International Journal of Control and Automation
Vol. 8, No. 2 (2015), pp. 105-116
http://dx.doi.org/10.14257/ijca.2015.8.2.11

De Novo Assembly and Characterization of the Transcriptome and Molecular Discovery in Capsicum Annuum L. R597

Xiaowan Xu, Tao Li, Ying Li*, Hengming Wang and Xiaomei Xu

Vegetable Research Institute, Guangdong Academy of Agricultural Science, Guangzhou, China
xxw7505@163.com

Abstract

Pepper (Capsicum spp.) is one of the most economically and agriculturally important vegetable crops on the globe. However, the genomic resources of Pepper (Capsicum spp.) were scarcity, and only a few nucleotide sequences could be obtained in public databases. In this study, we examined transcriptome in C.annuum R597. More than 66million of high quality reads were generated using Illumina paired-end sequencing technology. Approximately 6 Gbp of data were generated, and de novo assembled into 96, 340 unigenes, with an N50 of 989 bp. The average length of unigenes was 651 bp. Based on sequence similarity search with known protein database, 51, 053 (53.0%) showed significant similarity to know proteins in Nr database, and 33, 766 (35.0%) had BLAST hits in Swiss-Prot database. Among the annotated unigenes, 26, 050 of C.annuum R597 unigenes were assigned to GO term annotation, and 28, 428 were found to have COG classification. In addition, a total of 14, 279 unigenes were assigned to 270 KEGG pathways. Moreover, a total 5, 960 SNPs, 5, 426 SSRs, and 2, 199 InDels were detected as potential molecular markers in pepper. Our study enriches the genomic resources of Pepper (Capsicum spp.) and provides powerful information for future studies. The availability of this ample amount of information about the transcriptome and SSRs, SNPs and InDels in pepper could server as valuable for studies on the agronomic traits and identification of molecular mechanism such plant resistance.

Keywords: *Capsicum annuum, Transcriptome analysis, Molecular marker discovery*

Introduction

Pepper (*Capsicum* spp.) is one of the most economically and agriculturally important vegetable crops on the globe, with high consumption of fresh or processed products. As a member of the family *Solanaceae*, it has a modest-sized diploid genome (2n=24) [1]. *Capsicum* peppers includes C.*annuum*, C.*chinese*, C.*baccatum*, C.*frutescens*, and C.*pubescens* and is cultivated in different parts of the world [2]. Over the past several years, C.*annuum* pepper traits have been investigated, such as yield [3], male sterility [4, 5], disease resistance [6], fruit quality [7-9], and important secondary metabolites [9].

Molecular markers, including single nucleotide polymorphisms (SNPs) [10], simple sequence repeats (SSRs) [2], cleaved amplified polymorphic sequences (CAPs) [11], insertion-deletion (InDels) [12], play an important role in molecular genetic studies described above especially. The development of molecular markers depends mostly on the publicly genomic and cDNA libraries. Up to currently, expressed Sequence Tags (ESTs) are still valuable resources for markers mining. However, ESTs database provide comparable data for analysis of plants that lack comparable genomic resources [13]. Previously ESTs were short subsequences derived from randomly isolated cDNAs time consuming and cost effective [14].

As the pepper genome is not currently available, transcriptome data can provide valuable and comprehensive information on gene expression, gene regulation, and amino

ISSN: 2005-4297 IJCA

acid content of proteins. Nowadays, the development of inexpensive next-generation sequencing (NGS) technologies, including Illumina, Roche 454, and ABI SOLiD, have provided a novel method for the analyses of transcriptome. Such Illumina technology has been successfully applied in the Solanaceae family [15]. Transcriptome analysis offers an opportunity to discovery and annotate gene, and monitor the variation of gene expression [16]. Moreover, Transcriptome analysis also provides a platform to develope enormous polymorphism molecular makers , such as SSRs, SNPs and InDels [15].

Capsicum pepper is a major vegetable crop in the world, particularly tropics. Like many other crop plants, the main breeding objectives include improvement in fruit quality and yield, increase resistance to pests and diseases, and adaption to abiotic stress. Although that various molecular markers have been developed in Capsicum [2, 15], it still did not seem to meet the requirements of breeding and molecular genetic research [17]. In the present study, we acquired the detailed transcriptome profile of C.annuum L. R597 by utilizing Illumina paired-end sequencing technology, and developed SSR, SNP, and InDel markers based on the transcriptome sequences. And these molecular markers will facilitate the discovery of gene candidates for high temperature and air humidity resistance and shed new light on the underlying mechanism of resistance in subsequent research.

Materials and Methods

Plant Material and RNA Extraction

One accession of *C.annuum*, R597, was obtained from Guangdong Province of China. Under long-term pressures of natural selection, R597 now possesses high temperature and air humidity tolerance. In our research, the experimental material was grown in the research experiment field of Vegetable Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou, China. Leaves were collected from healthy plants in the budding stage and frozen in liquid nitrogen immediately and stored at -80℃ until use.

To isolate total NRA, approximately 100mg of pepper leaves were snap-frozen in liquid nitrogen, ground into a powder, and then extracted using the Trizol Kit (Promega, USA) according to the manufacturer's instructions. RNA sample was subjected to DNase digestion (Takara Bio, Japan) to remove any remaining DNA.

CDNA Library Preparation and Sequencing

At least 20 μg of total RNA (≥400 ng/μL) was used to prepare a standard cDNA library for sequence analysis analysis using Illumina HiSeq^{TM} 2000 (commercial service) at the Beijing Genomics Institute (Shenzhen, China). After the first read was completed, the templates were regenerated *in situ* to enable a second 75 bp read from the opposite end of the fragments. Once the original templates were cleaved and removed, the reverse strands underwent sequencing-by-synthesis.

De Novo Assembly and Gene Annotation of Illumina Reads

In our research, we used Trinity method [9]for *de novo* assembly of these reads to generate a non-redundant set of transcripts. A final data set composed of 66, 571, 508 sequencing reads with 75-mer length were obtained. All reads were deposited in the National Center for Biotechnology Information (NCBI) and be accessed in the Short Read Archive (SRA) under accession number: SRA107820. For optimizing the *de novo* assembly, the method of additive multiple-k [9] was used to combine the properties assemblies using 7 different k-mers(21-47). To evaluate the accuracy of the assembled sequences (transcripts), all the usable sequencing reads were realigned to the unigenes using SOAPaligne(http://soap.genomics.org.cn/soapaligner.html) [18]. BLASTN was used for comparison of transcriptome and the CDS sequences of *solanum lycopersicum* (E

value $<10^{-5}$). All of the unigenes were aligned to three public protein databased (Nr, Swiss-Prot and KEGG) by blastx, and the cut-off E-value was $1.0e^{-5}$. The best aligning results were chosen to decide the direction of unigenes. Based on the results of protein database annotation, Blast2GO [19] was employed to obtain GO annotation according to molecular function, biological process and cellular component ontologies. The unigenes were also aligned to the COG database to predict and classify possible functions.

Profiles of Molecular Markers

To survey the molecular markers present in the *C. annuum* R597 accession, all contigs and singletons from both transcriptomes were used to mine SSR motifs, SNP and InDel markers. The MIcroSAtellite (MISA, http://pgrc.ipk-gatersleben.de/misa/) was used for identification of SSRs. In this study, the SSRs were considered to contain motifs with two to six nucleotides in size and a minimum of 4 contiguous repeat units. SNP and InDel markers were detected by aligning individual reads against contigs from the assembly using SOAPaligner software (Release 2.21, 08-13-2009).

Results

Illumina Sequencing and *de novo* Assembly

The dataset of raw reads was deposited in NCBI database under SRA107820 accession number and contains about 66, 571, 508 high quality reads with 97.45% of Q20 bases (base quality more than 20). Based on the reads, a total of 108, 256 of contigs were assembled, with total nucleotides of 72, 614, 394 bp, and an N50 of 1, 047 bp (i.e. 50% of the assembled bases were incorporated into contigs of 1, 047 bp). The length of contigs ranged from 201 bp to 16, 251 bp, with an average of 671 bp. Using the Trinity assembling program, we generated 96, 340 unigenes, with an N50 of 989 bp. The average length of unigenes was 651 bp (Table 1).

Table 1. Overview of the Transcriptome of *C. Annuum* R597

	Total number	Total Nucleotides (bp)	Average length(bp)	N50
Reads	66,571,508	5, 991, 435, 720	90	-
Contigs	108, 256	72, 614, 394	671	1,047
Unigens	96, 340	62, 708, 675	651	989

To evaluate the accuracy of the assembled sequence, SOAPaligner, which allowed up to 2 base mismatches, was employed to realign all the usable sequencing reads onto the unigenes [20]. This assembly produced a substantial number of large unigenes: 37, 770 unigenes (25.55%) longer than 500 bp, 17, 991 unigenes (18.67%) longer than 1, 000 bp, and 4, 539 unigenes (4.71%) longer than 2, 000 bp (Figure 1).

In our present study, the *c.annuum* R597 unigenes to orthologous *solanum lycopersicum* coding sequences was compared. There are 2, 254 unigenes with the ratio greater than 1, and

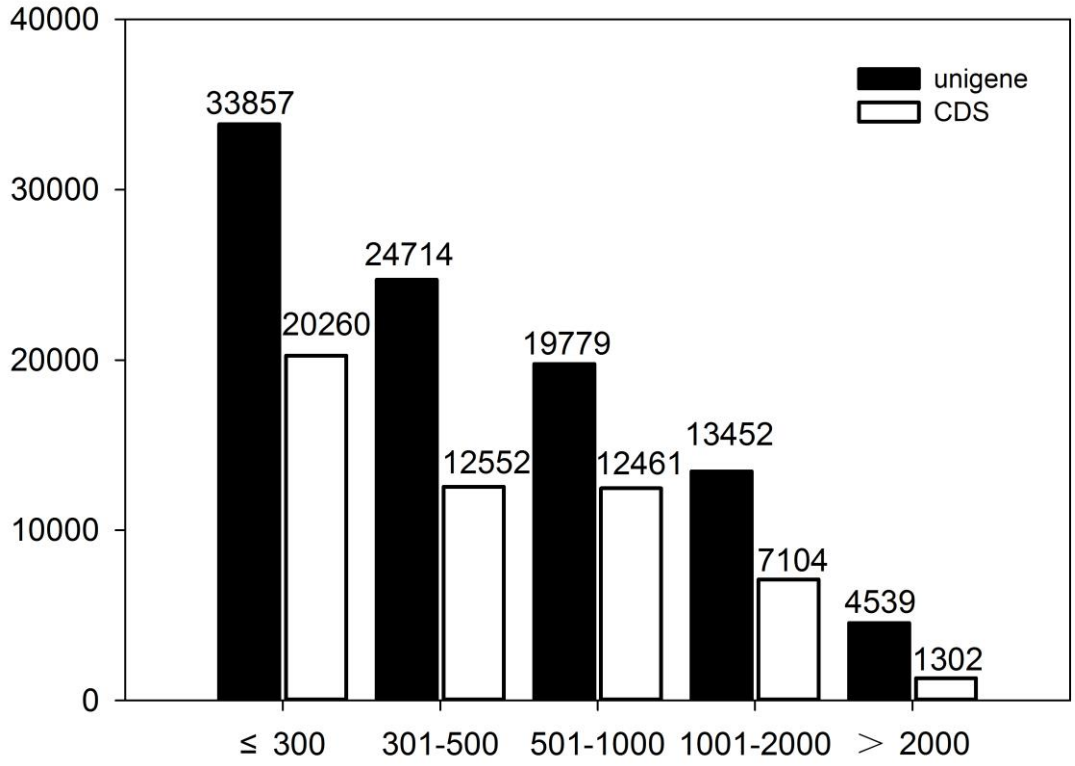

Figure 1. Size Distribution of the Unigenes and CDS. The Black and White Bars Indicate Unigene and CDS, Respectively

35, 723 unigenes with the ratio less than 1. In total, 6, 936 orthologs could be covered by unigenes with a percentage of more than 80%, and the coverage percentage of around 4, 205 orthologs ranged from 50% to 80%. Furthermore, 2, 045 orthologs were covered with only 20% or lower.

Annotation of Predicted Proteins

The unigenes were aligned to three public protein databases (Nr, Swiss-Prot, and KEGG). The results showed that out of 96, 340 unigenes, 51, 053 (53.0%) showed significant similarity to proteins in Nr database, and 33, 766 (35.0%) had BLAST hits in Swiss-Prot database. The E-value distribution of the top hits in the Nr and Swiss-Prot databases revealed that 44.0% and 37.0% of the mapped unigenes showed significant homology with the E-value less than 1E-50, respectively (Figure 2).

In addition, Based on the three public protein databases, we got a total of 53, 679 CDSs (50,306 CDSs predicted by blastx and 3, 373 by ESTScan), among which 1, 302 were over 2, 000 bp and 20, 867 were over 500 bp (Figure 1).

Gene Ontology (GO) Annotation

A total of 26, 050 unigenes were assigned to 42 functional groups using GO assignment. GO-annotated unigenes had three ontologies, including “biological process”, “cellular component”, and “molecular function”. The majority terms were “cellular components” (50, 575, 39.3%), followed by “biological process” (46, 990, 36.5%), and for “molecular function” (31, 147, 24.2%) (Table 2). “cell”, “cell part”, “binding”, and

"catalytic activity" were well represented. However, few genes were assigned to the terms "electron carrier activity", "cell killing", and "Locomotion".

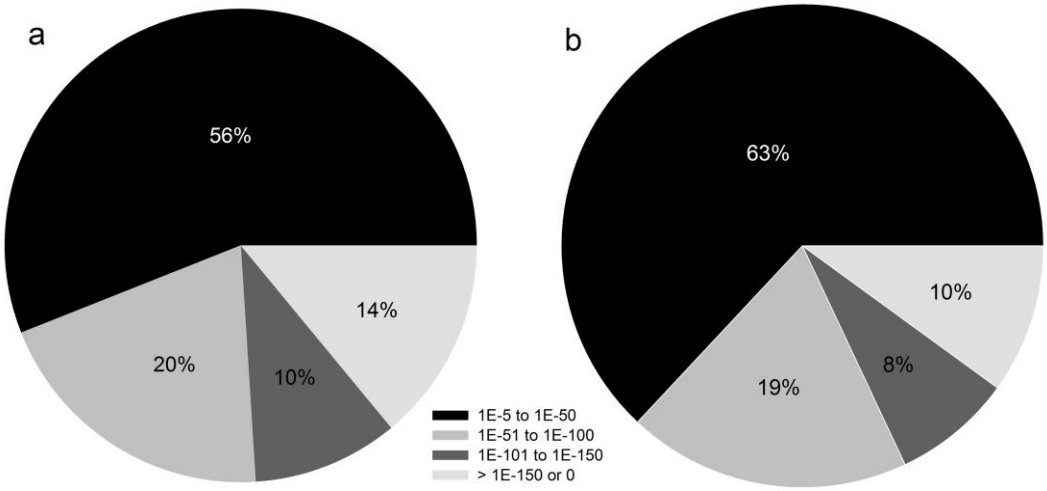

Figure 2. Characteristics of Similarity Search of Unigenes against Nr and Swiss-Prot Databases

COG Annotation

Furthermore, all unigenes were subjected to a search against the Cluster of Orthologous Groups (COG) database for functional prediction and classification. Overall, 28, 428 of the 51, 053 unigenes showing Nr hits were assigned to COG classifications (Figure 3). Among the 25

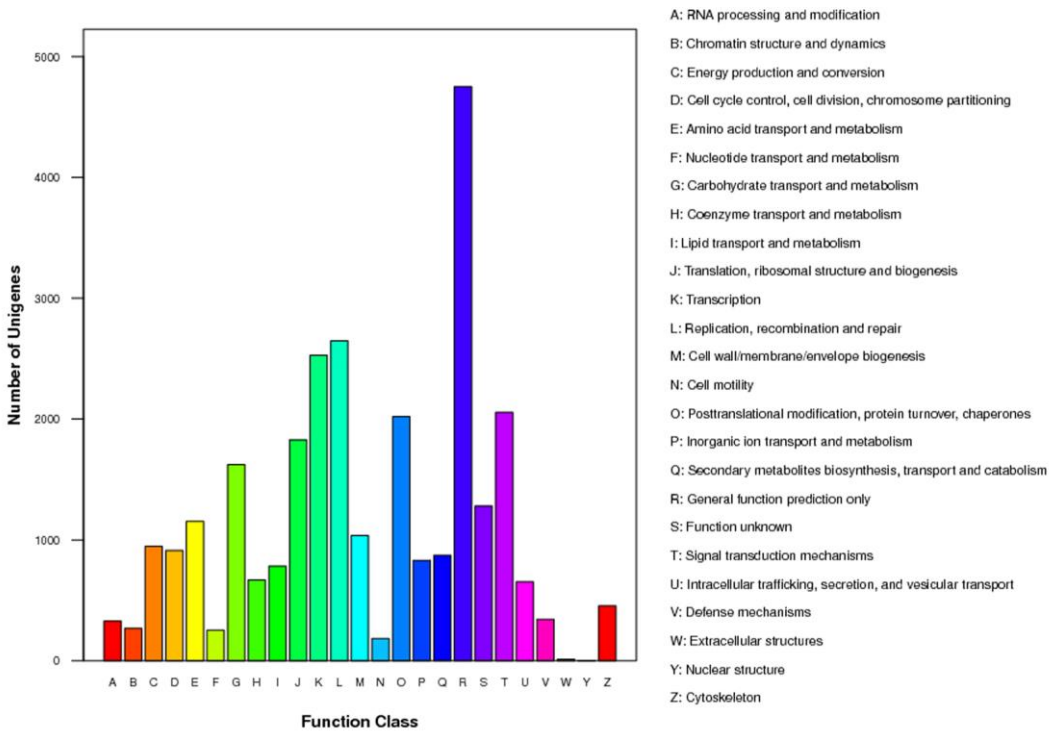

Figure 3. Clusters of Orthologous Group (COG) Classification of C.annuum Transcriptome

Table 2. Gene Ontology Classification of Assembled Unigenes

Ontology	Class	Number of unigene
Biological process	Anatomical structure formation	186
	Biological adhesion	15
	Biological regulation	2238
	Cell killing	12
	Cellular component biogenesis	306
	Cellular component organization	1436
	Cellular process	12265
	Death	341
	Developmental process	2260
	Establishment of localization	2280
	Growth	221
	Immune system process	144
	Localization	2344
	Locomotion	8
	Metabolic process	12926
	Multi-organism process	409
	Multicellular organismal process	1751
	Pigmentation	1602
	Reproduction	1112
	Reproductive process	1086
	Response to stimulus	3969
	Rhythmic process	64
	Viral reproduction	15
Cellular component	Cell	16421
	Cell part	16421
	Envelope	839
	Extracellular region	249
	Extracellular region part	27
	Macromolecular complex	1748
	Membrane-enclosed lumen	420
	Organelle	11442
	Organelle part	3008
Molecular function	Antioxidant activity	109
	Binding	14562
	Catalytic activity	13322
	Electron carrier activity	5
	Enzyme regulator activity	236
	Molecular transducer activity	736
	Structural molecule activity	405
	Transcription regulator activity	236
	Translation regulator activity	176
	Transporter activity	1360

The results were summarized in three main categories: biological process, cellular component and molecular function.

COG categories, “general function prediction only ” represented the largest group (4, 751, 16.71%), followed by “replication, recombination and repair” (2, 647, 9.31%), “transcription” (2, 529, 8.90%), “signal transduction mechanisms” (2, 055, 7.23%), “posttranslational modification, protein turnover, chaperones” (2, 021, 7.11%), “translation, ribosomal structure and biogenesis” (1, 828, 6.43%) , “carbohydrate

transport and metabolism" (1, 622, 5.71%), "function unknown" (1, 281, 4.51%), "amino acid transport and metabolism" (1, 154, 4.06%), and "cell wall/membrane/envelope biogenesis" (1, 037, 3.65%). Whereas, only a few unigenes assigned to "cell motility", "extracellular structures", and "nuclear structure" (183, 12 and 2, respectively).

Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Mapping

Functional classification and pathway assignment were performed by KEGG. In total 14, 279 unigenes were assigned to 270 KGEE pathways. The pathways with most representation by the unigenes were "spliceosome" (494), "plant hormone signal transduction" (491), "ribosome" (461), "RNA transport" (430), "protein processing in endoplasmic reticulum" (378), and "ribosome biogenesis in eukaryotes" (372).

Molecular Markers Discovery

SSRs are useful as molecular markers for genetics and biology researches. In this study, a total of 96, 340 unigenes from C. *annuum* R597 were used to mine potential microsatellites, which were defined as di- to hexa-nucleotide SSR with a minimum of four repeats for all motifs (except for di-nucleotide with a minimum of six repeats, and tri-nucleotide with a minimum of five repeats). Finally, 5, 426 microsatellites were detected in 4, 863 unigenes, out of which, 499 unigenes contained more than 1 SSRs (Table 3). Among the SSRs with a 2-6-nucleotide motif repeat, the majority of SSRs showed di-nucleotide (2, 184, 40.3%) or tri-nucleotide (2, 588, 47.7%) repeats. In contrast, the ratios for all other types of SSRs (tetra-, penta-, or hexa-nucleotide motifs) were relatively low (375, 6.9%, 114, 2.1%, 165, 3.0%, respectively). Furthermore, the length of SSRs was also analyzed, which was mainly distributed from 12 to 20 bp, accounting for 89.0% of the total SSRs (Table 4). The majority of di-nucleotide SSRs showed the AG/CT motif, followed by those with AT/AT motif, and finally the AC/GT motif. Among the tri-nucleotide SSRs, the AAC/GTT, AAG/CTT, and ATC/ATG motifs were prevalent (Figure 4).

Additionally, a total of 5, 960 putative SNPs, 2199 InDels, and 146 variants involving more than one nucleotide, existed in 6, 405 contigs. High-confidence differences were composed of 2621 SNPs, 1784 InDels and 12 variants involving more than one nucleotide. Statistical analysis suggested that , for both total and high-confidence SNPs, the proportion of transition nucleotide substitutions (63.71 and 65.39%) was greater than the proportion of transversions (36.29 and 34.61%, respectively)(Table 5).

Discussion

The lack of genomic information in hot pepper has impeded the research on this important vegetable crop at genetics and molecular biology. High-throughput sequencing is a superior technology for transcriptome analysis. During the past several years, the NGS technology has become a tremendous approach for high-throughput gene discovery on a genome-wide scale in non-model organisms whose genomic sequences are unknown. In this present study, we conducted a comprehensive study on the *de nove* assembly and characterization of the transcriptome of *Capsicum* pepper using Illumina platform and developed a large number of SSR and SNP markers based on the transcriptome information obtained.

In addition, transcriptome data will provide a valuable basis for the future studies on physiology, biochemistry, and molecular genetics on pepper.

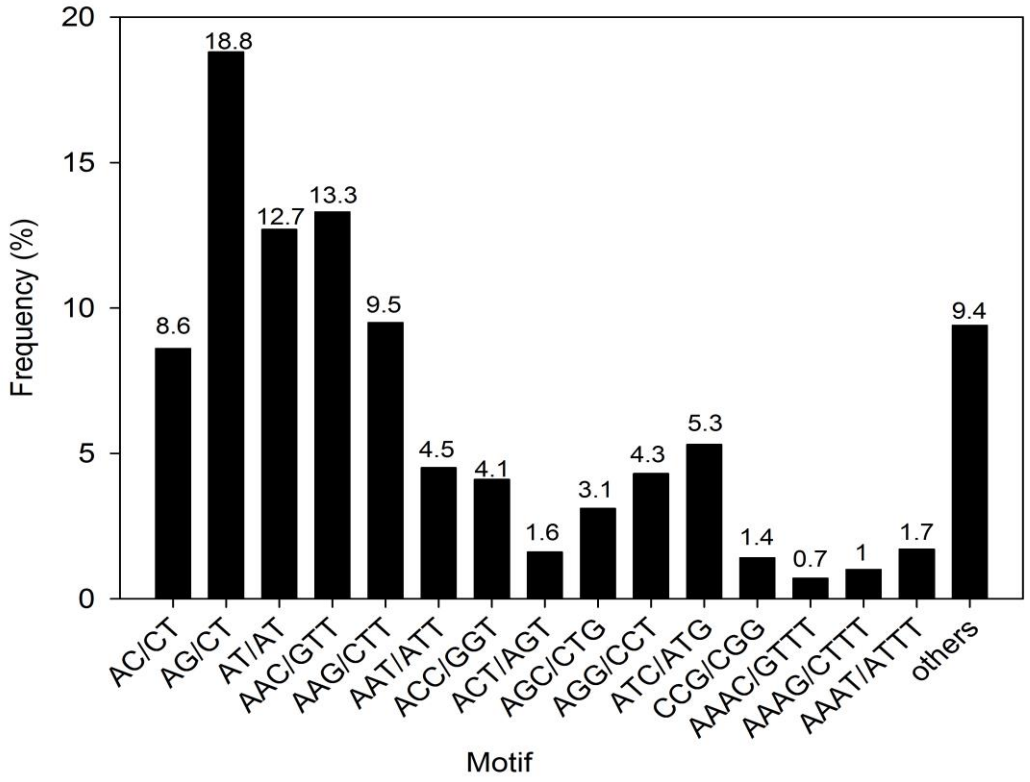

Figure 4. Frequency of Classified Repeat types of SSRs detected in our SSRs

Table 4. Summary of the Number of Repeat Units

Number of repeat unit	Di-	Tri-	Tetra-	Penta-	Hexa-
4	0	0	283	100	147
5	0	1520	68	14	6
6	917	644	23	0	5
7	464	380	1	0	1
8	310	36	0	0	3
9	199	4	0	0	1
10	182	0	0	0	1
11	105	0	0	0	0
≥12	7	4	0	0	1

There are many ways to deal with the possible sequencing errors, such as removal, trimming or correction to improve the assembly quality and decrease the amount of random access memory (RAM) required [21]. In order to obtain high-quality reads, we used a trimming strategy for filtering the reads in this study, and the Q20 is 97.45%. In our present study, we used the Trinity method to be *de nover* assemble, and 96, 340 unigenes with an N50 of 989 bp were yielded. The average length of unigenes was 651 bp, which was longer than those assemble in previous studies, for example, with butterfly (197 bp)[22], *Phoca largha* (392 bp) [23] , safflower (446 bp) [24], sweet potato (581 bp) [25], and *Arachis hypogaea* (619 bp)[18]. The above results indicated the high quality of our transcriptome sequencing and *de novo* assembly.

For sequence annotation, the sequence similarity search was performed against protein databases, including Nr, GO, Swiss-Prot, COG, and KEGG [26]. Most of our unigenes could match unique known proteins in public protein databases, indicating that the

transcriptome sequencing yielded a great number of unique genes in *Capsicum* pepper. Most of unigenes were assigned to a wide range of COG classification and gene ontology categories, which implied that our transcriptome data represented a broad diversity of transcripts in *C.annuum* pepper. Similar results were also reported in other species, such as *Benicasa hispida* [27], *Oryzias melastigma* [28], and *Piper nigrum* [29]. KEGG predictions identified many unigenes associated with the spliceosome", "plant hormone signal transduction" , "ribosome", and "RNA transport, which play an important role in the regulation of genetic activities and reactions to the changes of the environmental signals.Normally, through transcriptome sequencing and gene annotation, such large number of transcriptome sequences are of great importance for further research.

As demonstrated in this study, transcriptome sequencing is used for the development of DNA markers, not only because of the transcriptome data in which markers can be discovered, but also the discovered markers are gene-based [30,31]. Such markers are an important resource for determining functional variation and selecting the signature in genomic scans or association genetic studies [32,33]. Among the various molecular markers, simple sequence repeats (SSRs) are highly polymorphic, easier to develop, and serve as a rich resource of diversity [34]. In the present study, we predicted a total 5, 960 SNPs, 5, 426 SSRs, and 2199 InDels molecular markers (Figure 5). The large number of SSRs and SNPs detected here provide a wealth of potential markers that may prove useful in multiple applications, ranging from population genetics, linkage mapping, and comparative genomics, to gene-based association studies aimed at understanding the genetic control of plant resistance traits. However, all the predicted molecular markers need to rule out false positives and sequencing errors.

Table 5. SNP Statistical Information based on Mapping R597 Reads in Reference to *C.annuum* Unigenes

SNP-type	Number of all variations	Number of high-confidence variations
Transversion	2163(36.29%)	907(34.61%)
C-G	258	106
A-T	250	103
A-C	299	121
G-T	314	141
G-C	268	123
T-A	229	98
C-A	230	102
T-G	315	113
Transition	3797(63.71%)	1714(65.39%)
A-G	956	414
C-T	955	443
G-A	983	454
T-C	903	403
Total	5960	2621

Conclusion

In conclusion, the in this study, the transcriptome sequencing analysis of *C.annuum* R597 was conducted using Illumina paired-end sequencing technology. More than 66 million of high quality reads were generated, and approximately 6 Gbp data were generated, and *de novo* assemble into 96, 340 unigenes, with an N50 of 989 bp. Most of the unigenes have been sequence annotated. Moreover, a total 5, 960 SNPs, 5, 426 SSRs, and 2199 InDels were detected, these large set of transcribed SSRs, SNPs and InDels in pepper could server as valuable for studies on the agronomic traits and identification of molecular mechanism such plant resistance.

Acknowledgments

This work was financially supported by the presidential foundation of Guangdong academy of agricultural sciences (201108).

Reference

[1]. V. S. Govindarajan and M. N. Sathyanarayana, "Capsicum--production, technology, chemistry, and quality", Part V. Impact on physiology, pharmacology, nutrition, and metabolism; structure, pungency, pain, and desensitization sequences, Crit Rev Food Sci Nutr., vol. 29, **(1991)**, pp. 435-474.
[2]. H. J. Kim, K. H. Baek, S. W. Lee, J. Kim, B. W. Lee, H. S. Cho, W. T. Kim, D. Choi and C. G. Hur, "Pepper EST database: comprehensive in silico tool for analyzing the chili pepper (*Capsicum annuum*) transcriptome", BMC Plant Biol., vol. 8, **(2008)**, pp. 101.
[3]. G. U. Rao, A. Ben Chaim, Y. Borovsky and I. Paran, "Mapping of yield-related QTLs in pepper in an interspecific cross of *Capsicum annuum* and *C. frutescens"*, Theor Appl Genet, vol. 106, **(2003)**, pp. 1457-1466.
[4]. D. S. Kim, D. H. Kim, J. H. Yoo and B. D. Kim, "Cleaved amplified polymorphic sequence and amplified fragment length polymorphism markers linked to the fertility restorer gene in chili pepper (*Capsicum annuum* L.)", Mol Cells, vol. 21, **(2006)**, pp. 135-140.
[5]. C. Liu, N. Ma, P. Y. Wang, N. Fu and H. L. Shen, "Transcriptome sequencing and De Novo analysis of a cytoplasmic male sterile line and its near-isogenic restorer line in chili pepper (*Capsicum annuum* L.)", PLoS One, vol. 8, **(2013)**, pp. e65209.
[6]. E. A. Quirin, E. A. Ogundiwin, J. P. Prince, M. Mazourek, M. O. Briggs, T. S. Chlanda, K. T. Kim, M. Falise, B. C. Kang and M. M. Jahn, "Development of sequence characterized amplified region (SCAR) primers for the detection of Phyto.5.2, a major QTL for resistance to *Phytophthora capsici* Leon in pepper", Theor Appl Genet, vol. 110, **(2005)**, pp. 605-612.
[7]. Y. Borovsky and I. Paran, "Characterization of fs10.1, a major QTL controlling fruit elongation in *Capsicum"*, Theor Appl Genet, vol. 123, **(2011)**, pp. 657-665.
[8]. A. Ben-Chaim, Y. Borovsky, M. Falise, M. Mazourek, B. C. Kang, I. Paran and M. Jahn, "QTL analysis for capsaicinoid content in *Capsicum"*, Theor Appl Genet, vol. 113, **(2006)**, pp. 1481-1490.
[9]. S. Liu, W. Li, Y. Wu, C. Chen and J. Lei, "De novo transcriptome assembly in chili pepper (*Capsicum frutescens*) to identify genes involved in the biosynthesis of capsaicinoids", PLoS One, vol. 8, **(2013)**, pp. e48156.
[10]. K. J. Schmid, O. Torjek, R. Meyer, H. Schmuths, M. H. Hoffmann and T. Altmann, "Evidence for a large-scale population structure of *Arabidopsis thaliana* from genome-wide single nucleotide polymorphism markers", Theor Appl Genet, vol. 112, **(2006)**, pp. 1104-1114.
[11]. Y. Shu, Y. Li, Z. Zhu, X. Bai, H. Cai, W. Ji, D. Guo and Y. Zhu, "SNPs discovery and CAPS marker conversion in soybean", Mol Biol Rep, vol. 38, **(2011)**, pp. 1841-1846.
[12]. A. Heesacker, V. K. Kishore, W. Gao, S. Tang, J. M. Kolkman, A. Gingle, M. Matvienko, A. Kozik, R. M. Michelmore, Z. Lai, L. H. Rieseberg and S. J. Knapp, "SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility", Theor Appl Genet, vol. 117, **(2008)**, pp. 1021-1029.
[13]. W. A. Rensink, Y. Lee, J. Liu, S. Iobst, S. Ouyang and C. R. Buell, "Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts", BMC Genomics, vol. 6, **(2005)**, pp. 124.
[14]. W. R. McCombie, M. D. Adams, J. M. Kelley, M. G. FitzGerald, T. R. Utterback, M. Khan, M. Dubnick, A. R. Kerlavage, J. C. Venter and C. Fields, "Caenorhabditis elegans expressed sequence tags identify gene families and potential disease gene homologues", Nat Genet, vol. 1, **(1992),** pp. 124-131.
[15]. M. Nicolai, C. Pisani, J. P. Bouchet, M. Vuylsteke and A. Palloix, "Discovery of a large set of SNP and SSR genetic markers by high-throughput sequencing of pepper (*Capsicum annuum*)", Genet Mol Res, vol. 11, **(2012)**, pp. 2295-2300.
[16]. D. Cohen, M. B. Bogeat-Triboulot, E. Tisserant, S. Balzergue, M. L. Martin-Magniette, G. Lelandais, N. Ningre, J. P. Renou, J. P. Tamby, D. Le Thiec and I. Hummel, "Comparative transcriptomics of drought responses in Populus: a meta-analysis of genome-wide expression profiling in mature leaves and root apices across two genotypes", BMC Genomics, vol. 11, **(2010)**, pp. 630.
[17]. H. J. Kim, S. H. Nahm, H. R. Lee, G. B. Yoon, K. T. Kim, B. C. Kang, D. Choi, O. Y. Kweon, M. C. Cho, J. K. Kwon, J. H. Han, J. H. Kim, M. Park, J. H. Ahn, S. H. Choi, N. H. Her, J. H. Sung and B. D. Kim, "BAC-derived markers converted from RFLP linked to *Phytophthora capsici* resistance in pepper (*Capsicum annuum* L.)", Theor Appl Genet, vol. 118, **(2008)**, pp. 15-27.
[18]. K. Tanase, C. Nishitani, H. Hirakawa, S. Isobe, S. Tabata, A. Ohmiya and T. Onozaki, "Transcriptome analysis of carnation (*Dianthus caryophyllus* L.) based on next-generation sequencing technology", BMC Genomics, vol. 13, **(2012)**, pp. 292.
[19]. A. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research", Bioinformatics, vol. 21, **(2005)**, pp. 3674-3676.

[20]. R. Li, Y. Li, K. Kristiansen and J. Wang, "SOAP: short oligonucleotide alignment program", Bioinformatics, vol. 24, **(2008)**, pp. 713-714.
[21]. J. R. Miller, S. Koren and G. Sutton, "Assembly algorithms for next-generation sequencing data", Genomics, vol. 95, **(2010)**, pp. 315-327.
[22]. J. C. Vera, C. W. Wheat, H. W. Fescemyer, M. J. Frilander, D. L. Crawford, I. Hanski and J. H. Marden, "Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing", Molecular ecology, vol. 17, **(2008)**, pp. 1636-1647.
[23]. X. Gao, J. Han, Z. Lu, Y. Li and C. He, "Characterization of the spotted seal Phoca largha transcriptome using Illumina paired-end sequencing and development of SSR markers (Retraction of vol 7, pg 277, 2012)", COMPARATIVE BIOCHEMISTRY AND PHYSIOLOGY D-GENOMICS & PROTEOMICS, vol. 8, **(2013)**, pp. 163-163.
[24]. H. Lulin, Y. Xiao, S. Pei, T. Wen and H. Shangqin, "The first Illumina-based de novo transcriptome sequencing and analysis of safflower flowers", PloS one, vol. 7, **(2012)**, pp. e38653.
[25]. S. Guo, J. Liu, Y. Zheng, M. Huang, H. Zhang, G. Gong, H. He, Y. Ren, S. Zhong, Z. Fei and Y. Xu, "Characterization of transcriptome dynamics during watermelon fruit development: sequencing, assembly, annotation and gene expression profiles", BMC Genomics, vol. 12, **(2011)**, pp. 454.
[26]. A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research", Bioinformatics, vol. 21, **(2005)**, pp. 3674-3676.
[27]. B. Jiang, D. Xie, W. Liu, Q. Peng and X. He, "De Novo Assembly and Characterization of the Transcriptome, and Development of SSR Markers in Wax Gourd (*Benicasa hispida*)", PloS one, vol. 8, **(2013)**, pp. e71054.
[28]. Q. Huang, S. Dong, C. Fang, X. Wu, T. Ye and Y. Lin, "Deep sequencing-based transcriptome profiling analysis of *Oryzias melastigma* exposed to PFOS", Aquat Toxicol, vol. 120-121, **(2012)**, pp. 54-58.
[29]. S. Gordo, D. Pinheiro, E. Moreira, S. Rodrigues, M. Poltronieri, O. de Lemos, I. da Silva, R. Ramos, A. Silva and H. Schneider, "High-throughput sequencing of black pepper root transcriptome", BMC plant biology, vol. 12, **(2012)**, pp. 168.
[30]. W. B. Barbazuk, S. J. Emrich, H. D. Chen, L. Li and P. S. Schnable, "SNP discovery via 454 transcriptome sequencing", Plant J, vol. 51, **(2007)**, pp. 910-918.
[31]. D. L. Hyten, S. B. Cannon, Q. Song, N. Weeks, E. W. Fickus, R. C. Shoemaker, J. E. Specht, A. D. Farmer, G. D. May and P. B. Cregan, "High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence", Bmc Genomics, vol. 11, **(2010)**, pp. 38.
[32]. H.-R. Lee, I.-H. Bae, S.-W. Park, H.-J. Kim, W.-K. Min, J.-H. Han, K.-T. Kim and B.-D. Kim, "Construction of an integrated pepper map using RFLP, SSR, CAPS, AFLP, WRKY, rRAMP, and BAC end sequences", Molecules and cells, vol. 27, **(2009)**, pp. 21-37.
[33]. T. Parchman, K. Geist, J. Grahnen, C. Benkman and C. A. Buerkle, "Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery", BMC genomics, vol. 11, **(2010)**, pp. 180.
[34]. F.-H. Lu, M.-Y. Yoon, Y.-I. Cho, J.-W. Chung, K.-T. Kim, M.-C. Cho, S.-R. Cheong and Y.-J. Park, "Transcriptome analysis and SNP/SSR marker information of red pepper variety YCM334 and Taean", Scientia Horticulturae, vol. 129, **(2011)**, pp. 38-45.

