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 the roots, but then arrests further maturation. In order to provide researchers with a partial blueprint of its genetic make-up so that 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. Applying single-nucleotide polymorphism (SNP) analysis to the merged transcriptome indicated that this species has undergone a recent genome duplication and re-diploidization. Unexpectedly, 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 begins 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 instances, alternative methods using sequencing technologies, 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* 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 an Illumina and 454 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 and tetraploid potato between 3.31 to 3.86 pg per 2C. Therefore, SSI had 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.

The first attempt at generating an SSI transcriptome was done using a combination of Illumina and Roche 454 technologies. Each was used to compensate for the different types of mistakes inherent in the technology of the other, such as errors in sequence interpretation of homopolymer regions with 454, and in GGC motifs with Illumina (Loman *et al.,* 2012; Luo *et al.,* 2012). Plant material was harvested from leaves, roots, stems, and buds of uninfected plants growing in hydroponic solution, and RNA was extracted following protocols established in Casavant *et al.,* 2017). RNA from each organ was converted to cDNA and normalized. The normalized pools were then blended in equal amounts and sequenced in separate batches using Illumina MiSeq 2x300 and Roche 454 platforms. The 454 3' library consisted of 632,108 reads while the 454 5' library provided 326,451 reads. The Illumina library sequenced 31,310,146 reads (15,655,073 read pairs) which was digitally normalized to reduce redundant reads that could hinder efficient assembly. This normalization used a coverage cutoff of 30 and a kmer size of 20 to reduce the number of reads to 8,891,416 reads (4,446,708 read pairs). All remaining read pairs were then mapped to the chloroplast genome of *Solanum tuberosum*, and the unmapped reads that remained were used for assembly.

The assembly was performed using a Eurofins in-house script which included Velvet (v1.2.10) and Oases (v0.2.08). Further details of the assembly can be found in Materials and

114 Methods. This created a set of 351,982 contigs containing 148,324 transcripts with a GC
115 content of 39.1% which was then reduced at a 99% identity using CD-HIT-EST to produce
116 a final set of 102,226 unigenes that collectively had a GC content of 39.1% (Supplemental
117 Table 1). This GC content was slightly lower than expected when compared with other
118 Solanum species, such as potato (40.1%) and tomato (40.7%). This, added to the fact that
119 the number of unigene transcripts exceeded the expected number of genes for plants of this
120 family which has been estimated to range between 35,000−40,000 genes (Mueller *et al.,*
121 2009; Gálvez *et al.,* 2016) created a need to check the quality of the sequencing and assembly
122 processes.
123
124 **Bioinformatic assessment of the SSI transcriptome**
125 Our transcriptome represented a "snap-shot" of the metabolic pathways operating in
126 a single healthy plant growing in hydroponic medium. In order to judge whether the data we
127 obtained contained representatives of the sequences that have been found in other species,
128 we performed an annotation of it using Mercator (Lohse *et al.,* 2014). Mercator weights the
129 annotations in its database by the reliability of the source material. One of this program's
130 additional advantages is that it also bins the sequences based on the molecular pathway in
131 which they are most likely to participate (Supplemental Figure 2). In general, the percentage
132 of the SSI sequences binning to each functional class was similar to the percentages binned
133 using the much more intensively studied tomato and potato transcriptomes despite the fact
134 that we had sampled RNA from only 4 organs during a single phase of growth. Not only
135 were the percentages of "classifiable" genes in each of these species similar, but so too was
136 the percentage of "Not assigned" sequences as seen in Supplemental Figure 3. Finally, we
137 carried out one further commonly employed quality check using PfamScan (Finn *et al.,*
138 2008). PfamScan is a set of scripts that uses HMMer (Eddy, 1998) which is a domain
139 annotation program. This program in combination with protocols established by (Sarris
140 *et al.,* 2016), allowed for the annotation of domains found in the amino acid sequences
141 translated from the assembled transcriptome. Here, too, all three species showed very similar
142 percentages of sequences without an annotated domain, as can be seen in Supplemental Table
143 2.
144

145 **Evidence-based Quality Control of Illumina/454 Transcriptome**

146 The conventional sequence assembly analyses that we performed left nearly 50% of
147 the SSI sequences unannotated (Supplemental Figure 3). While this was not inconsistent
148 with previous studies (Yang *et al.,* 2014), it was high, and could obstruct future searches for
149 SSI genes. Thus, in addition to comparing our sequences with those of other species, we
150 performed an internal quality check by sequencing 45 randomly chosen clones from a cDNA
151 library using Sanger Dye Deoxy technology (ABI 3730, Applied Biosystems). Bowtie2
152 (Langmead and Salzberg, 2012) was then used to find the most likely equivalent of each
153 clone in the Illumina/454 assembly. A manual comparison of the Sanger-sequenced clone
154 and the assembled transcript was done using DNA Strider (Marck, 1988). This revealed that
155 out of 45 randomly selected clones, only 35 were found in the Illumina/454 assembled
156 transcripts (Figure 1A). Of these 35, only 19 of the clones were identical to at least 1
157 sequence in the transcriptome database. and more worryingly, 16 appeared to be chimeric
158 (Figure 1B). After consideration of alternative explanations, we concluded these chimeras
159 were generated through computational misassembly of short-read data.

160

161 **Establishing a SMRT sequenced transcriptome**

162 The chimericism found in many of the transcriptome sequences would
163 unquestionably hinder future attempts to use this database as a reference library for gene
164 identification. For this reason, we turned to a newer generation of sequencing, called Single-
165 Molecule Real Time (SMRT) sequencing by PacBio Sciences (Eid *et al.,* 2009) that does not
166 need to assemble short reads into one contiguous sequence. Rather than producing short
167 reads of up to 500 bp, SMRT provides reads up to 60,000 bp along a single molecule. This
168 allows the capture of entire genes with one read rather than chunking it into many small bits
169 that have to be assembled later. This sequencing strategy gives higher coverage than Sanger-
170 based reactions like those performed on Applied Biosystems™ gene analysis instruments,
171 and longer reads than Illumina/454. With SMRT technology, bases are checked with the read
172 coverage, and the sequence of entire genes can be obtained in single reads.

173 The Iso-Seq pipeline classifies the sequences as either full-length non-chimeric
174 (FLNC), or non-full length reads. Full length reads are those containing both 5' and 3'
175 adapters, in addition to the poly(A) tail. The reads containing these parts in the expected

176 order, i.e. 5' adapter–poly(A)–3' adapter, with no additional copies of these parts, are
177 classified as non- chimeric. The FLNC reads were corrected with the non-full length reads
178 using Iterative Cluster for Error correction and the Pacific Biosciences Quiver algorithm
179 (https://github.com/PacificBiosciences/SMRT-Analysis/wiki/ConsensusTools-v2.3.0-
180 Documentation).

181 In an attempt to improve our ability to detect differences in the suites of genes
182 expressed in different parts of the plant, we generated cDNA from 4 organs; leaves, stems,
183 roots, and unopened flower buds. We then independently carried out SMRT sequencing of
184 all 4 samples. Finally, all corrected FLNC reads were merged *in silico* and redundancy was
185 removed using CD-HIT-EST.

186 This SMRT sequencing strategy created 231,712 total corrected full-length non-
187 chimeric (FLNC) sequences (Table 1) using the aforementioned pipeline. These sequences
188 had a GC content of 41.2% which was slightly higher than the estimate of 40.9% that had
189 been based on Illumina/454 sequencing. CD-HIT-EST was used to reduce the redundant
190 sequences to sets with 100% identity. This lowered the number of sequences to 139,611 with
191 a GC content of 41.0%. The GC content continued to decrease further as the identity was
192 reduced using CD-HIT-EST. At 80% identity, there were 32,315 sequences, with an
193 estimated GC content of 39.7%. The decrease in GC content could be due to the methodology
194 of CD-HIT-EST that retains the longest sequence during the reduction process. Because of
195 this, reads that spanned untranslated regions of a transcript were expected to be favored over
196 those only consisting of coding regions.

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

202

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

204 We conducted the same quality checks on this provisional transcriptome that had
205 been performed on the previous Illumina/454 assembly. A search for *in silico* homologs of
206 the cDNA clones used previously revealed a much higher concordance then the assembled

207 short read sequences to which they had initially been compared. Firstly, all 45 cDNAs were
208 found in the SMRT transcriptome (Figure 1C). Secondly, only two of these SMRT-derived
209 sequences were chimeric (Figure 1D), and based on their short lengths, could have been
210 transcribed from different members of the same gene family rather than been created by
211 misassembly. During our analysis of the SMRT transcriptome, we did find entries that
212 consisted of inverted repeats of entire gene sequences. These inverted repeats likely occurred
213 during the preparation of the cDNA library prior to sequencing rather than during sequencing
214 or subsequent computational processing.

215 When the SMRT transcriptome was analyzed using Mercator, there was an increase
216 in annotated sequences compared to the Illumina/454 assembly. The SMRT transcriptome
217 had 38.6% unannotated sequences. This was markedly fewer than the percent unannotated
218 sequences of either potato (50.8%) or tomato (46.4%) transcriptomes processed in the same
219 way via Mercator (Supplemental Figure 3). Other than that, the binned profile of SSI was
220 very similar to the published transcriptomes (Supplemental Figure 2) of these plants. This
221 led us to believe that our transcriptome was at least of comparable quality with the working
222 transcriptomes of these two better studied species, and may have contained fewer chimeric
223 sequences than transcriptomes assembled only from Illumina/454 data.

224 PfamScan was also used to annotate the domains of the SMRT transcriptome. This
225 annotated 88.6% of the SMRT transcriptome with at least one recognizable domain. There
226 were fewer unannotated sequences in the SSI SMRT transcriptome than in the original
227 Illumina/454 version. Although the original number of unannotated sequences in the SMRT
228 data compared to STU and SLY transcriptomes (Supplemental Table 2), we now believe
229 many of those in the Illumina/454 database were chimeras. The reduced number of
230 unannotated sequences found in the SMRT transcriptome, compared to STU and SLY, might
231 reflect the fact that this set had undergone a conservative reduction to 90% identity.
232 Alternatively, the reduced number of unannotated SSI sequences could have resulted from
233 the fact that we had only sampled a "normally" growing plant, that is, plants manifesting a
234 physiological state which has been extensively studied in numerous species, while the STU
235 and SLY transcriptomes were compiled from plants sampled over a much broader range of
236 life-history stages and growth conditions ranging from fruit and tuber development to
237 exposure to biotic and abiotic stresses where the functions of many genes are still under

238 investigation.

239 To further test the quality and completeness of our transcriptome, BUSCO
240 benchmarking (Simão *et al.,* 2015) was performed. The BUSCO database was established to
241 allow researchers to assess the completeness of newly completed genome or transcriptomes
242 based on the detection of a set of universal, single-copy orthologs. We found 93% intact
243 BUSCO archetypes (889 genes) in the SSI SMRT transcriptome, 30.2% (289) of these were
244 found in multiple copies, while an additional 2.2% (21) of the BUSCO archetypes were
245 present in fragments, and 4.8% (46) were missing entirely (Table 2). These numbers
246 representing genes expressed during a single growth condition of SSI, were only somewhat
247 different than the numbers of BUSCO archetypes found in the entire SMRT sequenced
248 genome of *A. thaliana*.

249 SSI has the same number of chromosomes as most other Solanaceae, and does not
250 appear to be polyploid (data not shown), yet the BUSCO analysis showed that SSI had more
251 duplicate copy archetypes than diploid and tetraploid potato. This high number of similar
252 sequences could point to the fact that our transcriptome hasn't been reduced far enough, or
253 could be one line of evidence that SSI has undergone extensive genome duplication or
254 hybridization in the past. This latter hypothesis was evaluated by SNP analysis as has been
255 done with plants such as wheat (Krasileva *et al.,* 2013). When the program Freebayes
256 (Garrison and Marth, 2012) was run using a defined diploid setting, it output information
257 stating there were genes that had more than 2 alleles (or paralogues). We redid the analysis
258 using a defined tetraploid setting and found 44 genes with 4 alleles (Table 3). It is noteworthy
259 that no gene had more than 4 alleles. A simple explanation for these multiple alleles is that
260 SSI underwent a genome duplication followed by diploidization in the past and that over
261 time, some of the duplicated loci acquired additional mutations while other loci were lost.
262 To determine if this proposed duplication was restricted to one chromosome, or one
263 chromosomal arm, the 44 genes with 4 alleles were mapped onto SLY chromosomes
264 (Supplemental Table 3). There were 4–allele genes found on 11 of 12 SLY chromosomes
265 which indicated, assuming that genes dispersed in tomato did not become linked when the
266 two species diverged, that SSI has undergone a full genome duplication rather than a
267 segmental duplication within one chromosome.

268 Since SSI is not as well-known as other Solaneacae, we employed OrthoMCL v2.0.9

269 (Li *et al.,* 2003) to illustrate some of the common features its gene complement showed with
270 those of other plants. Protein sequences from our SSI transcriptome (translated using the
271 program ESTScan (Iseli *et al.,* 1999)), and protein sequences from tomato (SLY), potato
272 (STU), eggplant (SME), *Arabidopsis thaliana* (ATH), papaya (CPA), grapes (VVI), peaches
273 (PPE), black cottonwood (PTR), oranges (CSI), alfalfa (MTR), maize (ZMA) and rice (OSA)
274 were merged into 45,234 orthologous groups (gene families). In this set, 6097 orthologous
275 groups were shared by all 13 species, an overlap well within the range of previous studies
276 (Figure 2) (Yang *et al.,* 2014). Each species had many additional groups that were not shown
277 in this diagram because they were not shared with all members of this set of plants.
278 Interestingly, even closely related species like SSI, STU, SLY, and SME had hundreds of
279 groups not found in each other. When the annotations of the SSI unique set were compared
280 to the full transcriptome, several functional groups showed a disproportionate increase. It is
281 possible that these disproportionately expanded sets, that included photosynthetic genes, and
282 genes for amino acid and vitamin metabolism, diverged so much more than groups such as
283 those for cell wall composition, and hormone and secondary metabolism, was because
284 expansion of the former traits gave SSI a competitive edge over other species in their habitat
285 (Supplemental Figure 4). Overall, though, there were fewer groups of genes unique to SSI
286 than unique to STU and SME. As noted previously, this could merely reflect the fact that our
287 data came from a single-point snapshot of only 4 organs and so would have lacked those
288 transcripts specifically expressed during fruit and seed set, germination, senescence, abiotic
289 stress, pathogen attacks, and numerous other stages of a plant's lifecycle.

290 Using highly conserved orthologous genes, i.e. subunits of Rubisco, provisional
291 phylogenies were created for nuclear-encoded and chloroplast-encoded genes using the
292 aforementioned species (data not shown). In doing so, we concluded that nuclear SSI was
293 most closely related to eggplant, which has been noted previously (Särkinen *et al.,* 2015),
294 while chloroplast SSI was more closely related to tomato. This dichotomy has also been seen
295 by others (Miz *et al.,* 2008) and interpreted to indicate that SSI had undergone an ancient
296 hybridization and afterwards retained the chloroplast genome from one parent, and much of
297 the nuclear genome from another, however, many more SSI genes will have to be compared
298 with the genes of many more South American plants to confirm that this hybridization
299 occurred.

300

301 **Building a snapshot of organ-associated gene expression**

302 Since we had maintained separate cDNA pools from individual organs, it was
303 possible to backtrack each sequence within the final transcriptome to obtain a provisional
304 profile of gene expression throughout the plant (Figure 3A). This analysis showed that there
305 were 8019 sequences expressed solely in buds, 4957 solely in roots, 5349 solely in leaves,
306 4198 solely in stems, and 7212 sequences expressed in all tissues. That left 11,538 sequences
307 that were expressed in more than one organ but not in all 4.

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

312 A set of light-harvesting complex genes (LHC-I) were predictably found in aerial
313 organs with few exceptions (Figure 3B), demonstrating that the backtracking program could
314 extract biologically useful information about sequences with specified characteristics from
315 the merged transcriptome. In order to determine if this kind of analysis of SMRT sequences
316 could categorize the expression of very different sets of genes, we constructed an inter-organ
317 expression profile of genes that encoded both a leucine-rich repeat (LRR) domain together
318 and a nucleotide binding (NB-ARC) domain (Figure 3C), a pairing frequently found in
319 pathogen resistance genes (R-genes). A profile of R-gene prevalence in SSI, potato, and
320 tomato indicates there is a reduction of these genes in the SMRT transcriptome compared to
321 the other two species (Table 4). Three of these potential R-genes were then assayed by semi-
322 quantitative PCR (primers found in Supplemental Table 4) and quantified using a sample of
323 cDNA from the same pool that had been sequenced, and a sample from an independently-
324 prepared, unsequenced cDNA pool (Figure 4). In order to assess whether a SMRT data set
325 could be a reliable indicator of gene expression, both the *in silico* and PCR measurements of
326 gene expression were normalized in kind to an actin sequence (*ssAct7)*. The physical
327 measurements of expression of two of the three genes matched the expression snapshot
328 extremely well, but the third gene (Ssi038051) was more abundant in stems and buds than
329 expected based on its SMRT expression snapshot. This confirms that whole transcriptome
330 snapshots can provide a provisional picture of organ differences in gene expression, but

331 further shows that the expression of each gene of interest needs to be verified by independent
332 tests.

333 **Discussion**

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

336 We initially employed both Illumina and 454 sequencing in order to compensate for
337 the different kinds of errors to which each method was prone (Luo *et al.,* 2012). Our raw
338 data using these methods had an average coverage of 30, but after it was assembled *de novo*
339 using Velvet and Oases, many of the products were unexpectedly long, some being more
340 than 24.5 kb in length. We were concerned that such unusually long sequences were artifacts
341 of the assembly process. Further examination revealed that a number of randomly chosen
342 sequences lacked full-length open reading frames, and seemed to be chimeras consisting of
343 functionally disparate domains. More tellingly, 40% of the genes randomly selected from an
344 SSI cDNA library failed to match any of the assembled sequences in our short-read database
345 (Figure 1B). We do not know whether all *de novo* transcriptomes are prone to misassembly,
346 but the presence of misassembled sequences in a database is certainly not unique to the
347 current study (Yang and Smith, 2013). A number of factors are known to exasperate
348 misassembly including the presence of large gene families and of repeatedly occurring kmers
349 in the dataset (Moreton *et al.,* 2015). Even though we did not sequence the SSI genome, we
350 found 4 lines of evidence indicating that it might be complex enough to pre-dispose our
351 transcriptome to these kinds of mistakes. First, the nuclear DNA content of SSI was larger
352 than most diploid Solanaceae, roughly the same size as a tetraploid potato (Supplemental
353 Figure 1). Second, SNP analysis indicated that the SSI transcriptome was highly
354 heterologous and contained 3 and 4 distinguishable alleles for many genes (Table 3). Third,
355 putative R-genes, which contain a nucleotide binding domain (NB-ARC) and a leucine-rich
356 repeat domain (LRR), amounted to only 67 genes in the SMRT sequenced dataset compared
357 to the 301 genes appearing in the Illumina/454 (Table 4). Finally, an unusually high
358 percentage of the BUSCO gene set were present in multiple copies in SSI even though our
359 transcriptome could only consist of a portion of all the genes that are likely to be encoded in
360 its DNA (Table 2). One model consistent with these 4 facts was that SSI had, sometime in
361 the past, undergone a partial or complete genome duplication. Over time, as diploidy was re-
362 established, some of the duplicated alleles or paralogues diverged, while others were lost.
363 Nevertheless, enough of the expanded gene families remained to confound the alignment

364 programs that tried to differentiate between their members. While these kinds of errors might
365 be correctable with the use of other assembly programs, we chose, instead to create an
366 assembly-independent transcriptome using SMRT technology.

367 At the moment, SMRT technology does not provide the sequence coverage that can
368 be obtained with Illumina/454 sequencing. In order to increase our chances of sampling
369 uncommon organ-specific transcripts, we prepared independent cDNA pools from 4 parts of
370 the plant. Using an in-house script (https://github.com/AlexWixom/Transcriptome_scripts),
371 we were able to increase the value of the final library by generating expression snapshots for
372 genes of interest in each organ. These expression snapshots are no substitute for a more
373 thorough RNA-seq study, but they do provide a preliminary assessment of a plant's biology
374 at the time of harvest. Using these snapshots, we recognized different patterns of expression
375 of individual LHC-1 genes (Figure 3B) within the photosynthetic parts of the plant. We also
376 saw that 2 of the 3 R-genes re-examined by PCR showed the same expression pattern in two
377 independent RNA and cDNA preparations as found in the transcriptome itself (Figure 4).
378 Thus, in the absence of RNA-seq studies or experimental evidence for the role of a specific
379 locus, this kind of library assembly could be used to direct researchers to the subset of R-
380 genes most likely responsible for the resistance in a given organ.

381 R-genes coding for recognition proteins are commonly perceived as sentinels that
382 would be awaiting activation by molecules introduced during infection (Jones and Dangl,
383 2006). Therefore, the reduced number of R-genes found in SSI (67), compared to both potato
384 (309), or tomato (137) was unexpected based on the seemingly enriched protective responses
385 in SSI. This discrepancy could be explained in any one of several ways. First, SSI might be
386 using proteins with novel domain structures in place of classic R-genes. Second, sequencing
387 depth might simply have been inadequate to capture all R-genes that were actually being
388 expressed at low levels. Finally, SSI could be relying on rapidly inducing transcription of R-
389 genes after an infection has occurred. Any one of these hypotheses is worthy of continuing
390 analysis.

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

395 gleaned from a genomic sequence.

396 **Materials and Methods**

397

398 **Plant and Culture Conditions**

399 *S. sisymbriifolium* (SSI) seeds obtained from C. Brown (USDA-ARS, Prosser WA)
400 were germinated in soil. Nodes from a single plant were sterilized for 20 min using 10%
401 NaClO with 0.05% Tween20. Plant material was then washed 3x with sterile distilled H2O
402 and put into 120 mL baby food jars containing standard Murashige and Skoog salts, pH 5.6,
403 3% sucrose, 0.7% agar, 100 µg mL^{-1} myo-inositol, 2.0 µg mL^{-1} glycine, 1 µg mL^{-1} thiamine,
404 0.5 µg mL^{-1} pyridoxine, and 0.5 µg mL^{-1} nicotinic acid. A single plant was chosen as the
405 progenitor of all of the plants used in this study. All of its descendants were maintained at
406 25˚C in 16 h light, and subcultured vegetatively every 4 wk. Over the course of the project,
407 rooted clones with at least 4-6 leaves were put into 2 L of hydroponic medium (Yoshida
408 *et al.,* 1976), referred to here as Fake Field. Each container was diffusely aerated through an
409 aquarium stone, maintained at constant volume by the addition of distilled water, and
410 emptied and refilled with fresh hydroponic medium every 7 d. Hydroponic containers were
411 maintained at 22˚C, 16 h light with an irradiance level of 0.0006 W m^{-2}. Illumination was
412 provided by GE Lighting Fluorescent lamps (13781, F96T12/CW/1500). After a 2 wk lag-
413 time, plants began producing 1-3 new leaves each wk, and flowered continuously afterwards.
414 All experiments were performed on plants that had not been infected or wounded in any way
415 previously.

416

417 **RNA extraction**

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

421

422 **Genome size estimation by Flow Cytometry**

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

427 with a fresh razor blade in a clean Pyrex petri plate. The finely chopped suspension was then
428 filtered through a 50µm nylon mesh filter (Partec 04-0042-2317). This filtered suspension
429 was kept in the dark at 4˚C for between 15-90 min before it was analyzed.
430 Genome size estimations were made using a BD FACSARIA Flow Cytometer
431 (IBEST Imaging Core, University of Idaho, Moscow, ID, USA). A green laser at 488nm was
432 used to excite the propidium iodide stained cells and was then collected in the PE-A channel.
433 Thresholds for PE-A were set at 1,000 and FSC at 500. The voltages were set so the major
434 peak (2C) of the SSI samples were near 50,000 on the linear scale. Four suspensions were
435 made from separate donor plants once a day for three consecutive days. Two replicates of
436 two external standards were also used daily in addition to the 4 SSI samples. External
437 standards included *Solanum lycopersicum cv.* Stupicke polni tyckove rane (2C= 1.96 pg
438 DNA) and *Glycine max* cv. Polanka (2C= 2.50 pg DNA) (Doležel *et al.,* 1992; Doležel *et al.,*
439 1994) which were chosen because their genome sizes were in the expected range of SSI. One
440 repetition of internal standards was run using tomato and soybean. DNA content was
441 estimated using the equation described by Doležel *et al.,* 2007).
442
443 **Illumina and 454 Sequencing and *de novo* Transcriptome Assembly**
444 Extracted total RNA from each previously stated organ was sent to Eurofins
445 Genomics (Ebersberg, Germany) for library preparation and sequencing. Prior to library
446 preparation, quality control (QC) was performed on individual tubes of RNA and equal
447 aliquots of each preparation were blended into one pool. The Illumina library cDNA was
448 prepared using randomly-primed first and second strand synthesis, followed by gel sizing
449 and PCR amplification. The library was then physically normalized and found to have insert
450 sizes of 250-450 bp.
451 Two libraries were prepared for 454 sequencing. The 3'-fragment cDNA library was
452 pre- pared using polydT and synthesis, followed by gel sizing, PCR amplification, library
453 purification and QC. The 5'-fragment library for 454 was prepared by dephosphorylating the
454 mRNA, cleaving the 5'-cap structure and ligating a 5'-synthetic RNA adaptor. This was
455 followed by randomly-primed first strand synthesis, library purification, and QC.
456 Illumina sequencing was performed on a MiSeq v3 2x300. The read sequences were
457 clipped using Trimmomatic, version 0.32 (Bolger *et al.,* 2014), and bases with a Phred score

458 < 20 were removed. Trimmed reads shorter than 150 bp were removed; this step could
459 remove none, one, or both mates of a read-pair. Prior to assembly, digital normalization was
460 applied to the Illumina reads in order to reduce redundant information present in these large
461 datasets. A coverage cutoff of 30 and a kmer size of 20 decreased the data to 28% of the
462 initial 31,310,146 reads. BWA (Li and Durbin, 2009) was used to map read pairs to the
463 *Solanum tuberosum cultivar* Desiree chloroplast genome (GenBank accession DQ38616.2).
464 Only unmapped reads were retained for assembly. The 454 3' and 5' fragment cDNA libraries
465 were sequenced in 2x 1/2 segment of a full FLX++ run. These reads were converted to
466 FASTQ format using the software convert project (http://mira-assembler.sourceforge.net).
467 These datasets were assembled using a Eurofins in-house pipeline which included
468 Velvet (v1.2.10) and Oases (v0.2.08) (Zerbino and Birney, 2008; Schulz *et al.,* 2012). A
469 multi-kmer approach was applied using assemblies of 59, 69, 79, and 89 kmers that were
470 finally merged into a single assembly using a kmer 29. This assembly was then clustered at
471 an identity threshold of 0.99, corresponding to 1 mismatch in 100 bases, using the software
472 CD-HIT- EST v4.6 (Li and Godzik, 2006).
473
474 **Library Preparation for Iso-Seq**
475 SMRT library preparation and sequencing were performed by the National Center for
476 Genome Resources (Santa Fe, New Mexico). The Iso-Seq libraries for four organs, root,
477 stem, leaf and bud, were prepared for Isoform Sequencing (Iso-Seq) using the Clontech
478 SMARTer PCR cDNA Synthesis Kit and the BluePippin Size Selection System protocol as
479 described by Pacific Biosciences (https://goo.gl/ij71Hh) with the following modifications.
480 For cDNA conversion, 3 µg of total RNA was put into each Clonetech SMARTer reaction.
481 From the PCR optimization procedure specified in the protocol, it was determined that 14
482 cycles of PCR would be sufficient for amplification of each organ's cDNA. Amplification
483 was followed by size selection on each sample to obtain three size bins (0.5-2 kb, 1.5-3 kb
484 and 2.5-6 kb) using the Blue Pippin (Sage Science, Beverly, Massachusetts) instrument. The
485 amplified and size selected cDNA products were made into SMRTbell Template libraries per
486 the Isoform Sequencing protocol referenced above. Libraries were prepared for sequencing
487 by annealing a sequencing primer (component of the SMRTbell Template Prep Kit 1.0) and
488 then binding polymerase to this primer-annealed template. The polymerase-bound template

489 was bound to MagBeads (P/N 100-125-900) (https://goo.gl/wdZErU) and sequencing was
490 performed on a PacBio RS II instrument. 12 v3 SMRTcells were run for the root tissues, 14
491 for the leaf tissues, 9 for the stem tissues, and 12 for the bud tissues for a total of 47
492 SMRTcells (Pacific Biosciences, P/N 100-171-800). The libraries from each organ were
493 separately sequenced using P6C4 polymerase and chemistry and 240-minute movie times
494 (Pacific Biosciences, P/N 100-372-700, P/N 100-356-200).
495
496 **Single Molecule Real Time Sequencing**
497 All SMRTcells for a given organ were run through the Iso-Seq pipeline included in
498 the SMRT Analysis software package. First, reads of insert (ROIs, previously known as
499 circular consensus sequences, CCS) were generated using the minimum filtering
500 requirement of 0 or greater passes of the insert and a minimum read quality of 75. This
501 allowed for the high yields going into subsequent steps, while providing high accuracy
502 consensus sequences where possible. The pipeline then classified the ROI in terms of full-
503 length, nonchimeric and non-full length reads. This was done by identifying the 5' and 3'
504 adapters used in the library preparation as well as the poly(A) tail. Only reads that contained
505 all three in the expected arrangement and did not contain any additional copies of the adapter
506 sequence within the DNA fragment were classified as full-length non-chimeric copies.
507 Finally, all full-length non-chimeric reads were run through the Iterative Clustering for Error
508 correction algorithm then further corrected by the Pacific Biosciences Quiver algorithm
509 (https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-
510 transcriptome-data). Once the Iso-Seq pipeline result was available for each organ, the results
511 were combined into a single data set and redundant sequences were removed using CD-HIT-
512 EST (Li and Godzik, 2006).
513
514 **Annotation of Sequences**
515 Mercator sequence annotation was performed using the TAIR, PPAP, KOG, CDD,
516 IPR, BLAST CUTOFF of 80, and ANNOTATE options (Lohse *et al.,* 2014).
517
518 **Annotating Protein Domains of Translated Sequences**
519 PfamScan (Finn *et al.,* 2009) was run on the SSI transcriptomes following protocols

520 set forth by Sarris *et al.,* 2016).

521

522 **Biological Quality Check of *in silico* Sequences**

523 45 clones from a cDNA library (Express Genomics, Average insert size=1 kb,
524 Vector= pExpress 1) were randomly selected and sequenced via Sanger Dye-Deoxy DNA
525 Sequencing (ABI 3730). These sequences were then aligned to the transcriptomes using
526 Bowtie2 (Langmead and Salzberg, 2012) set for local alignment and best hit only. These
527 aligned sequences were then manually compared for possible chimeric features.

528

529 **Evolutionary Comparison of SSI to 13 Other Species**

530 The evolutionary clustering and comparison protocols were adapted from those set out in
531 Yang *et al.,* 2014). See Supplementary Table 5 for species used and online download sources.

532

533 **Single Nucleotide Polymorphism (SNP) Analysis to determine Ploidy**

534 Phasing of the SMRT transcriptome was completed using the unassembled Illumina
535 sequences adapted from protocols established by Krasileva *et al.,* 2013), with the addition of
536 an in-house Python script to quantify SNPs present per sequence
537 (https://github.com/AlexWixom/Transcriptome_scripts/freePloidy.py).

538

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

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

543

544 **Expression Snapshot Validation**

545 Sequence specific oligonucleotides were designed for several genes that were then
546 used to obtain semi-quantitative PCR expression snapshots on the same cDNA used to obtain
547 our SSI transcriptome (referred to as the "Sequenced" sample), as well as on a second cDNA
548 pool prepared from an independent RNA collection (referred to as the "Unsequenced"
549 sample). PCR fragment bands were quantified with a local background subtracted and
550 normalized to actin (following the procedure established by Casavant *et al.,* 2017)). The

551 primers for these genes can be found in Supplementary Table 4 with the proposed gene
552 description.
553
554 **Accession Numbers**
555 The SMRT sequenced transcriptome has been deposited at DDBJ/EMBL/GenBank under
556 the accession GGFC00000000. The version described in this paper is the first version,
557 GGFC01000000.

558 **Author Contributions**

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

564

Acknowledgments

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

572 Arumuganathan, K. and Earle, E. (1991). Nuclear dna content of some important plant
573 species. *Plant molecular biology reporter*, 9(3):208–218.
574 Bagalwa, J.-J. M., Voutquenne-Nazabadioko, L., Sayagh, C., and Bashwira, A. S. (2010).
575 Evaluation of the biological activity of the molluscicidal fraction of solanum sisymbriifolium
576 against non target organisms. *Fitoterapia*, 81(7):767–771.
577 Bolger, A., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina
578 sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–2120.
579 Casavant, N. C., Kuhl, J. C., Xiao, F., Caplan, A. B., and Dandurand, L.-M. (2017).
580 Assessment of globodera pallida rna extracted from solanum roots. *Journal of nematology*,
581 49(1):12.
582 Dandurand, L.-M. and Knudsen, G. (2016). Effect of the trap crop *Solanum sisymbriifolium*
583 and two biocontrol fungi on reproduction of the potato cyst nematode, *Globodera pallida*.
584 *Annals of Applied Biology*, 169(2):180–189.
585 Doležel, J., Binarová, P., and Lcretti, S. (1989). Analysis of nuclear dna content in plant cells
586 by flow cytometry. *Biologia plantarum*, 31(2):113–120.
587 Doležel, J., Doleželová, M., and Novák, F. (1994). Flow cytometric estimation of nuclear
588 dna amount in diploid bananas (musa acuminata andm. balbisiana). *Biologia plantarum*,
589 36(3):351–357.
590 Doležel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear dna content in plants
591 using flow cytometry. *Nature protocols*, 2(9):2233–2244.
592 Doležel, J., Sgorbati, S., and Lucretti, S. (1992). Comparison of three dna fluorochromes for
593 flow cytometric estimation of nuclear dna content in plants. *Physiologia plantarum*,
594 85(4):625–631.
595 Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, 14(9):755.
596 Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P.,
597 Bettman, B., *et al.* (2009). Real-time dna sequencing from single polymerase molecules.
598 *Science*, 323(5910):133–138.
599 Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L.,
600 Gunasekaran, P., Ceric, G., Forslund, K., *et al.* (2009). The pfam protein families database.
601 *Nucleic acids research*, 38(suppl_1):D211–D222.
602 Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G.,

603 Forslund, K., Eddy, S. R., Sonnhammer, E. L., *et al.* (2008). The pfam protein families
604 database. *Nucleic acids research*, 36(suppl 1):D281–D288.
605 Gálvez, J. H., Tai, H. H., Barkley, N. A., Gardner, K., Ellis, D., and Strömvik, M. V. (2016).
606 Understanding potato with the help of genomics.
607 Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read
608 sequencing. https://arxiv.org/abs/1207.3907v2.
609 Iseli, C., Jongeneel, C. V., and Bucher, P. (1999). Estscan: a program for detecting,
610 evaluating, and reconstructing potential coding regions in est sequences. In *ISMB*,
611 volume 99, pages 138–148.
612 Jones, J. D. and Dangl, J. L. (2006). The plant immune system. *Nature*, 444(7117):323–329.
613 Krasileva, K. V., Buffalo, V., Bailey, P., Pearce, S., Ayling, S., Tabbita, F., Soria, M., Wang,
614 S., Akhunov, E., Uauy, C., *et al.* (2013). Separating homeologs by phasing in the tetraploid
615 wheat transcriptome. *Genome Biology*, 14(6):R66.
616 Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature*
617 *methods*, 9(4):357–359.
618 Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler
619 transform. *Bioinformatics*, 25(14):1754–1760.
620 Li, L., Stoeckert, C. J., and Roos, D. S. (2003). Orthomcl: identification of ortholog groups
621 for eukaryotic genomes. *Genome research*, 13(9):2178–2189.
622 Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets
623 of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.
624 Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Tohge, T., Fernie, A. R.,
625 Stitt, M., and Usadel, B. (2014). Mercator: a fast and simple web server for genome scale
626 functional annotation of plant sequence data. *Plant, cell & environment*, 37(5):1250–1258.
627 Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and
628 Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing
629 platforms. *Nature biotechnology*, 30(5):434–439.
630 Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K. T. (2012). Direct
631 comparisons of illumina vs. roche 454 sequencing technologies on the same microbial
632 community dna sample. *PloS one*, 7(2):e30087.
633 Marck, C. (1988). 'dna strider': a 'c'program for the fast analysis of dna and protein

634 sequences on the apple macintosh family of computers. *Nucleic acids research*, 16(5):1829–
635 1836.
636 Meyre-Silva, C., Niero, R., Bolda Mariano, L. N., Gomes do Nascimento, F., Vicente Farias,
637 I., Gazoni, V. F., dos Santos Silva, B., Giménez, A., Gutierrez-Yapu, D., Salamanca, E., *et*
638 *al.* (2013). Evaluation of antileishmanial activity of selected brazilian plants and
639 identification of the active principles. *Evidence-Based Complementary and Alternative*
640 *Medicine*, 2013.
641 Miz, R. B., Mentz, L. A., and Souza-Chies, T. T. (2008). Overview of the phylogenetic
642 relationships of some southern brazilian species from section torva and related sections of
643 "spiny solanum"(solanum subgenus leptostemonum, solanaceae). *Genetica*, 132(2):143–
644 158.
645 Moreton, J., Izquierdo, A., and Emes, R. D. (2015). Assembly, assessment, and availability
646 of de novo generated eukaryotic transcriptomes. *Frontiers in genetics*, 6.
647 Mueller, L. A., Lankhorst, R. K., Tanksley, S. D., Giovannoni, J. J., White, R., Vrebalov, J.,
648 Fei, Z., van Eck, J., Buels, R., Mills, A. A., *et al.* (2009). A snapshot of the emerging tomato
649 genome sequence. *The Plant Genome*, 2(1):78–92.
650 Ocwieja, K. E., Sherrill-Mix, S., Mukherjee, R., Custers-Allen, R., David, P., Brown, M.,
651 Wang, S., Link, D. R., Olson, J., Travers, K., *et al.* (2012). Dynamic regulation of hiv-1 mrna
652 populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic*
653 *acids research*, 40(20):10345–10355.
654 Särkinen, T., Barboza, G. E., and Knapp, S. (2015). True black nightshades: Phylogeny and
655 delimitation of the morelloid clade of solanum. *Taxon*, 64(5):945–958.
656 Sarris, P. F., Cevik, V., Dagdas, G., Jones, J. D., and Krasileva, K. V. (2016). Comparative
657 analysis of plant immune receptor architectures uncovers host proteins likely targeted by
658 pathogens. *BMC biology*, 14(1):8.
659 Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: Robust *de novo*
660 rna-seq assembly across the dynamic range of expression levels. *Bioinformatics*,
661 28(8):1086–1092.
662 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M.
663 (2015). Busco: assessing genome assembly and annotation completeness with single-copy
664 orthologs. *Bioinformatics*, 31(19):3210.

Timmermans, B. G. (2005). *Solanum sisymbriifolium (Lam.): a trap crop for potato cyst nematodes*.

Yang, X., Cheng, Y.-F., Deng, C., Ma, Y., Wang, Z.-W., Chen, X.-H., and Xue, L.-B. (2014). Comparative transcriptome analysis of eggplant (solanum melongena l.) and turkey berry (solanum torvum sw.): phylogenomics and disease resistance analysis. *BMC genomics*, 15(1):412.

Yang, Y. and Smith, S. A. (2013). Optimizing de novo assembly of short-read rna-seq data for phylogenomics. *BMC genomics*, 14(1):328.

Yoshida, S., Douglas, A. F., James, H. C., and Kwanchai, A. G. (1976). Laboratory manual for physiological studies of rice. *Manila: IRRI*, 56:69–77.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829.

Zhang, W., Ciclitira, P., and Messing, J. (2014). Pacbio sequencing of gene families-a case study with wheat gluten genes. *Gene*, 533(2):541–546.

680 **Figure Legends**

681

682 **Figure 1: Correspondence between 45 randomly selected Sanger-sequenced SSI cDNAs**
683 **with Illumina/ 454 and SMRT transcriptomes.** A) Bowtie2 was used to determine the
684 presence of 45 cDNA clones in the Illumina/454 transcriptome. B) Alignment of matched
685 sequences to Illumina/454 SSI transcriptome was performed using DNA Strider and manual
686 evaluation of alignments yielded equivalent or chimeric assignments. C) Bowtie2 was used
687 to determine the presence of 45 cDNA clones in the SMRT SSI transcriptome. D) Alignment
688 of matched sequences to SMRT SSI transcriptome was performed using DNA Strider and
689 manually evaluated as either equivalent or chimeric. More Sanger sequenced clones were
690 found in the SMRT dataset and a lower percentage were found to be chimeric than in the
691 Illumina/454 assembled dataset.

692

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

701

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

711 organ.

712

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

723

724 **Tables**

725 Table 1: Summary of the SSI transcriptome derived using SMRT technology. Organ sub-
726 transcriptomes were sequenced and combined from 33,170 root, 99,924 bud, 50,825 leaf,
727 and 47,793 stem reads. See Supplemental Figure 2 and Supplemental Figure 3 for gene
728 ontology bins of this transcriptome.

729

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

Table 2: BUSCO assessment for completeness of 3 transcriptomes and one genome. The SSI transcriptome appears to be nearly complete, but contains a disproportion number of duplicated sequences. See Supplementary Table 5 for sources of datasets.

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

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

739 Table 4: R-gene profile of potato (STU), tomato (SLY), and both provisional SSI transcriptomes.
740 The SSI (SMRT) database had fewer assigned R genes (based on the presence of nucleotide-
741 binding domains and leucine-rich repeats within the same open reading frame) than either SLY or
742 STU genomes. Refer to Supplementary Table 2 for full domain annotation statistics.

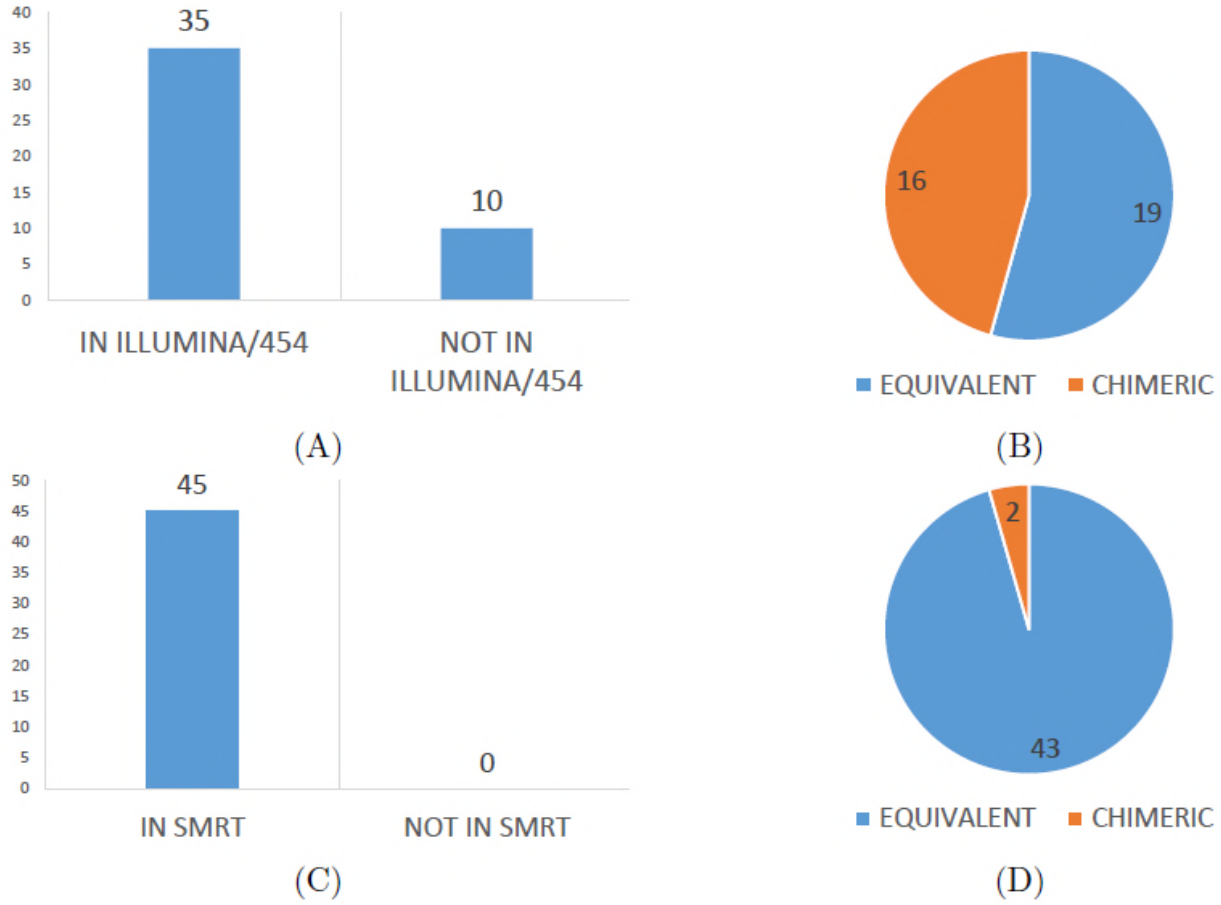
	R-genes
SSI (SMRT)	67
SSI (Illumina/454)	301
STU	309
SLY	137

743

744

745 **Figures**

746



747

748 **Figure 1: Correspondence between 45 randomly selected Sanger-sequenced SSI cDNAs with**
749 **Illumina/ 454 and SMRT transcriptomes.** A) Bowtie2 was used to determine the presence of 45
750 cDNA clones in the Illumina/454 transcriptome. B) Alignment of matched sequences to
751 Illumina/454 SSI transcriptome was performed using DNA Strider and manual evaluation of
752 alignments yielded equivalent or chimeric assignments. C) Bowtie2 was used to determine the
753 presence of 45 cDNA clones in the SMRT SSI transcriptome. D) Alignment of matched sequences
754 to SMRT SSI transcriptome was performed using DNA Strider and manual evaluated as either
755 equivalent or chimeric. More Sanger sequenced clones were found in the SMRT dataset and a
756 lower percentage were found to be chimeric than in the Illumina/454 assembled dataset.

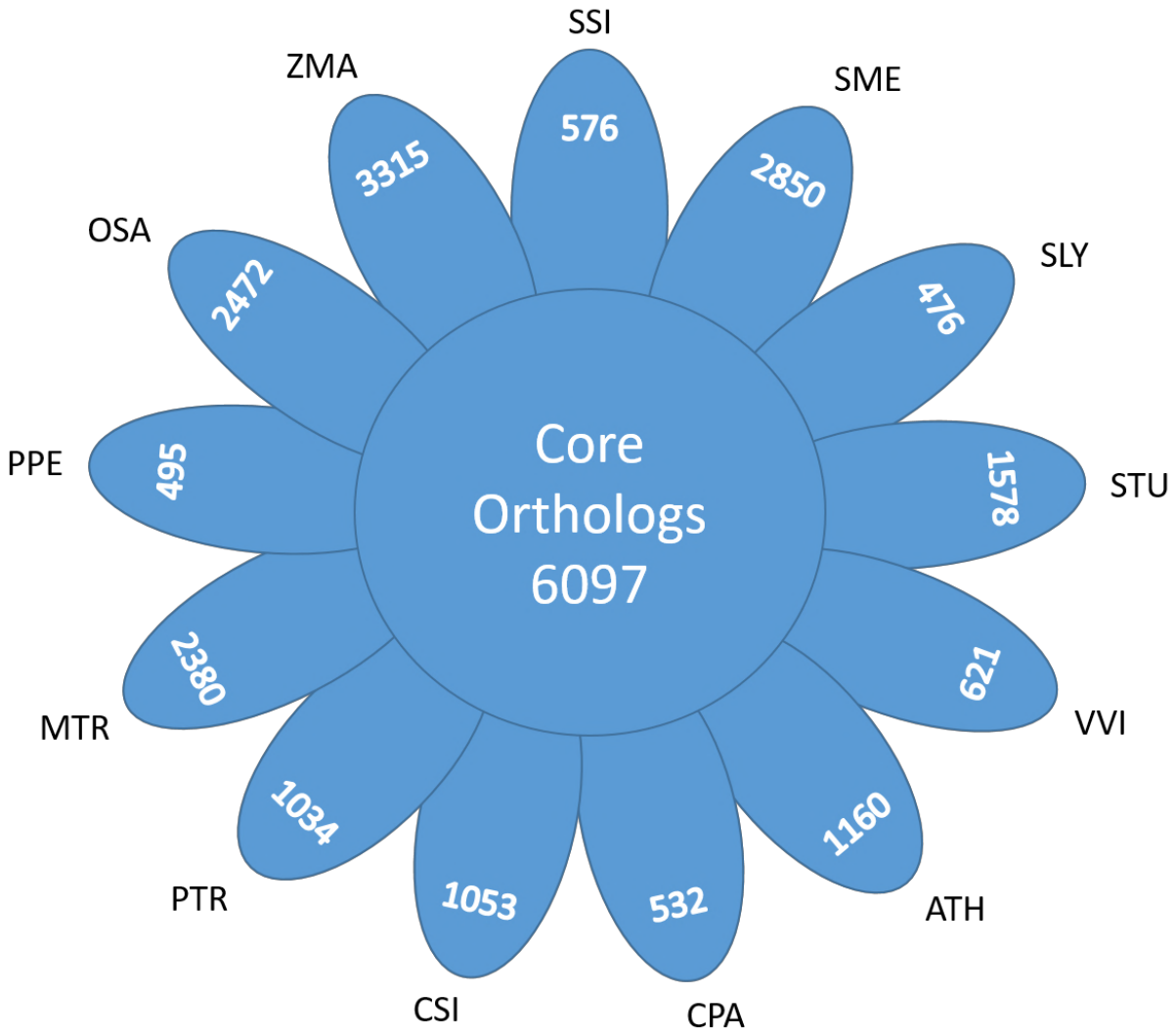


Figure 2: Shared and restricted orthologous genes among 13 species. All species shown here shared 6067 core orthologs. Each petal shows the number of gene groups unique to each species. Not shown are groups shared by only 2–12 species. *Solanum sisymbriifolium*, SSI; *Solanum tuberosum*, STU; *Solanum 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 4 for gene ontology bins for the SSI unique groups, and Supplementary Table 5 for sources of datasets.

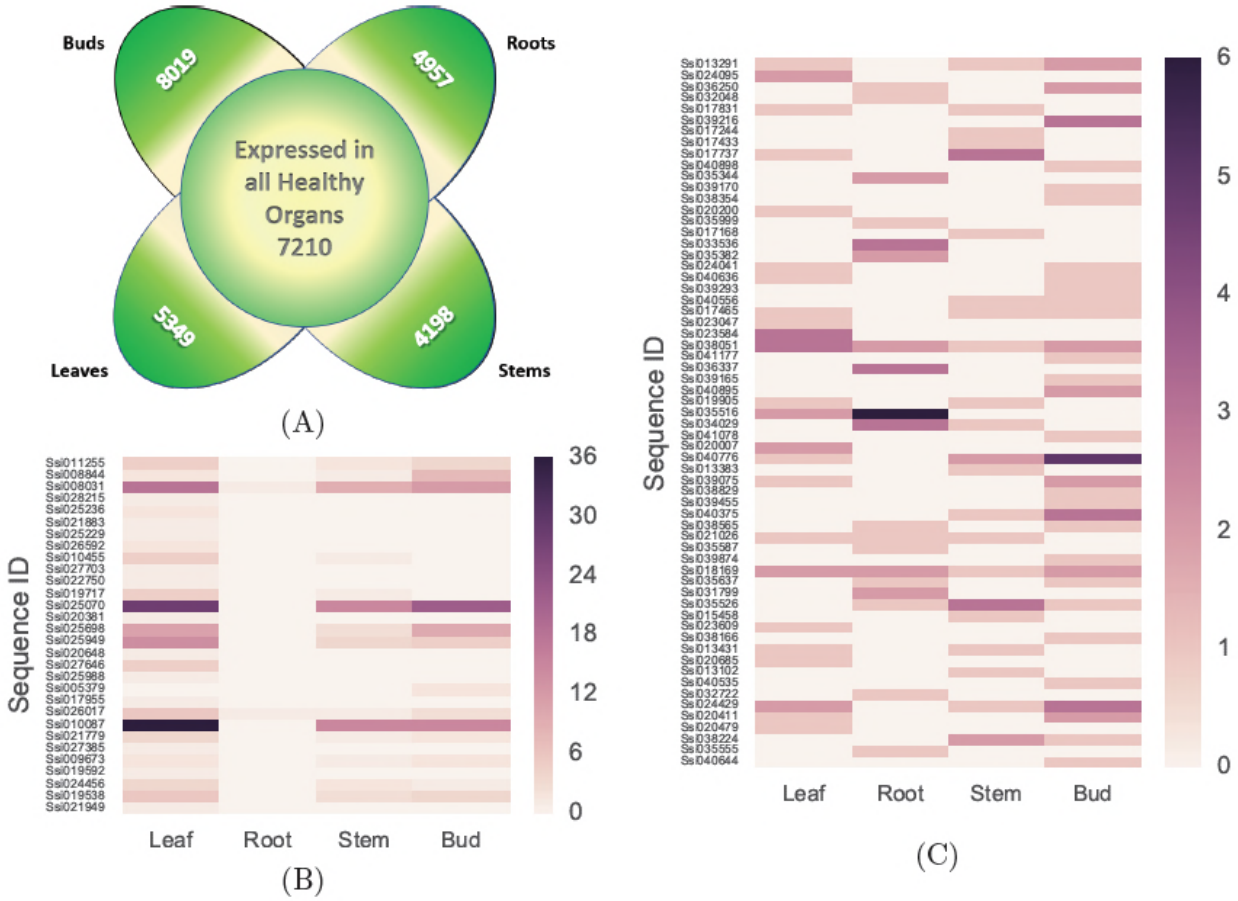
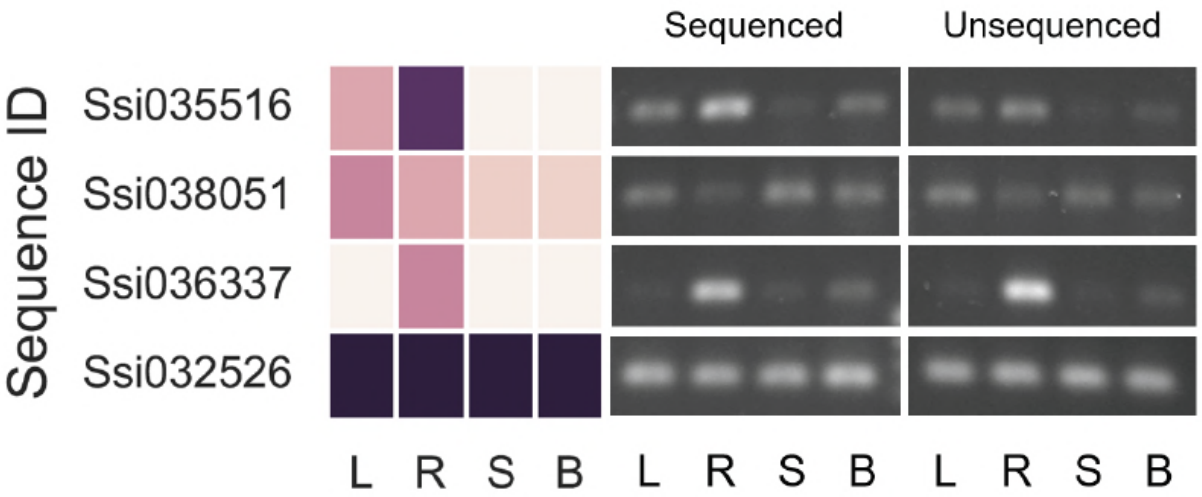


Figure 3: Final SMRT transcriptome sequences were backtracked through the de-redundification process to the organ sub-transcriptomes. A) Flower plot of genes expressed in all, or in only one, organ. Each petal shows the number of genes only expressed in one organ. In the center are the number of genes 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 4 for primers used.