Title

Assessment of an organ-specific *de novo* transcriptome of the nematode trap-crop, *Solanum sisymbriifolium.*

Authors/Affiliations

Alexander Q Wixom[a], N Carol Casavant[a], Joseph C Kuhl[a], Fangming Xiao[a], Louise-Marie Dandurand[b], and Allan B Caplan[a*]

[a]Department of Plant Sciences, University of Idaho, Moscow, ID 83844-2333

[b]Department of Entomology, Plant Pathology, and Nematology, University of Idaho, Moscow, ID 83844-2329

Contact

*Corresponding author e-mail: acaplan@uidaho.edu

Running Title

Running title: A *de novo S. sisymbriifolium* transcriptome

Abstract

Solanum sisymbriifolium, also known as "Litchi Tomato" or "Sticky Nightshade," is an undomesticated and poorly researched plant related to potato and tomato. Unlike the latter species, *S. sisymbriifolium* induces eggs of the cyst nematode, *Globodera pallida*, to hatch and migrate into its roots, but then arrests further nematode maturation. In order to provide researchers with a partial blueprint of its genetic make-up so that the mechanism of this response might be identified, we used single molecule real time (SMRT) sequencing to compile a high quality *de novo* transcriptome of 41,189 unigenes drawn from individually sequenced bud, root, stem, and leaf RNA populations. Functional annotation and BUSCO analysis showed that this transcriptome was surprisingly complete, even though it represented genes expressed at a single time point. By sequencing the 4 organ libraries separately, we found we could get a reliable snapshot of transcript distributions in each organ. A divergent site analysis of the merged transcriptome indicated that this species might have undergone a recent genome duplication and re-diploidization. Further analysis indicated that the plant then retained a disproportionate number of genes associated with photosynthesis and amino acid metabolism in comparison to genes with characteristics of R-proteins or involved in secondary metabolism. The former processes may have given *S. sisymbriifolium* a bigger competitive advantage than the latter did.

Introduction

Solanum sisymbriifolium (SSI), otherwise known as "litchi tomato", "*morelle de Balbis"*, or "sticky nightshade", is an undomesticated relative of potato and tomato. For more than a decade, SSI has been investigated as a trap-crop (a plant that attracts nematodes but kills them before they can reproduce) for nematodes such as *Globodera pallida* that normally parasitize potatoes and tomatoes (Timmermans, 2005; Dandurand and Knudsen, 2016). It is also a potential source of anti-protozoan (Meyre-Silva *et al.*, 2013) and anti-molluscan (Bagalwa *et al.*, 2010) metabolites. If the genetic basis for these protective processes could be identified, it might be possible to transfer these traits, either through cross-breeding or through modern transgenic technologies, from this weed to its domesticated relatives. However, while the genomes of potato and tomato have been studied extensively, spiny solanums, like SSI, have not (Yang *et al.*, 2014). Only 54 SSI nucleotide sequences have been submitted to NCBI as of 2016. This ignorance about the biology and genetics of the spiny solanums could be masking a wealth of genetic resources that could be used to protect agriculturally important crops.

Most bioinformatic analyses of a species begin with the assembly and annotation of a complete genome. Once assembled, these data can be searched for genes encoding a particular protein or RNA sequence. For those working on a species that has not been studied extensively in the past, and which is only being studied now in order to conduct a limited number of experiments, whole genome sequencing can be more expensive and time consuming than can be justified. In these circumstances, alternative methods using sequencing technologies that are generally referred to as next-generation sequencing (NGS), have allowed researchers to by-pass whole genome sequencing in favor of generating a smaller database, one depleted of the silent regions of the genome and of genes that are not contributing to the phenotypes of interest. Most commonly, this is done using Illumina or 454 platforms that generate 10's and 100's of millions of short reads from cDNA copies of all of the mRNAs expressed during a given moment of time. Once obtained, these sequences can then be merged *in silico* into full length protein coding sequences. However, this *de novo* transcriptome can sometimes prove problematic. Short reads derived from highly conserved coding domains and repetitively organized genes can potentially be aligned and joined into chimeric assemblies that cannot be verified or removed because there is no independently

sequenced genome available to serve as an extended template or scaffold to ensure that the merged sequences are indeed co-linear (Yang and Smith, 2013). A recent technical improvement, Pacific Biosciences' single-molecule real-time (SMRT) "sequencing by synthesis" strategy, has become sufficiently accurate and attainably priced to be utilized by small research groups. The benefit of using SMRT sequencing is that it produces vastly longer reads than previous methodologies, although with lower coverage (Eid *et al.*, 2009). The longer reads allow researchers to establish a transcriptome consisting of nearly complete open reading frames free of the kinds of errors possible when sequences must be assembled *in silico* from short reads (Ocwieja *et al.*, 2012; Zhang *et al.*, 2014).

The specific goal of the current project was to establish a four organ (bud, leaf, stem and root) *de novo* transcriptome of SSI. In doing so, we wanted to ensure that the final sequences were high-quality and consisted of genes that were biologically relevant and not artifacts of some *in silico* assembly process. This transcriptome will provide a reference library to be used in future RNA-seq experiments to identify genes for nematode and other pathogen resistances in SSI.

Results

Establishing a SMRT sequenced transcriptome

Before any sequencing was attempted, the genome size of SSI was estimated using flow cytometry (Supplemental Figure 1). This showed that the genome mass of SSI was approximately 4.73 pg per 2C, or 2,315 mega-base pairs per 1C. By comparison, Arumuganathan and Earle, 1991, using the same technology estimated that the tomato genome massed between 1.88 to 2.07 pg per 2C while tetraploid potato massed between 3.31 to 3.86 pg per 2C. Thus, these initial measurements gave SSI a genome size greater than tetraploid potato. Despite their unusual length (Paul and Banerjee, 2015), SSI has 24 chromosomes, like diploid potatoes and many other Solanaceae (data not shown). Due to the size of this genome, and our interest in generating a database of protein-coding genes, we elected to sequence the SSI transcriptome rather than its genome.

Generating an SSI transcriptome was done using Single-Molecule Real Time (SMRT) sequencing by PacBio Sciences (Eid *et al.*, 2009) that does not need to assemble short reads into one contiguous sequence. Rather than producing only short reads, SMRT technology can provide reads up to 60,000 bp along a single molecule of DNA. This allows the capture of entire genes with one read rather than chunking it into many small bits that have to be assembled later. This sequencing strategy gives higher coverage than Sanger-based reactions like those performed on Applied Biosystems™ gene analysis instruments, and longer reads than Illumina or Roche 454.

The Iso-Seq pipeline classifies the sequences as either full-length non-chimeric (FLNC), or non-full length reads. Full length reads are those containing both 5' and 3' adapters, in addition to the poly(A) tail. The reads containing these parts in the expected order, i.e. 5' adapter–poly(A)–3' adapter, with no additional copies of these parts, are classified as non- chimeric. The FLNC reads in the present dataset were corrected with the non-full length reads using Iterative Cluster for Error correction and the Pacific Biosciences Quiver algorithm (https://github.com/PacificBiosciences/SMRT-Analysis/wiki/ConsensusTools-v2.3.0-Documentation).

In an attempt to improve our ability to detect differences in the suites of genes expressed in different parts of the plant, we generated cDNA from 4 organs; leaves, stems,

114 roots, and unopened flower buds. We then independently carried out SMRT sequencing of
115 all 4 samples. Finally, all corrected FLNC reads were merged *in silico* and redundancy was
116 removed using CD-HIT-EST (Li and Godzik, 2006).

117 This SMRT sequencing strategy created 231,712 total corrected FLNC sequences
118 (Table 1) using the aforementioned pipeline. These sequences had a GC content of 41.2%.
119 CD-HIT-EST was then used to reduce the redundant sequences to sets with 100% identity.
120 This lowered the number of sequences to 139,611 with a GC content of 41.0%. The GC
121 content continued to decrease as the identity was reduced using CD-HIT-EST. At 80%
122 identity, there were 32,315 sequences, with an estimated GC content of 39.7%. The decrease
123 in GC content could be due to the methodology of CD-HIT-EST that retains the longest
124 sequence during the reduction process. Because of this, reads that spanned untranslated
125 regions of a transcript were expected to be favored over those only consisting of coding
126 regions.

127 In the end, we chose to work with a final SMRT dataset that had been reduced to 90%
128 identity and consisted of a set of 41,189 sequences with a GC content of 39.9%. We judged
129 that this estimate of the number of transcripts present in the 4 organs would most likely err
130 on the high side, yet still retain most splice variants within a gene, as well as many paralogs
131 and single nucleotide polymorphisms between alleles of this obligate outbreeder.

132

133 **Evidence based Quality Control of the SMRT Transcriptome**

134 We performed an internal quality check by sequencing 45 randomly chosen clones
135 from a cDNA library using Sanger Dye Deoxy technology (ABI 3730, Applied Biosystems).
136 Bowtie2 (Langmead and Salzberg, 2012) was then used to find the most likely equivalent of
137 each clone in our SSI transcriptome. A manual comparison of the Sanger-sequenced clone
138 and the assembled transcript was done using DNA Strider (Marck, 1988). Firstly, all 45
139 cDNAs were found in the SSI transcriptome (Figure 1A). Secondly, only two of these SMRT-
140 derived sequences appeared to be chimeric (Figure 1B), and based on the length of non-
141 homologous stretches, could have been transcribed from different members of the same gene
142 family rather than been created by misassembly. During our analysis of the SSI
143 transcriptome, we did find entries that consisted of inverted repeats of entire gene sequences.
144 These inverted repeats likely occurred during the preparation of the cDNA library prior to

145 sequencing rather than during sequencing or subsequent computational processing as can be
146 found in Illumina or 454 assemblies (Loman *et al.*, 2012; Luo *et al.*, 2012).

147 When the SSI transcriptome was analyzed using Mercator (Lohse *et al.*, 2014), it
148 contained 38.6% unannotated sequences. This was markedly fewer than the percent
149 unannotated sequences of either potato (50.8%) or tomato (46.4%) transcriptomes processed
150 in the same way via Mercator (Supplemental Figure 4). Other than that, the binned profile
151 of SSI was very similar to the published transcriptomes (Supplemental Figure 3) of these
152 plants. This led us to believe that our transcriptome was at least of comparable quality with
153 the working transcriptomes of these two better studied species.

154 PfamScan (Finn *et al.*, 2008) was also used to annotate the domains of the
155 transcriptome. This program uses HMMer (Eddy, 1998) domain annotations, and used in
156 combination with protocols established by Sarris *et al.*, 2016, allowed for the annotation of
157 domains found in the amino acid sequences translated from the assembled transcriptome.
158 This annotated 84.7% of the transcriptome with at least one recognizable domain. There
159 were fewer unannotated sequences in the SSI transcriptome than in the STU and SLY
160 transcriptomes (Supplemental Table 1). The reduced number of unannotated sequences
161 found in the SMRT transcriptome might reflect the fact that this set had undergone a
162 conservative reduction to 90% identity. Alternatively, the reduced number of unannotated
163 SSI sequences could have resulted from the fact that we had only sampled the four most
164 frequently studied organs of a "normally" growing plant, that is, plants manifesting a
165 physiological state which has been extensively studied in numerous species, while the STU
166 and SLY transcriptomes were compiled from plants sampled over a much broader range of
167 life-history stages and growth conditions ranging from fruit and tuber development to
168 exposure to biotic and abiotic stresses where the functions of many genes are still under
169 investigation.

170 To further test the quality and completeness of our transcriptome, BUSCO
171 benchmarking (Simão *et al.*, 2015) was performed. The BUSCO database was established to
172 allow researchers to assess the completeness of new genomes or transcriptomes based on the
173 detection of a set of universal, single-copy orthologs. We found 93% intact BUSCO
174 archetypes (889 genes) in the SSI SMRT transcriptome, 30.2% (289) of these were found in
175 multiple copies, while an additional 2.2% (21) of the BUSCO archetypes were present in

176 fragments, and 4.8% (46) were missing entirely (Table 2). These numbers representing genes
177 expressed during a single growth condition of SSI, were only 4.7 percentages different from
178 the numbers of BUSCO archetypes found in the entire SMRT sequenced genome of *A.*
179 *thaliana*.

180 SSI has the same number of chromosomes as most other Solanaceae, and does not
181 appear to be polyploid (data not shown), yet the BUSCO analysis showed that SSI had more
182 duplicate copy archetypes than diploid and tetraploid potato. This high number of similar
183 sequences could point to the fact that our transcriptome has not been reduced far enough, or
184 could be one line of evidence that SSI has undergone extensive genome duplication or
185 hybridization in the past. This latter hypothesis was evaluated by divergent gene analysis as
186 has been done with plants such as wheat (Krasileva *et al.*, 2013). When the program
187 Freebayes (Garrison and Marth, 2012) was run using a defined diploid setting, it output
188 information stating there were genes that had more than 2 alleles or paralogues. We redid the
189 analysis using defined triploid and tetraploid settings and found that even after merging
190 sequences with more than 90% identity using CD-HIT-EST, the SSI transcriptome contained
191 1,348 genes with 3 distinguishable alleles or paralogues and furthermore, 44 genes with 4
192 distinguishable copies (Table 3). It was noteworthy that no gene had more than 4 alleles or
193 paralogues. A simple explanation for these multiple gene variants, that would be consistent
194 with the BUSCO analysis, was that SSI underwent a genome duplication followed by
195 diploidization in the past and that over time, some of the duplicated loci acquired additional
196 mutations while other loci were lost. To determine if this proposed duplication was restricted
197 to one chromosome, or one chromosomal arm, the 44 genes with 4 alleles were mapped onto
198 SLY chromosomes (Supplemental Table 2). There were "4–allele" genes found on 11 of 12
199 SLY chromosomes which indicated, assuming that genes dispersed in tomato were not linked
200 when the two species diverged, that SSI has undergone a full genome duplication rather than
201 a segmental duplication within one chromosome.

202 Since SSI is not as well-known as other Solaneacae, we employed OrthoMCL v2.0.9
203 (Li *et al.*, 2003) to illustrate some of the common features its gene complement showed with
204 those of other plants. Protein sequences from our SSI transcriptome (translated using the
205 program ESTScan (Iseli *et al.*, 1999)), and protein sequences from tomato (SLY), potato
206 (STU), eggplant (SME), *Arabidopsis thaliana* (ATH), papaya (CPA), grapes (VVI), peaches

207 (PPE), black cottonwood (PTR), oranges (CSI), alfalfa (MTR), maize (ZMA) and rice (OSA)
208 were merged into 45,234 orthologous groups (gene families). In this set, 6097 orthologous
209 groups were shared by all 13 species (Figure 2), an overlap well within the range of previous
210 studies (Yang *et al.*, 2014). Each species had many additional groups that were not shown in
211 this diagram because they were not shared with all members of this set of plants.
212 Interestingly, even closely related species like SSI, STU, SLY, and SME had hundreds of
213 groups not found in each other. When the annotations of the SSI unique set were compared
214 to the full transcriptome, several functional groups showed a disproportionate increase. It is
215 possible that these disproportionately expanded sets, that included photosynthetic genes, and
216 genes for amino acid and vitamin metabolism (Supplemental Figure 5), diverged so much
217 more than groups such as those for cell wall composition, and hormone and secondary
218 metabolism, because expansion of the former traits gave SSI a competitive edge over other
219 species in their habitat. Overall, though, there were fewer groups of genes unique to SSI
220 than unique to STU and SME. As noted previously, this could merely reflect the fact that our
221 data came from a single-point snapshot of only 4 organs and so would have lacked those
222 transcripts specifically expressed during fruit and seed set, germination, senescence, abiotic
223 stress, pathogen attacks, and numerous other stages of a plant's lifecycle.
224 Using highly conserved orthologous genes, i.e. subunits of Rubisco, provisional
225 phylogenies were created for nuclear-encoded and chloroplast-encoded genes using the
226 aforementioned species (data not shown). In doing so, we concluded that nuclear SSI was
227 most closely related to eggplant, which has been noted previously (Särkinen *et al.*, 2015),
228 while chloroplast SSI was more closely related to tomato. This dichotomy has also been seen
229 by others (Miz *et al.*, 2008) and interpreted to indicate that SSI had undergone an ancient
230 hybridization and afterwards retained the chloroplast genome from one parent, and much of
231 the nuclear genome from another. However, many more SSI genes will have to be compared
232 with the genes of many more South American plants to confirm that this hybridization
233 occurred.
234
235 **Building a snapshot of organ-associated gene expression**
236 Since we had maintained separate cDNA pools from individual organs, it was
237 possible to backtrack each sequence within the final transcriptome to obtain a provisional

238 profile of gene expression throughout the plant (Figure 3A). This analysis showed that there
239 were 8019 sequences expressed solely in buds, 4957 solely in roots, 5349 solely in leaves,
240 4198 solely in stems, and 7212 sequences expressed in all tissues. That left 11,538 sequences
241 that were expressed in more than one organ but not in all 4.

242 This backtracking allowed us to construct an expression snapshot that showed how
243 different genes were being expressed at the time the organs were harvested. Using several
244 in-house Python scripts, we recorded the number of reads for genes that had common
245 annotations for several different physiological processes.

246 A set of light-harvesting complex genes (LHC-I) were predictably found in aerial
247 organs with few exceptions (Figure 3B), demonstrating that the backtracking program could
248 extract biologically useful information about sequences with specified characteristics from
249 the merged transcriptome. In order to determine if this kind of analysis of SMRT sequences
250 could categorize the expression of very different sets of genes, we constructed an inter-organ
251 expression profile of genes that encoded both a leucine-rich repeat (LRR) domain together
252 and a nucleotide binding (NB-ARC) domain (Figure 3C), a pairing frequently found in
253 pathogen resistance genes (R-genes). This profile of R-gene prevalence in SSI, potato, and
254 tomato indicated that there was a reduction of these genes in the SMRT transcriptome
255 compared to the other two species (Table 4). Three of these potential R-genes were then
256 assayed by semi-quantitative PCR (primers found in Supplemental Table 3) and quantified
257 using a sample of cDNA from the same pool that had been sequenced, and a sample from an
258 independently-prepared, unsequenced cDNA pool (Figure 4). In order to assess whether a
259 SMRT data set could be a reliable indicator of gene expression, both the *in silico* and PCR
260 measurements of gene expression were normalized in kind to an actin sequence (Ssi032526).
261 The physical measurements of expression of two of the three genes matched the expression
262 snapshot extremely well, but the third gene (Ssi038051) was more abundant in stems and
263 buds than expected based on its SMRT expression snapshot. This confirms that whole
264 transcriptome snapshots can provide a provisional picture of organ differences in gene
265 expression, but further shows that the expression of each gene of interest needs to be verified
266 biologically, most usefully by multiple independent tests.

267 **Discussion**

268 The creation of a *de novo* transcriptome necessitates massive amounts of follow-up
269 analyses, both *in silico* and biologically, to estimate its reliability.

270 We initially employed both Illumina and 454 sequencing (data not shown) in order
271 to compensate for the different kinds of errors to which each method was prone (Luo *et al.*,
272 2012). Screening this assembly with genes randomly selected from an SSI cDNA library
273 revealed that 20% of these genes failed to match any of the assembled sequences in this
274 database (Supplemental Figure 2A), and of those that matched, 40% appeared to be chimeric
275 (Supplemental Figure 2B). In contrast, all of these cDNAs were found in our SMRT
276 sequenced transcriptome and few were patently chimeric (Figure 1).

277 A number of factors are known to exasperate misassembly including the presence of
278 large gene families and of repeatedly occurring kmers in the dataset (Moreton *et al.*, 2015).
279 Even though we did not sequence the SSI genome, we found 4 lines of evidence indicating
280 that it might be complex enough to pre-dispose our transcriptome to these kinds of assembly
281 mistakes. First, the nuclear DNA content of SSI was larger than most diploid Solanaceae,
282 roughly the same size as a tetraploid potato (Supplemental Figure 1). Second, divergent gene
283 analysis indicated that the SSI transcriptome was unusually complex and contained 3 and 4
284 distinguishable alleles for many genes (Table 3). Third, there were only 67 putative R-genes,
285 that is, genes containing a nucleotide binding domain (NB-ARC) and a leucine-rich repeat
286 domain (LRR), in the SMRT sequenced dataset compared to the 309 in STU, and 137 in
287 STU (Table 4). Finally, an unusually high percentage of the BUSCO gene set were present
288 in multiple copies in SSI even though our transcriptome could only consist of a portion of
289 all the genes that are likely to be encoded in its DNA (Table 2). One model consistent with
290 these 4 facts was that SSI had, sometime in the past, undergone a partial or complete genome
291 duplication. Over time, as diploidy was re-established, some of the duplicated alleles or
292 paralogues diverged, while others were lost. Nevertheless, enough of the expanded gene
293 families remained to confound the alignment programs that tried to differentiate between
294 their members. While these kinds of errors might be correctable with the use of other
295 assembly programs, we chose, instead to create an assembly-independent transcriptome
296 using SMRT technology.

297 At the moment, SMRT technology does not provide the sequence coverage or depth

298 that can be obtained with Illumina or 454 sequencing. In order to increase our chances of
299 sampling uncommon organ-specific transcripts, we prepared independent cDNA pools from
300 4 organs of the plant. Using an in-house script
301 (https://github.com/AlexWixom/Transcriptome_scripts), we were able to increase the value
302 of the final library by generating expression snapshots for genes of interest in each organ.
303 These expression snapshots are no substitute for a more thorough RNA-seq study, but they
304 do provide a preliminary assessment of a plant's biology at the time of harvest. Using these
305 snapshots, we recognized different patterns of expression of individual LHC-1 genes (Figure
306 3B) within the photosynthetic parts of the plant. We also saw that 2 of the 3 R-genes re-
307 examined by PCR showed the same expression pattern in two independent RNA and cDNA
308 preparations as found in the transcriptome itself (Figure 4). Thus, in the absence of RNA-
309 seq studies or experimental evidence for the role of a specific locus, this kind of library
310 assembly could be used to direct researchers to the subset of R-genes most likely responsible
311 for the resistance in a given organ.

312 R-genes coding for recognition proteins are commonly perceived as sentinels that are
313 awaiting activation by molecules introduced during infection (Jones and Dangl, 2006).
314 Therefore, the reduced number of R-genes found in SSI (67), compared to both potato (309),
315 or tomato (137) was unexpected. This discrepancy could be explained in any one of several
316 ways. First, SSI might be using proteins with novel domain structures in place of classic R-
317 genes. Second, sequencing depth might simply have been inadequate to capture all R-genes
318 that were actually being expressed at low levels. Finally, SSI could be relying on rapidly
319 inducing transcription of R-genes after an infection has occurred. Any one of these
320 hypotheses is worthy of continuing analysis.

321 With this transcriptome as an example, we have established a protocol that opens the
322 door to further genetic mining of previously uncharacterized species. The completeness of
323 our database indicates that *de novo* transcriptomes not only provide an economical and time-
324 saving way to study a new species, but can also provide expression data that could not be
325 gleaned from a genomic sequence.

326 **Materials and Methods**

327

328 **Plant and Culture Conditions**

329 *S. sisymbriifolium* (SSI) seeds obtained from C. Brown (USDA-ARS, Prosser WA)
330 were germinated in soil. Nodes from a single plant were sterilized for 20 min using 10%
331 NaClO with 0.05% Tween20. Plant material was then washed 3x with sterile distilled H_2O
332 and put into 120 mL baby food jars containing standard Murashige and Skoog salts, pH 5.6,
333 3% sucrose, 0.7% agar, 100 µg mL^{-1} myo-inositol, 2.0 µg mL^{-1} glycine, 1 µg mL^{-1} thiamine,
334 0.5 µg mL^{-1} pyridoxine, and 0.5 µg mL^{-1} nicotinic acid. A single plant was chosen as the
335 progenitor of all of the plants used in this study. All of its descendants were maintained at
336 25˚C in 16 h light, and subcultured vegetatively every 4 wk. Over the course of the project,
337 rooted clones with at least 4-6 leaves were put into 2 L of hydroponic medium (Yoshida
338 *et al.*, 1976), referred to here as Fake Field. Each container was diffusely aerated through an
339 aquarium stone, maintained at constant volume by the addition of distilled water, and
340 emptied and refilled with fresh hydroponic medium every 7 d. Hydroponic containers were
341 maintained at 22˚C, 16 h light with an irradiance level of 0.0006 W m^{-2}. Illumination was
342 provided by GE Lighting Fluorescent lamps (13781, F96T12/CW/1500). After a 2 wk lag-
343 time, plants began producing 1-3 new leaves each wk, and flowered continuously afterwards.
344 All experiments were performed on plants that had not been infected or wounded in any way
345 previously.

346

347 **RNA extraction**

348 RNA was extracted from SSI bud, stem, leaf, and root and infected root organs
349 adapted from the protocol in Casavant *et al.*, 2017. Adaptions included use of a coffee grinder
350 to homogenize tissue with the addition of dry ice to maintain RNA integrity.

351

352 **Genome size estimation by Flow Cytometry**

353 Healthy green leaf tissues were collected from SSI plantlets growing *in vitro*.
354 Roughly 1 cm^2 (0.01g or less) of leaf was chopped in 1mL ice cold LB01 buffer for 1.5-2
355 min (Doležel *et al.*, 1989). The LB01 buffer contained 50µg mL^{-1} RNase stock and 50µg
356 mL^{-1} propidium iodide (25% PI stock in DMSO) per mL of LB01. Each sample was chopped

357 with a fresh razor blade in a clean Pyrex petri plate. The finely chopped suspension was then
358 filtered through a 50µm nylon mesh filter (Partec 04-0042-2317). This filtered suspension
359 was kept in the dark at 4˚C for between 15-90 min before it was analyzed.
360 Genome size estimations were made using a BD FACSARIA Flow Cytometer
361 (IBEST Imaging Core, University of Idaho, Moscow, ID, USA). A green laser at 488nm was
362 used to excite the propidium iodide stained cells and was then collected in the PE-A channel.
363 Thresholds for PE-A were set at 1,000 and FSC at 500. The voltages were set so the major
364 peak (2C) of the SSI samples were near 50,000 on the linear scale. Four suspensions were
365 made from separate donor plants once a day for three consecutive days. Two replicates of
366 two external standards were also used daily in addition to the 4 SSI samples. External
367 standards included *Solanum lycopersicum cv.* Stupicke polni tyckove rane (2C= 1.96 pg
368 DNA) and *Glycine max* cv. Polanka (2C= 2.50 pg DNA) (Doležel *et al.*, 1992; Doležel *et al.*,
369 1994) which were chosen because their genome sizes were in the expected range of SSI. One
370 repetition of internal standards was run using tomato and soybean. DNA content was
371 estimated using the equation described by Doležel *et al.*, 2007.
372
373 **Library Preparation for Iso-Seq**
374 SMRT library preparation and sequencing were performed by the National Center for
375 Genome Resources (Santa Fe, New Mexico). The Iso-Seq libraries for four organs, root,
376 stem, leaf and bud, were prepared for Isoform Sequencing (Iso-Seq) using the Clontech
377 SMARTer PCR cDNA Synthesis Kit and the BluePippin Size Selection System protocol as
378 described by Pacific Biosciences (https://goo.gl/ij71Hh) with the following modifications.
379 For cDNA conversion, 3 µg of total RNA was put into each Clonetech SMARTer reaction.
380 From the PCR optimization procedure specified in the protocol, it was determined that 14
381 cycles of PCR would be sufficient for amplification of each organ's cDNA. Amplification
382 was followed by size selection on each sample to obtain three size bins (0.5-2 kb, 1.5-3 kb
383 and 2.5-6 kb) using the Blue Pippin (Sage Science, Beverly, Massachusetts) instrument. The
384 amplified and size selected cDNA products were made into SMRTbell Template libraries per
385 the Isoform Sequencing protocol referenced above. Libraries were prepared for sequencing
386 by annealing a sequencing primer (component of the SMRTbell Template Prep Kit 1.0) and
387 then binding polymerase to this primer-annealed template. The polymerase-bound template

388 was bound to MagBeads (P/N 100-125-900) (https://goo.gl/wdZErU) and sequencing was
389 performed on a PacBio RS II instrument. 12 v3 SMRTcells were run for the root tissues, 14
390 for the leaf tissues, 9 for the stem tissues, and 12 for the bud tissues for a total of 47
391 SMRTcells (Pacific Biosciences, P/N 100-171-800). The libraries from each organ were
392 separately sequenced using P6C4 polymerase and chemistry and 240-minute movie times
393 (Pacific Biosciences, P/N 100-372-700, P/N 100-356-200).
394
395 **Single Molecule Real Time Sequencing**
396 All SMRT cells for a given organ were run through the Iso-Seq pipeline included in
397 the SMRT Analysis software package. First, reads of insert (ROIs, previously known as
398 circular consensus sequences or CCS) were generated using the minimum filtering
399 requirement of 0 or greater passes of the insert and a minimum read quality of 75. This
400 allowed for the high yields going into subsequent steps, while providing high accuracy
401 consensus sequences where possible. The pipeline then classified the ROI in terms of full-
402 length, nonchimeric and non-full length reads. This was done by identifying the 5' and 3'
403 adapters used in the library preparation as well as the poly(A) tail. Only reads that contained
404 all three in the expected arrangement and did not contain any additional copies of the adapter
405 sequence within the DNA fragment were classified as full-length non-chimeric copies.
406 Finally, all full-length non-chimeric reads were run through the Iterative Clustering for Error
407 correction algorithm then further corrected by the Pacific Biosciences Quiver algorithm
408 (https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-
409 transcriptome-data). Once the Iso-Seq pipeline result was available for each organ, the results
410 were combined into a single data set and redundant sequences were removed using CD-HIT-
411 EST (Li and Godzik, 2006).
412
413 **Illumina Sequencing for divergent gene analysis**
414 Extracted total RNA from each previously stated organ was sent to Eurofins
415 Genomics (Ebersberg, Germany) for library preparation and sequencing. Prior to library
416 preparation, quality control (QC) was performed on individual tubes of RNA and equal
417 aliquots of each preparation were blended into one pool. The Illumina library cDNA was
418 prepared using randomly-primed first and second strand synthesis, followed by gel sizing

419 and PCR amplification. The library was then physically normalized and found to have insert
420 sizes of 250-450 bp.
421 Illumina sequencing was performed on a MiSeq v3 2x300. The read sequences were
422 clipped using Trimmomatic, version 0.32 (Bolger *et al.*, 2014), and bases with a Phred score
423 < 20 were removed. Trimmed reads shorter than 150 bp were removed; this step could
424 remove none, one, or both mates of a read-pair. Digital normalization was applied to the
425 Illumina reads in order to reduce redundant information present in these large datasets. A
426 coverage cutoff of 30 and a kmer size of 20 decreased the data to 28% of the initial
427 31,310,146 reads. BWA (Li and Durbin, 2009) was used to map read pairs to the *Solanum*
428 *tuberosum cultivar* Desiree chloroplast genome (GenBank accession DQ38616.2). Only
429 unmapped reads were retained.
430
431 **Annotation of Sequences**
432 Mercator sequence annotation was performed using the TAIR, PPAP, KOG, CDD,
433 IPR, BLAST CUTOFF of 80, and ANNOTATE options (Lohse *et al.*, 2014).
434
435 **Annotating Protein Domains of Translated Sequences**
436 PfamScan (Finn *et al.*, 2009) was run on the SSI transcriptomes following protocols
437 set forth by Sarris *et al.*, 2016.
438
439 **Biological Quality Check of *in silico* Sequences**
440 45 clones from a cDNA library (Express Genomics, Average insert size=1 kb,
441 Vector= pExpress 1) were randomly selected and sequenced via Sanger Dye-Deoxy DNA
442 Sequencing (ABI 3730). These sequences were then aligned to the transcriptomes using
443 Bowtie2 (Langmead and Salzberg, 2012) set for local alignment and best hit only. These
444 aligned sequences were then manually compared for possible chimeric features.
445
446 **Evolutionary Comparison of SSI to 13 Other Species**
447 The evolutionary clustering and comparison protocols were adapted from those set
448 out in Yang *et al.*, 2014. See Supplemental Table 4 for species used and online download
449 sources.

450

451 **Divergent Gene Analysis to determine Ploidy**

452 Phasing of the SMRT transcriptome was completed using unassembled Illumina
453 sequences adapted from protocols established by Krasileva *et al.*, 2013, with the addition of
454 an in-house Python script to quantify single-nucleotide polymorphisms present per sequence
455 (https://github.com/AlexWixom/Transcriptome_scripts/freePloidy.py).

456

457 **Creation of Expression Snapshots Using only SMRT Sequences**

458 In-house Python scripts were used to backtrack final transcriptome sequences to each
459 organ using CD-HIT-EST cluster files
460 (https://github.com/AlexWixom/Transcriptome_scripts).

461

462 **Expression Snapshot Validation**

463 Sequence specific oligonucleotides were designed for several genes that were then
464 used to obtain semi-quantitative PCR expression snapshots on the same cDNA used to obtain
465 our SSI transcriptome (referred to as the "Sequenced" sample), as well as on a second cDNA
466 pool prepared from RNA collected from independently grown plants (referred to as the
467 "Unsequenced" sample). PCR fragment bands were quantified with a local background
468 subtracted and normalized to actin (following the procedure established by Casavant *et al.*,
469 2017). The primers for these genes can be found in Supplemental Table 3 with the proposed
470 gene description.

471

472 **Accession Numbers**

473 The SMRT sequenced transcriptome has been deposited at DDBJ/EMBL/GenBank under
474 the accession GGFC00000000. The version described in this paper is the first version,
475 GGFC01000000.

476

477 **Data availability**

478 File S1 contains detailed descriptions of all supplemental files. Sequence data are available
479 at GenBank and the accession numbers are GGFC00000000. In-house code used to generate
480 data can be found at https://github.com/AlexWixom/Transcriptome_scripts.

481 **Author Contributions**

482 NCC prepared many of the RNA samples, and all of the PCR analyses. AQW performed all
483 remaining research, statistical analysis, and figure assemblies. AQW and ABC contributed
484 to the experimental design. AQW wrote the manuscript with assistance from ABC. JCK, FX,
485 and L-MD provided equipment and facilities, bacterial clones and plant material, and
486 assisted in data interpretation. All authors have read and approved the manuscript.

487

Acknowledgments

The authors would like to thank Dr. R. Tripepi for providing us with vegetatively propagated *S. sisymbriifolium*. We are also extremely grateful to the National Center for Genomics Research for preparation and sequencing of SMRT libraries. This research was funded by the Idaho Potato Commission, the Northwest Potato Research Consortium, and USDA-APHIS Farm Bill 10201, 10007.

Arumuganathan, K. and Earle, E. (1991). Nuclear dna content of some important plant species. *Plant molecular biology reporter*, 9(3):208–218.

Bagalwa, J.-J. M., Voutquenne-Nazabadioko, L., Sayagh, C., and Bashwira, A. S. (2010). Evaluation of the biological activity of the molluscicidal fraction of solanum sisymbriifolium against non target organisms. *Fitoterapia*, 81(7):767–771.

Bolger, A., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–2120.

Casavant, N. C., Kuhl, J. C., Xiao, F., Caplan, A. B., and Dandurand, L.-M. (2017). Assessment of globodera pallida rna extracted from solanum roots. *Journal of nematology*, 49(1):12.

Dandurand, L.-M. and Knudsen, G. (2016). Effect of the trap crop *Solanum sisymbriifolium* and two biocontrol fungi on reproduction of the potato cyst nematode, *Globodera pallida*. *Annals of Applied Biology*, 169(2):180–189.

Doležel, J., Binarová, P., and Lcretti, S. (1989). Analysis of nuclear dna content in plant cells by flow cytometry. *Biologia plantarum*, 31(2):113–120.

Doležel, J., Doleželová, M., and Novák, F. (1994). Flow cytometric estimation of nuclear dna amount in diploid bananas (musa acuminata andm. balbisiana). *Biologia plantarum*, 36(3):351–357.

Doležel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear dna content in plants using flow cytometry. *Nature protocols*, 2(9):2233–2244.

Doležel, J., Sgorbati, S., and Lucretti, S. (1992). Comparison of three dna fluorochromes for flow cytometric estimation of nuclear dna content in plants. *Physiologia plantarum*, 85(4):625–631.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, 14(9):755.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.* (2009). Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., *et al.* (2009). The pfam protein families database. *Nucleic acids research*, 38(suppl_1):D211–D222.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G.,

526 Forslund, K., Eddy, S. R., Sonnhammer, E. L., *et al.* (2008). The pfam protein families
527 database. *Nucleic acids research*, 36(suppl 1):D281–D288.
528 Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read
529 sequencing. https://arxiv.org/abs/1207.3907v2.
530 Iseli, C., Jongeneel, C. V., and Bucher, P. (1999). Estscan: a program for detecting,
531 evaluating, and reconstructing potential coding regions in est sequences. In *ISMB*,
532 volume 99, pages 138–148.
533 Jones, J. D. and Dangl, J. L. (2006). The plant immune system. *Nature*, 444(7117):323–329.
534 Krasileva, K. V., Buffalo, V., Bailey, P., Pearce, S., Ayling, S., Tabbita, F., Soria, M., Wang,
535 S., Akhunov, E., Uauy, C., *et al.* (2013). Separating homeologs by phasing in the tetraploid
536 wheat transcriptome. *Genome Biology*, 14(6):R66.
537 Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature*
538 *methods*, 9(4):357–359.
539 Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler
540 transform. *Bioinformatics*, 25(14):1754–1760.
541 Li, L., Stoeckert, C. J., and Roos, D. S. (2003). Orthomcl: identification of ortholog groups
542 for eukaryotic genomes. *Genome research*, 13(9):2178–2189.
543 Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets
544 of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.
545 Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Tohge, T., Fernie, A. R.,
546 Stitt, M., and Usadel, B. (2014). Mercator: a fast and simple web server for genome scale
547 functional annotation of plant sequence data. *Plant, cell & environment*, 37(5):1250–1258.
548 Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and
549 Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing
550 platforms. *Nature biotechnology*, 30(5):434–439.
551 Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K. T. (2012). Direct
552 comparisons of illumina vs. roche 454 sequencing technologies on the same microbial
553 community dna sample. *PloS one*, 7(2):e30087.
554 Marck, C. (1988). 'dna strider': a 'c'program for the fast analysis of dna and protein
555 sequences on the apple macintosh family of computers. *Nucleic acids research*, 16(5):1829–
556 1836.

557 Meyre-Silva, C., Niero, R., Bolda Mariano, L. N., Gomes do Nascimento, F., Vicente Farias,
558 I., Gazoni, V. F., dos Santos Silva, B., Giménez, A., Gutierrez-Yapu, D., Salamanca, E., *et al.*
559 (2013). Evaluation of antileishmanial activity of selected brazilian plants and identification
560 of the active principles. *Evidence-Based Complementary and Alternative Medicine*, 2013.
561 Miz, R. B., Mentz, L. A., and Souza-Chies, T. T. (2008). Overview of the phylogenetic
562 relationships of some southern brazilian species from section torva and related sections of
563 "spiny solanum"(solanum subgenus leptostemonum, solanaceae). *Genetica*, 132(2):143–
564 158.
565 Moreton, J., Izquierdo, A., and Emes, R. D. (2015). Assembly, assessment, and availability
566 of de novo generated eukaryotic transcriptomes. *Frontiers in genetics*, 6.
567 Ocwieja, K. E., Sherrill-Mix, S., Mukherjee, R., Custers-Allen, R., David, P., Brown, M.,
568 Wang, S., Link, D. R., Olson, J., Travers, K., *et al.* (2012). Dynamic regulation of hiv-1 mrna
569 populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic*
570 *acids research*, 40(20):10345–10355.
571 Paul, A. and Banerjee, N. (2015). Phylogenetic relationship of some species of *SOLANUM*
572 based on morphological, biochemical and cytological parameters. *Indian Journal of*
573 *Fundamental and Applied Life Sciences*, 5(3):51–56.
574 Särkinen, T., Barboza, G. E., and Knapp, S. (2015). True black nightshades: Phylogeny and
575 delimitation of the morelloid clade of solanum. *Taxon*, 64(5):945–958.
576 Sarris, P. F., Cevik, V., Dagdas, G., Jones, J. D., and Krasileva, K. V. (2016). Comparative
577 analysis of plant immune receptor architectures uncovers host proteins likely targeted by
578 pathogens. *BMC biology*, 14(1):8.
579 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M.
580 (2015). Busco: assessing genome assembly and annotation completeness with single-copy
581 orthologs. *Bioinformatics*, 31(19):3210.
582 Timmermans, B. G. (2005). *Solanum sisymbriifolium (Lam.): a trap crop for potato cyst*
583 *nematodes*.
584 Yang, X., Cheng, Y.-F., Deng, C., Ma, Y., Wang, Z.-W., Chen, X.-H., and Xue, L.-B. (2014).
585 Comparative transcriptome analysis of eggplant (solanum melongena l.) and turkey berry
586 (solanum torvum sw.): phylogenomics and disease resistance analysis. *BMC genomics*,
587 15(1):412.

588 Yang, Y. and Smith, S. A. (2013). Optimizing de novo assembly of short-read rna-seq data
589 for phylogenomics. *BMC genomics*, 14(1):328.
590 Yoshida, S., Douglas, A. F., James, H. C., and Kwanchai, A. G. (1976). Laboratory manual
591 for physiological studies of rice. *Manila: IRRI*, 56:69–77.
592 Zhang, W., Ciclitira, P., and Messing, J. (2014). Pacbio sequencing of gene families-a case
593 study with wheat gluten genes. *Gene*, 533(2):541–546.
594

595 **Figure Legends**

596

597 **Figure 1: Correspondence between 45 randomly selected Sanger-sequenced SSI cDNAs**
598 **and the SMRT transcriptome.** A) Bowtie2 was used to determine the presence of 45 cDNA
599 clones in the SMRT SSI transcriptome. B) Alignment of matched sequences to SMRT SSI
600 transcriptome was performed using DNA Strider and manually evaluated as either equivalent
601 or chimeric. All Sanger sequenced clones were found in the SMRT dataset and a small
602 percentage were found to be chimeric. See Supplemental Figure 2 for corresponding analysis
603 of Illumina sequenced, Velvet/Oases transcriptome.

604

605 **Figure 2: Shared and restricted orthologous genes among 13 species.** All species shown
606 here shared 6067 core orthologs. Each petal shows the number of gene groups unique to each
607 species. Not shown are groups shared by only 2–12 species. *Solanum sisymbriifolium*, SSI;
608 *Solanum tuberosum*, STU; *Solanum lycopersicum*, SLY; *Solanum melongena*, SME;
609 *Arabidopsis thaliana*, ATH; *Carica papaya*, CPA; *Vitis vinifera*, VVI; *Prunus persica*, PPE;
610 *Populus trichocarpa*, PTR; *Citrus sinensis*, CSI; *Medicago truncatula*, MTR; *Zea mays*,
611 ZMA; and *Oryza sativa*, OSA. See Supplemental Figure 5 for gene ontology bins for the SSI
612 unique groups, and Supplemental Table 4 for sources of datasets.

613

614 **Figure 3: Final SMRT transcriptome sequences were backtracked through the de-**
615 **redundification process to the organ sub-transcriptomes.** A) Flower plot of genes
616 expressed in all, or in only one, organ. Each petal shows the number of genes only expressed
617 in one organ. In the center are the number of genes expressed in all 4 organs. Not shown are
618 genes expressed in 2 or 3 organs. B) Green-tissue specificity of sequences annotated as genes
619 involved in the light harvesting complex-I pathway via Mercator. C) Sequences annotated as
620 putative resistance genes because they contained nucleotide binding (NB-ARC) and leucine
621 rich repeat (LRR) domains showed varied expression patterns. As shown on the scales on
622 the right of (B) and (C), the darker the color, the more times the sequence was found in that
623 organ.

624

625 **Figure 4: Comparison of expression of 3 putative R-gene sequences in the SMRT**

626 **database to semi-quantitative PCR from 2 cDNA preparations.** A) The expression of
627 three genes with LRR and NB-ARC domains characteristic of the R-genes and an actin
628 isoform is shown in the heat map at the left and compared on the right to semi-quantitative
629 PCR of those same genes in two independently prepared cDNA pools, one from the pool
630 used to generate the transcriptome (Sequenced) and one prepared independently, and not
631 used to make the transcriptome (Unsequenced). B) The expression of each PCR product from
632 each pool was quantified and then normalized to the expression of an actin isoform
633 (Ssi032526). Data (biological replicates, n=2) are represented as mean ± STD. See
634 Supplemental Table 3 for primers used.
635

636 **Tables**

637 **Table 1: Summary of the SSI transcriptome derived using SMRT technology.** Organ
638 sub-transcriptomes were sequenced and combined from 33,170 root, 99,924 bud, 50,825
639 leaf, and 47,793 stem reads. See Supplemental Figure 3 and Supplemental Figure 4 for gene
640 ontology bins of this transcriptome.

641

SMRT Assembly		SSI
Total raw reads		231,712
Read lengths		300-7883
Total raw reads size (bp)		362,086,346
GC content		41.17
Transcripts	Number	139,611
	Total length	237,865,670
	N50	2,050
	Max length	7,883
	GC content	40.97
Unigenes	Number	41,189
	Total length	74,642,518
	N50	2,158
	Max length	7,883
	GC content	39.90

642 **Table 2: BUSCO assessment for completeness of 3 transcriptomes and one genome.** The SSI
643 transcriptome appears to be nearly complete, but contains a disproportion number of duplicated
644 sequences. See Supplemental Table 4 for sources of datasets.
645

	Transcriptome			Genome
	SSI (%)	Tomato (%)	Potato (%)	ATH (%)
Complete BUSCOs	93	96.2	86.7	97.7
Complete Single-copy BUSCOs	62.8	94.2	64.1	N/A
Complete Duplicated BUSCOs	30.2	2.0	22.6	N/A
Fragmented BUSCOs	2.0	0.8	4.7	0.6
Missing BUSCOs	4.1	3.0	8.6	1.7

Table 3: Divergent gene assessment of allele and/or paralog number in the SSI transcriptome. 4-allele genes were mapped to tomato chromosomes, see Supplemental Table 2.

Allele or Paralog	# of genes
Homozygous	17,773
2	22,098
3	1,358
4	44

651 **Table 4: R-gene profile of potato (STU), tomato (SLY), and the SSI transcriptome.** The SSI
652 database had fewer assigned R genes (based on the presence of nucleotide-binding domains and
653 leucine-rich repeats within the same open reading frame) than either SLY or STU genomes. Refer to
654 Supplemental Table 1 for full domain annotation statistics.

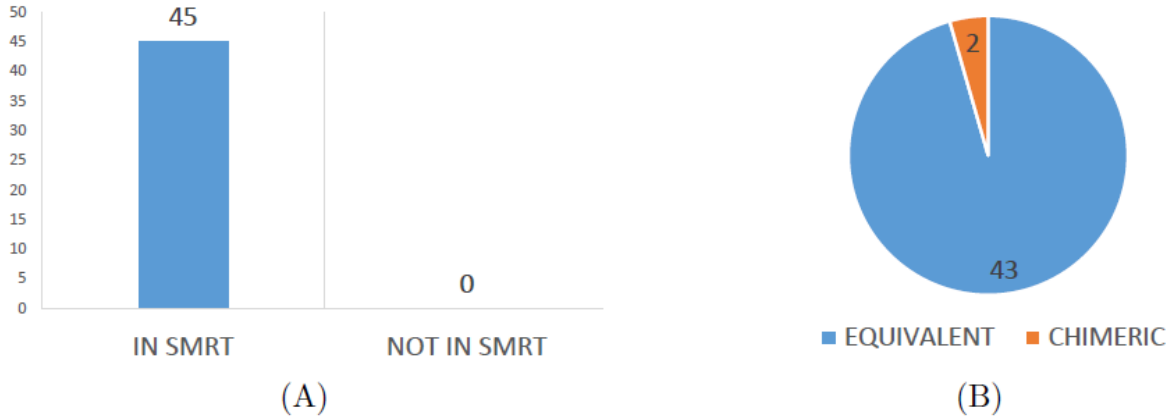
	R-genes
SSI	67
STU	309
SLY	137

655

656

657 **Figures**

658



659

660 **Figure 1: Correspondence between 45 randomly selected Sanger-sequenced SSI cDNAs and**
661 **the SMRT transcriptome.** A) Bowtie2 was used to determine the presence of 45 cDNA clones in
662 the SMRT SSI transcriptome. B) Alignment of matched sequences to SMRT SSI transcriptome was
663 performed using DNA Strider and manually evaluated as either equivalent or chimeric. All Sanger
664 sequenced clones were found in the SMRT dataset and a small percentage were found to be chimeric.
665 See Supplemental Figure 2 for corresponding Illumina sequenced, Velvet/Oases assembled
666 transcriptome.

667

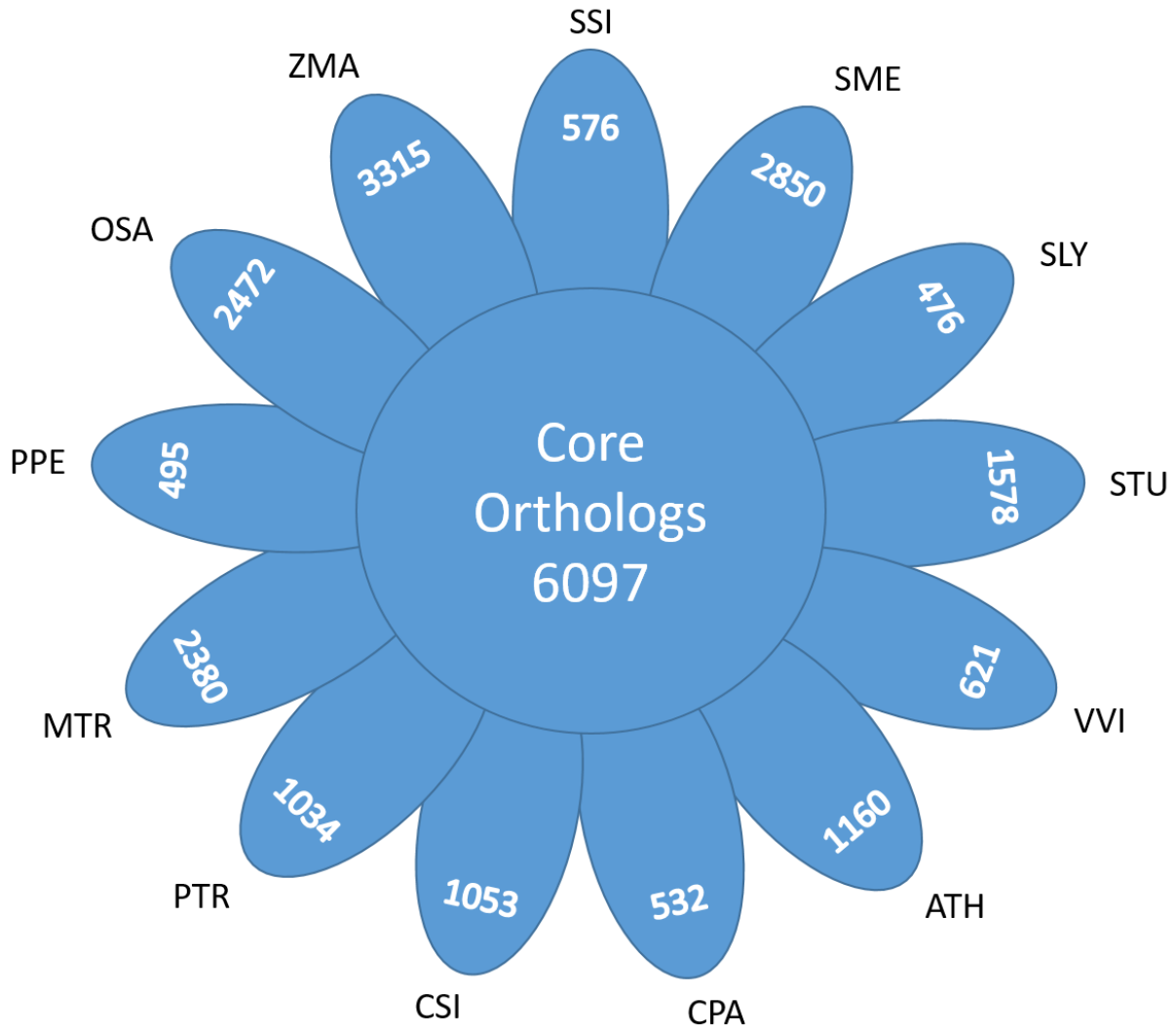


Figure 2: Shared and restricted orthologous genes among 13 species. All species shown here shared 6067 core orthologs. Each petal shows the number of gene groups unique to each species. Not shown are groups shared by only 2–12 species. *Solanum sisymbriifolium*, SSI; *Solanum tuberosum*, STU; *Solanum lycopersicum*, SLY; *Solanum melongena*, SME; *Arabidopsis thaliana*, ATH; *Carica papaya*, CPA; *Vitis vinifera*, VVI; *Prunus persica*, PPE; *Populus trichocarpa*, PTR; *Citrus sinensis*, CSI; *Medicago truncatula*, MTR; *Zea mays*, ZMA; and *Oryza sativa*, OSA. See Supplemental Figure 5 for gene ontology bins for the SSI unique groups, and Supplemental Table 4 for sources of datasets.

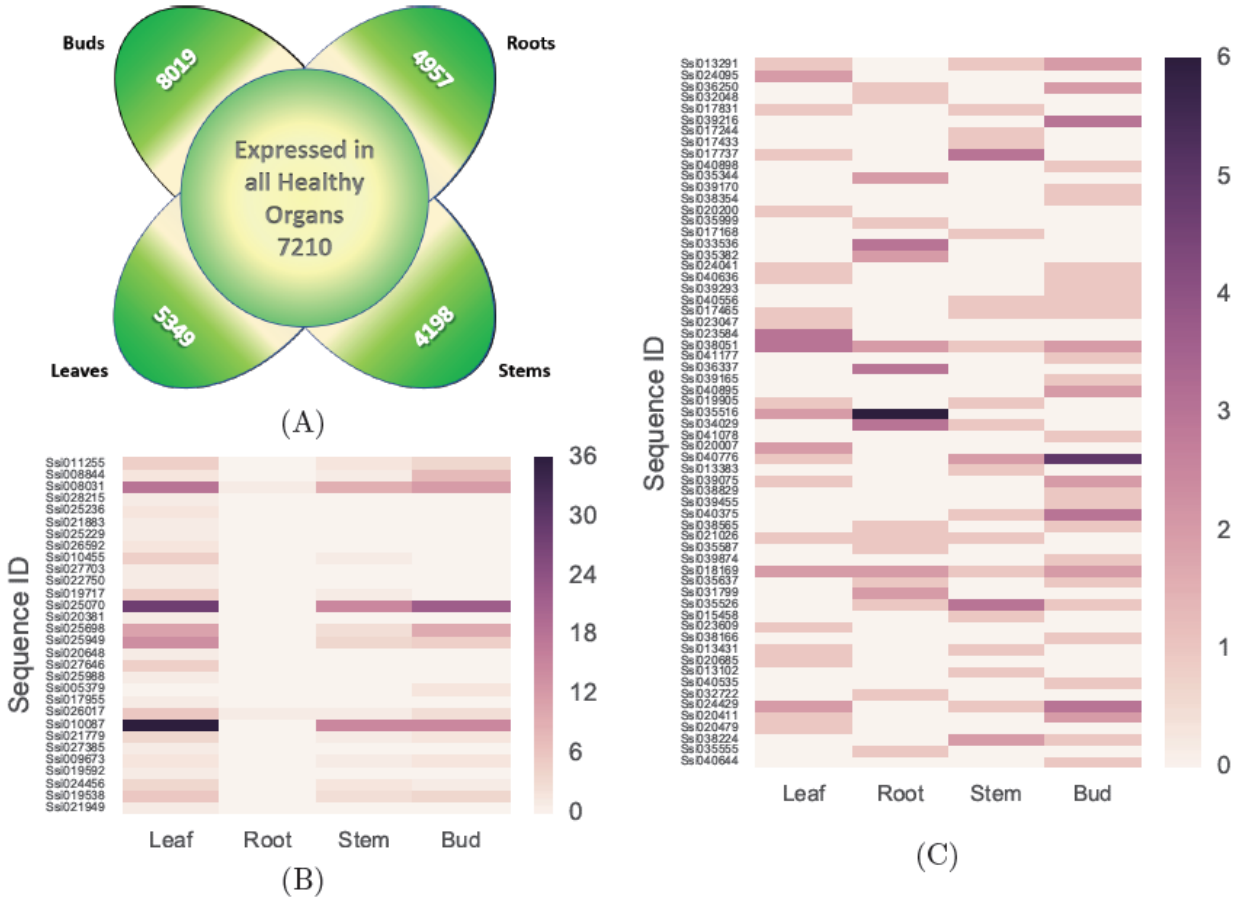
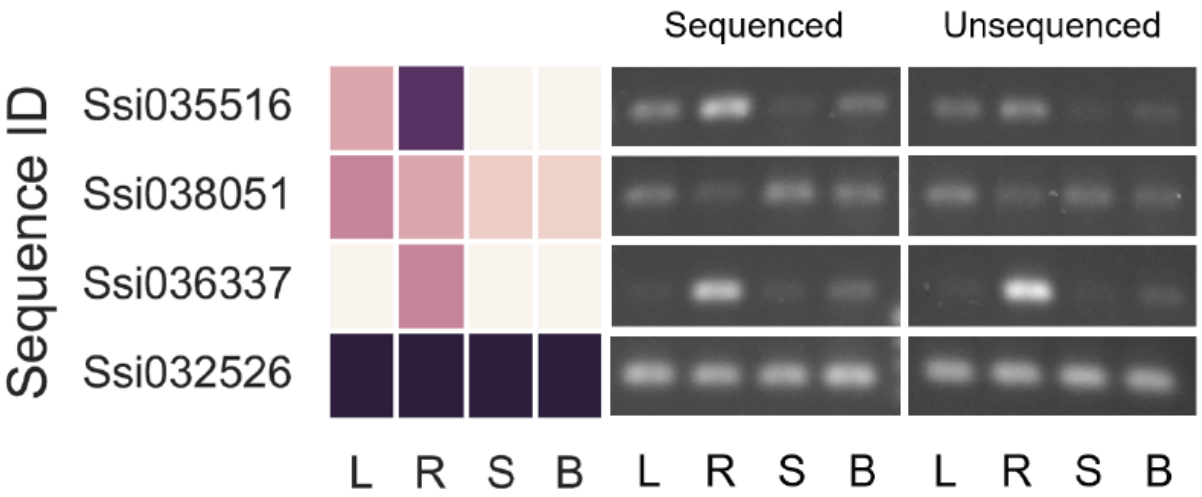


Figure 3: Final SMRT transcriptome sequences were backtracked through the de-redundification process to the organ sub-transcriptomes. A) Flower plot of genes expressed in all, or in only one, organ. Each petal shows the number of genes only expressed in one organ. In the center are the number of genes expressed in all 4 organs. Not shown are genes expressed in 2 or 3 organs. B) Green-tissue specificity of sequences annotated as genes involved in the light harvesting complex-I pathway via Mercator. C) Sequences annotated as putative resistance genes because they contained a nucleotide binding domain (NB-ARC) and leucine rich repeat (LRR) domains show varied expression patterns. As shown on the scales on the right of (B) and (C), the darker the color, the more times the sequence was found in that organ.



(A)

SSI sequence ID	*A. thaliana* homologue	Leaf	Root	Stem	Bud
Ssi035516	ADR1-L1 (at4g33300)	0.64 ± 0.22	0.99 ± 0.13	0.20 ± 0.02	0.39 ± 0.00
Ssi038051	NB-ARC domain-containing disease resistance protein (at1g50180)	0.88 ± 0.40	0.43 ± 0.24	0.69 ± 0.06	0.64 ± 0.13
Ssi036337	RPM1 (at3g07040)	0.22 ± 0.02	1.56 ± 0.56	0.19 ± 0.05	0.36 ± 0.10

(B)

Figure 4: Comparison of expression of 3 putative R-gene sequences in the SMRT database to semi-quantitative PCR from 2 cDNA preparations. A) The expression of three genes with LRR and NB-ARC domains characteristic of the R-genes and an actin isoform is shown in the heat map at the left and compared on the right to semi-quantitative PCR of those same genes in two independently prepared cDNA pools, one from the pool used to generate the transcriptome (Sequenced) and one prepared independently, and not used to make the transcriptome (Unsequenced). B) The expression of each PCR product from each pool was quantified and then normalized to the expression of an actin isoform (Ssi032526). Data (biological replicates, n=2) are represented as mean ± STD. See Supplemental Table 3 for primers used.