Name |
|---|---|---|---|---|
| cluster 4369 | 191 | Biological Process | GO:0006278 | RNA-dependent DNA replication |
| cluster 4369 | 191 | Molecular Function | GO:0003676 | nucleic acid binding |
| cluster 4369 | 191 | Molecular Function | GO:0003723 | RNA binding |
| cluster 4369 | 191 | Molecular Function | GO:0003964 | RNA-directed DNA polymerase activity |
| cluster 4369 | 191 | Molecular Function | GO:0004523 | ribonuclease H activity |
| cluster 4368 | 158 | Biological Process | GO:0006278 | RNA-dependent DNA replication |
| cluster 4368 | 158 | Molecular Function | GO:0003723 | RNA binding |
| cluster 4368 | 158 | Molecular Function | GO:0003964 | RNA-directed DNA polymerase activity |
| cluster 4368 | 158 | Molecular Function | GO:0016787 | hydrolase activity |
| cluster 4364 | 138 | Molecular Function | GO:0003676 | nucleic acid binding |
| cluster 4364 | 138 | Molecular Function | GO:0004523 | ribonuclease H activity |
| cluster 4363 | 121 | Biological Process | GO:0006278 | RNA-dependent DNA replication |
| cluster 4363 | 121 | Molecular Function | GO:0003676 | nucleic acid binding |
| cluster 4363 | 121 | Molecular Function | GO:0003723 | RNA binding |
| cluster 4363 | 121 | Molecular Function | GO:0003964 | RNA-directed DNA polymerase activity |
| cluster 4363 | 121 | Molecular Function | GO:0008270 | zinc ion binding |
| cluster 4362 | 119 | Molecular Function | GO:0004386 | helicase activity |
| cluster 4362 | 119 | Molecular Function | GO:0005524 | ATP binding |
| cluster 4360 | 98 | Biological Process | GO:0006278 | RNA-dependent DNA replication |
| cluster 4360 | 98 | Molecular Function | GO:0003676 | nucleic acid binding |
| cluster 4360 | 98 | Molecular Function | GO:0003964 | RNA-directed DNA polymerase activity |
| cluster 4359 | 67 | Molecular Function | GO:0008270 | zinc ion binding |
| cluster 4355 | 60 | Molecular Function | GO:0003676 | nucleic acid binding |
| cluster 4354 | 57 | Biological Process | GO:0006278 | RNA-dependent DNA replication |
| cluster 4354 | 57 | Molecular Function | GO:0003723 | RNA binding |
| cluster 4354 | 57 | Molecular Function | GO:0003964 | RNA-directed DNA polymerase activity |
| cluster 4354 | 57 | Molecular Function | GO:0004523 | ribonuclease H activity |
| cluster 4350 | 52 | Biological Process | GO:0051252 | regulation of RNA metabolic process |
| cluster 4350 | 52 | Molecular Function | GO:0003676 | nucleic acid binding |
| cluster 4350 | 52 | Molecular Function | GO:0004523 | ribonuclease H activity |

**Supplemental Table 15**. Enrichment of functional categories among differentially expressed genes in nonacclimated (NAC, *) and acclimated (AC, **) conditions,

| Term ID | AC | | NAC | | Description |
|---|---|---|---|---|---|
| | No | % | No | % | |
| GO:0006418* | 0 | 0.00 | 213 | 22.35 | tRNA aminoacylation for protein translation |
| GO:0048528* | 0 | 0.00 | 125 | 13.12 | post-embryonic root development |
| GO:0043543* | 0 | 0.00 | 121 | 12.70 | protein acylation |
| GO:0046470* | 0 | 0.00 | 119 | 12.49 | phosphatidylcholine metabolic process |
| GO:0019321 | 0 | 0.00 | 65 | 6.82 | pentose metabolic process |
| GO:0006401 | 0 | 0.00 | 54 | 5.67 | RNA catabolic process |
| GO:0051788 | 0 | 0.00 | 51 | 5.35 | response to misfolded protein |
| GO:0006084 | 0 | 0.00 | 35 | 3.67 | acetyl-CoA metabolic process |
| GO:0030243 | 0 | 0.00 | 35 | 3.67 | cellulose metabolic process |
| GO:0009855 | 0 | 0.00 | 34 | 3.57 | determination of bilateral symmetry |
| GO:0010817 | 0 | 0.00 | 21 | 2.20 | regulation of hormone levels |
| GO:0007292 | 0 | 0.00 | 15 | 1.57 | female gamete generation |
| GO:0042445 | 0 | 0.00 | 15 | 1.57 | hormone metabolic process |
| GO:0048610 | 0 | 0.00 | 14 | 1.47 | cellular process involved in reproduction |
| GO:0051789 | 0 | 0.00 | 13 | 1.36 | response to protein stimulus |
| GO:0010027 | 0 | 0.00 | 12 | 1.26 | thylakoid membrane organization |
| GO:0048532 | 0 | 0.00 | 12 | 1.26 | anatomical structure arrangement |
| GO:0009626 | 0 | 0.00 | 11 | 1.15 | plant-type hypersensitive response |
| GO:0008202 | 0 | 0.00 | 10 | 1.05 | steroid metabolic process |
| GO:0048585 | 0 | 0.00 | 10 | 1.05 | negative regulation of response to stimulus |
| GO:0043248 | 0 | 0.00 | 9 | 0.94 | proteasome assembly |
| GO:0006499 | 0 | 0.00 | 7 | 0.73 | N-terminal protein myristoylation |
| GO:0042157 | 0 | 0.00 | 7 | 0.73 | lipoprotein metabolic process |
| GO:0046417 | 0 | 0.00 | 7 | 0.73 | chorismate metabolic process |
| GO:0031365 | 0 | 0.00 | 6 | 0.63 | N-terminal protein amino acid modification |
| GO:0051604 | 0 | 0.00 | 6 | 0.63 | protein maturation |
| GO:0007020 | 0 | 0.00 | 5 | 0.52 | microtubule nucleation |
| GO:0016117 | 5 | 1.18 | 86 | 9.02 | carotenoid biosynthetic process |
| GO:0030001 | 5 | 1.18 | 19 | 1.99 | metal ion transport |
| GO:0045036 | 5 | 1.18 | 0 | 0.00 | protein targeting to chloroplast |
| GO:0018130* | 6 | 1.42 | 329 | 34.52 | heterocycle biosynthetic process |
| GO:0006220 | 6 | 1.42 | 10 | 1.05 | pyrimidine nucleotide metabolic process |
| GO:0048589 | 7 | 1.66 | 26 | 2.73 | developmental growth |
| GO:0009657 | 7 | 1.66 | 21 | 2.20 | plastid organization |
| GO:0019637 | 9 | 2.13 | 28 | 2.94 | organophosphate metabolic process |
| GO:0048519 | 10 | 2.37 | 36 | 3.78 | negative regulation of biological process |
| GO:0033554 | 11 | 2.61 | 83 | 8.71 | cellular response to stress |
| GO:0042440 | 11 | 2.61 | 17 | 1.78 | pigment metabolic process |
| GO:0065008 | 12 | 2.84 | 60 | 6.30 | regulation of biological quality |
| GO:0048518 | 13 | 3.08 | 27 | 2.83 | positive regulation of biological process |
| GO:0016579** | 13 | 3.08 | 0 | 0.00 | protein deubiquitination |
| GO:0016052* | 15 | 3.55 | 127 | 13.33 | carbohydrate catabolic process |

| | | | | | |
|---|---|---|---|---|---|
| GO:0016192 | 17 | 4.03 | 31 | 3.25 | vesicle-mediated transport |
| GO:0051186 | 18 | 4.27 | 39 | 4.09 | cofactor metabolic process |
| GO:0009308 | 19 | 4.50 | 45 | 4.72 | amine metabolic process |
| GO:0009314 | 19 | 4.50 | 39 | 4.09 | response to radiation |
| GO:0044085 | 20 | 4.74 | 54 | 5.67 | cellular component biogenesis |
| GO:0015031 | 24 | 5.69 | 139 | 14.59 | protein transport |
| GO:0010038** | 26 | 6.16 | 0 | 0.00 | response to metal ion |
| GO:0034641* | 28 | 6.64 | 203 | 21.30 | cellular nitrogen compound metabolic process |
| GO:0006629 | 29 | 6.87 | 79 | 8.29 | lipid metabolic process |
| GO:0009628 | 33 | 7.82 | 75 | 7.87 | response to abiotic stimulus |
| GO:0051641** | 37 | 8.77 | 52 | 5.46 | cellular localization |
| GO:0043687 | 39 | 9.24 | 97 | 10.18 | post-translational protein modification |
| GO:0051649** | 40 | 9.48 | 51 | 5.35 | establishment of localization in cell |
| GO:0005975 | 41 | 9.72 | 97 | 10.18 | carbohydrate metabolic process |
| GO:0032501 | 42 | 9.95 | 118 | 12.38 | multicellular organismal process |
| GO:0009056 | 44 | 10.43 | 89 | 9.34 | catabolic process |
| GO:0043412 | 50 | 11.85 | 123 | 12.91 | macromolecule modification |
| GO:0016070 | 52 | 12.32 | 105 | 11.02 | RNA metabolic process |
| GO:0010467 | 62 | 14.69 | 133 | 13.96 | gene expression |
| GO:0051179 | 69 | 16.35 | 140 | 14.69 | localization |
| GO:0065007 | 71 | 16.82 | 200 | 20.99 | biological regulation |
| GO:0044281 | 71 | 16.82 | 171 | 17.94 | small molecule metabolic process |
| GO:0050896 | 83 | 19.67 | 197 | 20.67 | response to stimulus |
| GO:0019538 | 88 | 20.85 | 205 | 21.51 | protein metabolic process |
| GO:0006807 | 90 | 21.33 | 205 | 21.51 | nitrogen compound metabolic process |
| GO:0009058 | 104 | 24.64 | 261 | 27.39 | biosynthetic process |
| GO:0044260 | 138 | 32.70 | 315 | 33.05 | cellular macromolecule metabolic process |
| GO:0043170 | 149 | 35.31 | 342 | 35.89 | macromolecule metabolic process |
| GO:0044237 | 195 | 46.21 | 455 | 47.74 | cellular metabolic process |
| GO:0044238 | 195 | 46.21 | 452 | 47.43 | primary metabolic process |
| GO:0009987 | 239 | 56.64 | 562 | 58.97 | cellular process |
| GO:0008152 | 241 | 57.11 | 553 | 58.03 | metabolic process |

**Supplemental Table 16**. Number of non-redundant protein families annotated with the Gene Ontology term *cold acclimation* (CA)*, cellular response to cold* (CRTC), and *response to cold* (RTC) and related number of proteins in *A. thaliana* and *S. tuberosum*.

| GO category | Protein families | *A. thaliana* proteins | *S. tuberosum* proteins |
|---|---|---|---|
| CA | 17 | 177 | 239 |
| RTC | 146 | 1,429 | 2,833 |
| CRTC | 10 | 96 | 208 |
| Total | 173 | 1,702 | 3,280 |

## Supplemental Methods

### Genetic background of sequenced material

We sequenced the genome of clone cmm1t of *Solanum commersonii*. It derived from a single seed from accession PI243503 obtained from the Inter-Regional Potato Introduction Station, Sturgeon Bay, Wis (Supplemental Figure 2). To produce plant material for this study, one-month old plants were transferred from *in vitro* cultures into styrofoam trays filled with sterile soil and acclimated to *ex vitro* conditions in a growth chamber at 18-20°C (day/night). After two weeks, they were transferred to 5-cm-diameter plastic pots and grown in a temperature-controlled (20–24°C) greenhouse. DNA from leaves was purified using DNeasy Plant Maxi Kit (Qiagen, Valencia, CA USA) according to manufacturer's instructions.

### Library construction, sequencing, and quality control

A total amount of 2.5 µg of genomic DNA was sonicated with a Covaris S2 instrument (Covaris, inc., Woburn, MA) to obtain fragments ranging from 200bp to 1000bp in length. Preparation of *S. commersonii* DNA libraries was carried out using the TruSeq DNA Sample Prep Kit v2 (Illumina, San Diego, CA) according to manufacturer's instructions. Libraries were size selected at 400bp, 550bp and 700bp on 1.5% agarose gel cassettes using a Pippin Prep instrument (Sage Science, Beverly, MA). Preparation of *S. commersonii* cDNA libraries was carried out starting from 2.5 µg of total RNA extracted from leaf tissue grown under the conditions specified above. cDNA libraries were prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina, San Diego, CA) accordingly to manufacturer's instructions. Mate-pair libraries of 3Kb, 5Kb and 10Kb target insert sizes were constructed by Fasteris SA (Geneva, Switzerland) using an in-house modified Roche MP protocol.

Quality control of libraries was performed using High Sensitivity DNA Kit (Agilent, Wokingham, UK). Libraries were quantified using qPCR with a KAPA Library Quantification kit (KapaBiosystems, USA). Libraries were sequenced using Illumina HiSeq 1000 with TruSeq SBS Kit v3-HS and TruSeq PE Cluster Kit v3-cBot-HS kits (Illumina, USA) generating 100-bp paired-end sequences. Sequencing depth was estimated according to Varshney et al. (2012a).

### Read filtering

Sequence reads were pre-processed by first discarding reads with more than 10% of undetermined bases or with more than 50 bases of qualities lower than 7. Duplicated reads were discarded as well. Sequencing adapters were clipped using scythe (https://github.com/vsbuffalo/scythe). After clipping, the 3' ends of reads were quality trimmed with a threshold of 20 over a window of 10 bases using sickle (https://github.com/najoshi/sickle). Mate-pair reads were further filtered with Deloxer (Van Nieuwerburgh et al., 2012) (http://genomes.sdsc.edu/downloads/deloxer/) to identify and discard unpaired and paired-end reads.

### Genome size estimation

We estimated the genome size of *S. commersonii* using flow cytometry. *S. commersonii* and *Glycine max* nuclei were isolated, propidium iodide-stained and analyzed simultaneously (Doležel et al., 1998). Soybean (*G. max* 'Polanka', 2C= 2.50 pg DNA) served as an internal reference standard. The absolute DNA amount of *S. commersonii* was calculated on the values of G1 peak means as follows: (G1 peak means *S. commersonii*/G1 peak means of *G. max*) × *G. max* DNA content.

### Genome assembly and SNP calling

High quality reads from the paired-end libraries were assembled into contigs using SOAPdenovo v2.04 (Luo et al., 2012), with multiple k-mers between 79 and 99. Paired-end and mate-pair libraries were used for scaffolding by increasing library size. Gaps were closed using GapCloser v1.12 (a SOAP suite tool) and sequences shorter than 1,000 bp bases were discarded from the final assembly. The gene space of the assembled genome was assessed by aligning Core Eukaryotic Genes (CEGs) (Parra et al., 2009) to the assembly using Blast (Altschul et al., 1990) with a 65% identity threshold. Reads were aligned to the assembled genome using SOAPaligner v2.21 (a SOAP suite tool) with standard parameters but "-r 0" parameter. We called the SNPs by

aligning and comparing S. commersonii reads to the assembled S. commersonii genome, using SOAPsnp v1.03 (a SOAP suite tool) with "-u" and "-n" options enabled to give better accuracy for heterozygous SNP detection. Heterozigosity was then calculated by estimating the number of heterozygous calls over the total of the callable bases (Zheng L-Y et al., 2011; Varshney et al., 2012b). Variant calls were filtered for a sequencing depth higher than 10 and lower than 300, a quality scores higher than 20, and mapped best and second-best bases supported by at least four unique reads. Finally, sites with best base calling read count less than four times second-best base calling read count were identified as heterozygous sites.

**Genome annotation**

The assembled masked genome of *S. commersonii* was annotated using the MAKER pipeline (Cantarel et al., 2008). To investigate the nature of repetitive DNA in *S. commersonii*, we annotated repeat clusters using similarity to known repetitive DNA, using a RepBase library (Jurka et al., 2005), RepeatMasker (RepeatMasker Open-3.0. URL http://www.repeatmasker.org) and RepeatRunner (Smith et al., 2007). The RepeatMasker (http://www.repeatmasker.org) suite (Smit et al. 2004) was run with the public Solanaceae libraries using default parameters. RepeatRunner was run using the database of transposable elements encoded proteins included by default in MAKER pipeline installation. Putative SINEs were identified using the SINE-Finder tool and were used to search against published SINE sequences of *S. tuberosum* and other *Solanaceae* using FASTA (E-value ≤ 1e-10) (Wenke et al. 2011; ftp://ftp.ebi.ac.uk/ pub/software/unix/fasta/fasta36/). Different E-value thresholds at increasing stringency were tested without significant differences. Members of each family detect in *S. commersonii* were multiple aligned with MUSCLE (Edgar 2004) and consensus sequences were calculated with the Cons program from EMBOSS suite (Rice et al. 2000). Following evidences were used for protein coding gene models annotation: (i) alignments to amino acid sequences of *A. thaliana* (35,386 sequences, TAIR10), *S. tuberosum* (56,218 sequences, PGSC v. 3.4), *S. lycopersicum* (34,727 sequences, ITAG 2.3), Swiss-Prot Plants protein database (36,104 sequences, 13/04/2013); (ii) nucleotide alignments to 548,500 EST sequences of *S. commersonii* (67 sequences, NCBI, 17/04/2013), *S. tuberosum* (250,127 sequences, NCBI, 17/04/2013) and *S. lycopersicum* (298,306 sequences, NCBI, 17/04/2013); (iii) nucleotide alignments to 117,816 contigs *de novo* assembled from RNA-seq reads of *S. commersonii* using Trinity release 2013/02/25 (Grabherr et al., 2011) with a minimum contig length of 300 bp and at least two independent reads covering each contig; and (iv) predictions from SNAP (Korf, 2004) and Augustus (Stanke and Waack, 2003), all trained with gene models obtained from a first iteration of MAKER run using previously established evidence (i, ii and iii) and standard parameters, and predictions from GeneMark (Lukashin and Borodovsky, 1998), trained using randomly selected scaffolds covering about 40 Mbps, in accordance with author's instructions. In total, two MAKER annotation interations were carried out. Gene models with an Annotation Edit Distance (AED) (Yandell and Ence, 2012) higher than 0.5 were discarded from the final annotation. Predicted open reading frames (ORFs) were aligned against the NR database (06/2012 release) with Blast (BlastP, e-value < 10-5) and functionally annotated by automatic annotations performed with Blast2GO (Conesa and Götz, 2008).

**Evaluation of repeated elements content from unassembled reads**

Unassembled filtered reads from 3 millions of fragments were random sampled from 700bp insert libraries, transformed to fasta interleaved format, uploaded into RepeatExplorer (Novak et al., 2013) public server and analyzed using the “Clustering” module of RepeatExplorer using default parameters.

**Comparative genome analyses**

The OrthoMCL pipeline (Li et al., 2003) was used to identify and estimate the number of paralogous and orthologous gene clusters between *S. commersonii*, *S. tuberosum* and *S. lycopersicum*. Standard settings (BlastP, e-value < 10-5) were used to compute the all-against-all similarities. Syntenic blocks (≥5 genes per block) between *S. commersonii* and *S. tuberosum* were identified using MCScanX (Wang et al., 2013) based on the orthologous and co-orthologous gene pairs found by OrthoMCL pipeline.

**Long non-coding RNA and miRNA annotation**

Raw sequencing reads from RNA-seq experiments performed on root, stolons, tuber, leaf and flower samples were checked for quality using FastQC v0.10.1 (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimming and removal of adapters were performed with AdapterRemoval 1.5.2 (Lindgreen 2012) and FASTX Toolkit 0.0.13.2 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Trimmed reads were then mapped against the *S. commersonii* genome sequence with TopHat v2.0.11 (Kim et al., 2013). Duplicated reads were removed with Picard Tools 1.110 (http://picard.sourceforge.net) and the resulting files were used to annotate new transcripts with Cufflinks v2.2.0 (Trapnell et al., 2010). Removing the isoforms contained in other isoforms created a new annotation file comprising those belonging to the class "s" as reported by Cuffmerge. Long non-coding RNAs (lncRNAs) were identified using the approach described by Boerner and McGinnis (2012). In order to distinguish lncRNA from precursors of other ncRNA, the set of lncRNAs was first analysed with cmscan (e-value 0.01) from Infernal 1.1 (Nawrocki and Eddy, 2013) against the database of covariate models of Rfam 11.0. Non-coding transcripts were blasted as well against a database of plant mature miRNA sequences in miRBase (http://www.mirbase.org/) to identify homologous miRNAs. MIReNA (Mathelier and Carbone, 2010) was then used to check if the identified hits corresponded to miRNAs. The transcripts annotated as rRNA, tRNA, miRNA, or other ncRNA by cmscan and those validated positively by MIReNA were excluded. The remaining transcripts were analyzed with MIReNA without providing any genomic position in order to identify novel putative pre-miRNAs. The remaining transcripts were considered lncRNAs. Cufflinks v2.2.0 (Trapnell et al., 2010) was used to obtain RPKM expression values. miRNA target prediction was performed by using psRNATarget (Dai and Zhao, 2011) with default settings.

**Cold resistance gene analysis**

To annotate putative cold resistance genes in *S. commersonii*, a set of reference proteins were selected from *A. thaliana*. In detail, 58 proteins annotated with the Gene Ontology term *cold acclimation* (CA), 28 proteins annotated as *cellular response to cold* (CRTC), and 619 proteins as *response to cold* (RC) were selected. INTERPROSCAN was used to identify the domains of the proteins included in each gene family (Supplemental Table 16). The proteins showing the same domain composition were grouped and aligned using MUSCLE 3.6 (Edgar, 2004) and a consensus sequence was calculated. For each protein group, the generated consensus sequence was used to interrogate the proteome of *A. thaliana* (ftp://ftp.arabidopsis.org/home/tair/Proteins/TAIR10_protein_lists/TAIR10_pep_20101214) with a BlastP threshold of e-value ≤ 10-3, in order to identify all proteins with the same domain composition. The same analysis was carried out for *S. tuberosum* (http://potato.plantbiology.msu.edu/data/PGSC_DM_v3.4_pep.fasta.zip) (Supplemental Table 16). For each protein family, specific hidden Markov models (HMMs) were created using a modified version of Matrix-R (Supplemental Dataset 8 ). The obtained HMM modules were used to identify putative cold responsive proteins in *S. tuberosum* and *S. commersonii.* Several filtering steps were then performed to remove false positives. First, protein Blast searches were performed against the proteins used to create the HMM modules, with filtering conditions set as e-value ≤ 10-5 and the alignment length as at least 90% of the query length. Second, a promoter analysis was performed to identify genes having putative promoter binding sites for transcription factors related to response to cold, as reported by (Maruyama et al., 2004; 2012).

**R-Genes analysis**

Matrix-R was used to screen the proteomes of *S. commersonii* and *S. tuberosum* (37,662 and 39,031 proteins, respectively). Protein sequences corresponding to annotated genes (39,031) from the PGSC whole genome annotation of DM assembly were used (PGSC_DM_v3_superscaffolds.fasta.zip; http://potatogenomics.plantbiology.msu.edu/index.html). The set of predicted proteins identified via HMM profiling was further analyzed using INTERPROSCAN software version 5.0 (http://www.ebi.ac.uk/Tools/pfa/iprscan5/) to verify the presence of conserved domains and motifs characteristic of R-proteins (Nucleotide Binding Sites, NBS; Leucine Rich Repeats, LRR; Toll-Interleukin receptor, TIR; KINASE; SERINE/ THREONINE). To identify *S. tuberosum R1* orthologues in *S. commersonii,* we used the orthology relationships

among *S. commersonii* genes and we used a phylogenetic approach to define orthologues. Then, selected homologous sequences were aligned using two different programs: MUSCLE v3.8 (Edgar, 2004) and MAFFT v6.712b (Katoh and Toh, 2008), and were further analyzed using INTERPROSCAN software version 5.0 (http://www.ebi.ac.uk/Tools/pfa/iprscan5/) to verify the presence of conserved domains and motifs characteristic of R1 proteins.

**Transcriptional analysis**

Twelve clonally propagated plants from cmm1t (PI243503) were cultured in a growth chamber under cool white fluorescent lamps (350-400 mmol $m^{-2}s^{-1}$) at 24°C and then exposed to -2°C for 6 hours to test their resistance to low temperature under non acclimated (NAC) conditions. To evaluate cold resistance following acclimation (AC), six plants were first transferred from a 24°C growth chamber to a cold room (4°C) under cool white fluorescent lamps (100 mmol $m^{-2}s^{-1}$) for two weeks and then exposed to -2°C for 6 hours. For each test, RNA was isolated from 100 mg of leaf tissue pooled from five plants. Pooled tissue was homogenized (TissueLyzer by Qiagen) using a TRIZOL reagent (Life Technologies) and RNA was extracted following TRIZOL Life Technologies protocols. The concentration and purity of extracted RNAs were estimated using the NanoDrop spectrophotometer (Thermo Fisher Scientific). The quality and integrity of RNA were checked after electrophoresis of 1 mg of RNA samples on 1% agarose gel stained with SYBR® Safe (Life Technologies). The synthesized and labeled antisense-RNA (aRNA) was generated using the Kreatech's kit RNA ampULSe: Amplification and Labeling Kit for CombiMatrix (Kreatech Biotechnology, Amsterdam, The Netherlands) arrays with Cy5 dye. The purified, labeled aRNA was quantified by spectrophotometer and 4 mg were hybridized to the Combimatrix array (described below) according to manufacturer's directions. Pre-hybridization, hybridization, washing and imaging were performed according to manufacturer's protocols (http://www.combimatrix.com/support_docs.htm). Imaging of array slides was performed using a GenePix® 4400A Microarray Scanner controlled by the GENEPIX PRO V.7 software (Molecular Devices) at 5ìm resolution. The GENEPIX PRO v.7 software was also used for densitometry analysis and raw data extraction. Probe signals higher than negative control values plus twice the standard deviation were considered as 'present'.

The analysis was performed on a Combimatrix *S. tuberosum* chip produced by the Plant Functional Genomics Center at the University of Verona. The chip contained 27,234 non-redundant 35-40-mer oligo probes in triplicate. Probes were designed on tentative consensus sequences (TCs; 23.453 probes) and singletons with a 3' poly(A) tail (46 probes) derived from the SolEST database (D'Agostino et al., 2009) using Oligoarray 2.1 (Rouillard et al., 2003). Oligo probes were designed to identify the 3'-UTR region of genes. Results from Blastx comparisons against the UniPortKB/Swiss-Prot database were exploited to determine the correct open reading frame and to define forward/reverse TC orientation. 13,207 TC sequences had forward orientation, while 2,027 had reverse orientation. In the case of 9,000 TC sequences, no Blast hits were found and it was not possible to assess where the 3'-UTR region was located for these sequences. As a consequence, we filtered out 6,000 TCs generated by assembling the largest number of ESTs and considered both the orientations for probe design. Nine bacterial oligonucleotide sequences provided by CombiMatrix, 40 probes designed based on seven Ambion spikes and 11 additional probes based on *Bacillus anthracis*, *Haemophilus ducreyi* and *Alteromonas phage* sequences were used as negative controls. Three to four replicates of each probe were randomly distributed across the array. Three technical and three biological replicates were used for each hybridization experiment. Data analysis was performed using the R package limma (Smyth, 2005). The median of the signal was used for the analysis. Replicate agreement was checked by hierarchical clustering of resulting data based on Euclidean distances between samples. Samples not clustering with their corresponding replicates were discarded. Maximum likelihood normexp was used for background correction and the arrays were normalized by quantile normalization. Identification of differentially expressed probes was performed by fitting a linear model including the correlation between replicated probes followed by a Bayesian test. Raw p-values were adjusted for multiple correction via the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Adjusted p-values <=0.05 were considered statistically significant. The TC sequences used to design the Combimatrix probes were blasted (Blastn, e-value < 0.01) against the transcriptome of *S. commersonii* in order to determine matches between probes and annotated loci. To validate

the microarray data, we performed real time PCR analysis for three cold-regulated genes. These included COR413 (SOLTUB01G046490), Histone demethylase (SOTUB05G023460.1.1), and Histone-lysine N-methyltransferase (SOTUB10G019470.1.1). The qPCR results showed that the three genes are all cold regulated, with expression kinetics very similar to those obtained from microarray analysis (Supplemental Figure 14).

**Phylome reconstruction**

We reconstructed the complete collection of gene evolutionary histories (i.e. the phylome) for the wild potato transcriptome, and 11 other plant genomes (Supplemental Table 11). For this, Smith-Waterman (Smith and Waterman, 1981) searches were used to retrieve homologs (cut-offs: 1e-5 e-value, alignments covering 50% of the query). Homologous sequences were aligned using three different programs: MUSCLE v3.8 (Edgar, 2004), MAFFT v6.712b (Katoh and Toh, 2008), and Kalign v2.04 (Lassmann et al., 2009) and in forward and reverse direction (Landan and Graur, 2007). The six resulting alignments were combined using M-Coffee (Wallace et al., 2006) and trimmed with trimAl v1.4 (Capella-Gutiérrez et al., 2009), using a consistency score cutoff of 0.1667 and a gap score cutoff of 0.1. Maximum Likelihood (ML) phylogenetic trees were inferred using the best fitting evolutionary model as described elsewhere (Huerta-Cepas et. al. 2011) and the NNI tree search approach, and a gamma distribution with four rate categories and a fraction of invariant positions inferred from the data. Branch supports were computed using an aLRT (approximate likelihood ratio test) a chi-square distribution, as implemented in PhyML (Guindon et al., 2010).

**Phylogeny-based prediction of orthology and paralogy**

Orthology and paralogy relationships were inferred from the phylome using a phylogenetic approach (Gabaldón, 2008), using a species-overlap algorithm implemented in ETE v2 (Huerta-Cepas et al., 2010). The resulting orthology and paralogy predictions can be accessed through phylomeDB.org (Huerta-Cepas et al., 2014), and have been used in subsequent analyses such as orthology-based functional annotation, identification of gene expansions, or duplication dating.

**Species tree reconstruction and shared genomic content**

A phylogeny for the species included in the phylome was inferred using two complementary approaches, which rendered identical topologies. First, a super tree was inferred from the 34,633 trees in the phylome, using a Gene Tree Parsimony approach as implemented in the dup-tree algorithm (Wehe et al., 2008). Secondly, 454 gene families with a clear, phylogeny-based, one-to-one orthology present in at least 11 of the 12 species included in the analyses were used to perform a multi-gene phylogenetic analysis by concatenating all the corresponding alignment into a super-matrix. Species relationships were inferred from this alignment using PhyML (Guindon et al., 2010), with JTT as the evolutionary model, which best fitted 357 out of 454 gene families, SPR search mode, and computation of aLRT per branch.

**Phylostratigraphic dating of duplication events**

We scanned the phylome to detect and date duplication events, using a previously described algorithm (Huerta-Cepas and Gabaldón, 2011). We focused on events assigned to three different relative evolutionary periods: 1) *S. commersonii* lineage, 2) Potato ancestor, 3) *Solanum* ancestor, and 4) Basal to Asterids. Individual trees were scanned and all duplication events of in lineages leading to *S. commersonii* genes were dated. Enrichment analyses for overrepresented GO terms were performed using FatiGO (Medina et al., 2010). A Fisher exact test looking for overrepresented terms in specific sets of proteins against the whole annotated genome was used with a e-value cutoff of 0.01. Then, GO terms redundancy was reduced using the REViGO webserver (Supek et al., 2011), setting a similarity threshold of 0.5, using as quality score the ratio of log odds values, and SimRel as the semantic similarity algorithm.

We focused on lineage-specific genome expansions. In-Paralogs groups were grouped into clusters with at least 50% of overlap of shared genes. **Figure 8** shows the number of clusters detected according to their size. Only clusters comprising 10 or more genes were considered in this analysis and inspected for enriched functional terms as indicated above.

**Functional annotation**

*S. commersonni* predicted protein-coding genes were functionally annotated using two complementary approaches, one based on protein signatures and the other based on orthology relationships. In the first approach InterProScan (Zdobnov and Apweiler 2001) was used to annotate proteins. Using this approximation, 91,566 gene ontology (GO) terms were assigned to 21,352 proteins. In the second approach we used phylogeny-based analyses, 12,435 one-to-one orthology relationships among *S. commersonii* genes and genes from species used in the phylome with some GO annotation were found. Using these predictions 39,574 non-redundant GO terms were transferred to *S. commersonii* genes. Supplemental Figure 8 shows the overlap between the two approaches.

**Additional references**

**Akaike, H.** (1974). A new look at the statistical model identification. IEEE T. Automat. Contr. **19**: 716–723.

**Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J.** (1990). Basic local alignment search tool. J. Mol. Biol. **215**: 403–410.

**Benjamini, Y. and Hochberg, Y.** (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B Met. **57**: 289–300.

**Boerner, S. and McGinnis, K.M.** (2012). Computational identification and functional predictions of long noncoding RNA in *Zea mays*. PLoS ONE **7**: e43047.

**Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T.** (2009). TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**: 1972–1973.

**Conesa, A. and Götz, S.** (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genomics **2008**: 1–12.

**Cantarel B.L., Korf I., Robb S.M.C., Parra G., Ross E., Moore B., Holt C., Sánchez Alvarado A., Yandell M.** (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. **18**: 188–196.

**D'Agostino, N., Traini, A., Frusciante, L., and Chiusano, M.L.** (2009). SolEST database: a "one-stop shop" approach to the study of *Solanaceae* transcriptomes. BMC Plant Biol. **9**: 142.

**Dai, X. and Zhao, P.X.** (2011). psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res. **39**: W155–9.

**Doležel, J., Greilhuber, J., Lucretti, S., and Meister, A.** (1998). Plant Genome Size Estimation by Flow Cytometry: Inter-laboratory Comparison. Ann. Bot-London **82**: 17–26.

**Edgar, R.C.** (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32** (5): 1792–97.

**Gabaldón, T.** (2008). Large-scale assignment of orthology: back to phylogenetics? Genome Biol. **9**: 235.

**Gascuel, O.** (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of

sequence data. Mol. Biol. Evol. **14**: 685–695.

**Grabherr, M.G. et al.** (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**: 644–652.

**Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O.** (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**: 307–321.

**Holt, C. and Yandell, M.** (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics **12**: 491.

**Huerta-Cepas, J. and Gabaldón, T.** (2011). Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics **27**: 38–45.

**Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L.P., Marcet-Houben, M., and Gabaldón, T.** (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. **42**: D897–902.

**Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J.** (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110**: 462–467.

**Katoh, K. and Toh, H.** (2008). Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics **9**: 286–298.

**Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L.** (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. **14**: R36.

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

**Landan, G. and Graur, D.** (2007). Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments. Mol. Biol. Evol. **24**: 1380–1383.

**Lassmann, T., Frings, O., and Sonnhammer, E.L.L.** (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res. **37**: 858–865.

**Li, L., Stoeckert, C.J., and Roos, D.S.** (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13**: 2178–2189.

**Lindgreen, S.** (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res. Notes b:337

**Lukashin, A.V, and Borodovsky, M.** (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. **26(4)**: 1107–1115.

**Luo, R. et al.** (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience **1**: 18.

**Maruyama, K. et al.** (2012). Identification of cis-acting promoter elements in cold- and dehydration-induced transcriptional pathways in *Arabidopsis*, rice, and soybean. DNA Res. **19**: 37–49.

**Maruyama, K., Sakuma, Y., Kasuga, M., Ito, Y., Seki, M., Goda, H., Shimada, Y., Yoshida, S., Shinozaki, K., and Yamaguchi-Shinozaki, K.** (2004). Identification of cold-inducible

downstream genes of the *Arabidopsis* DREB1A/CBF3 transcriptional factor using two microarray systems. Plant J. 38: 982–993.

**Mathelier, A. and Carbone, A.** (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics **26**: 2226–2234.

**Medina, I. et al.** (2010). Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. **38**: W210–W213.

**Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J.** (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. Bioinformatics **29**: 792-793.

**Nawrocki, E.P. and Eddy, S.R.** (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics **29**: 2933–2935.

**Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I.** (2009). Assessing the gene space in draft genomes. Nucleic Acids Res. **37**: 289–297.

**Rice, P., Longden, I., Bleasby, A.** (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. **16**(6): 276—277.

**Rouillard, J.-M., Zuker, M., and Gulari, E.** (2003). OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. **31**: 3057–3062.

**Smit, A., Hubley, R., and Green, P.** (2004) RepeatMasker Open-3.0 at http://repeatmasker.org.

**Smith, T.F. and Waterman, M.S.** (1981). Identification of common molecular subsequences. J. Mol. Biol. 147: 195–197.

**Smith, C. D., Edgar, R. C., Yandell, M. D., Smith, D. R., Celniker, S. E., Myers, E. W., & Karpen, G. H.** (2007). Improved repeat identification and masking in Dipterans. Gene **389**: 1–9.

**Smyth, G.K.** (2005). limma: Linear Models for Microarray Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V. Carey, R.A. Irizarry, and S. Dudoit, eds (Springer-Verlag: New York), pp. 397–420.

**Stanke, M. and Waack, S.** (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl **2**: ii215–25.

**Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T.** (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE **6**: e21800.

**Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M., and Paterson, A.H.** (2008). Unravelling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. **18**: 1944–1954.

**Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., and Borodovsky, M.** (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. **18**: 1979–1990.

**Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L.** (2010). Transcript assembly and quantification by RNA-Seq

reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**: 511–515.

**Van Nieuwerburgh, F., Thompson, R.C., Ledesma, J., Deforce, D., Gaasterland, T., Ordoukhanian, P., and Head, S.R.** (2012). Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. Nucleic Acids Res. **40**: e24.

**Varshney, R.K., Ribaut, J.-M., Buckler, E.S., Tuberosa, R., Rafalski, J.A., and Langridge, P.** (2012a). Can genomics boost productivity of orphan crops? Nat. Biotechnol. **30**: 1172–1176.

**Varshney, R.K. et al.** (2012b). Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat. Biotechnol. **30**: 83–89.

**Wallace, I.M., O'Sullivan, O., Higgins, D.G., and Notredame, C.** (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. **34**: 1692–1699.

**Wang, Y., Li, J., and Paterson, A.H.** (2013). MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. Bioinformatics (Oxford, England) **29**: 1458–1460.

**Wehe, A., Bansal, M.S., Burleigh, J.G., and Eulenstein, O.** (2008). DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics **24**: 1540–1541.

**Wenke, T., Dobel, T., Sorensen, T.R., Junghans, H., Weisshaar, B., Schmidt, T.** (2011) Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. Plant Cell **23**: 3117–3128.

**Yandell, M. and Ence, D.** (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. **13**: 329–342.

**Zdobnov, E.M., and Apweiler, R.** (2001). InterProScan—An integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848.

**Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., Liu, C.-M., and Jing, H.-C.** (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome Biol. **12**: R114.