Running title: regulatory networks for tomato organ development

Comprehensive regulatory networks for tomato organ development based on the genome and RNAome of microTom tomato

Jia-Yu Xue[1,#], Hai-Yun Fan[1,#], Zhen Zeng[2,#], Yu-Han Zhou[1], Shuai-Ya Hu[1], Sai-Xi Li[2], Ying-Juan Cheng[1], Xiang-Ru Meng[1], Fei Chen[3,4,*], Zhu-Qing Shao[2,*], Yves Van de Peer[1,5,6,*]

[1]College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China

[2]State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, 210023, China

[3]College of Tropical Crops, Sanya Nanfan Research Institute, Hainan University, Haikou 570228, China

[4]Hainan Yazhou Bay Seed Laboratory, Sanya 572025, China

[5]Department of Plant Biotechnology and Bioinformatics, Ghent University, VIB-UGent Center for Plant Systems Biology, B-9052 Ghent, Belgium

[6]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa

[#]These authors contributed equally to this work.

*To whom correspondence should be addressed: yvpee@psb.vib-ugent.be (Y.V.d.P.), zhuqingshao@nju.edu.cn (Z.Q.S) or feichen@hainanu.edu.cn (F.C.).

Abstract:

MicroTom tomato has a short growth cycle and high transformation efficiency, and is a prospective model plant for studying organ development, metabolism, and plant-microbe interactions. Here, with a newly assembled reference genome for this tomato cultivar and abundant RNA-seq data derived from tissues of different organs/developmental stages/treatments, we constructed multiple gene co-expression networks, which will provide valuable clues for the identification of important genes involved in diverse regulatory pathways during plant growth, e.g., arbuscular mycorrhizal symbiosis and fruit development. Additionally, non-coding RNAs, including miRNAs, lncRNAs and circRNAs were also identified, together with their potential targets. Interacting networks between different types of non-coding RNAs (miRNA-lncRNA), and non-coding RNAs and genes (miRNA-mRNA and lncRNA-mRNA) were constructed as well. Our results and data will provide valuable information for the study of organ differentiation and development of this important fruit. Lastly, we established a database (http://eplant.njau.edu.cn/microTomBase/) with genomic and transcriptomic data, as well as details of gene co-expression and interacting networks on microTom, and this database should be of great value to those who wants to adopt microTom as a model plant for research.

Keywords: microTom genome and RNAome, co-expression, non-coding RNA, interaction, database

45 **Introduction**

46 Tomato (*Solanum lycopersicum*) is one of the most popular fruits (although usually
47 referred to as a vegetable) in the world. In 2020, its annual worldwide production
48 was estimated at about 186 millions of tons and has been increasing every year
49 (http://www.fao.org/faostat). Tomato is also an emerging model plant system for
50 developmental biology and, in some cases, can be an better option than *Arabidopsis*
51 *thaliana*, e.g., for studies of fruit development [1,2], metabolism [3-5], plant-pathogen
52 interactions [6,7], and arbuscular mycorrhizal (AM) symbiosis [8,9].

53 MicroTom is a tomato cultivar and currently a widely applied experimental
54 model plant for lab studies. This cultivar has a smaller size than regular tomato
55 cultivars (e.g., Heinz 1706 and M82), and together with a shorter growth cycle and
56 higher transformation efficiency makes it one of the best choices for a lab model
57 among tomato cultivars. However, a high quality genome has been lacking for this
58 model cultivar, despite the fact that hundreds of tomato cultivars/accessions have
59 been (re-)sequenced already [4,10-12]. Nowadays, developmental biologists rely on the
60 Heinz tomato genome as a reference, whereas functional experiments are mainly
61 performed using microTom as transformation system. Many re-sequencing studies
62 have indicated considerable sequence diversity among cultivar/accession genomes
63 [4,12-14] and the newly developed pan-genome strategy revealed the existence of
64 specific genes only belonging to certain cultivars/accessions [11,15,16]. Such
65 observations suggest that the different genetic background between the reference
66 genome and others might complicate experiments, for instance in cloning target
67 genes, potentially leading to failure of the entire experimental design. Therefore, the
68 availability of a high quality microTom genome was badly needed.

69 In this study, we provide a high-quality genome of microTom and conducted
70 comparative genomic analysis with the previously published 'Heinz' tomato genome.
71 Additionally, together with large amounts of RNA-seq data sequenced by this study
72 and collected from public databases, we present the RNAome landscape of
73 microTom across different organ/developmental stages/treatments and performed
74 comprehensive analyses of the transcriptome of microTom protein-coding genes,

with respect to (aspects of) gene expression and alternative splicing (AS). With the reference genome and abundant gene expression data, we constructed co-expression networks for tomato organ development, which will provide valuable information for the regulation of organ development in the life cycle. Non-coding RNAs were also identified by combining different sequence strategies, and their interaction networks were predicted. Finally, we constructed a database (microTomBase, http://eplant.njau.edu.cn/microTomBase), available for data download, online search, and demonstration of analytical results. This database will be invaluable for researchers who employ the microTom tomato as the lab model plant.

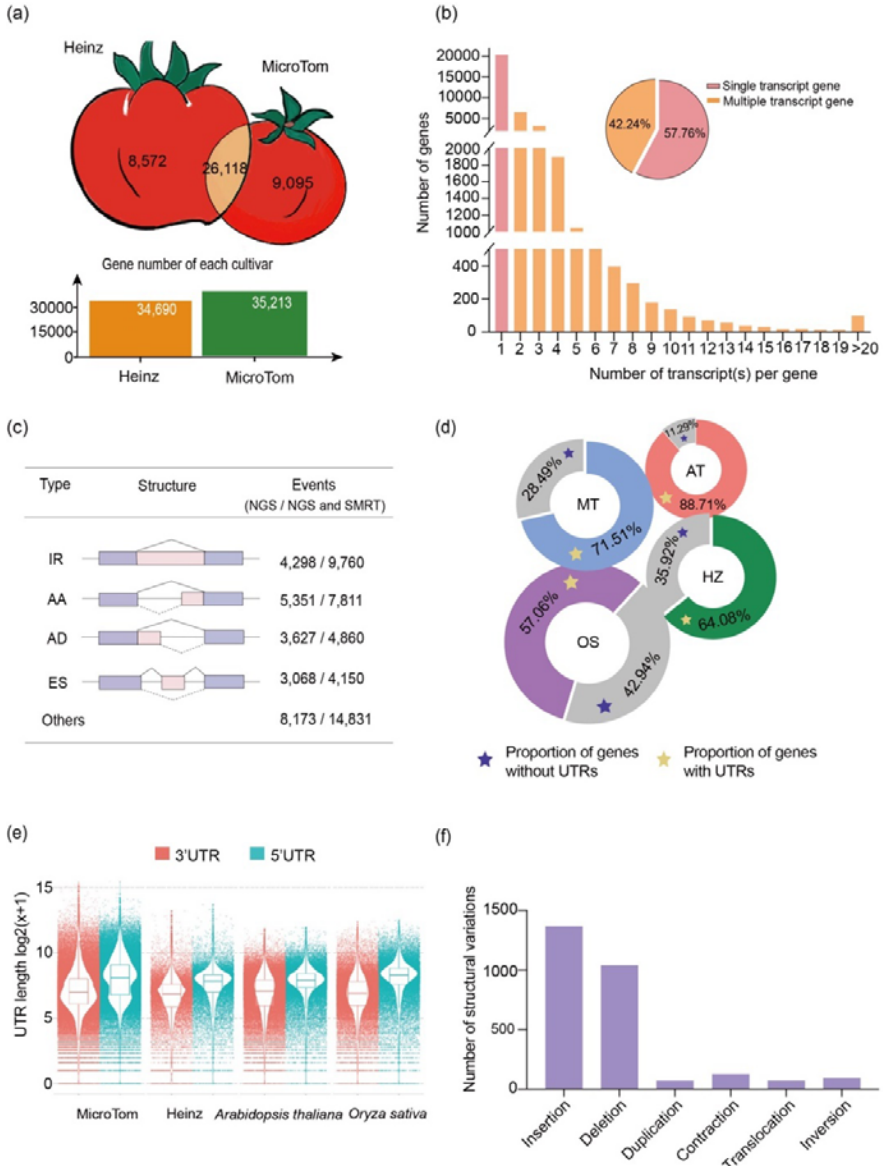
Results

Assembly of the microTom genome and comparative genomics between microTom and Heinz

Using a combination of 92.40 Gb Nanopore data and 54.38 Gb Illumina data, we first obtained a microTom genome assembly of 799 Mb with 60 contigs (contig N50 = 41.37 Mb). Then, using the genome of Heinz as a reference, 58 microTom contigs were further merged to 12 corresponding pseudo-chromosomes. Protein-coding gene annotation of the microTom assembly captures 98.57 % of the embryophyta BUSCO (odb10) genes, with 97.89 % single copy genes and 0.68% of duplicates. These results indicate that our assembled microTom genome has reached a high standard of quality and completeness (**Table 1**).

With the assistance of three full-length transcriptome RNA-seq, 12 ribo-minus RNA-seq data, and 69 poly-A enriched RNA-seq datasets derived from different organs/tissues under different developmental stages and/or treatments (See Methods and Materials) from this and previous studies [3,17], 35,213 protein-coding genes were annotated in microTom. Among those, 31,891 genes (90.57%) received transcriptomic data support from at least one RNA-seq sample. Although microTom has a similar number of protein-coding genes than Heinz's (34,690), these two cultivars share only 2/3 one-to-one orthologous genes according to synteny analysis

105 and bi-directional blast (**Figure 1a**), yet each cultivar encodes thousands of specific
106 genes, comprising real specific genes, paralogs by duplications, and genes missed by
107 annotation (mainly in Heinz probably due to inadequate RNA-seq data)
108 (**Supplementary Figure S1**). For the two cultivars, 4,449 microTom genes were
109 "hiding" in the Heinz genome (gene length coverage and DNA sequence identity >
110 90%), only not annotated; while the corresponding number of Heinz genes not
111 annotated in microTom dropped to a half. This comparison indicates that adequate
112 and diverse transcriptomic data are critical to a more comprehensive genome
113 annotation.



114

115

116 **Figure 1**. **Genome annotation of microTom and comparison with Heinz and other species. (a)**
117 Protein-coding gene numbers for Heinz and microTom varieties. Shared one-to-one orthologs in
118 microTom and Heinz genomes are presented in the cross section of the two tomato cartoon
119 pictures. **(b)** Proportion of genes with multiple transcripts and genes with different transcript
120 numbers. **(c)** Categories and frequency of AS events in microTom. IR, intron retention; AA,
121 alternative adaptor; AD, alternative donor; ES, exon skipping. **(d)** Proportion of genes with
122 annotated UTRs in microTom (MT), Heinz (HZ), *Arabidopsis thaliana* (AT), and *Oryza sativa* (OS).
123 **(e)** Dot plot of annotated UTR lengths in microTom, Heinz, *Arabidopsis thaliana* and *Oryza sativa*.
124 **(f)** Detected structural variation events of microTom, compared with Heinz. The minimum size of
125 insertion, deletion, duplication and contraction is 150 bp, the minimum size of inversion is 1 Kb,
126 and the minimum size of translocation is 10 Kb.
127
128 While alternative splicing (AS) has been widely detected in eukaryotes to
129 generate functionally divergent transcripts from a common parental gene [18], the
130 Heinz genome did not provide multi-transcript annotation for protein-coding genes.
131 By incorporating dozens of RNA-seq datasets from PacBio and Illumina platforms,
132 14,873 (42.24%) microTom protein-coding genes were identified to have multiple
133 transcripts (**Figure 1b**), with an average of two transcripts per gene. Among all AS
134 transcripts, intron retention accounts for the largest proportion (23.57%), followed
135 by alternative acceptor (18.86%), and alternative donor (11.74%) (**Figure 1c**). The
136 expression patterns of all transcripts were examined based on the transcriptomic
137 data (**Supplementary Table S1**). Among the AS transcripts, the majority (> 70%)
138 showed obvious differential expression in different organs and/or developmental
139 stages (**Supplementary Table S1**), suggesting that these AS transcripts may be
140 functionally distinct and regulate organ differentiation and development in a more
141 prevalent manner than we previously thought.
142 Due to the large amount of transcriptomic data, more comprehensive
143 information for the untranslated regions (UTRs) was also available. About 71.28% of
144 microTom genes have annotated UTRs, showing an obviously higher percentage than
145 that in Heinz and the rice and genomes (**Figure 1d**), and the medium lengths of 5'

146 and 3' UTRs are 125 bp and 271 bp, respectively (**Figure 1e**). The annotation of UTRs
147 should provide useful information for gene regulation and expression studies.
148 Comparing the genomes of microTom and Heinz, over one million single
149 nucleotide polymorphism (SNP) sites and 40 thousand insertion-deletions (indels) (<
150 150 bp) were detected. Additionally, a great number of larger structural variations
151 were also identified (**Figure 1f**), such as bigger-sized genomic insertions, deletions,
152 duplications, contractions, translocations, and inversions. These small and large
153 structural variations indicate genomic divergence between the two tomato cultivars,
154 which may affect experimental results that demand high accuracy, e.g. SNPs and
155 indels in the coding regions would greatly affect the design of CRISPR guide RNAs.
156
157 **Co-expression networks**
158 Co-expression networks underlying the development of organs and response to
159 stimuli were inferred using transcriptome data derived from 22 different
160 tissues/organs and at different developmental stages and/or different treatments
161 (three replicates for each sample). All expressed genes were classified into 123
162 modules (**Figure 2a**, **Supplementary Figure S2** and **Table S2**), whereby each module
163 contains genes that share a similar expression pattern, and possibly function in the
164 same regulatory network, defining and/or regulating a specific phenotype. When a
165 specific phenotype is defined, the corresponding module can be identified, and the
166 hub genes, as well as the entire network linked to the phenotype can be extracted.
167 This way, unknown genes in the pathway can be identified, potentially providing
168 important clues of targets and gene interactions.

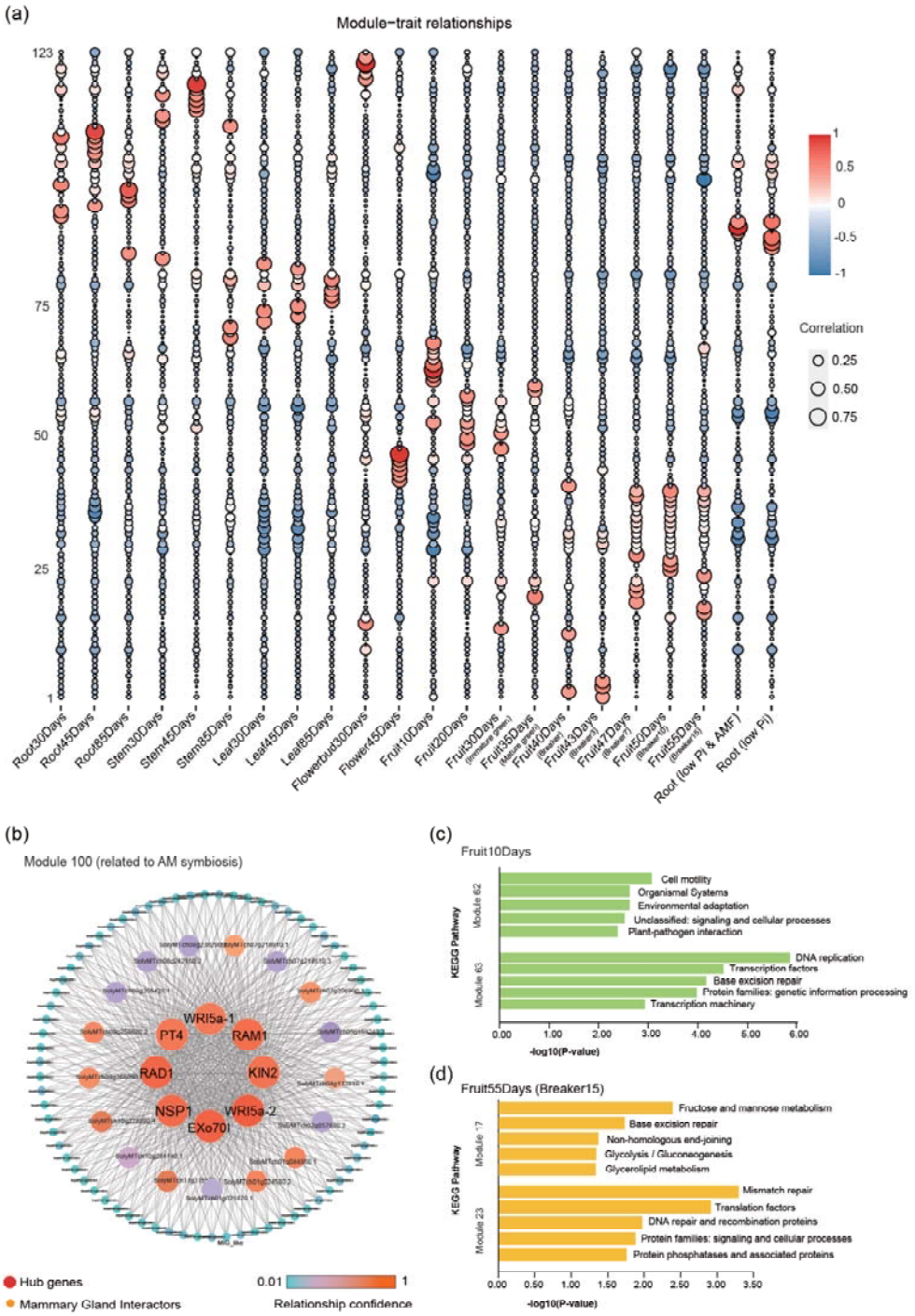


Figure 2. **MicroTom co-expression networks established based on 22 transcriptome samples of different organs/developmental stages/treatments. (a)** 123 modules and their correlation with 22 phenotypes or molecular processes. **(b)** Module 100 and its network, significantly correlated with AM symbiosis. Gene names (instead of IDs) indicate that these genes have been functionally

175 characterized and involved in the AM symbiosis. Threads indicate correlation between genes. **(c)**
176 Top five enriched KEGG pathways in the early stage of fruit development (10 days after anthesis,
177 "Fruit10Days" on Figure 2a) in the most correlated Module 62 and 63. **(d)** Top five enriched KEGG
178 pathways in the fully mature stage of fruit development (55 days after anthesis, "Fruit55Days
179 (Breaker15) on Figure 2a) in the most correlated Module 17 and 23.
180
181 For instance, our results indicate that the AM symbiosis in microTom (Root, low
182 Pi & AMF) is significantly associated with Module 100 (**Figure 2a** and **2b**,
183 **Supplementary Table S2**). It could be noticed that several important genes of known
184 functions in the AM symbiosis pathway are present, and the majority of them are
185 even hub genes in the network, including PT4, NSP1, RAD1, RAM1, Exo70I, KIN2, and
186 two WRI5a orthologous (designated as WRI5a-1 and WRI5a-2). It is well-known that
187 RAM1, RAD1 and NSP1 are regulators relatively upstream in the AM symbiosis
188 pathway [19-23], and WRI5a is even considered as a master regulator [24], so they should
189 have broad associations with other genes in the same pathway. Therefore, this
190 co-expression result makes good sense, not only in identifying the right (expected)
191 genes, but likely also in suggesting other critical, potentially even hub, genes. In this
192 case, other unknown genes in the pathway can also be identified through such
193 expression association (guilt by association), and co-expression analysis should serve
194 as an effective means to explore unknown functional genes in pathways.
195 Tomato is one of the, if not the most important model plant for fruit
196 development studies. Therefore, we collected transcriptomes derived from fruits of
197 nine different developing stages (10, 20, 30, 35, 40, 43, 47, 50, 55 days after anthesis
198 and three replicates each), and observed fruits at different developing stages
199 showing specific associations with different transcriptional modules (**Figure 2a**). For
200 instance, fruits of 10 days after anthesis are associated with modules 62 and 63
201 (**Figure 2a** and **Supplementary Table S2**), the genes of which are functionally
202 enriched in the categories of "Cell motility", "Organismal Systems", "DNA replication"
203 and "Transcription factors", suggesting that early development of tomato fruits
204 mainly involves cell division, cell proliferation, and regulatory network development

205 of this organ (**Figure 2c**); whereas fruits of 55 days after anthesis are associated with
206 modules 17 and 23 (**Figure 2a** and **Supplementary Table S2**), the genes of which are
207 functionally enriched in the categories "Fructose and mannose metabolism", "Base
208 excision repair", "Mismatch repair" and "Translation factors", suggesting
209 biosynthesis of metabolites, which mainly seems to occur in fully mature fruits
210 (**Figure 2d**). These different modules indicate distinct major regulatory priorities at
211 different developmental stages of tomato fruits.
212
213 **Diverse non-coding RNAs, their targets and interactions**
214 Non-coding RNAs, including miRNA, lncRNA and circRNA, were comprehensively
215 annotated using multiple RNA-seq data sets. These datasets were generated from six
216 development stages/organs of microTom tomato plants (root, stem, leaf, flower,
217 green fruit and red fruit), and under phosphorus deficiency and AM symbiosis status.
218 A total of 210 miRNAs were identified from the sRNA datasets, of which 164 belong
219 to 48 known tomato miRNA families in miRBase (**Figure 3a**). We identified 19 tomato
220 miRNA families previously documented in the miRBase with increased family sizes,
221 with one to seven more members discovered for each family. Notably, five miRNA
222 families that have not been documented in tomato (only documented in other plants)
223 were identified in this study, namely miR157, miR1446, miR1886, miR2111 and
224 miR3627. Additionally, 46 novel miRNAs were identified (**Figure 3a**). The lengths of
225 microTom miRNAs range from 20-nt to 24-nt, with 21-nt miRNAs occupying the
226 largest proportion. Expression analysis revealed that most of the identified miRNAs
227 were organ-/developmental stage-specific or were only expressed upon stimulation
228 by Pi deficiency and/or AM symbiosis. Only a few of them could be detected in
229 multiple samples, e.g. miR403-3p, miR9472-5p and miR166a/g.
230

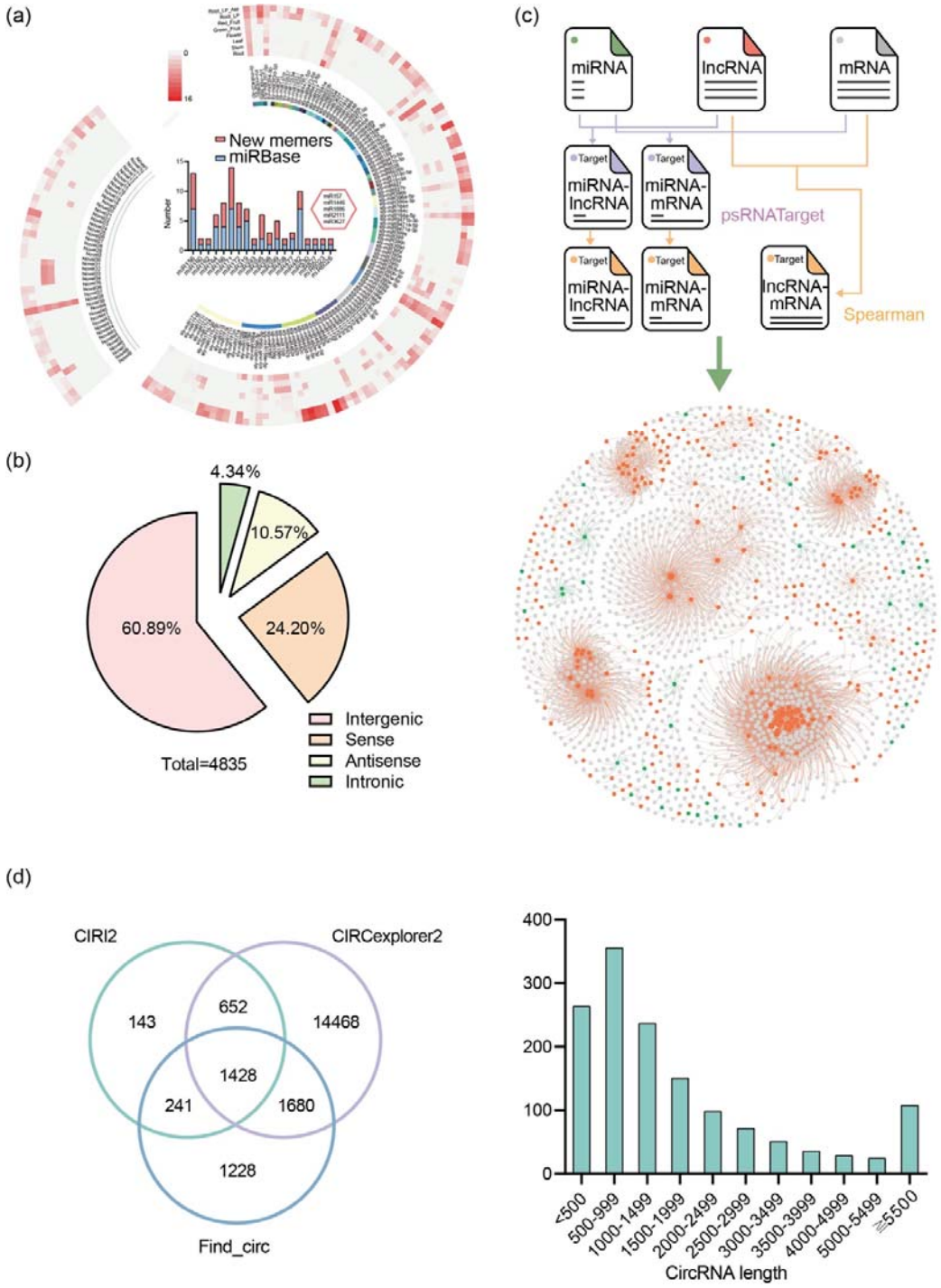


Figure 3. **Non-coding RNAs and interaction networks in microTom genome. (a)** Identification of the microTom miRNAs. The right panel represents known miRNA families and their members identified in this study and their expression, and the left panel represents novel miRNA families. The circle heatmap shows miRNA expression (outer to inner): 1. in roots symbiotic with AM fungi

237 for six weeks under low phosphorus condition, 2. roots without AM fungi under low phosphorus
238 condition, 3. red fruits, 4. green fruits, 5. flowers, 6. leaves, 7. stems, and 8. roots under normal
239 phosphorus condition. The histogram represents the new members identified in the known
240 miRNA families in microTom, the red part is the newly discovered members, and the blue part
241 indicates the existing members of miRBase. The hexagon represents miRNA families previously
242 not identified in tomatoes but identified in other plants. **(b)** Proportion of four different types of
243 lncRNAs. **(c)** An integrated interaction network between miRNAs, lncRNAs and mRNAs. The
244 purple line in the flow chart represents miRNA target gene prediction using psTargetscan, and the
245 orange line represents spearman analysis. The green, red and gray dots represent miRNAs,
246 lncRNAs and mRNAs, respectively. **(d)** The Venn diagram represents the comparison of circRNAs
247 predicted using three software. **(e)** Distribution of circRNA lengths.
248
249 Altogether, 4,835 lncRNAs were annotated from microTom transcripts, and over
250 2,944 of them (60.89%) were transcribed from intergenic regions, whereas 720, 511
251 and 210 lncRNAs, were transcribed from the genomic regions overlapping with
252 protein-coding genes at the sense/antisense strand, or from the intronic sequences,
253 respectively (**Figure 3b**). The length of microTom lncRNAs ranges from 200 to 15,073
254 nt, with an average length of 986 nt. Similar to miRNAs, the majority of lncRNAs
255 showed development stages/tissue-specific or stimulus induced expression. To
256 explore the potential regulatory roles of miRNAs and lncRNAs, pair-wise interaction
257 analyses were performed for combinations of miRNA-mRNA, miRNA-lncRNA and
258 lncRNA-mRNA. The results showed that 266 mRNAs and seven lncRNAs could be
259 predicted as targets of miRNAs with a significant negative expression correlation.
260 Significant expression correlation was also observed for 8,672 lncRNA-mRNA pairs,
261 suggesting potential regulatory relationships. Putting these predicted interactions
262 together, networks integrating miRNAs, lncRNAs, and mRNAs were constructed
263 (**Figure 3c** and **Supplementary Table S3**), and these networks should serve as a basic
264 resource for elucidating the regulatory roles of miRNAs and lncRNAs in tomato
265 development and response to stimuli.
266 A comprehensive annotation of circRNA was performed by incorporating the

267 sequencing results from this study and our previous study [17]. A total of 19,840
268 circRNAs, supported by three independent software tools were identified by
269 mapping the sequencing reads onto the microTom genome. The lengths of identified
270 1,428 high-confidence circRNAs ranged from 197 to 27,820 nt, with an average of
271 1,428 nt.
272
273 **Construction of microTom genomic and RNA-omic database**
274 To make our microTom genomic data and analytical results conveniently accessible,
275 we constructed a database (microTomBase,
276 http://eplant.njau.edu.cn/microTomBase), providing online search and download
277 possibilities of our data and results (**Figure 4**). Through gene ID or sequence blast
278 searches, interesting genes can be found, together with detailed information,
279 including genomic positions, gene sequence with detailed structure information
280 (UTRs and CDSs), functional annotation, multi-transcripts and expressional profiles.
281 Information of non-coding RNAs and their targets as well as interacting networks is
282 also provided under the 'non-coding RNA' section, which can also be accessed
283 through blast search. All microTom protein-coding genes are ascribed to 123
284 co-expression modules and can be found under the "Co-expression" section.
285 Genome assemblies and annotation files of the microTom tomato are available for
286 download. The microTom database will facilitate comparative genomics,
287 transcriptional regulation and functional studies for tomatoes.
288

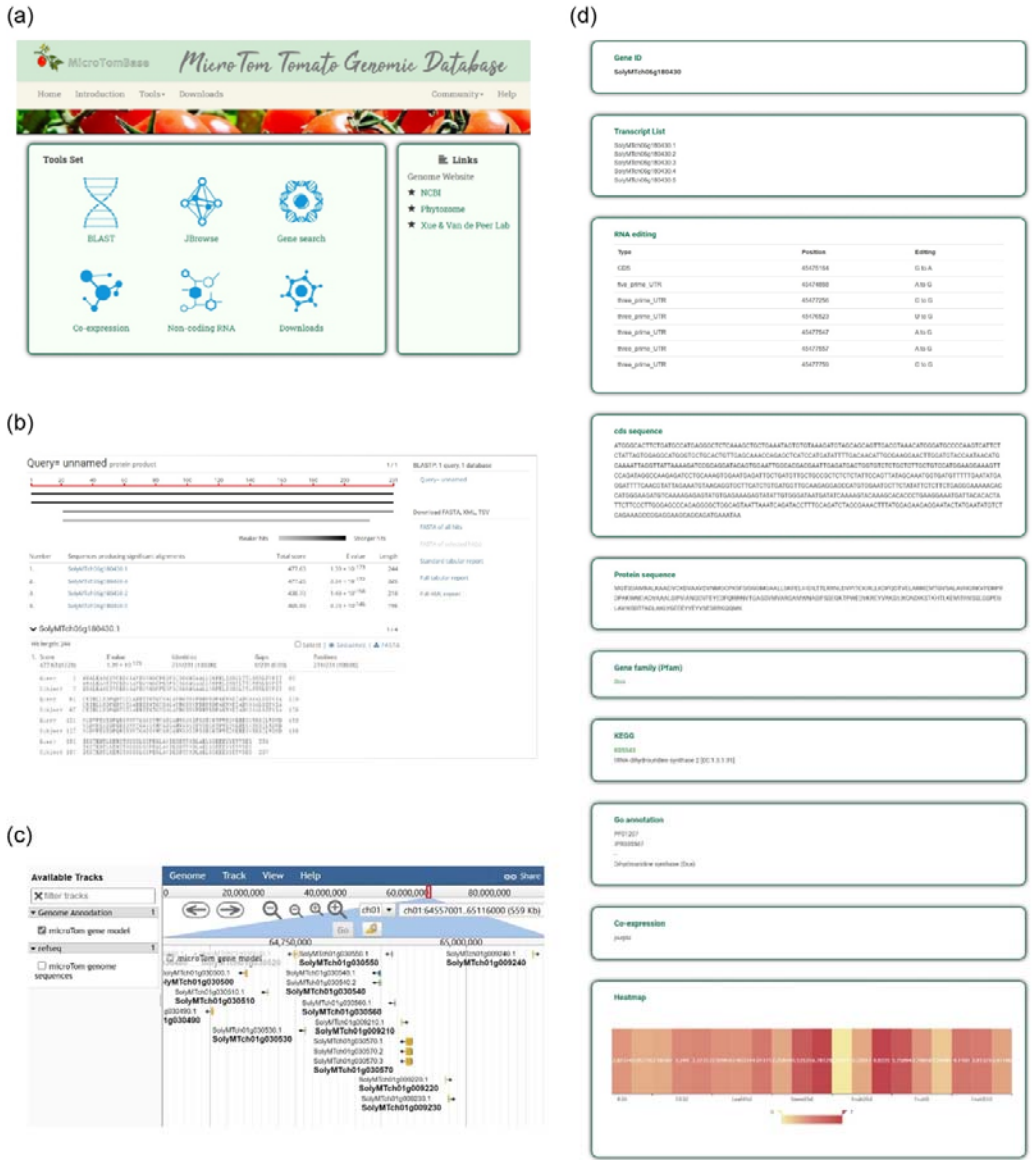


Figure 4. **Demonstration of the database MicroTomBase.** (a) Homepage of MicroTomBase and the main functions, providing service/information of sequence blast, genomic position anchoring, gene ID search, co-expression modules, non-coding RNAs and data download. **(b)** The BLAST tool of MicroTomBase. Users can input genomic, PEP, or CDS sequences as the query sequence and the resulting alignment scores are ranked from high to low. **(c)** The JBrowse tool for visualization of microTom genomic details, including gene visualization interface and detailed data on individual genes. **(d)** Detailed information about genes, including protein and CDS sequence, RNA editing sites, KEGG and GO annotation, gene family, expression profile and co-expression module.

300 **Discussion**

301 Tomato has been playing an increasingly important role as a model plant for studies
302 of fruit development, plant disease resistance, and symbiosis. Despite being the
303 most employed cultivar, and larges amounts and wide diversities of multi-omic data
304 being available, microTom tomato until now lacked a high quality genome. This
305 study fills this gap by providing a high quality microTom genome assembly, obtained
306 by a combination of Nanopore and Illumina sequencing. Moreover, we integrated a
307 diversity of transcriptome data (obtained from both public resources and our own
308 data) to achieve a comprehensive annotation and profound analyses for this genome,
309 reporting more genes, more and longer UTRs, multiple AS transcripts, co-expression
310 networks, non-coding RNAs and their interaction networks, results which should all
311 serve as a great resource for further genetic and functional studies.

312 AS is considered as a mechanism to generate multiple transcripts with different
313 functions [18], and a number of studies have reported its extensiveness in organismal
314 life cycles, e.g. organ differentiation and development [25,26], disease ontogenesis [27,28]
315 and response to abiotic and biotic stimuli [29-31]. However, the Heinz tomato genome
316 did not provide information for AS. The microTom genome annotation fills this gap
317 by including multiple transcripts generated by AS. Over 40% of microTom genes are
318 alternatively spliced and the majority of AS transcripts shows differential expression
319 under different conditions, suggesting a broad functional role of AS.

320 Other than the assist at in-depth genome annotation, the large amounts of
321 transcriptomes also provide data for our co-expression network analyses, which
322 classified microTom transcripts into different expressional modules, and accordingly
323 established diverse networks associated with different phenotypes (different organs,
324 developmental stages and treatments). Potential genes with important functions
325 that have not been recognized can thus be identified based on their association and
326 applied for further functional characterization. Therefore, our co-expression
327 networks should provide useful information for a more precise targeting of
328 candidate genes related to specific phenotypes in the tomato life cycle.

329 Non-coding RNAs are extensively involved in the post-transcriptional regulation

330 of gene expression through different mechanisms, and play important roles in a
331 variety of life processes on plant growth and development, stress resistance, and
332 interaction with pathogenic or beneficial microbes [32,33]. In tomato, high-throughput
333 sequencing of miRNAs of microTom have been reported [17,34-36], and functional roles
334 of miRNAs in fruit developmental regulation have been characterized, e.g.,
335 miRNA156 [37] and miRNA159 [38]. The profiles of lncRNAs and circRNAs of microTom
336 have also been investigated, but by only a few studies [17,39]. Moreover, these studies
337 merely focused on non-coding RNAs at certain specific developmental stages or
338 treatments, while a full picture of tomato non-coding RNAs is still lacking. In this
339 study, we comprehensively annotated miRNAs, lncRNAs and circRNAs in the
340 microTom genome by integrating multiple datasets generated from diverse organs,
341 developmental stages and treatments (**Supplementary Table S3**), and our results
342 greatly extended the non-coding RNA list of tomato. Furthermore, the expression
343 profile of non-coding RNAs among different samples indicate that most of the
344 identified miRNAs and lncRNAs were specifically expressed in different organ- and
345 developmental stage or under AM symbiosis, suggesting a conditional induction of
346 these non-coding RNAs. Furthermore, non-coding RNAs can regulate the expression
347 of protein-coding genes, directly or coordinatively affecting the translation and
348 homeostasis of mRNAs [40]. By constructing the interaction networks of the microTom
349 miRNAs, lncRNAs and mRNAs, we uncovered highly complicated regulatory
350 relationships between non-coding RNAs and their target mRNAs, which provides
351 novel insights into the post-transcriptional regulation of protein-coding genes in
352 microTom.

353 In summary, this study presents a high-quality genome for the microTom tomato,
354 and a comprehensive annotation for both coding and non-coding genes, including
355 splice variants and UTRs. Taking advantage of large amounts of transcriptomes, the
356 gene expression profile and the co-expression networks are constructed, which will
357 provide clues for the identification of novel genes involved in diverse pathways.
358 Non-coding RNAs, including miRNAs, lncRNAs and circRNAs, as well as their potential
359 targets, were identified and interacting networks were constructed, which will

360 provide valuable information for future studies dedicated to exploring the
361 post-transcriptional regulation of plant development by non-coding RNAs. All these
362 findings suggest diverse and complicated modifications and regulations at the RNA
363 level. All these results have been integrated into our online resource, which we hope,
364 could provide an invaluable resource for researchers studying tomatoes and
365 employing tomato as a model plant.
366

367 **Materials and methods**

368 **Plant material and data sources**

369 Seedlings of the *Solanum lycopersicum* (cv. 'microTom') were grown in plastic pots
370 filled with a mixture of sterilized sand/gravel (1:1 ratio) in a climate-controlled
371 growth room with 16 h light at 24°C and 8 h dark at 22°C. Roots, stems, leaves,
372 flower, immature green fruits, and mature red fruits were collected from plants
373 grown in nutrient-rich soil for six weeks, and used for RNAome sequencing. Green
374 and red fruits were classified by color chart "USDA Visual Aid TM-L-1" (USDA,
375 Agricultural Marketing Service, 1975).

376 We also downloaded 75 Illumina RNA-seq datasets and two PacBio RNA-seq
377 datasets involving different tissues/organs and different developmental stages
378 and/or different treatments from our previous study [17], and a study performed by
379 others [3]. Twelve microTom tomato proteomic datasets were retrieved from the
380 PRIDE Archive database (Proteomics identifications database,
381 https://www.ebi.ac.uk/pride/archive/).
382

383 **Library construction and sequencing**

384 Fresh leaves of four-week old microTom that were grown in half-strength Murashige
385 and Skoog basal medium with sucrose and phytagel were collected, and the genomic
386 DNA was sequenced on the Nanopore platform using a MinION R9 flow cell, and the
387 Illumina NovaSeq6000 platform with an insert size of 450bp at Benagen (Wuhan,
388 China).

389 Fresh tissues of roots, stems, leaves, flower, immature green fruits and mature

390 red fruits of plants grown in nutrient-rich soil were subjected to total RNA extraction
391 using TRIzol (Invitrogen, Carlsbad, USA) according to the manufacturer's protocol,
392 which were used for subsequent RNA sequencing using different strategies. Samples
393 from each tissue were subjected to construct a ribo-minus RNA library and small
394 RNA library for RNA sequencing, whereas a mixture of the six samples were used to
395 construction a PacBio SMRT library and a circular RNA library. The ribo-minus RNA
396 libraries, small RNA libraries and circular RNA library were sequenced using the
397 Illumina HiSeq platform at Novogene Co., Ltd. (Tianjin, China), and the PacBio SMRT
398 library was sequenced using the PacBio Sequel System at Novogene Co., Ltd. (Tianjin,
399 China).
400
401 **Genome assembly and assessment**
402 Firstly, the raw Nanopore reads were corrected and assembled into contigs by
403 NextDenovo-v2.5.0 (https://github.com/Nextomics/NextDenovo) with default
404 parameters. After that, Nanopore reads were further mapped into primary contigs
405 by software Minimap2-v2.17 [41] with the parameter '-ax map-ont', followed by
406 Nextpolish-v1.4.1 [42] to polish contigs. Meanwhile, the Illumina paired-end reads
407 were processed to remove adaptor and low quality sequences using
408 Trimmomatic-v0.38 [43]. Then, the clean Illumina short reads were mapped to the
409 polished contigs using BWA-MEM-v0.7.17 [44] with default parameters, based on three
410 iterative rounds of polishing with parameters: '--fix all' by Pilon-v1.23 [45] (read length
411 N50 = 41.37 Mb). Finally, the polished contigs were aligned to the Heinz genome to
412 form pseudomolecules, by the software RaGOO-v1.1 [46]. Embryophyta BUSCO-v5.4.2
413 (odb10) [47] was used to evaluate the integrity of the final assembly.
414
415 **PacBio transcriptome processing**
416 The SMRTlink-v6.0.0 pipeline was applied to PacBio *Iso-seq* raw data. Firstly, the
417 circular consensus sequence reads (CCSs) were extracted from subreads of BAM files
418 with parameters: minLength= 300, minZScore= -999, minPasses= 1,
419 maxDropFraction= 0.8, minPredictedAccuracy= 0.8, minSnr= 4, noPolish= TURE.

420 Secondly, CCS reads were divided into full-length non-chimeric (FLNC) reads and
421 non-full-length (nFL) reads, by identifying whether 5′ and 3′ adapters and the
422 poly(A) tail were present. Then, we obtained consensus isoforms using the ICE
423 algorithm from FLNC reads (Iterative Clustering for Error Correction), which were
424 further polished with nFL reads to get high-quality consensus FLNC reads based on
425 Quiver (parameters: hq_quiver_min_accuracy 0.99, others with default parameters).
426 Finally, the nonredundant FLNC reads were obtained by CD-HIT-v4.8.1 [48] with an
427 identity cutoff of 0.95. We also used the software LoRDEC-v0.9 [49] to correct
428 mismatch and nucleotide indels in FLNC reads.
429
430 **Repetitive DNA identification and Genome annotation**
431 We identified the repetitive elements in the microTom genome based on a
432 combination of *de novo* and homology-based strategies. A *de novo* library was first
433 built through RepeatModeler-v2.0.1 [50] and LTR_retriever-v2.9.0 [51]. The repeat region
434 in the genome was further masked by scanning *de novo* repeat library. Next, we used
435 RepeatMasker-v4.0.9 [52] to classify the repeat sequences using the parameters: -poly
436 -html -gff -lib -no_is -xsmall.
437 The combination of *ab initio*, homology-based and RNA-Seq-based annotation
438 strategy was applied to predict and annotate the protein-coding genes. Firstly, the
439 masked genome was trained by BRAKER2-v2.1.5 [53] and GeneMark-ET-v4.69 [54] for *ab*
440 *initio* gene prediction. Secondly, for finding homologous genes, related species like
441 *Solanum tuberosum* (686_v6.1) and *Solanum lycopersicum* Heinz 1706 (691_ITAG4.0),
442 which downloaded from Phytozome v13, were loaded into GenomeThreader-v1.7.1
443 [55] with the parameters: -species arabidopsis -gff3out -intermediate. Thirdly, 81
444 RNA-Seq (**Supplementary Figure S3 and Table S4**) data sets were used to assist the
445 annotation. These were evaluated by FastQC-v0.11.9 and Trimmomatic-v0.38 [43] to
446 trim low-quality and adaptor sequences. Then, these high-quality reads were aligned
447 to the genome by HISAT2-v2.2.1 [56]. We assembled transcripts with StringTie-v2.1.5 [57]
448 and TransDecoder-v5.5.0 (https://github.com/TransDecoder/TransDecoder). The
449 PacBio *Iso-seq* was also applied to forecast complete coding sequences (CDS) by

450 PASA-v2.5.2 pipeline [58]. Finally, the EVidenceModeler-v1.1.1 pipeline was applied to
451 integrate the above-mentioned prediction strategies [59]. The GTF annotation file,
452 assembled from the Illumina and PacBio data, was used for identifying AS events
453 using the AStalavista v3.2 tool [60].
454
455 **Functional annotation**
456 Several public databases (PFAM, GO and KEGG) were used to map GO terms for
457 searching functional motifs and domains by InterProScan-v5.50-84.0 [61]. When the
458 Gene Ontology terms for each gene were obtained, functions of protein-coding
459 genes in microTom genome were annotated by SWISS-PROT database, only retaining
460 the best alignment.
461
462 **Identification of SVs and SNPs between reference genomes**
463 Structural variations (SVs) were identified by genome-wide alignment between the
464 microTom and Heinz 1706 (ITAG4.0) genomes using MUMandCo-v3.8 [62], which can
465 detect over 1 kb structural variations in length. We also used the MUMmer
466 package-v4.0.0 [63] to align these two genomes. The reciprocal best alignments were
467 found using the delta-filter program. Then, the uniquely aligned fragments were used
468 to identify single nucleotide polymorphism (SNP) sites and indels with the show-snp
469 tool.
470
471 **Gene expression and construction of co-expression networks**
472 The RNA-seq dataset for gene co-expression analysis was obtained from samples
473 with 22 various tissues/organs and at different developmental stages and/or
474 different treatments. The normalized expression of RNA-seq data was calculated with
475 FPKM (fragments per kb per million reads) from all samples. FPKM values lower than
476 1 were filtered out. We used the R package WGCNA (v1.66) [64] to construct
477 co-expression networks, and the soft threshold value was calculated by
478 pickSoftThreshold integrated in WGCNA. We used cytoHubba and MCODE to identify
479 candidate hub genes, and used Cytoscape-v3.9.1 [65] to visualize the co-expression

480 networks.

481

482 **Identification and annotation of Non-coding RNAs**

483 The prediction of protein-coding potential for transcripts generated from Illumina
484 and SMRT data was conducted by integrating CPC2 v0.1, CNCI v2, PLEK v1.2, and
485 PFAM v32[66-69]. Candidate lncRNAs were defined as transcripts with the longest
486 representative sequence lacking any open reading frame (ORF) exceeding 100 amino
487 acids, while having a minimum nucleotide sequence length of 200 nts. MiRNAs were
488 identified using miRador (https://github.com/rkweku/miRador) with slight
489 modification [70]. Specifically, the sequences with the highest abundance of each
490 miRNA locus were added as candidate miRNAs. Candidate miRNAs were annotated
491 as known or novel miRNAs through referencing the latest miRBase version 22.1 [71].
492 CIRI2 (v2.0.6), CIRIexplorer2 (v2.4.0), and Find_ciric (v1.2) were combined to detect
493 potential back-splice sites of candidate circRNAs [72-74]. Reconstruction of full-length
494 circRNAs was achieved using CIRI-full [75].

495

496 **Identification of targets of miRNAs/lncRNAs**

497 Potential miRNA targets (mRNAs and lncRNAs) were firstly predicted by using the
498 psRNATarget v2 program. Then, the expression values of resultant
499 miRNA-mRNA/lncRNA pairs for six tissues (roots, stems, leaves, flower, immature
500 green fruits, and mature red fruits) were subjected to the spearman correlation
501 coefficient analysis. MiRNA-mRNA/lncRNA pairs were considered to be a candidate
502 interactional miRNA-target when the spearman correlation coefficient was <-0.80
503 and the p value was <0.01.

504 To predict lncRNA-mRNA co-expressed pairs, the spearman correlation analysis
505 was performed between lncRNAs and mRNAs in 75 RNA-seq samples. For more
506 accurate prediction of lncRNA-mRNA correlations, we increased the threshold of
507 spearman correlation coefficient| to >0.9, and the p value <0.01. We also screened
508 the results that met the above requirements based on six or more samples with
509 FPKM greater than 1, considering these lncRNA-mRNA pairs to be co-expressed.

510 **Data availability**

511 The data supporting the findings of this work are available within the paper and its
512 Supplementary information files. The datasets generated and analyzed during this
513 study are available from the corresponding author upon request. All sequencing data
514 generated have been deposited in National Genomics Data Center
515 (https://bigd.big.ac.cn/) with BioProject ID: PRJCA012325 under GSA: CRA008482.
516 The genome assembly, along with the gene models, the functional annotation, and
517 some supplementary information files are available on the FigShare at the link:
518 (https://doi.org/10.6084/m9.figshare.21509817). All these information can also be
519 accessed under GWH: WGS029763 and microTomBase
520 (http://eplant.njau.edu.cn/microTomBase).

521

522 **Author contributions**

523 J.Y.X., Z.Q.S., F.C. and Y.V.d.P. conceived the study. Z.Q.S. and Z.Z. collected plant
524 samples. H.Y.F. conducted genome assembly and annotation. H.Y.F., Z.Z., S.Y.H. Y.J.C.,
525 and S.X.L conducted analyses of RNA-seq data. Y.H.Z. constructed the database. J.Y.X.
526 and Z.Q.S. drafted the manuscript. All authors read the manuscript and participated
527 in the revision of the manuscript. All authors approved the final manuscript.

528

529 **Acknowledgments**

530 This work was supported by grants from the Fundamental Research Funds for the
531 Central Universities (KYCXJC2022003), the National Natural Science Foundation of
532 China (32070243) and the Outstanding Young Teacher of "QingLan Project" of
533 Jiangsu Province. Y.V.d.P. acknowledges funding from the European Research
534 Council (ERC) under the European Union's Horizon 2020 research and innovation
535 program (No. 833522) and from Ghent University (Methusalem funding,
536 BOF.MET.2021.0005.01). The computational resources and services were provided
537 by the Bioinformatics Center of Nanjing Agricultural University.

538

539 **Conflict of interest statement**

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

544 **Supplementary information**

545 Supplementary Fig. S1. Statistics the number of duplicate genes, specific genes and
546 missed genes due to annotation in Heinz and microTom.

547 Supplementary Fig. S2. The expressed transcripts classified into 123 modules.

548 Supplementary Fig. S3. Information about transcriptome datasets and sequencing
549 strategies.

550 Supplementary Fig. S4. Evolutionary and divergence time estimation. Phylogeny of
551 eleven species. The numbers on the nodes indicated the divergence time of species
552 (Mya, million years ago), with the confidence range in brackets.

553 Supplementary Fig. S5. Genomic collinearity analysis between microTom and Heinz.

554 Supplementary Fig. S6. Density of synonymous substitutions per synonymous site (Ks)
555 of collinear paralogous genes.

556 Supplementary Table S1. Quantitative expression of transcriptome data (Fpkm).

557 Supplementary Table S2. Gene information in the model block.

558 Supplementary Table S3. A dataset of significantly correlated combinations between
559 non-coding RNAs and their target RNAs.

560 Supplementary Table S4. Summary of transcriptome mapping for microTom tissues.

561

562

563 **References**

564 1. Klee, H.J. & Giovannoni, J.J. Genetics and control of tomato fruit ripening and quality
565 attributes. *Annu Rev Genet* **45**, 41-59
566 http://dx.doi.org/10.1146/annurev-genet-110410-132507 (2011).
567 2. Azzi, L. *et al.* Fruit growth-related genes in tomato. *J Exp Bot* **66**, 1075-86
568 http://dx.doi.org/10.1093/jxb/eru527 (2015).
569 3. Li, Y. *et al.* MicroTom Metabolic Network: Rewiring Tomato Metabolic Regulatory Network
570 throughout the Growth Cycle. *Mol Plant* **13**, 1203-1218
571 http://dx.doi.org/10.1016/j.molp.2020.06.005 (2020).
572 4. Zhu, G. *et al.* Rewiring of the Fruit Metabolome in Tomato Breeding. *Cell* **172**, 249-261 e12
573 http://dx.doi.org/10.1016/j.cell.2017.12.019 (2018).
574 5. Li, J. *et al.* Biofortified tomatoes provide a new route to vitamin D sufficiency. *Nat Plants* **8**,
575 611-616 http://dx.doi.org/10.1038/s41477-022-01154-6 (2022).
576 6. Lozano-Torres, J.L. *et al.* Dual disease resistance mediated by the immune receptor Cf-2 in
577 tomato requires a common virulence target of a fungus and a nematode. *Proc Natl Acad Sci U*
578 *S A* **109**, 10119-24 http://dx.doi.org/10.1073/pnas.1202867109 (2012).
579 7. Weiberg, A. *et al.* Fungal small RNAs suppress plant immunity by hijacking host RNA
580 interference pathways. *Science* **342**, 118-23 http://dx.doi.org/10.1126/science.1239705
581 (2013).
582 8. Chitarra, W. *et al.* Insights on the Impact of Arbuscular Mycorrhizal Symbiosis on Tomato
583 Tolerance to Water Stress. *Plant Physiol* **171**, 1009-23 http://dx.doi.org/10.1104/pp.16.00307
584 (2016).
585 9. Liao, D. *et al.* SlSPX1-SlPHR complexes mediate the suppression of arbuscular mycorrhizal
586 symbiosis by phosphate repletion in tomato. *Plant Cell*
587 http://dx.doi.org/10.1093/plcell/koac212 (2022).
588 10. Alonge, M. *et al.* Major Impacts of Widespread Structural Variation on Gene Expression and
589 Crop Improvement in Tomato. *Cell* **182**, 145-161 e23
590 http://dx.doi.org/10.1016/j.cell.2020.05.021 (2020).
591 11. Gao, L. *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit
592 flavor. *Nat Genet* **51**, 1044-1051 http://dx.doi.org/10.1038/s41588-019-0410-2 (2019).
593 12. Zhou, Y. *et al.* Graph pangenome captures missing heritability and empowers tomato
594 breeding. *Nature* **606**, 527-534 http://dx.doi.org/10.1038/s41586-022-04808-9 (2022).
595 13. Wang, W. *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*
596 **557**, 43-49 http://dx.doi.org/10.1038/s41586-018-0063-9 (2018).
597 14. Ma, Z. *et al.* Resequencing a core collection of upland cotton identifies genomic variation and
598 loci influencing fiber quality and yield. *Nat Genet* **50**, 803-813
599 http://dx.doi.org/10.1038/s41588-018-0119-7 (2018).
600 15. Song, J.M. *et al.* Eight high-quality genomes reveal pan-genome architecture and ecotype
601 differentiation of Brassica napus. *Nat Plants* **6**, 34-45
602 http://dx.doi.org/10.1038/s41477-019-0577-7 (2020).
603 16. Liu, Y. *et al.* Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162-176 e13
604 http://dx.doi.org/10.1016/j.cell.2020.05.023 (2020).
605 17. Zeng, Z. *et al.* The RNAome landscape of tomato during arbuscular mycorrhizal symbiosis
606 reveals an evolving RNA layer symbiotic regulatory network. *Plant Commun* 100429

607 http://dx.doi.org/10.1016/j.xplc.2022.100429 (2022).
608 18. Ule, J. & Blencowe, B.J. Alternative Splicing Regulatory Networks: Functions, Mechanisms,
609 and Evolution. *Mol Cell* **76**, 329-345 http://dx.doi.org/10.1016/j.molcel.2019.09.017 (2019).
610 19. Rey, T. *et al.* The Medicago truncatula GRAS protein RAD1 supports arbuscular mycorrhiza
611 symbiosis and Phytophthora palmivora susceptibility. *Journal of Experimental Botany* **68**,
612 5871-5881 http://dx.doi.org/10.1093/jxb/erx398 (2017).
613 20. Pimprikar, P. *et al.* A CCaMK-CYCLOPS-DELLA Complex Activates Transcription of RAM1 to
614 Regulate Arbuscule Branching (vol 26, pg 987, 2016). *Curr Biol* **26**, 1126-1126
615 http://dx.doi.org/10.1016/j.cub.2016.04.021 (2016).
616 21. Xue, L. *et al.* Network of GRAS Transcription Factors Involved in the Control of Arbuscule
617 Development in Lotus japonicus. *Plant Physiology* **167**, 854-+
618 http://dx.doi.org/10.1104/pp.114.255430 (2015).
619 22. Rich, M.K. *et al.* The Petunia GRAS Transcription Factor ATA/RAM1 Regulates Symbiotic Gene
620 Expression and Fungal Morphogenesis in Arbuscular Mycorrhiza. *Plant Physiology* **168**, 788-+
621 http://dx.doi.org/10.1104/pp.15.00310 (2015).
622 23. Gobbato, E. *et al.* A GRAS-Type Transcription Factor with a Specific Function in Mycorrhizal
623 Signaling. *Curr Biol* **22**, 2236-2241 http://dx.doi.org/10.1016/j.cub.2012.09.044 (2012).
624 24. Jiang, Y.N. *et al.* Medicago AP2-Domain Transcription Factor WRI5a Is a Master Regulator of
625 Lipid Biosynthesis and Transfer during Mycorrhizal Symbiosis. *Molecular Plant* **11**, 1344-1359
626 http://dx.doi.org/10.1016/j.molp.2018.09.006 (2018).
627 25. Park, J.W., Yang, J. & Xu, R.H. Paired Box Protein 6 Alternative Splicing and Corneal
628 Development. *Stem Cells Dev* **27**, 367-377 http://dx.doi.org/10.1089/scd.2017.0283 (2018).
629 26. Baralle, F.E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity.
630 *Nat Rev Mol Cell Bio* **18**, 437-451 http://dx.doi.org/10.1038/nrm.2017.27 (2017).
631 27. Zhang, Y.J., Qian, J.J., Gu, C.Y. & Yang, Y. Alternative splicing and cancer: a systematic review.
632 *Signal Transduct Tar* **6**, http://dx.doi.org/ARTN 78
633 10.1038/s41392-021-00486-7 (2021).
634 28. Lukas, J. *et al.* Alternative and aberrant messenger RNA splicing of the mdm2 oncogene in
635 invasive breast cancer. *Cancer Res* **61**, 3212-3219 <Go to ISI>://WOS:000167967100066
636 (2001).
637 29. Laloum, T., Martin, G. & Duque, P. Alternative Splicing Control of Abiotic Stress Responses.
638 *Trends Plant Sci* **23**, 140-150 http://dx.doi.org/10.1016/j.tplants.2017.09.019 (2018).
639 30. John, S., Olas, J.J. & Mueller-Roeber, B. Regulation of alternative splicing in response to
640 temperature variation in plants. *Journal of Experimental Botany* **72**, 6150-6163
641 http://dx.doi.org/10.1093/jxb/erab232 (2021).
642 31. Zhang, X.C. & Gassmann, W. Alternative splicing and mRNA levels of the disease resistance
643 gene RPS4 are induced during Defense responses(1[W][OA]). *Plant Physiology* **145**,
644 1577-1587 http://dx.doi.org/10.1104/pp.107.108720 (2007).
645 32. Song, X., Li, Y., Cao, X. & Qi, Y. MicroRNAs and Their Regulatory Roles in Plant-Environment
646 Interactions. *Annu Rev Plant Biol* **70**, 489-525
647 http://dx.doi.org/10.1146/annurev-arplant-050718-100334 (2019).
648 33. Yu, Y., Zhang, Y., Chen, X. & Chen, Y. Plant Noncoding RNAs: Hidden Players in Development
649 and Stress Responses. *Annu Rev Cell Dev Biol* **35**, 407-431
650 http://dx.doi.org/10.1146/annurev-cellbio-100818-125218 (2019).

651 34. Dong, F. *et al.* Differential expression of microRNAs in tomato leaves treated with different
652 light qualities. *BMC Genomics* **21**, 37 http://dx.doi.org/10.1186/s12864-019-6440-4 (2020).
653 35. Cao, H. *et al.* miRNA expression profiling and zeatin dynamic changes in a new model system
654 of in vivo indirect regeneration of tomato. *PLoS One* **15**, e0237690
655 http://dx.doi.org/10.1371/journal.pone.0237690 (2020).
656 36. Feng, J., Liu, S., Wang, M., Lang, Q. & Jin, C. Identification of microRNAs and their targets in
657 tomato infected with Cucumber mosaic virus based on deep sequencing. *Planta* **240**, 1335-52
658 http://dx.doi.org/10.1007/s00425-014-2158-3 (2014).
659 37. Ferreira e Silva, G.F. *et al.* microRNA156-targeted SPL/SBP box transcription factors regulate
660 tomato ovary and fruit development. *Plant J* **78**, 604-18 http://dx.doi.org/10.1111/tpj.12493
661 (2014).
662 38. da Silva, E.M. *et al.* microRNA159-targeted SlGAMYB transcription factors are required for
663 fruit set in tomato. *Plant J* **92**, 95-109 http://dx.doi.org/10.1111/tpj.13637 (2017).
664 39. Lamin-Samu, A.T., Zhuo, S., Ali, M. & Lu, G. Long non-coding RNA transcriptome landscape of
665 anthers at different developmental stages in response to drought stress in tomato. *Genomics*
666 **114**, 110383 http://dx.doi.org/10.1016/j.ygeno.2022.110383 (2022).
667 40. Fabian, M.R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA translation and stability by
668 microRNAs. *Annu Rev Biochem* **79**, 351-79
669 http://dx.doi.org/10.1146/annurev-biochem-060308-103103 (2010).
670 41. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572-4574
671 http://dx.doi.org/10.1093/bioinformatics/btab705 (2021).
672 42. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for
673 long-read assembly. *Bioinformatics* **36**, 2253-2255
674 http://dx.doi.org/10.1093/bioinformatics/btz891 (2019).
675 43. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
676 data. *Bioinformatics* **30**, 2114-20 http://dx.doi.org/10.1093/bioinformatics/btu170 (2014).
677 44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
678 *Bioinformatics* **25**, 1754-60 http://dx.doi.org/10.1093/bioinformatics/btp324 (2009).
679 45. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and
680 genome assembly improvement. *PLoS One*. Preprint at
681 http://europepmc.org/abstract/MED/25409509
682 https://doi.org/10.1371/journal.pone.0112963
683 https://europepmc.org/articles/PMC4237348
684 https://europepmc.org/articles/PMC4237348?pdf=render (2014).
685 46. Alonge, M. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes.
686 *Genome Biology* **20**, 224 http://dx.doi.org/10.1186/s13059-019-1829-6 (2019).
687 47. Manni, M., Berkeley, M.R., Seppey, M. & Zdobnov, E.M. BUSCO: Assessing Genomic Data
688 Quality and Beyond. *Curr Protoc* **1**, e323 http://dx.doi.org/10.1002/cpz1.323 (2021).
689 48. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation
690 sequencing data. *Bioinformatics* **28**, 3150-3152
691 http://dx.doi.org/10.1093/bioinformatics/bts565 (2012).
692 49. Salmela, L. & Rivals, E. LoRDEC: accurate and efficient long read error correction.
693 *BIOINFORMATICS* **30**, 3506-3514 http://dx.doi.org/10.1093/bioinformatics/btu538 (2014).
694 50. Bao, Z. & Eddy, S.R. Automated de novo identification of repeat sequence families in

695 sequenced genomes. *Genome research* **12**, 1269-1276 http://dx.doi.org/10.1101/gr.88502
696 (2002).
697 51. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of
698 Long Terminal Repeat Retrotransposons. *Plant Physiol* **176**, 1410-1422
699 http://dx.doi.org/10.1104/pp.17.01310 (2018).
700 52. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in
701 genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4.10
702 http://dx.doi.org/10.1002/0471250953.bi0410s25 (2009).
703 53. Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic
704 eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein
705 database. *NAR Genomics and Bioinformatics* **3**, http://dx.doi.org/10.1093/nargab/lqaa108
706 (2021).
707 54. Lomsadze, A., Burns, P.D. & Borodovsky, M. Integration of mapped RNA-Seq reads into
708 automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**, e119
709 http://dx.doi.org/10.1093/nar/gku557 (2014).
710 55. Gremme, G., Brendel, V., Sparks, M.E. & Kurtz, S. Engineering a software tool for gene
711 structure prediction in higher organisms. *Information and Software Technology* **47**, 965-978
712 http://dx.doi.org/https://doi.org/10.1016/j.infsof.2005.09.005 (2005).
713 56. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and
714 genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915
715 http://dx.doi.org/10.1038/s41587-019-0201-4 (2019).
716 57. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq
717 reads. *NATURE BIOTECHNOLOGY* **33**, 290-+ http://dx.doi.org/10.1038/nbt.3122 (2015).
718 58. Haas, B.J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript
719 alignment assemblies. *NUCLEIC ACIDS RESEARCH* **31**, 5654-5666
720 http://dx.doi.org/10.1093/nar/gkg770 (2003).
721 59. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and
722 the program to assemble spliced alignments. *GENOME BIOLOGY* (Article) **9**, R7
723 http://dx.doi.org/10.1186/gb-2008-9-1-r7 (2008).
724 60. Foissac, S. & Sammeth, M. ASTALAVISTA: dynamic and flexible analysis of alternative splicing
725 events in custom gene datasets. *Nucleic Acids Res* **35**, W297-9
726 http://dx.doi.org/10.1093/nar/gkm311 (2007).
727 61. Zdobnov, E.M. & Apweiler, R. InterProScan - an integration platform for the
728 signature-recognition methods in InterPro. *BIOINFORMATICS* **17**, 847-848
729 http://dx.doi.org/10.1093/bioinformatics/17.9.847 (2001).
730 62. O'Donnell, S. & Fischer, G. MUM&Co: accurate detection of all SV types through
731 whole-genome alignment. *Bioinformatics* **36**, 3242-3243
732 http://dx.doi.org/10.1093/bioinformatics/btaa115 (2020).
733 63. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *GENOME BIOLOGY*
734 **5**, R12 http://dx.doi.org/10.1186/gb-2004-5-2-r12 (2004).
735 64. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis.
736 *BMC BIOINFORMATICS* **9**, 559 http://dx.doi.org/10.1186/1471-2105-9-559 (2008).
737 65. Shannon, P. *et al.* Cytoscape: A software environment for integrated models of biomolecular
738 interaction networks. *Genome Research* **13**, 2498-2504 http://dx.doi.org/10.1101/gr.1239303

739 (2003).
740 66. Li, A., Zhang, J. & Zhou, Z. PLEK: a tool for predicting long non-coding RNAs and messenger
741 RNAs based on an improved k-mer scheme. *BMC Bioinformatics* **15**, 311
742 http://dx.doi.org/10.1186/1471-2105-15-311 (2014).
743 67. Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features
744 and support vector machine. *Nucleic Acids Res* **35**, W345-9
745 http://dx.doi.org/10.1093/nar/gkm391 (2007).
746 68. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-30
747 http://dx.doi.org/10.1093/nar/gkt1223 (2014).
748 69. Sun, L. *et al.* Utilizing sequence intrinsic composition to classify protein-coding and long
749 non-coding transcripts. *Nucleic Acids Res* **41**, e166 http://dx.doi.org/10.1093/nar/gkt646
750 (2013).
751 70. Hammond, R.K., Gupta, P., Patel, P. & Meyers, B.C. miRador: a fast and precise tool for the
752 prediction of plant miRNAs. *bioRxiv* http://dx.doi.org/doi:
753 https://doi.org/10.1101/2021.03.24.436803 (2021).
754 71. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to
755 function. *Nucleic Acids Res* **47**, D155-D162 http://dx.doi.org/10.1093/nar/gky1141 (2019).
756 72. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency.
757 *Nature* **495**, 333-8 http://dx.doi.org/10.1038/nature11928 (2013).
758 73. Gao, Y., Zhang, J. & Zhao, F. Circular RNA identification based on multiple seed matching. *Brief*
759 *Bioinform* **19**, 803-810 http://dx.doi.org/10.1093/bib/bbx014 (2018).
760 74. Zhang, X.O. *et al.* Diverse alternative back-splicing and alternative splicing landscape of
761 circular RNAs. *Genome Res* **26**, 1277-87 http://dx.doi.org/10.1101/gr.202895.115 (2016).
762 75. Zheng, Y., Ji, P., Chen, S., Hou, L. & Zhao, F. Reconstruction of full-length circular RNAs enables
763 isoform-level quantification. *Genome Med* **11**, 2
764 http://dx.doi.org/10.1186/s13073-019-0614-1 (2019).
765
766
767
768

769 **Table 1. Statistics and comparison of microTom and Heinz genomes**

	MicroTom				Heinz 1706 (SL4.0)			
	Contigs		Pseudomolecules		Contigs		Pseudomolecules	
	Size(bp)	Number	Size(bp)	Number	Size(bp)	Number	Size(bp)	Number
N50	41,367,923	8	66,788,879	6	6,007,830	37	65,269,487	6
Maximum length	68,578,620		94,893,795		26,291,688		90,863,682	
Total number	60		12		448		12	
Gaps number	47				435			
Assembled length	798,935,606				782,475,302			
BUSCO completeness	98.57%				97.40%			
Gene models number	35,213				34,690			

770