A DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Multiple Reference Pepper Genome Analysis Provides Insights into Genome Evolution and Speciation of *Capsicum spp.*

다중 표준 유전체 분석을 통한 고추 속 식물의 유전체 진화 및 종 분화 연구

AUGUST 2015

SEUNGILL KIM

INTERDISCIPLINARY PROGRAM IN AGRICULTURAL BIOTECHNOLOGY

COLLEGE OF AGRICULTURE AND LIFE SCIENCES

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Multiple Reference Pepper Genome Analysis Provides Insights into Genome Evolution and Speciation of *Capsicum* spp.

UNDER THE DIRECTION OF DR. DOIL CHOI
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF
SEOUL NATIONAL UNIVERSITY

BY
SEUNGILL KIM

MAJOR IN HORTICULTURAL CROP GENOMICS
INTERDISCIPLINARY PROGRAM IN AGRICULTURAL BIOTECHNOLOGY

AUGUST 2015

APPROVED AS A QUALIFIED DISSERTATION OF SEUNGILL KIM
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
BY THE COMMITTEE MEMBERS

CHAIRMAN	Yong-Hwan Lee, Ph.D.
VICE-CHAIRMAN	Doil Choi, Ph.D.
MEMBER	Ryan W. Kim, Ph.D.
MEMBER	Tae-Jin Yang, Ph.D.
MEMBER	Byoung-Cheorl Kang, Ph.D.

Multiple Reference Pepper Genome Analysis Provides Insights into Genome Evolution and Speciation of *Capsicum spp.*

SEUNGILL KIM

Interdisciplinary Program in Agricultural Biotechnology,

Seoul National University

ABSTRACT

Pepper (*Capsicum* spp.) is widely cultivated spice crop and a major ingredient of worldwide cuisines. In this study, a high-quality reference genome of hot pepper (*Capsicum annuum* CM334) was constructed containing genome assembly, annotation and chromosome pseudomolecules. A total of 87.9 % (3.06 Gb) of 3.48 Gb genome was assembled and 2.63 Gb of the assembled sequence was integrated into twelve chromosomes. By comparing to the completed BAC sequences, the assembled genome was validated and the validation result showed that the identities between the assembled genome and BAC sequences were more than 99%. A total of 34,903 protein-coding genes were predicted and their biological functions were annotated. To

construct multiple reference genomes of the genus *Capsicum*, high-quality *de novo* genome sequences of two domesticated *Capsicum* spp. (*C. chinense* and *C. baccatum*) were obtained and comparative analysis with *C. annuum* CM334 genome was performed. The phylogenetic analysis revealed that the speciation of the *Capsicum* species has occurred at approximately between one and two million years ago. The genome size variation between *C. baccatum* and the other *Capsicum* spp. was caused by excessive accumulation of athila subgroup of LTR/Gypsy retrotransposon. The insertion pattern of LTR-retrotransposons was represented that the accumulation of LTRs for *C. chinense* and *C. baccatum* were distinctly increased around their speciation time. Comparative and evolutionary analysis of gene families unveiled that the expanded gene families in each pepper genome were actively increased by recent gene duplication event occurred around the speciation times. The high-quality multiple reference genomes of the genus *Capsicum* will provide a versatile platform for genetic and genomic improvements as well as essential resources to promote population and comparative genomic studies in the genus *Capsicum*.

Keywords: genome, *Capsicum annuum*, *Capsicum baccatum*, *Capsicum chinense*, next generation sequencing (NGS), assembly, annotation, speciation, genome evolution, transposable elements (TEs), LTR-retrotransposons, compartive genomic analysis

Student number: 2009-21228

CONTENTS

LIST OF TABLES

CHAPTER 1

CHAPTER 2

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

LIST OF ABBREVIATIONS

NGS	next generation sequencing
WGS	whole genome shotgun
TEs	transposable elements
LTR	long terminal repeat
BAC	bacterial artificial chromosome
FAO	food and agriculture organization
PE	paired-end
MP	mate-pair
RIL	recombinant inbred line
PGA	pepper gene annotation
EVM	evidence modeler
COS	conserved orthologue set
OrthoMCL	ortholog markov cluster
BEAST	bayesian evolutionary analysis sampling trees
MYA	million years ago

GENERAL INTRODUCTION

Since the first plant genome sequencing project, *Arabidopsis thaliana*, was completed in 2000, a wide variety of plant genome sequences were reported including plants with different genome size ranging from 125Mb (*Arabidopsis thaliana*) to 20.0 Gb (*Pinus taeda*) [1-3]. Compared to other eukaryotic genomes, plant genomes are usually larger in size and highly enriched with repeat sequences. These features have been barriers of the genome assembly, and required enormous amount of sequencing data. For these reasons, huge amount of sequencing cost and human resources have been required to perform plant genome project. Subsequently, most of the plant genome projects were conducted by consortium or collaboration scale[4-8]. However, in the middle of 2000, the emergence of next-generation-sequencing (NGS) technology has enabled generation of high-throughput data and accelerated the completion of genome projects. Starting with the cucumber genome in 2009, plant genomes have been actively sequenced using NGS technology and publications of *de novo* sequenced plant genomes are exponentially increased[9].

A total of 72 publications of *de novo* genomes from 73 plant species were reported comprising fifty dicot, seventeen monocot and eight other plant genomes by 2013. Until 2010, thirteen genomes were sequenced mainly by BAC tiling method including major crops and model plants such as *Arabidopsis*, rice, and maize[2,8,10,11]. Thereafter, the draft genome of cucumber was finished firstly via NGS

technology[9], and followed by sixteen genomes were sequenced including soybean, brachypodium, medicago, potato and cacao from 2010 to 2011[6,7,12-14]. In the 2011, a genome of *A. lylata* was published as first multi-species reference genome for the genus *Arabidopsis*[15]. Furthermore, eighteen *de novo* genomes of *A. thaliana* accessions were reported as the multiple reference genomes for the intra-species of *A. thaliana*[16].

Mostly, small plant genomes had been sequenced except 2 Gb of maize genome until 2012. Larger genome sequencing has been launched by advancement of sequencing and assembly technologies with the falling down of sequencing cost. In 2012, seventeen genomes were sequenced containing wheat, watermelon, banana, tobacco, and cotton genome[17-21]. Especially, wheat genome (*Triticum aestivum*) with 17 Gb of genome size was the largest genome solely constructed using NGS technology by this time. However, its low assembly quality resulted from short fragments and low coverage for whole genome has been barriers for wheat genomic studies.

In 2013, total twenty-seven genomes were sequenced and published including current largest sequenced genome of *P. glauca* (20.0 Gb), and various inter-species of existing genomes[3,22-29]. Especially, as enormous genome sequencings become eligible, genome construction has been no more recognized as unchallengeable work in this time. Therefore, multiple *de novo* genome projects have been performed by single groups to construct reference genomes for genome-wide

association studies or for comparative genomic studies among genomes for closely related species[30,31]. As a representative example, three crucifer genomes (*Leavenworthia alabamica*, *Sisymbrium irio* and *Aethionema arabicum*) were constructed and compared with pre-existing crucifer genomes[27]. In addition, two tobacco genomes (*Nicotiana sylvestris* and *Nicotiana tomentosiformis*) were assembled to obtain reference genomes for ancestral species of allotetraploid tobacco (*N. tabacum*)[22].

Until now, all of enormous plant genomes over 4 Gb were assembled via NGS technology except barley genome. However, due to the limitation of NGS technology in terms of short-read length, a lot of the genomes are still remaining as low quality genomes. Specifically, only 21 % of assembled genomes (6 out of 28 genomes) generated solely by NGS provided pseudomolecules with N50 value over 1Mb. In contrast, 62 % of genomes (28 out of 45) constructed by other methods were completed pseudomolecule construction with N50 value over 1Mb. Even though, genome construction using NGS technology enables easy and rapid completion of a genome project, a solution is still required to assure the quality like the classical method.

For over a decade, the trend of plant genome sequencing has been dramatically changed. Initial plant genome sequencings were focused on the construction of genome sequence and published papers reported genomic contents such as gene repertories and type of repeats. As the completed genome sequences were

accumulated, researchers have started to compare newly sequenced genomes to related existing genome sequences to discover factors to lead the differences[15,32-36]. Furthermore, comparative genomic analysis among the species level is appeared. For example, the study by Hu et al. (2011) revealed genome size variation and change of gene repertories between *A. lylata* and *A. thaliana*[15]. In addition, Gan et al. (2011) reported that eighteen *de novo* genomes of *A. thaliana* accessions are extremely diverged, even though all of the sequenced genomes were the same species[16]. They insisted that a single genome could not play a role as a reference genome of the genus. In addition, Huang et al. (2010) referred that more assembled landrace genomes of rice will be effective for the comprehensive identification of structural variation events in genome wide association study[37]. Thus, sequencing projects for inter-species genomes of rice have been performed to construct multiple reference genomes of *Oryza* spp. [28].

Even though a number of plant genomes have been sequenced, there are a great deal more plant species which has no reference genome including chilli peppers (*Capsicum* spp.). Chilli pepper is widely cultivated food plant in the world, and a major vegetable crop used for worldwide cuisines[38]. According to FAO statistics, worldwide production of chilli pepper in the top 20 producing countries was reached to 14.4 billion US dollars in 2011[39]. Furthermore, the pepper production has rised 40% during the last decade. Most importantly, pepper provides many nutrients containing essential vitamins and minerals that have great importance for

human health[40-43]. However, despite of its importance as one of major vegetable crops for nutritional and medicinal values, a reference genome of pepper has not been sequenced due to enormous size of its genome.

In this study, I report a high quality reference pepper genome containing assembly, annotation and pseudomolecules. In addition, this study contains construction of two additional reference pepper genomes via *de novo* genome sequencing with their evolutionary history in *Capsicum* spp. Studies focused on the following topics:

Chapter 1: Construction of a reference pepper genome (*Capsicum annuum* landrace CM334)

Chapter 2: Multiple reference pepper genome sequencing and evolutionary history of the genus *Capsicum*.

REFERENCES

1. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant Journal* **53**, 661-673 (2008).
2. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).
3. Birol, I. *et al.* Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. *Bioinformatics* **29**, 1492-1497 (2013).
4. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641 (2012).
5. International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711-716 (2012).
6. Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-195 (2011).
7. Vogel, John P., *et al*. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**, 763-768 (2010).
8. Goff, S.A. *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**, 92-100 (2002).
9. Huang, S. *et al.* The genome of the cucumber, Cucumis sativus L. *Nat. Genet.* **41**, 1275-1281 (2009).
10. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115 (2009).
11. Yu, J. *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp indica). *Science* **296**, 79-92 (2002).
12. Young, N.D. *et al.* The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**, 520-524 (2011).
13. Argout, X. *et al.* The genome of Theobroma cacao. *Nat. Genet.* **43**, 101-118 (2011).
14. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183 (2010).

15. Hu, T.T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476-481 (2011).
16. Gan, X. *et al.* Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**, 419-423 (2011).
17. Guo, S.G. *et al.* The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**, 51-58 (2013).
18. Wang, K.B. *et al.* The draft genome of a diploid cotton Gossypium raimondii. *Nat. Genet.* **44**, 1098-1103 (2012).
19. D'Hont, A. *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213-217 (2012).
20. Brenchley, R. *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705-710 (2012).
21. Bombarely, A. *et al.* A Draft Genome Sequence of Nicotiana benthamiana to Enhance Molecular Plant-Microbe Biology Research. *Molecular Plant-Microbe Interactions* **25**, 1523-1530 (2012).
22. Sierro, N. *et al.* Reference genomes and transcriptomes of Nicotiana sylvestris and Nicotiana tomentosiformis. *Genome Biol.* **14**, R60 (2013).
23. Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579-584 (2013).
24. Motamayor, J.C. *et al.* The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol.* **14**, r53 (2013).
25. Ling, H.Q. *et al.* Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature* **496**, 87-90 (2013).
26. Jia, J.Z. *et al.* Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91-95 (2013).
27. Haudry, A. *et al.* An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* **45**, 891-898 (2013).
28. Chen, J. *et al.* Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. *Nat. Commun.* **4**, 1595 (2013).

29. Albert, Victor A., *et al*. The Amborella genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
30. Garcia-Mas, J. *et al*. The genome of melon (Cucumis melo L.). *Proc. Natl. Acad. Sci. USA* **109**, 11872-7 (2012).
31. Guo, S. *et al*. The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**, 51-8 (2013).
32. Wang, X. *et al*. The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet.* **43**, 1035-1039 (2011).
33. Shulaev, V. *et al*. The genome of woodland strawberry (Fragaria vesca). *Nat. Genet.* **43**, 109-116 (2011).
34. Dassanayake, M. *et al*. The genome of the extremophile crucifer Thellungiella parvula. *Nat. Genet.* **43**, 913-918 (2011).
35. Al-Dous, E.K. *et al*. De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). *Nat. Biotechnol.* **29**, 521-527 (2011).
36. Velasco, R. *et al*. The genome of the domesticated apple (Malus x domestica Borkh.). *Nat. Genet.* **42**, 833-839 (2010).
37. Huang, X. *et al*. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961-967 (2010).
38. Bosland, P.W.E.J.V.P. *Vegetable and Spice Capsicums* (CABI, Wallingford, UK, 2012).
39. FAO. Food and Agricultural Commodities Production. FAO, Rome http://faostat.fao.org (2013).
40. Marin, A., Ferreres, F., Tomas-Barberan, F.A. & Gil, M.I. Characterization and quantitation of antioxidant constituents of sweet pepper (*Capsicum annuum* L.). *J. Agric. Food Chem.* **52**, 3861-3869 (2004).
41. Matsufuji, Hiroshi, et al. "Anti-oxidant content of different coloured sweet peppers, white, green, yellow, orange and red (Capsicum annuum L.)." *Int. J. Food Sci. Technol.* **42**, 1482-1488 (2007).
42. Matus, Z., Deli, J. & Szabolcs, J.J. Carotenoid composition of yellow pepper during ripening-isolation of beta-cryptoxanthin 5, 6-epoxide. *J. Agric. Food Chem.*

39, 1907-1914 (1991).

43. Mejia, L.A., Hudson, E., deMejia, E.G. & Vazquez, F. Carotenoid content and vitamin-a activity of some common cultivars of Mexican peppers (*Capsicum annuum*) as determined by HPLC. *J. Food Sci.* **53**, 1448-1451 (1998).

CHAPTER 1

Construction of a reference pepper genome

(*Capsicum annuum* landrace CM334)

The research described in this Chapter has been published in Nature Genetics Vol 46, No. 3, 2014, pp.270- 278.

ABSTRACT

Pepper (*Capsicum annuum* L.) is one of the oldest domesticated and the most widely grown spice crop in the world. Here, I report the high-quality genome sequence of hot pepper (Mexican landrace CM334, 2n=2X=24). A total of 87.6% (3.06 Gb) of 3.48 Gb genome was assembled and 85.9% (2.63 Gb) of the assembled sequence was constructed as 12 chromosome pseudomolecules. I predicted 34,903 protein coding genes and their functional description in the genome. The high-quality reference genome of hot pepper will serve as not only a platform for improving nutritional and medicinal values of *Capsicum* spp. but also an important resource for functional, comparative, and evolutionary studies in plants.

INTRODUCTION

Hot pepper, a member of the Solanaceae family, is a diploid, facultative self-pollinating crop and is related to potato, tomato, eggplant, tobacco and petunia. Solanaceae plants belong to the asterid clade of eudicot that include over 3000 diverse species on the planet and many members have the same number of chromosomes (X=12), yet they differ drastically in genome size. The hot pepper provides a wide variety of uses for pharmaceuticals, natural coloring agents, cosmetics, as well as ornamental plants and the source for the active ingredient in most defense repellants. Most importantly, hot peppers provide many essential vitamins, minerals and nutrients that have great importance for human health[1-4]. In 2011, fresh and dried hot pepper production of the top 20 producing countries was 33.3 million tons planted on 3.8 Mha[5]. In the last decade, world hot pepper production has increased 40%.

The pungency (heat) of the hot pepper is due to the accumulation of capsaicinoids, a group of alkaloids that are unique to the *Capsicum* genus. The heat sensation these capsaicinoids create is such a defining aspect of this crop that the genus name, *Capsicum*, comes from the Greek “kapto”, meaning "to bite". Capsaicin, dihydrocapsaicin and nordihydrocapsaicin constitute the primary capsaicinoids, which have an organ-specific biosynthesis and are produced exclusively in glands on the placenta of the fruit. The organoleptic sensation of heat

caused when capsaicin binds to the mammalian TRPV1 receptor in the pain pathway[6], can be argued to be a sixth taste along with sweet, sour, bitter, salty, and umami (tasty). Many of the enzymes involved in capsaicinoid biosynthesis are not well characterized, and the regulation of the pathway is not fully understood. With more than 22 capsaicinoids isolated from hot peppers, this genus provides an excellent example for exploring the evolution of secondary metabolites in plants[7]. Capsaicinoids have been found in nature to have antifungal and antibacterial properties, act as a deterrent to animal predation when ingested and have inherent properties that aid avian seed dispersal. Moreover, capsaicinoids have many health benefits for humans; they are effective at inhibiting the growth of several forms of cancer[8-10], are an analgesic for arthritis and other pain[11] and reduce appetite, promoting weight loss[12-14]. It is surprising that a complete understanding of the capsaicinoid pathway is lacking, especially at the molecular level, considering the economic and cultural importance of capsaicinoids.

Here, I report the high-quality genome sequence of hot pepper (Mexican landrace, Criollo de Morelos 334, 2n=2X=24, USDA PI636424). Criollo de Morelos 334 (CM334) a landrace collected from the Mexican state Morelos has consistently exhibited high levels of resistance to diverse pathogens such as *Phytophthora capsici*, Pepper Mottle Virus, and Root-Knot Nematodes. It has been used extensively in hot pepper research and cultivar breeding.

MATERIALS AND METHODS

Plant materials and high molecular weight DNA preparation

Capsicum annuum 'CM334' (Criollo de Morelos 334, a landrace collected from the Mexican state Morelos) was used for sequencing the genome. CM334 have multiple disease resistance traits and extensively used as a major breeding source for disease resistance against *Phytophthora spp.*, Tobacco Mosaic Virus (TMV), Potato Virus Y (PVY) and root-knot nematode (*Meloidogyne* spp.)[15]. The plants were grown in greenhouse and fresh meristemic expanding leaves were harvested and frozen in liquid nitrogen for isolation of genomic DNA.

High molecular weight genomic DNA for paired-end and mate-pair libraries was extracted from isolated leaf nuclei, then the quality and size range of isolated DNA was confirmed by agarose gels following separation by pulsed field gel electrophoresis. The libraries for NGS sequencing were constructed with > 100 kb genomic DNA according to the manufacturers instruction (Illumina, USA). Before sequencing, quality of each library was validated with KAPA SYBR FAST Master Mix Universal 2X qPCR Master Mix (Kapa Biosystems, USA).

Genome sequencing

Paired-end and mate-pair libraries for sequencing were prepared with paired-end (PE) or mate-pair (MP) library kits (Illumina, San-Diego, CA) following

manufacturer's instructions and validated with KAPA SYBR FAST Master Mix Universal 2X qPCR Master Mix (Kapa Biosystems, Woburn, MA). Constructed libraries were sequenced on Illumina platforms (GA2 and Hiseq 2000) using standard protocols.

Genome assembly

Before genome assembly, short reads sequences from each library were pre-processed using in-house pre-processing pipelines to increase accuracy of genome assembly. Contaminant of bacterial sequences and duplicated short reads were removed as well as low-quality bases in each short read sequence. The pre-processed short reads were error corrected using Quake[16]. Then the remaining sequence was assembled using SOAPdenovo[17] with optimal K-mer of each library.

Construction of genetic linkage map and pseudomolecules

High density genetic map of pepper was constructed with 120 recombinant inbred lines (RIL) population derived from an intraspecific cross between *C. annuum* cv. Dempsey and Perenial using SNP markers. Then markers were aligned to the scaffolds using blastn (identity≥98% and coverage≥70%).

Genome annotation

Annotation of the pepper genome was performed using the Pepper Genome

Annotation (PGA) pipeline. The PGA pipeline consisted of repeat masking, mapping of different protein sequence sets and mapping of evidence transcripts with pepper ESTs and putative full-length cDNA generated by TopHat[18] and Cufflinks[19] with RNA-Seq reads. Independent *ab initio* predictions were performed with GENEID[20], AUGUSTUS[21], and GlimmerHMM[22], all specifically trained for pepper. Fgensh is a commercial gene prediction program sold by Softberry. Fgensh gene prediction was carried with a training set for tomato annotation. The EVidenceModeler[23] (EVM) software combines *ab initio* gene predictions with protein and transcript alignments into weighted consensus gene structures. Single automated gene structures were determined from an intuitive framework for combining diverse evidence types and manual curation of these single automated gene structures was performed with the Apollo program[24].

RNA-seq assembly

Putative full-length enriched cDNA pools were constructed with RNA-seq from mixed stages of 4 tissues (flowers, root, leaf and fruit from CM334) using TopHat and Cufflinks. Redundant cDNAs were removed by BLASTn (Ver 2.2.6). A previous study reported that some transcribed regions detected by RNA-seq data were discarded by the PASA pipeline. This problem was caused by noise from individual transcript assemblies, such as artifacts of the sequence alignment process, unspliced intronic pre-mRNA, and genomic DNA contamination[25]. Thus, RNAs

assembled using TopHat and Cufflinks were validated with tomato, potato, and *Arabidopsis* proteins by tBLASTn. Predicted cDNAs from redundancy-removed pools (query) were aligned against the tomato and *Arabidopsis* protein sets (subject). Pepper query sequences whose best hit alignment was at least 50% identical to their best hits and whose length-ratio of subject and query were > 0.70, were used for determination of consensus gene models with EVM. The output file of Cufflinks (gtf file) was converted into GFF3 format with a custom Perl script for EVM analysis

RESULTS

Whole genome sequencing and preprocessing analysis

Pepper genome sequences were generated by Illumina platforms (GA2 and Hiseq 2000). PE libraries of 180, 300 and 400 bp were generated with read length of 2x101 bp, 2x76 bp and 2x101 bp, respectively. In case of MP, 2 kb to 20 kb libraries ranging from 2x36 bp to 2x101 bp were sequenced. Total sequences generated by PE and MP sequencing were 350.3 Gb and 284.6 Gb, respectively (Table 1).

To avoid assembly error, we have constructed a pipeline through three-step processes to filter out the low quality and error candidates sequences of pepper genome before assembly. In the first phase, we removed the sequence reads that matched to the released 1,045 bacterial genomes (identity > 98% and coverage > 50%) using CLC Assembly Cell (CLCBio, Denmark). As a second phase, duplicated reads generated by PCR amplification[26] and low-quality sequences were eliminated by in-house script (Cut-off value: 25 and 20 for PE and MP sequences, respectively). Finally, Quake[16] was performed for error correction through the frequency information derived from k-mer graph of sequences using Jellyfish[27].

Table 1. Hot pepper genome sequence generated in this study

Sequencing data	Insert size	Total length (Gb)	Sequencing depth (X)	Read length (bp)
Illumina reads	180 bp	40.1	11.5	101
	300-400 bp	310.2	89.0	76,101
	2 kb	13	3.7	36
	3 kb	4.2	1.2	54
	4 kb	43.5	12.5	101
	5 kb	14.5	4.2	101
	8 kb	27.8	8.0	101
	10 kb	64.6	18.5	101
	15 kb	43.4	12.5	100
	20 kb	73.6	21.1	101
Total		634.9	182.2	

I performed 19-mer analysis for estimation of genome size. All PE reads (180, 300 and 400 bp) were divided into sliding short sequences of 19 bp, overlapping 18 bp except the first base-pair. Using Jellyfish[27], the frequency of 19-mers in consideration of both strand was calculated. The 19-mer depth distribution was generated and error candidates with low frequencies of 19-mers were decided. The genome size was estimated as 3.48 Gb by the total volume except low frequencies over the peak of distribution. Low frequency data, indicated as red bar in the Figure 1, were also removed for further assembly.

Insert size of each library is a very important prerequisite for accurate scaffolding. To avoid exaggeration of insert distance and confirm quality of raw data before scaffolding, estimation of insert size for PE and MP was accomplished by reference mapping to the initial contigs from SOAPdenovo[17]. As shown in Figure. 2, the calculated insert distance of 180, 300, and 400 bp was shown normal distributions indicating the quality of raw data fulfill for genome assembly. In case of MP, ranging from 2kb to 20kb, distribution of insert size for MP also evaluated the raw data as unbiased (Figure 2).

De novo **genome assembly and validation**

The pre-processed raw sequences were assembled by SOAPdenovo[17]. To make longer initial contigs, paired reads of the smaller left summit of 180 bp library to each single read were merged by FLASH[28]. The longer single reads and remained

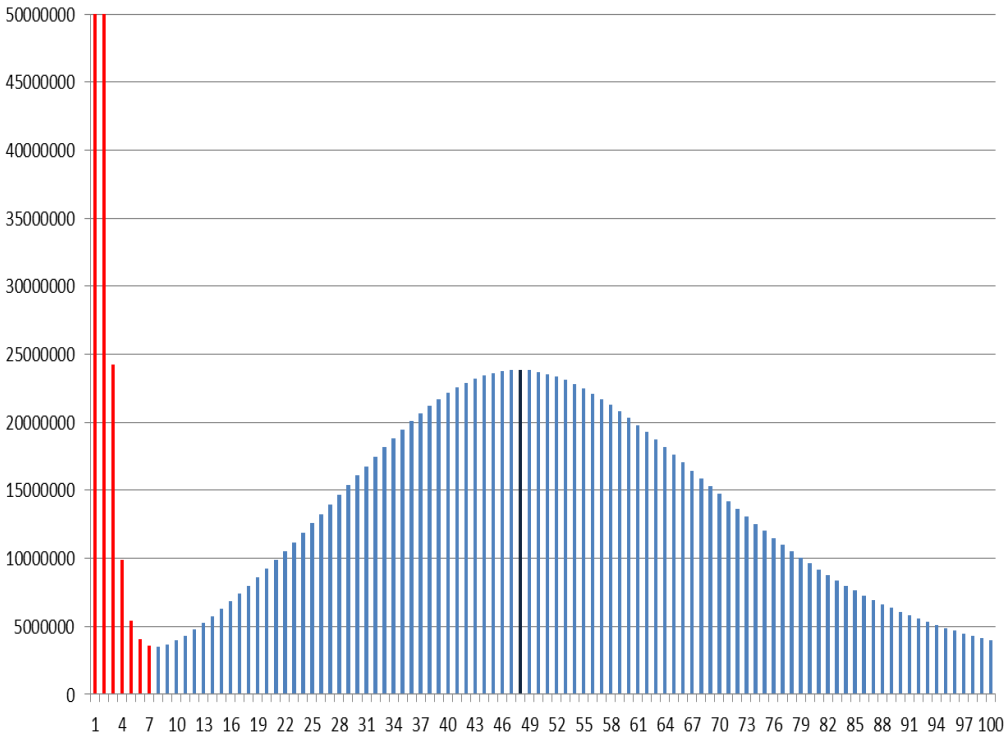


Figure 1. The 19-mer depth distribution for the raw sequence data of pepper genome. The 19-mer depth distribution shows the information for low frequencies, depth of sequencing and degree of heterozygosity. The X-axis indicates frequency of 19-mer and the y-axis means volume of 19-mer. The red bars in low frequency region are error candidates and the black bar representing sequencing depth is the peak of distribution.

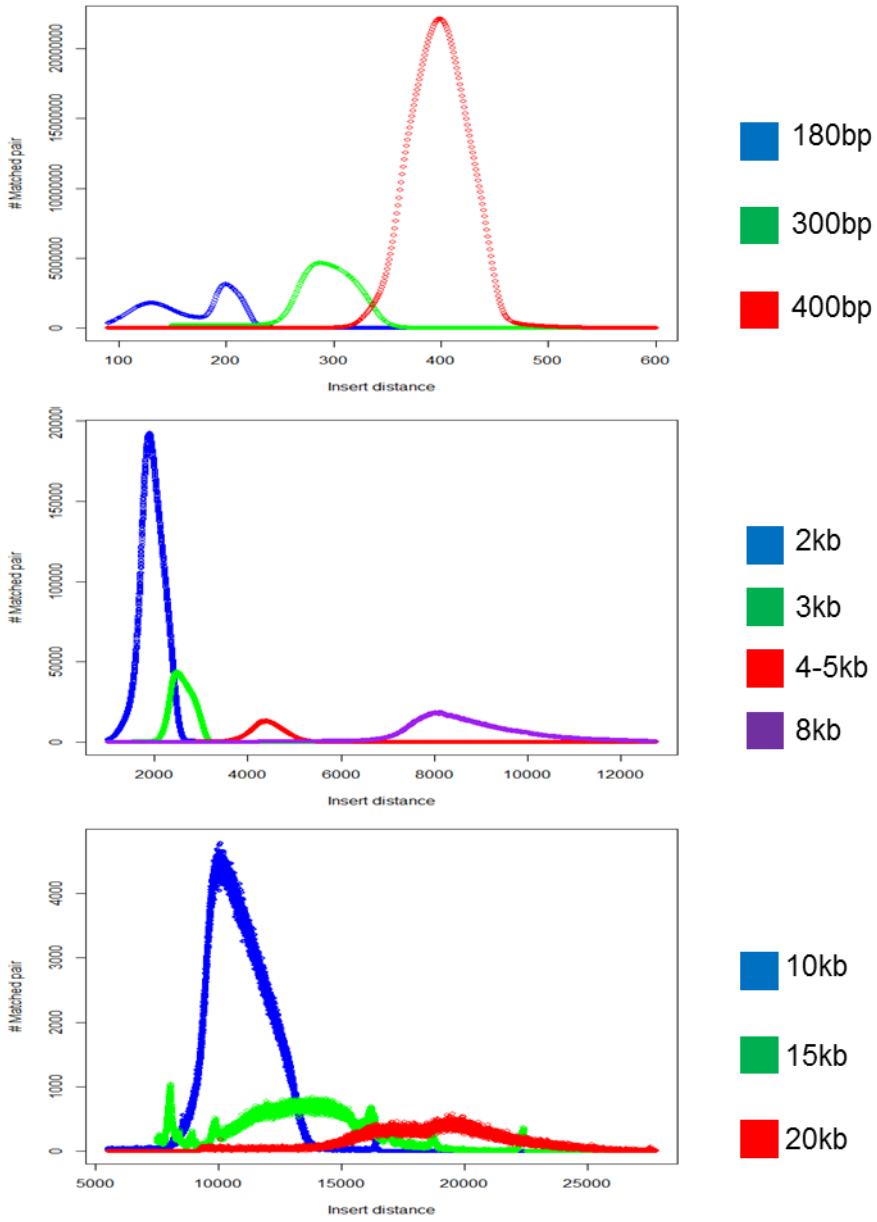


Figure 2. Distribution of insert size for pepper genome.

paired reads of PE were assembled as initial contigs using SOAPdenovo without information of the pair. To search optimized k-mer for scaffolding, k-mer size determination was performed using in-house Perl scripts and 'map' of SOAPdenovo repeatedly. The best k-mer size of each library for scaffolding was decided independently. Using optimized k-mer of each library, scaffolding was carried out serially until reach to the maximum length of scaffolds. In order to make longer scaffolds, serial scaffolding using SSPACE[29] with strict parameters was performed with the scaffolds generated by SOAPdenovo. Finally, gaps in the scaffolds were filled by Gapcloser (http://soap.genomics.org.cn/down/GapCloser_release_2011.tar.gz). Total 3.06 Gb (87.9% of the 3.48-Gb total) were assembled into 37,989 scaffolds (N50=2.47 Mb), and 90% of the genome assembly was in 1,276 scaffolds (Table 2 and Table 3). After removing gap sequences, 2.96 Gb of contig sequences were remained (N50=30kb).

The assembled genome was validated using 27 bacterial artificial chromosome (BAC) single contigs (> 70 kb) generated by ABI and 454 sequencing. BLAST analyses were performed with each BAC clone as a query and scaffolds as subject. The matched regions were selected with a cut-off value of 98% identity and an e-value of 1e-30. Twenty-six BAC sequences were covered by single scaffolds and the assembled scaffolds were matched to all BAC sequences with more than 99.9% identity (Table 3). Although our scaffolds covered most of the regions of BAC

Table 2. Details of CM334 genome assembly

Step	Software	Insert size (kbp)	Raw data coverage (X)	N50 (bp)	Total number	Total size (Gb)
Initial contig	SOAPdenovo			6,369	866,287	2.53
Scaffold 1	SOAPdenovo	0.18-0.4	100.5	39,971	239,579	2.87
Scaffold 2	SOAPdenovo	2.0	3.7	128,039	123,841	2.90
Scaffold 3	SOAPdenovo	3.0	1.2	167,112	104,553	2.91
Scaffold 4	SOAPdenovo	4.0-5.0	16.7	304,994	74,984	2.94
Scaffold 5	SOAPdenovo	8-10	26.5	562,642	68,781	2.99
Scaffold 6	SOAPdenovo	15-20	38.0	1,300,626	65,751	3.04
Scaffold 7	SSPACE	PE + MP		2,472,394	37,989	3.06
Final	All (gap filled)			2,472,394	37,989	3.06

Table 3. Statistics of CM334 genome assembly

	Contig (bp)	Scaffold (bp)
N10	112,552 (1,927th)	7,426,011 (31th)
N20	76,075 (5,182th)	4,993,279 (82th)
N30	55,055 (9,787th)	3,957,205 (151th)
N40	40,567 (16,087th)	3,068,199 (240th)
N50	29,995 (24,618th)	2,472,394 (352th)
N60	21,778 (36,237th)	1,947,990 (491th)
N70	15,269 (52,522th)	1,469,508 (672th)
N80	9,796 (76,680th)	1,083,843 (915th)
N90	4,797 (119,071th)	628,252 (1,276th)
Max / Min	442,125 / 71	18,549,843 / 264
Total Length / Number	2.96 Gb/337,328 ea	3.06 Gb/37,989 ea

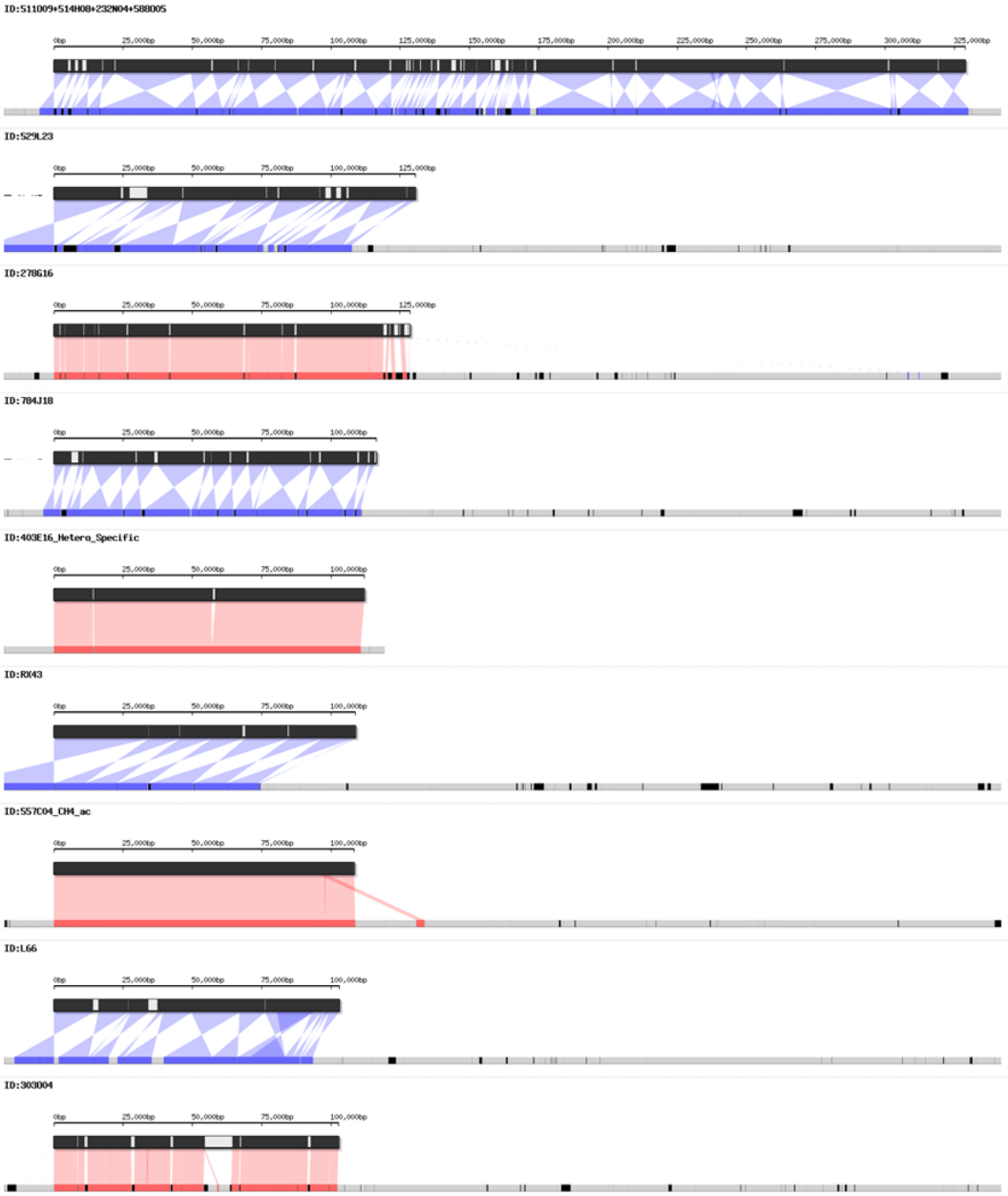
sequences, several low coverage regions were also detected. To confirm the BLAST results for low coverage regions, detailed BLAST results were visualized using an in-house script (Table 4 and Figure 3). As a result, most uncovered regions were confirmed as gap sequences or flanking sequences; however, a few unaligned and inserted regions of the scaffolds were also detected (Figure 3).

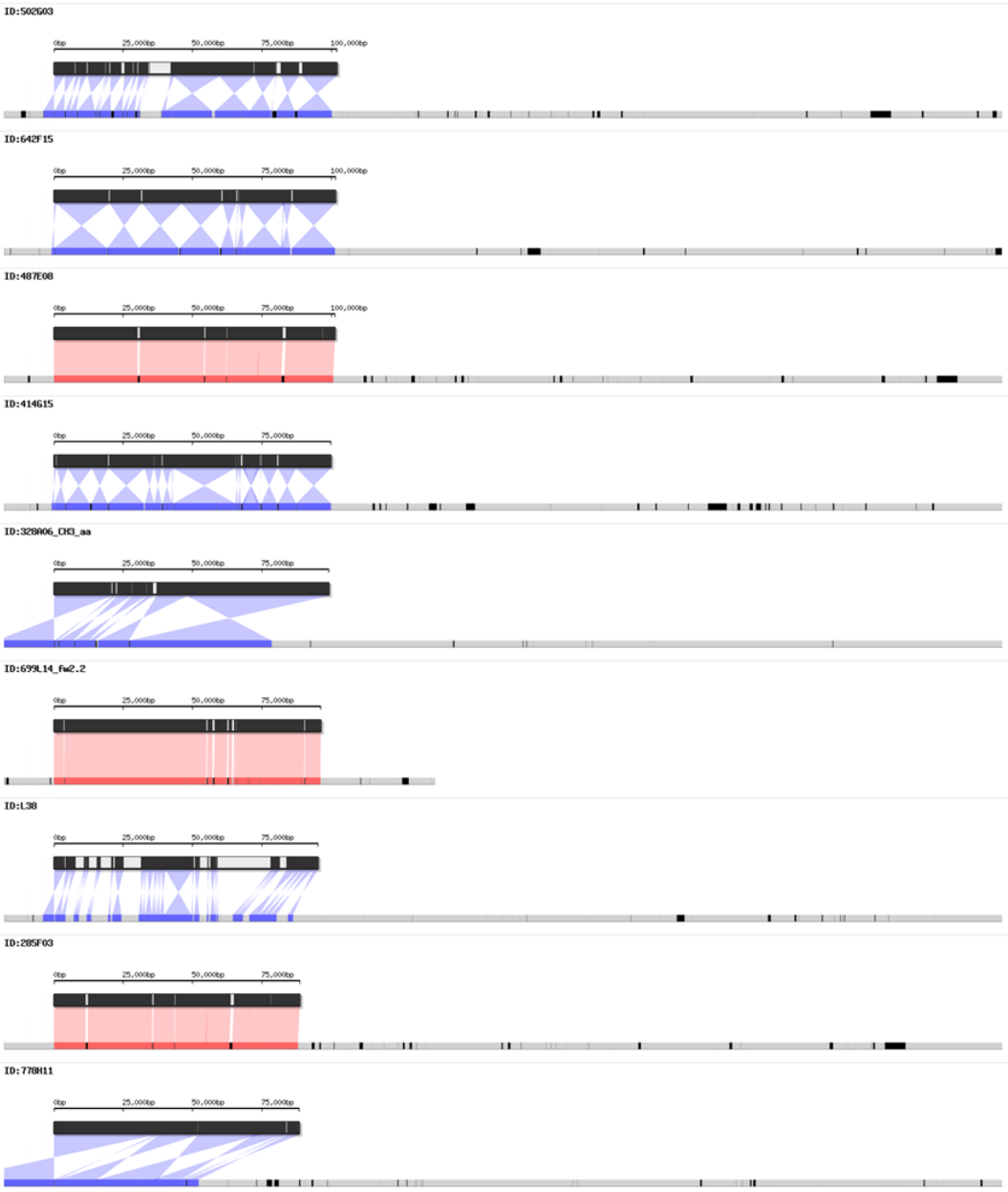
Construction and validation of chromosome pseudomolecules

To construct the chromosome pseudomolecules, a high-density SNP map was generated by low-depth (approximately 1X coverage of the genome) whole genome sequencing (WGS) of 120 individual RILs from a cross between Perennial and Dempsey. WGS data was aligned to contigs of the CM334 genome using the Burrows-Wheeler Aligner program (BWA, ver. 0.6.1-r04)[30]. SNPs were detected using modified default parameters (seed length = 28, maximum differences in the seed =2) and positions of SNPs among aligned reads were identified using SAMtools (ver. 0.1.16) [31]. SNPs were filtered with parameters (mapping quality = 30, minimum read depth = 3 and maximum read depth = 100). A total of 3,146,751 SNPs between CM334 and Dempsey and 3,160,966 SNP loci between CM334 and Perennial were identified. Qualified SNPs among the three cultivars were determined by filtering non-polymorphic SNPs and heterozygote type loci between Dempsey and Perennial. As a result, 1,798,058 SNPs were identified from 267,173 contigs. Finally, SNPs present in the contigs of less than 10 kb were removed and

Table 4. Summary of validation of the CM334 genome assembly using 27 BAC clones

BAC ID	Scaffold ID	BAC length (bp)	Scaffold length (bp)	Coverage (identity) of BAC region (%)
511O09+514H08+232N04 +588O05	809	365,416	1,223,123	96.6 (100)
529L23	497	144,956	1,933,764	90.9 (99.98)
278G16	660	142,903	1,497,426	94.5 (100)
784J18	411	129,419	2,247,124	93.7 (100)
403E16_Hetero_Specific	740	124,468	1,345,326	99.1 (100)
RX43	80	120,973	5,021,183	98.8 (100)
557C04_CH4_ac	2	120,459	16,242,286	100 (99.99)
L66	309	114,328	2,660,594	100 (99.85)
303O04	358	114,092	2,448,025	87.7 (100)
502G03	551	113,462	1785210	88.6 (100)
642F15	112	112,803	4,577,148	99.6 (100)
487E08	650	112,609	1,515,983	97.7 (100)
414G15	890	110,990	1,120,672	98.9 (99.99)
328A06_CH3_aa	428	110,265	2,159,027	98.2 (100)
699L14_fw2.2	836	106,982	1,188,378	97.4 (100)
L38	309	105,856	2,660,594	58.3 (99.97)
285F03	650	98,468	1,515,983	97.4 (100)
778H11	225	98,130	3,150,599	100 (100)
779F05	890	97,798	1,120,672	99.1 (100)
402I14-1_CH1_aa	1,379 2401	93,598	523,271 26,725	72.4 (99.97) 13,1 (99.92)
329G02_KAS	567	91,964	1,738,180	98.8 (99.98)
388I15_CH12_aa	881	91,831	1,124,753	91.2 (100)
409N01_CH11_aa	27	91,232	7,607,120	88.4 (100)
L12	309	90.142	2,660,594	86.5 (99.99)
456G17	809	85.597	1,223,123	93.2 (100)
583K07_CCS	919	80.891	1,081,997	87.7 (99.99)
409M06-1_CH10_aa	867	74.310	1,143,798	100 (99.95)


ID:511009+514H08+232N04+588D05
ID:529L23
ID:278G16
ID:784J18
ID:403E16_Hetero_Specific
ID:RX43
ID:557C04_CM4_ac
ID:L66
ID:303D04


ID:502G03
0bp
25,000bp
50,000bp
75,000bp
100,000bp
ID:642F15
ID:487E08
ID:414G15
ID:320A06_CH3_aa
ID:699L14_fw2.2
ID:L38
ID:285F03
ID:778H11

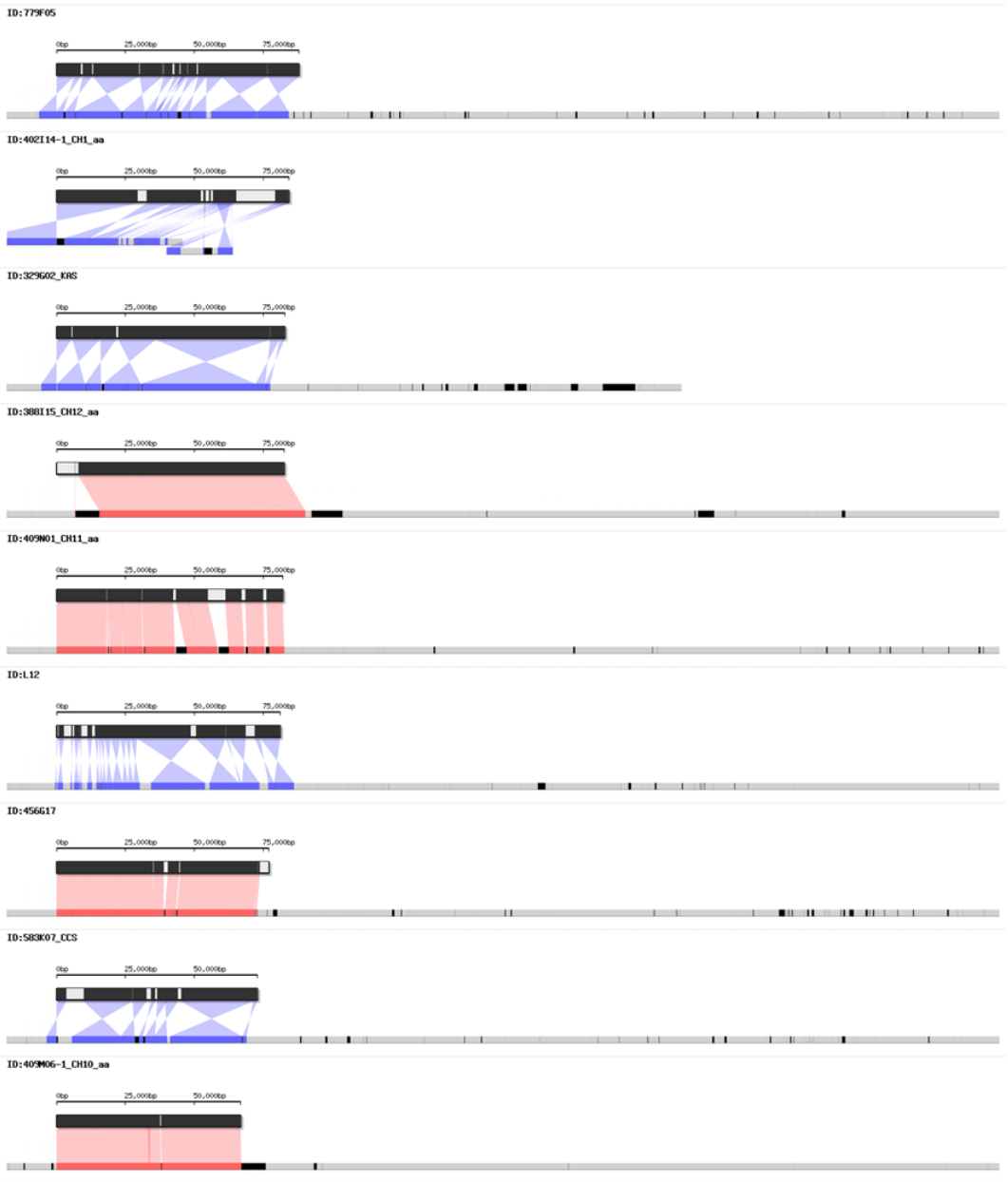


Figure 3. Detailed visualization of validation of CM334 genome assembly against 27 BAC sequences. The upper bars indicate each BAC sequence and the lower bars represent scaffold(s). In BAC sequences, black and white indicate regions matched and unmatched to the scaffold, respectively. In scaffolds, red and blue shading indicate regions matched to the plus and minus strands, respectively. Black intervals represent gap sequences and gray represents unmatched regions of scaffold.

771,587 SNP markers, representing 40,727 CM334 contigs, were used for map construction.

Of 40,272 genetic markers from 40,727 contigs, low-quality markers that contain more than 10% of missing genotype information (10 individuals of 120 RILs) were also removed. Finally, the remaining 21,121 markers were divided into 20 subgroups for genetic map construction (Table 5). To combine each individual genetic map, 180 markers were randomly selected and added to each subgroup, resulting in 1,180 markers in each subgroup. Genetic maps were constructed using Joinmap4 and each genetic map was composed of 11-14 linkage groups (Table 5). To construct a high-density genetic map, 12 identical linkage groups in each genetic map were selected from 20 subgroups based on assignments of 180 commonly used markers in each genetic map. Each linkage group was constructed using selected markers from 12 identical linkage groups, and a high-density genetic linkage map was constructed by combining the individual linkage groups. Finally, a 6,281 marker-containing genetic map, covering 3,796 cM in Kosambi function, was constructed (Table 6).

The pepper chromosome pseudomolecules were built by anchoring scaffolds to a newly constructed high-density genetic map. The pepper genetic map constructed earlier[32] was used to increase the accuracy of the pseudomolecule. Markers of the new genetic map were assigned by BLAST to scaffolds with cut-off values of 98% and 80% coverage. The assigned positions of markers were validated

Table 5. Summary of linkage groups from subdivided markers groups

Subgroup	# of linkage groups	# of markers in linkage group
subgroup01	11	1,092
subgroup02	13	1,046
subgroup03	12	1,042
subgroup04	12	1,057
subgroup05	14	1,065
subgroup06	12	1,053
subgroup07	12	1,022
subgroup08	13	1,079
subgroup09	12	992
subgroup10	12	1,008
subgroup11	13	1,059
subgroup12	13	1,093
subgroup13	12	1,065
subgroup14	12	937
subgroup15	13	1,077
subgroup16	13	1,126
subgroup17	13	1,064
subgroup18	14	1,105
subgroup19	13	1,096
subgroup20	12	1,043
Total		21,121

Table 6. Summary of the integrated 12 linkage groups

Chromosome	# markers	# selected markers	Genetic distance	# recombination breakpoint
1	1,620	527	468	532
2	1,485	509	493	343
3	1,821	593	375	630
4	1,504	560	427	315
5	1,455	531	254	294
6	2,000	557	223	286
7	1,559	538	287	258
8	540	134	75	54
9	2,143	536	228	242
10	2,826	676	204	224
11	2,652	596	317	350
12	1,516	524	439	333
Total	21,121	6,281	3,796	3,861

using markers of the earlier genetic map. A total of 4,562 markers from the newly constructed genetic map satisfied the cut-off value, and we were able to anchor 86.0% of scaffolds (2.63 Gb; 1,357 scaffolds) to 12 chromosome pseudomolecules (Figure 4 and Table 7). The direction of scaffolds in 12 chromosomes was determined by considering the ratio of markers shown in identical order in genetic and physical maps using an in-house Perl script. The majority of scaffolds (75.6%) were assigned a direction, and we further assigned the direction of 36 more scaffolds by manual alignment of the unique markers generated from the earlier genetic map.

Accuracy of the linkage map was a crucial factor in constructing pseudomolecules. Validation of the genetic map was carried out by checking map reproducibility and by comparison with a previously constructed conserved ortholog set (COS) map [33]. Total 360 COS markers were used for validation except 10 missed markers and 257 markers were detected in assembled CM334 genome (90% identity, 50% coverage and over 30 bp). As a result, 58 out of 257 markers were appeared in chromosome *0* and 196 out of 199 markers were correctly assigned in 12 chromosomes compared to genetic information (Table 8).

Repeat annotation

Prior to gene prediction, all transposable element-related repeats were masked using REPEATMASKER by masking the genome sequences with the custom library of pepper. The results showed that pepper genome contain approximately

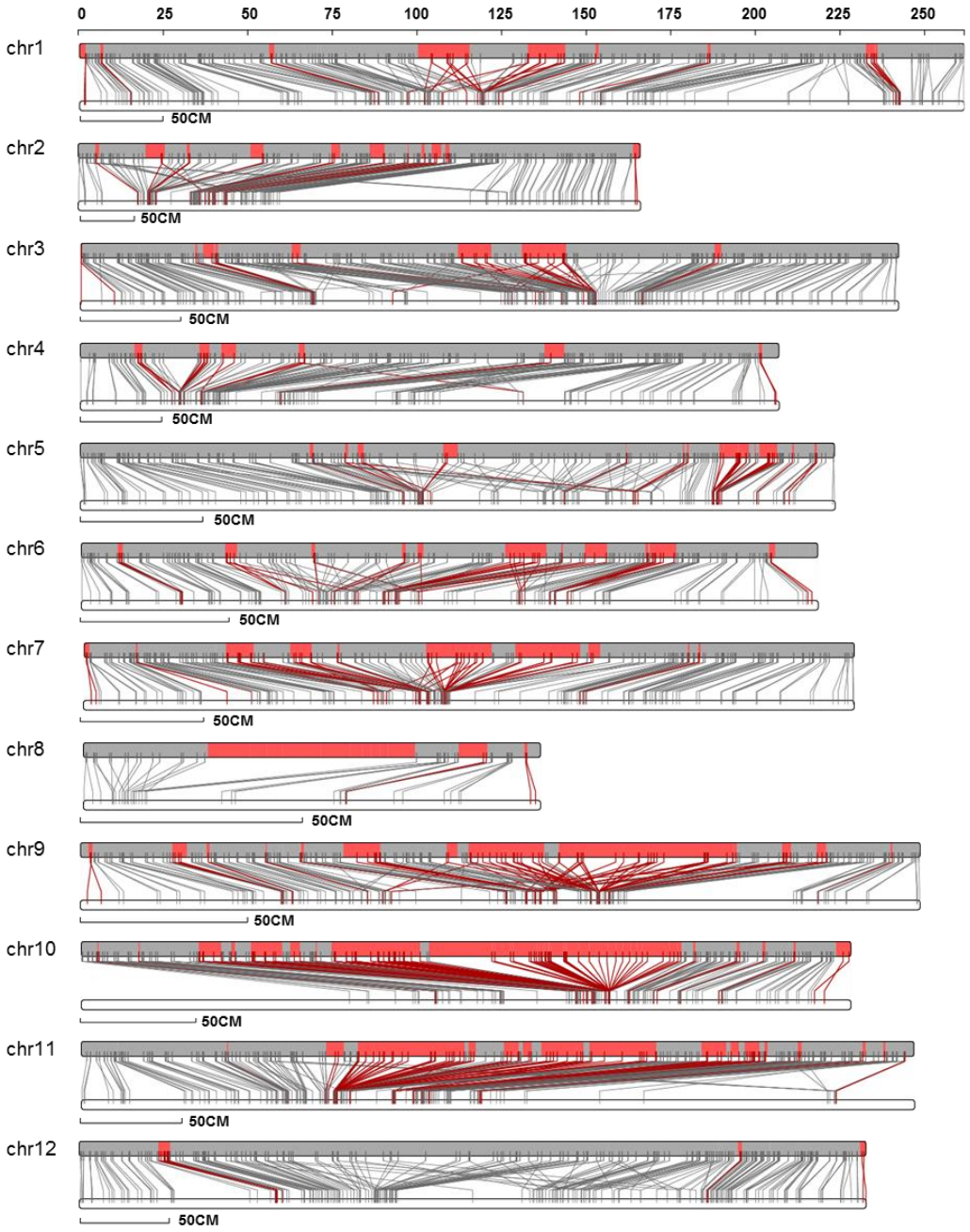


Figure 4. Genetic and physical maps of CM334 genome. The upper bars indicate pseudomolecules, and the lower bars represent linkage groups corresponding to 12 chromosomes. The grey blocks and lines indicate ordered regions and markers, and the red blocks and lines represent unordered but matched scaffolds to the genetic map.

Table 7. Details of the 12 chromosome pseudomolecules

Chromosome	# anchored scaffolds	Length of pseudomolecule (bp)	# markers	# ordered scaffolds (coverage)
1	138	261,560,226	412	120 (86.3%)
2	99	166,118,313	374	85 (84.4%)
3	147	241,745,451	472	130 (86.5%)
4	91	206,470,299	417	81 (91.3%)
5	113	223,151,943	413	91 (89.0%)
6	104	217,864,955	424	83 (82.5%)
7	111	227,551,634	420	85 (73.1%)
8	35	134,909,690	142	21 (47.7%)
9	127	247,983,219	351	81 (58.4%)
10	137	227,301,773	305	65 (43.5%)
11	129	246,428,986	374	89 (60.7%)
12	126	232,591,935	458	117 (97.3%)
Total	1,357	2,633,678,424	4,562	1,048 (75.6%)

Table 8. Validation of pseudomolecules using COS markers

Chromosome	# total markers	# markers in same (different) chromosome	# unassigned markers in chromosome	# unassigned markers in scaffold
1	63	28 (1)	8	23
2	40	31 (0)	0	8
3	38	22 (0)	5	10
4	31	10 (0)	10	10
5	23	10 (1)	9	4
6	27	10 (0)	8	9
7	30	23 (0)	4	4
8	5	7 (1)	1	1
9	21	10 (0)	4	7
10	20	8 (0)	2	10
11	29	15 (0)	5	9
12	33	23 (0)	2	8
Total	360	196 (3)	58	103

2.34 Gb (76.4%) of TEs and Table 9 reports detail information of transposable elements (TEs) in pepper genome.

Transposable elements (TEs) have multiple roles in driving genome evolution in eukaryotes[34]. A total of 2.34 Gb (76.4%) of sequences in the assembled CM334 genome was identified as TEs (Table 9). The major type of TE was long terminal repeats (LTRs), which represented approximately 1.7 Gb (more than 70%) of the total number of TEs in the pepper genome. Most of the LTRs were Gypsy elements, which accounted for 67.0% of TEs in CM334, respectively. A large number of Caulimoviridae elements were unique in pepper genome (Table 9). The TEs were widely dispersed throughout the pepper genome and often led to the conversion of euchromatin to heterochromatin. The distribution of the TEs was inversely correlated with the density of genes (Figure 5).

Structural gene annotation

Some gene families containing repetitive sequences (e.g., leucine-rich-repeat (LRR)) sequences such as NB-LRR resistance genes, receptor-like proteins, and receptor-like kinases, had a probability of being recognized as repeat sequences and masked by RepeatMasker. tBLASTn analyses were carried out to prevent masking of genic sequences by repeat masking. Before repeat masking, tBLASTn analyses were carried with protein-coding genes from eight genomes (*Arabidopsis thaliana, Solanum lycopersicum, Solanum phureja, Nicotiana benthamiana, Ricinus*

Table 9. Classification summary of transposable elements (TEs) of pepper genome

Type	Length (bp)	Coverage in repeat (%)	Coverage in genome (%)
DNA/En-Spm	7,415,218	0.31	0.25
DNA/MuDR	65,925,363	2.74	2.19
DNA/hAT-Ac	7,793,596	0.32	0.26
DNA/hAT-Tip100	1,235,551	0.05	0.04
LINE/L1	26,983,050	1.12	0.89
LINE/RTE-BovB	29,369,266	1.22	0.97
LTR/Caulimovirus	28,190,570	1.17	0.93
LTR/Copia	157,291,528	6.55	5.21
LTR/Gypsy	1,116,409,907	46.48	37.01
Low_complexity	3,034,446	0.13	0.1
RC/Helitron	1,888,884	0.08	0.06
Simple_repeat	17,919,823	0.75	0.59
Unknown	938,500,119	39.07	31.11
Total	2,401,957,321	100	79.62

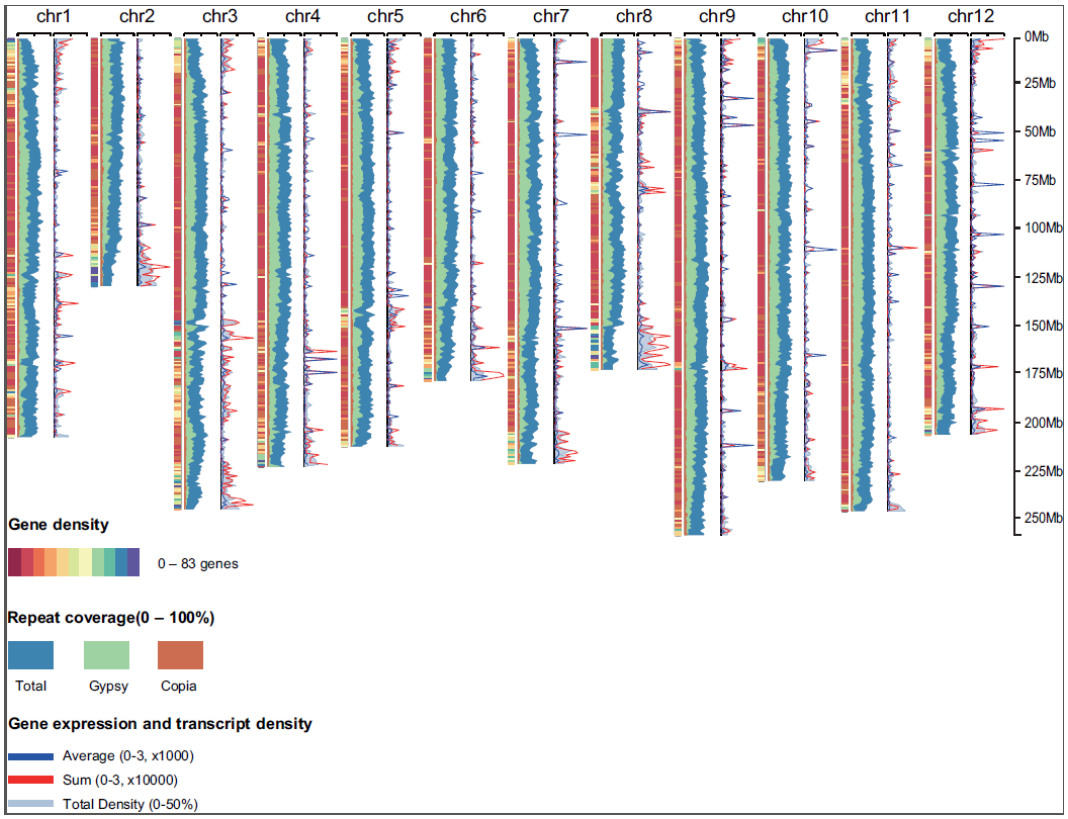


Figure 5. Global overview of the pepper genome. The global overview of the pepper genome shows the gene density, coverage of repeats, expression of genes and transcript density at intervals of 1 Mb. Ranging from 0 to 83 genes, and with a genome-wide distribution, gene-rich regions and gene-poor regions are discriminated via colour coding in the 12 chromosomes. The coverage of the repeats represents the proportion of the total TEs and of the Gypsy and Copia elements of LTRs. Gene expression values for total transcriptome except placenta tissues are calculated by the sum of the RPKM (Reads per kb per million reads) and average of RPKM. The transcript density represents the coverage of the total transcriptome, ranging from 0 to 50%.

communis, Vitis vinifera, Glycine max and *Populus trichocarpa*) to detect putative genic regions. The genome sequences of protein-matched regions was saved as genic regions and recovered as their original sequences instead of being masked. The fraction of the pepper repeat-masked genome with the custom library was 63.28%. No simple repeats, rRNA, tRNA, or microsatellites were masked. This masked genome sequence was used for the pepper genome annotation.

The protein sequences of tomato (iTAG 2.3), potato (PGSC 3.4), grape (*Vitis vinifera* ver 2.0), *Arabidopsis* (TAIR10), tobacco (*Nicotiana benthamiana*, ver 0.4.4), castor bean (*Ricinus communis*, ver 0.1), soybean (*Glycine max*, Gmax 109), poplar (*Populus trichocarpa*, Ptrichocarpa 156), and pepper were mapped using Genewise[35] to generate protein-based gene models to determine a consensus gene model. Genewise allows mapping of proteins taking intron-exon boundaries into account. Although the Genewise-predicted pepper gene model was more precise than *ab initio* programs, Genewise had problems predicting the duplicated genes in pepper the genome. For the annotation of duplicated genes or families of genes, mapping regions of a reference protein in the pepper genome were determined from tBLASTn results using custom Perl scripts. These steps prevented mis-annotation of duplicated or family genes by lacking mapping data of evidence or reference proteins from parsing of single best-matched region from the tBLASTn results in the pepper genome. The Genewise output was reformatted into GFF3 and used as highly reliable data for determining consensus gene model using EVM.

AUGUSTUS[21], GENEID[20], GlimmerHMM[22], and Fgenesh were used for gene prediction in pepper. AUGUSTUS was trained on the pepper genome (release 1.0) using assembled RNA sequences from various tissues (Table 10) and full-length cDNA (GenBank, March 2012). As for preparing the training set for *ab initio* programs, only predicted pepper cDNAs from TopHat[18] and Cufflinks[19] whose best hit alignment was at least 50% identical to their best hits and whose length-ratio of subject and query was ≥0.90 were retained as possible full-length pepper proteins. Among these cDNAs, 2000 models were selected and used for generating training sets for gene prediction with GeneID and GlimmerHMM (Table 11).

EVM was used to integrate the various gene models from assembled transcript data, protein data, and *ab initio* data for determining consensus gene model. To run the EVM, the transcript data, protein data, and *ab initio* data were converted to EVM-compatible GFF3 format. For conversion, scripts provided by EVM or in-house Perl scripts were used. Genome sequences were then partitioned into smaller data sets to reduce memory requirements, and the sizes of segments were less than 1 Mb. Overlap sizes of partitioned segments were determined using PGA 0.9 version and determined as at least two standard deviations greater than expected gene length. EVM was then run on each of the data partitions, and results of corresponding to partitioned segments were joined into single outputs. Raw output provided by EVM describes the consensus gene structures in a tab-delimited format, listing each exon with the evidence sets that fully support each exon structure.

Table 10. Properties of transcriptome used for gene annotation

Tissue	Read type	Total length (Gb)	Read length (bp)
Leaf	Single	2,2	73
Stem	Single, Paired	9.4	73,121
Root	Paired	7.1	101
Fruit (6 DPA)	Single	1.9	73
Fruit (16 DPA)	Single	1.5	73
Fruit (25 DPA)	Single	1.8	73
Fruit (MG)	Single, Paired	8.7	73,101
Fruit (B)	Single, Paired	10.3	73,121
Fruit (B5)	Single	1.3	73
Fruit (B10)	Single	1.6	73
Flower	Single	1.5	73
Total	Single, Pair	45.1	

Table 11. Training set for *ab initio* gene prediction

Average transcript length	1,159 bp
Median transcript length	1,106 bp
Average coding length	886 bp
Median coding length	540 bp
Total length of transcript sequence	1,457,666 bp
Total length of coding sequence	1,279,640 bp
Total exons	1,443 (1.14/gene)
Average exon length	818bp
Median exon length	540 bp
Number of initial exons	256
Average length of initial exons	318 bp
Number of internal exons	748 bp
Average length of internal exon	125 bp
Average length of terminal exons	231 bp

Finally, this raw output was converted to standard GFF3 format. Consensus gene models were validated with tomato and potato proteins (Figure 6). The putative gene models showing longer or shorter gene length compared to tomato or potato genes were validated with the NCBI non-redundant database (NR-DB).

Consensus gene models determined by automatic annotation pipelines mis-annotated genes. To correct these mis-annotated genes, the genomic annotation viewer and editor Apollo (http://apollo.berkeleybop.org/current/userguide.html) was used for error correction. The Apollo program visualizes each gene model used for automatic annotation, such as TopHat/Cufflinks, Genewise gene models from eight genomes, and *ab initio* gene models, and the user can edit the consensus gene model. Missing or mis-annotated genes were selected from the characterization of each gene family and curated using Apollo. Of the original 35,258 predicted genes, approximately 86 genes were reclassified as pseudogenes and approximately 1,789 genes were discarded in favor of better gene models built by the expert. These manually curated genes were transferred to PGA 1.5 annotation, and the current status reports 34,903 predicted genes, of which 1,789 (3.8%) genes underwent an expert intervention to correct erratic/incomplete gene models (Table 12).

Functional annotation of protein coding genes

To infer functions for the protein-coding genes, we used INTERPROSCAN version 4.8[36] to scan protein sequences against the protein signatures from InterPro

Table 12. Metrics of pepper gene models

	Protein coding Loci (no.)	Total CDS length (bp)	Avg CDS length (bp)	Avg exon length (bp)	Avg intron length (bp)
Pepper[a]	34,903	35,247,975	1,009.9	286	541
Tomato[a]	34,771	35,972,459	1,057	179	533
Potato[b]	35,004	31,678,620	905	356	592
Arabidopsis[c]	27,206	24,861,465	1,212	265	164
Rice[d]	28,236	78,281,992	1,081	312	414

[a]The ITAG pre-2.3 pre-release data were used .
[b]Only protein-coding nuclear transcripts were used.
[c]All protein-coding transcripts were included, with the exception of TEs and pseudogenes.
[d]All protein-coding transcripts (MSU Release 6.3) were included, with the exception of TEs, pseudogenes, organellar insertions, and small genes.

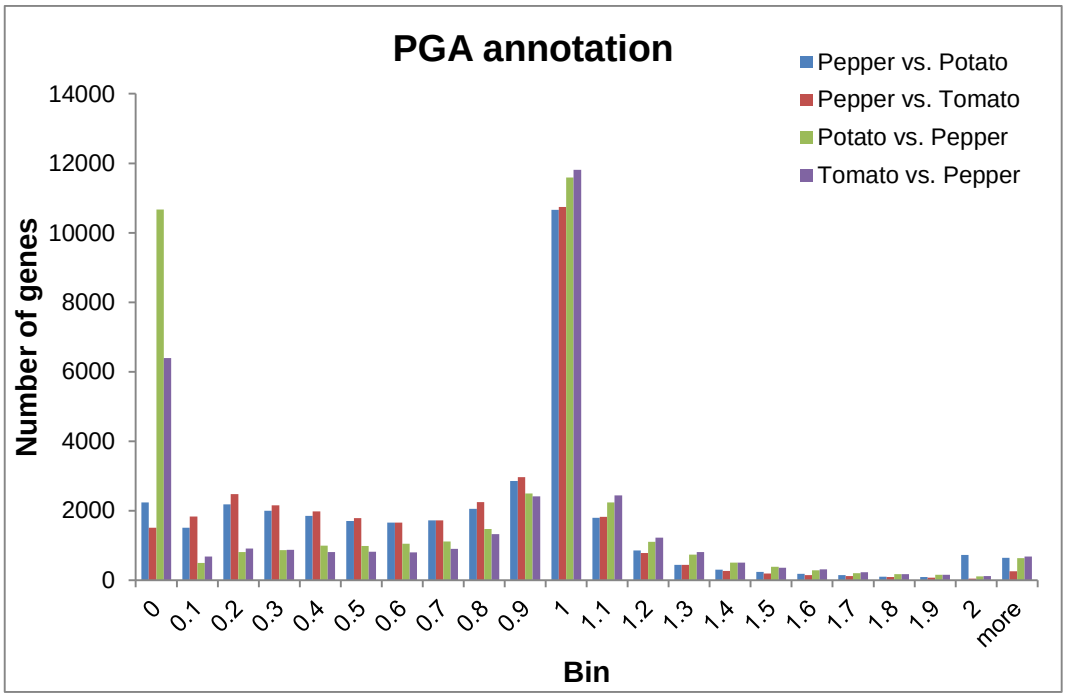


Figure 6. Histogram of pairwise comparison of protein length ratios with their best blast hit relative to the reference pepper and tomato versus potato. Proteins with a query coverage ratio of 0.9–1.1 can be considered highly confident predictions (43%) because the protein sequences match at the sequence level as well as at the protein length level. Bins of ratios <0.9 are proteins predicted to be shorter than their tomato (iTAG 2.3) reference, while bins of ratios >1.1 indicate proteins predicted to be longer. The zero-bin indicates the number of genes not reporting a hit to the reference given the threshold used. This bin will collect all real species-specific genes and predicted small genes, but also prediction artifacts (false positives = overprediction) or gene models resulting from genome/assembly issues.

(version 31.0). InterPro integrates protein families, domains and functional sites from different databases, such as Pfam (Ver 24.0), PROSITE (Ver 20.66), PRINTS (Ver 41.1), ProDom (Ver 2006.1), SMART (Ver 6.1), TIGRFAMs (Ver 9.0), PIRSF (Ver 2.74), SUPERFAMILY (Ver 1.73), Gene3D (Ver 3.3.0), and PANTHER (Ver 7.0). INTERPROSCAN integrates the search algorithms of all these databases. INTERPROSCAN identified 193,478 protein domains of 5,544 distinct domain types. Ninety-one percent of the genes (31,566 of a total of 34,903 genes) have been assigned at least one domain. The top 20 SUPERFAMILY domains are plotted in Figure 7.

AGF used similarity searches and lexical analysis for Assignment of Gene Function to protein sequences. It utilized (1) BLASTP[37] search results against the Swissprot[38], TAIR[39] and TrEMBL[38] databases and (2) domain search results from INTERPROSCAN. Query genes matching to genes associated with terms such as “hypothetical protein”, “similarity to” or “predicted protein” were assigned to gene function “unknown”. The highest-scoring description was assigned to the query, and the database accession of the hit protein was added to enable evidence tracking. If INTERPROSCAN results were available, the domain names were extracted and appended to the description line. In case of multiple InterPro matches, the InterPro Parent-Child-tree was used to reduce the number of reported domains by selecting the most specific child and excluding its parents. The descriptions contain, in brackets between the transferred AGF and InterPro domain results.

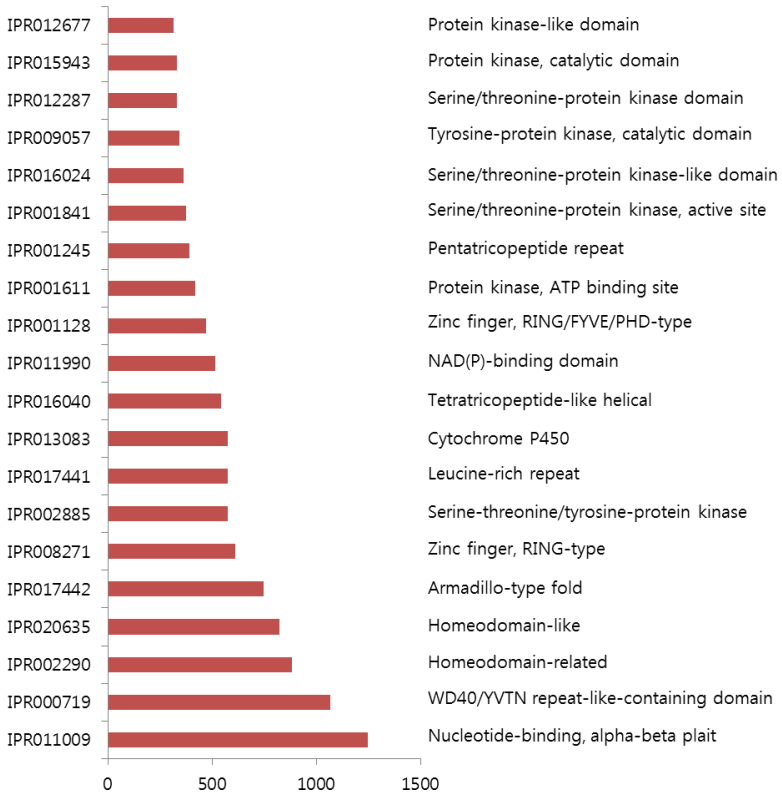


Figure 7. The top 20 INTERPRO domains in PGA version.1.5.

DISCUSSION

In 2011, the world production value of pepper was 14.4 billion US dollars which was 40 times increased since 1980[5]. Pepper consumption is keep growing every year because of its high nutritional values for human beings. Despite of its increased value, lack of a reference genome sequence has obstructed genomic or genetic research of pepper. In this study, I constructed *de novo* genome sequence of the pepper as a first reference pepper genome through various processes for preprocessing analysis, assembly, validation and annotation.

Prior to genome assembly, extraction of accurate sequences in huge amount of NGS data is an essential process to minimize assembly error. I performed preprocessing analysis consisted of three steps (1) elimination of prokaryotic genome sequences in the generated raw data, (2) removal of low quality and duplicated reads, and (3) deletion of low frequency region (Figure 8). After preprocessing, as preparation of optimal data for genome assembly and scaffolding, insert sizes of all PE and MP libraries were estimated to construct accurate genomes. Additionally, reads of PE were merged to single reads to make longer initial contigs. Resulting from preprocessing analysis and preparation of optimal data, purified and optimized sequences were prepared for genome assembly.

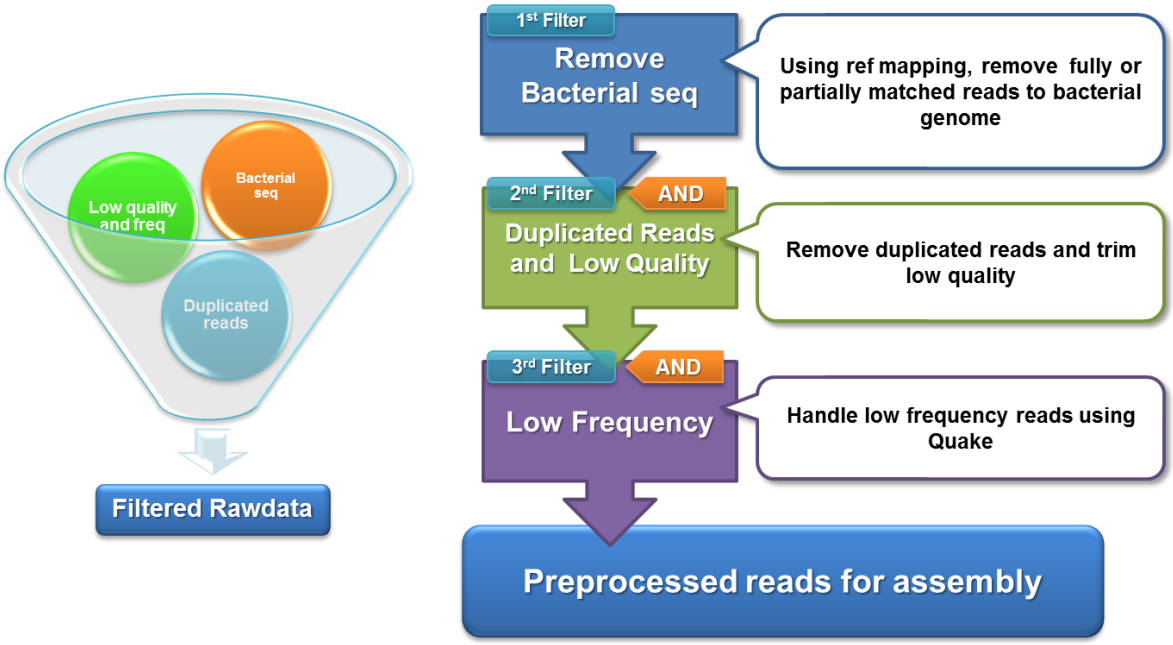


Figure 8. Scheme of raw data processing pipeline.

Until now, general process of assembly and scaffolding have been performed using single assembly tool[40-44]. In contrast, pepper genome assembly and scaffolding were conducted using firstly SOAPdenovo and then additional scaffolding was carried out using SSPACE to make longer scaffolds. In addition, due to differential account of each library, I performed iterative mapping for each library before scaffolding to obtain optimal values of k-mer. Using the optimal k-mers for each library, serial scaffolding was conducted. As results, the serial scaffolding process effectively contributed to extend length of scaffolds (Table 2). Furthermore, The genome assembly was validated using 27 bacterial artificial chromosome (BAC) sequences from CM334; 26 BAC sequences were fully covered by each scaffold(s), and showed identities of more than 99.9% (Figure 3 and Table 4).

A distinct difference between completed and draft genome is that whether pseudomolecule construction was performed or not. To construct pseudomolecules, the scaffolds were anchored to a SNP map that was constructed by the sequencing of parental lines and the low depth sequencing of 120 recombinant inbred lines derived from *C. annuum* cv. Dempsey and *C. annuum* cv. Perennial. We anchored 86.0% (2.63 Gb, 1,357 scaffolds) of the assembly as 12 chromosome pseudomolecules and ordered them (75.6%) on the basis of genetic distance (Figure x and Table x). Comparing to a previously constructed conserved ortholog set (COS) map[33], the validation result revealed that the genetic map correctly assigned the

scaffolds in 12 chromosomes (Table 8).

Although sequencing has become easy, genome-wide annotation of gene structures has become more challengeable due to inaccurate genome assembly generated by short read sequencing[45]. Based on the high quality pepper genome, a total of 34,903 protein-coding genes were predicted in the PGA pipeline (Pepper Genome Annotation). This gene number is approximately the same as that for tomato (iTAG v2.3, 34,771)[46] and potato (PGSC v3.4, 39,031)[47], which suggests a similar gene number in Solanaceae plants. Overall, 93.2% of the predicted coding sequences were supported by Illumina data, demonstrating the high accuracy of the PGA gene prediction. To validate and improve the gene models, the inaccurately annotated genes were manually curated; 335 genes were manually added and 86 genes were reclassified as pseudogenes. This manual inspection and curation resulted in the replacement of 1,789 genes with better gene models.

Consequently, the pepper genome sequence can serve as an important genomic resource for improving nutritional and pharmaceutical values of pepper as well as support the evolutionary and comparative genomic studies of one of the world-most diversified plant family Solanaceae. The pepper genome sequence provides an excellent tool for exploring the evolution of secondary metabolites in plants. Combined with the recently published tomato[46] and potato genomes[44] , the pepper genome will elucidate the evolution, diversification, and adaptation of more than

3,000 Solanaceae species, which are adapted to a wide range of geo-ecological habitats ranging from the driest deserts to tropical rainforests.

REFERENCES

1. Marin, A., Ferreres, F., Tomas-Barberan, F.A. & Gil, M.I. Characterization and quantitation of antioxidant constituents of sweet pepper (*Capsicum annuum* L.). *J. Agric. Food Chem.* **52**, 3861-9 (2004).
2. Matsufuji, H., Ishikawa, K., Nunomura, O., Chino, M. & Takeda, M. Oxidant content of different coloured sweet peppers, white, green, yellow, orange and red (*Capsicum annuum* L.). *Int. J. Food Sci. Technol* **42**, 1482-1488 (2007).
3. Matus, Z., Deli, J. & Szabolcs, J.J. Carotenoid composition of yellow pepper during ripening-isolation of beta-cryptoxanthin 5, 6-epoxide. *J. Agric. Food Chem.* **39**, 1907-1914 (1991).
4. Mejia, L.A., Hudson, E., deMejia, E.G. & Vazquez, F. Carotenoid content and vitamin-a activity of some common cultivars of Mexican peppers (*Capsicum annuum*) as determined by HPLC. *J. Food Sci.* **53**, 1448-1451 (1998).
5. FAO. Food and Agricultural Commodities Production. FAO, Rome http://faostat.fao.org (2013).
6. Caterina, M.J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816-824 (1997).
7. Bosland, P.W.E.J.V.P. *Vegetable and Spice Capsicums* (CABI, Wallingford, UK, 2012).
8. Mori, A. *et al.* Capsaicin, a component of red peppers, inhibits the growth of androgen-independent, p53 mutant prostate cancer cells. *Cancer Res.* **66**, 3222-3229 (2006).
9. Surh, Y.J. More than spice: capsaicin in hot chili peppers makes tumor cells commit suicide. *J. Natl. Cancer Inst.* **94**, 1263-1265 (2002).
10. Ito, K. *et al.* Induction of apoptosis in leukemic cells by homovanillic acid derivative, capsaicin, through oxidative stress: implication of phosphorylation of p53 at Ser-15 residue by reactive oxygen species. *Cancer Res.* **64**, 1071-1078 (2004).

11. Fraenkel, L., Bogardus, S.T., Jr., Concato, J. & Wittink, D.R. Treatment options in knee osteoarthritis: the patient's perspective. *Arch. Intern. Med.* **164**, 1299-1304 (2004).
12. Lejeune, M.P., Kovacs, E.M. & Westerterp-Plantenga, M.S. Effect of capsaicin on substrate oxidation and weight maintenance after modest body-weight loss in human subjects. *Br. J. Nutr.* **90**, 651-659 (2003).
13. Westerterp-Plantenga, M.S., Smeets, A. & Lejeune, M.P.G. Sensory and gastrointestinal satiety effects of capsaicin on food intake. *Int. J. Obes. (Lond.)* **29**, 682-688 (2004).
14. Ludy, M.-J., Moore, G.E. & Mattes, R.D. The effects of capsaicin and capsiate on energy balance: critical review and meta-analyses of studies in humans. *Chem. Senses* **37**, 103-121 (2012).
15. Pegard, A. *et al.* Histological Characterization of Resistance to Different Root-Knot Nematode Species Related to Phenolics Accumulation in Capsicum annuum. *Phytopathology* **95**, 158-165 (2005).
16. Kelley, D.R., Schatz, M.C. & Salzberg, S.L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **11**, R116 (2010).
17. Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311-317 (2010).
18. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
19. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protoc.* **7**, 562-578 (2012).
20. Blanco, E., Parra, G. & Guigo, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics* **Chapter 4,** Unit 4-3, (2007).
21. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435-439 (2006).
22. Allen, J.E., Majoros, W.H., Pertea, M. & Salzberg, S.L. JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biol.* **7 Suppl 1**, S9 1-13 (2006).

23. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
24. Lewis, S.E. *et al.* Apollo: a sequence annotation editor. *Genome Biol.* **3**, RESEARCH0082 (2002).
25. Xu, Q. *et al.* The draft genome of sweet orange (Citrus sinensis). *Nat. Genet.* **45**, 59-66 (2013).
26. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6**, 291-295 (2009).
27. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
28. Schatz, M.C., Witkowski, J. & McCombie, W.R. Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.* **13**, 243 (2012).
29. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579 (2011).
30. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
31. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
32. Yarnes, S.C. *et al.* Identification of QTLs for capsaicinoids, fruit quality, and plant architecture-related traits in an interspecific Capsicum RIL population. *Genome* **56**, 61-74 (2013).
33. Wu, F. *et al.* A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus Capsicum. *Theor. Appl. Genet.* **118**, 1279-1293 (2009).
34. Feschotte, C., Jiang, N. & Wessler, S.R. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329-341 (2002).
35. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988-995 (2004).

36. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.*, 59-70 (2007).
37. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).
38. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-370 (2003).
39. Huala, E. *et al.* The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* **29**, 102-105 (2001).
40. Guo, S.G. *et al.* The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**, 51-58 (2013).
41. Chen, J. *et al.* Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. *Nat. Commun.* **4**, 1595 (2013).
42. D'Hont, A. *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213-217 (2012).
43. Wang, X. *et al.* The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet.* **43**, 1035-1039 (2011).
44. The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-195 (2011).
45. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329-342 (2012).
46. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641 (2012).

CHAPTER 2

Multiple reference pepper genome sequencing and evolutionary history of the genus *Capsicum*

ABSTRACT

A diversity of *Capsicum* species has been derived from their evolutionary process. Here, I report high-quality *de novo* genome sequences of two *Capsicum* species (*C. chinense* and *C. baccatum*) and comparative analysis with the existing pepper reference genome (*C. annuum*). A total of 95 % (3.07 Gb) of 3.21 Gb of *C. chinense* genome and 76 % (3.22 Gb of 4.2 Gb) of *C. baccatum* genome were assembled. Of these genome assemblies, 87.3 % (2.63 Gb) and 87.2 % (2.79 Gb) were anchored to 12 chromosome pseudomolecules, respectively. The phylogenetic analysis revealed that speciation of *C. baccatum* and *C. chinense* has occurred at 1.7 and 1.1 million years ago, respectively. Athila, a subgroup of *Gypsy* family, was excessively accumulated in *C. baccatum* and caused genome size increase in this species comparing to the other species. The distribution for insertion time of LTR-retrotransposons showed that the accumulation of LTR elements in *C. chinense* and *C. baccatum* was distinctly increased during their speciation time. Comparative gene family analysis unveiled that the coincidence of gene duplication events actively occurred at the same time with estimated speciation. These results provide insights for speciation and genome size variation among the pepper species. Furthermore, the multiple pepper genomes will serve as important resources for comparative and population genomics as well as evolutionary studies of the genus *Capsicum*.

INTRODUCTION

Capsicum species originated in Americas have been globally cultivated after Columbus[1]. It is known that five domesticated and twenty-two wild species are belonged to the genus *Capsicum*[2,3]. The genus *Capsicum* is divided into three complexes based on species characterization studies considering morphology, genome size, and hybridization[4-6]. The three domesticated species of pepper grown mostly in the world, *C. annuum*, *C. chinense*, and *C. frutescens*, are included the *C. annuum* complex. The *C. baccatum* and *C. pubescens*, which have been grown in Latin America[7] are another domesticated peppers contained in distinct complexes.

In general, it is known that genomic variations such as changes of genome structure, gene repertories, and gene collinearity are mostly derived from evolutionary process[8]. As *de novo* genome sequences of many different organisms have been accumulated over a decade, comparative genomic analysis have enabled to provide insight into genome evolution within or between species[9-14]. Most importantly, transposable elements (TEs) which play multiple roles in leading genome evolution are major regulators causing variations of genome sequence, structure, and size[15]. TEs can also influence gene disruption and differentiate gene expression. In eukaryotes, changes of gene expressions have played an essential role in genome evolution and in some case, lead speciation[16,17]. As one of the controlling elements of genome evolution, TEs act to alter gene expression pattern

by being inserted into flanking regions of genes and regulate gene activity[17].

A diversity of the genus *Capsicum* obtained by evolutionary process has led separation of the peppers into distinct complexes and hybridization incompatibility among the complexes[18]. However, due to the lack of information for evolutionary process acting on the genus *Capsicum*, the genomic studies of peppers have not yet clearly explained a basis for understanding their diversity. Even though the completion of the reference pepper genome[19] has started to lead the pepper genomic studies, single reference genome is still insufficient to understand diversity and evolution of the genus *Capsicum*.

Here, I report *de novo* genome sequences of the two domesticated peppers, *C. chinense* and *C. baccatum*, including genome assembly, annotation, and chromosome pseudomolecules. Comparative genomic analysis between the newly obtained pepper genomes and the pre-existing reference genome provides insight into genomic changes, divergence and evolutionary history of the peppers mainly regulated by TEs. The multiple *de novo* genomes of the *Capsicum* species will serve as an excellent system for comparative and population genomics as well as pepper breeding.

MATERIALS AND METHODS

Plant materials and genome sequencing

Capsicum chinense 'PI159236' (hereafter *C. chinense*) and *Capsicum baccatum* 'PBC81' (hereafter *C. baccatum*) were used for genome sequencing. *C. chinense* has shown phenotypes resistant to tomato spotted wilt virus (TSWV)[20]. *C. baccatum* is an important resistance source to anthracnose caused by *Colletotrichum* species[21]. The plants were grown in a greenhouse and fresh expanding meristermatic leaves were harvested and then frozen in liquid nitrogen. The high molecular weight DNA was extracted according to the previous protocol[19]. Paired-end (200, 400, and 600bp) and mate-pair (2kb to 10kb) libraries of the each genome were constructed according to manufacturer's instructions with validation via KAPA SYBR FAST Master Mix Universal 2X qPCR Master Mix (Kapa Biosystems, Woburn, MA). The constructed libraries were sequenced by Illumina Hiseq 2000 with read length for ranging from 101 to 151 bp.

Preprocessing analysis

To remove unnecessary data, the raw sequences of the *C. chinense* and *C. baccatum* were filtered out through preprocessing analysis using in-house pipeline[19,22]. Firstly, the raw sequences were mapped to public prokaryotic sequences downloaded from GenBank using Bowtie2 v2.0.0-beta7[23] (--local -D 15 -R 2 -N 0 -

L 20 -i S,1,0.65) and then the mapped prokaryotic sequence candidates were removed. Secondly, duplicated reads generated during the sequencing process and low quality reads under Q20 were eliminated. Thirdly, low frequency fragments were assigned and deleted using Jellyfish[24] and Quake[25]. After the preprocessing analysis, reads of paired-end libraries were merged to single reads using FLASH[26]. The preprocessed reads were used for genome assembly of *C. chinense* and *C. baccatum*, respectively.

Transcriptome sequencing and assembly

For *C. chinense* and *C. baccatum*, RNA extraction and strand-specific transcriptome library preparations for 5 tissues (root, stem, leaf, flower, and fruit) were carried out as described[27]. In all tissues, three independent biological replications were performed and the prepared libraries were sequenced using Illumina Hiseq 2000. The raw sequences were preprocessed using the previously constructed preprocessing pipeline[22] to obtain purified raw sequences. For the each species, transcriptome assembly was performed for combined libraries of the 5 tissues using the pepper transcriptome assembly pipeline[19]. The assembled transcriptome for *C. chinense* and *C. baccatum* was used for gene annotation of the each genome.

Structural gene annotation

Structural gene annotation of the two pepper genomes was performed using in-

house annotation pipeline. Firstly, *de novo* and reference assembly for whole transcriptome were performed and then protein coding genes in the transcriptome were obtained using the previously constructed transcriptome annotation pipeline[23]. The obtained genes were mapped using Exonerate version 2.2.0[28] to the assembled genomes and then, full-length genes were extracted as initial coding gene set. Secondly, public proteins of *Arabidopsis*[29] (TAIR 10), tomato[30] (iTAG 2.3), pepper[19] (PGA 1.5), and plant refSeq[31] downloaded from NCBI were mapped using Exonerate to the genomes. Thirdly, *ab initio* prediction was performed by AUGUSTUS[32] using custom training set generated from the initial coding gene set. Lastly, except the regions of initial coding gene set, EvidenceModeler[33] (EVM) was performed for extraction of consensus genes considering gene regions from protein mapping and *ab initio* prediction. The final annotated genes were obtained by merging the initial coding gene set and the consensus genes (Figure 1).

Annotation and evolution analysis of repeat sequences

To build custom library of transposable elements (TEs) related repeat in pepper genomes, the published pepper genomes and the assembled genomes were used for construction of initial repeat library using RepeatModeler. In addition, intact LTR elements were predicted and classified using LTRHarvest and LTRDigest. Using the predicted intact LTR elements, substitution rates were calculated by baseml in PAML package[34]. Comparing the initial repeat library to the intact LTRs, sub families of

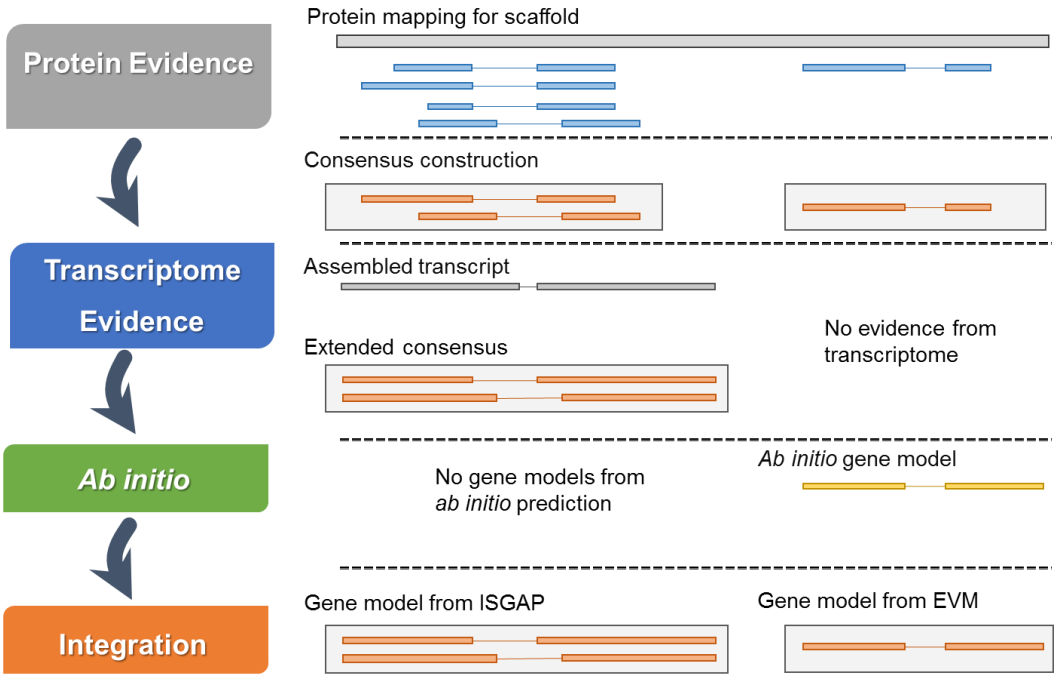


Figure 1. Gene annotation scheme for the pepper genomes.

LTRs were assigned, and then final repeat library was constructed. The final repeat library was used for repeat masking through REPEATMASKER.

OrthoMCL analysis

Orthologous gene clusters were assigned from OrthoMCL[35] with its standard parameters of 8 species to identify gene families expanded or contracted in pepper genomes. Tomato (v2.3), potato[36] (v3.4), Arabidopsis[37] (TAIR10), grape[38] (VvGDB v2.0), rice[39] (MSU RGAP 7), and the pepper gene sets containing *C. annuum*, *C. baccatum* and *C. chinense* were used to infer putative orthologous gene families. The TEs-related gene models in the genomes were removed, and then an all-by-all comparison was performed using BLASTP with an Evalue of $1e \times 10^{-5}$

Calculation of divergence time

For estimation of speciation time among the 8 species, single copy genes, present in all the tested 8 genomes, were obtained by OrthoMCL analysis. Multiple alignments for the single copy were performed to calculate substitution rate using PRANK [40]. The alignment outputs resulted in NEXUS format from PRANK were used as input data of Bayesian Evolutionary Analysis Utility[41] (BEAUti) and BEAUti optimized the input data for accurate tree construction. Bayesian Evolutionary Analysis Sampling Trees[41] (BEAST v1.8.2) estimated divergence time of the eight species and then constructed species tree.

RESULTS

Sequencing, assembly, and annotation

Genomic sequence (425.7 Gb) was generated solely using Illumina platforms (Hiseq 2000) with PE and MP libraries ranging from 180 bp to 10 kb for *C. chinense* with read lengths of 101 bp (Table 1). A total of 418.4 Gb of genomic sequence data was also generated for *C. baccatum* with read lengths of 101 and 151 bp through Illumina Hiseq 2500 (Table 2). Based on 19-mer distribution, the estimated genome size of *C. chinense* and *C. baccatum* were 3.2 and 4.2 Gb, respectively (Figure 2). After preprocessing analysis, *de novo* genome assembly of *C. chinense* and *C. baccatum* was performed using SOAPdenovo2 and SSPACE. As results, 3.07 Gb of *C. chinense* genome (95% of 3.21 Gb) was assembled (3.64 Mb of N50) and 3.2 Gb of *C. baccatum* genome (76 % of 4.2 Gb) was also assembled into 25,206 scaffolds with 2.1Mb of N50 value (Table 3). In addition, 90% of the assembled *C. chinense* and *C. baccatum* genomes was covered by 968 and 1,645 scaffolds, respectively (Table 3). To construct chromosome pseudomolecules, *C. annuum* genome were used to assign chromosome numbers and genomic order of the newly assembled genomes as shown in Figure 3. A total of 2.68 Gb and 2.79 Gb (87.3 % and 87.2 % of genome assembly) were anchored into 12 chromosomes of *C. chinense* and *C. baccatum*, respectively, and ordered by genomic distance of

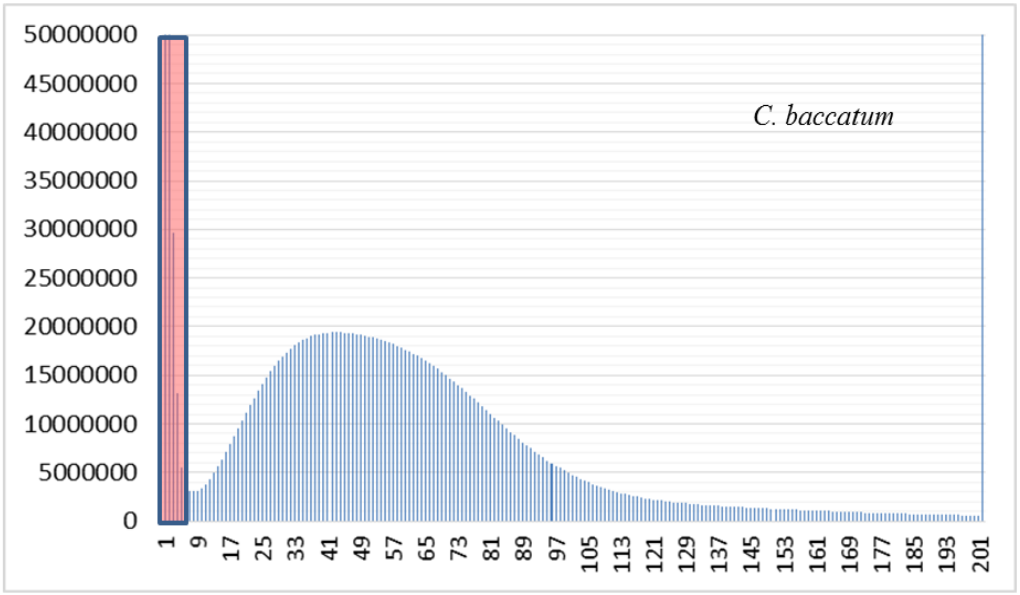
Table 1. Generated *C. chinense* genome sequences in this study

Sequencing data	Insert size	Total length(Gb)	Sequencing depth(X)	Length (bp)
Illumina reads	200 bp	98.9	30.0	101
	400 bp	94.7	28.7	101
	600 bp	96.0	29.1	101
	2 kb	41.0	12.4	101
	5 kb	27.2	8.2	101
	10 kb	67.9	20.6	101
Total		425.7	129.0	

Table 2. Generated *C. baccatum* genome sequences in this study

Sequencing data	Insert size	Total length(Gb)	Sequencing depth(X)	Length (bp)
Illumina reads	200 bp	112.8	32.2	151
	400 bp	73.6	21.0	151
	600 bp	72.1	16.5	151
	2 kb	52.4	8.2	151
	5 kb	52.8	15.1	151
	10 kb	54.7	15.6	101
Total		418.4	119.5	

A



B

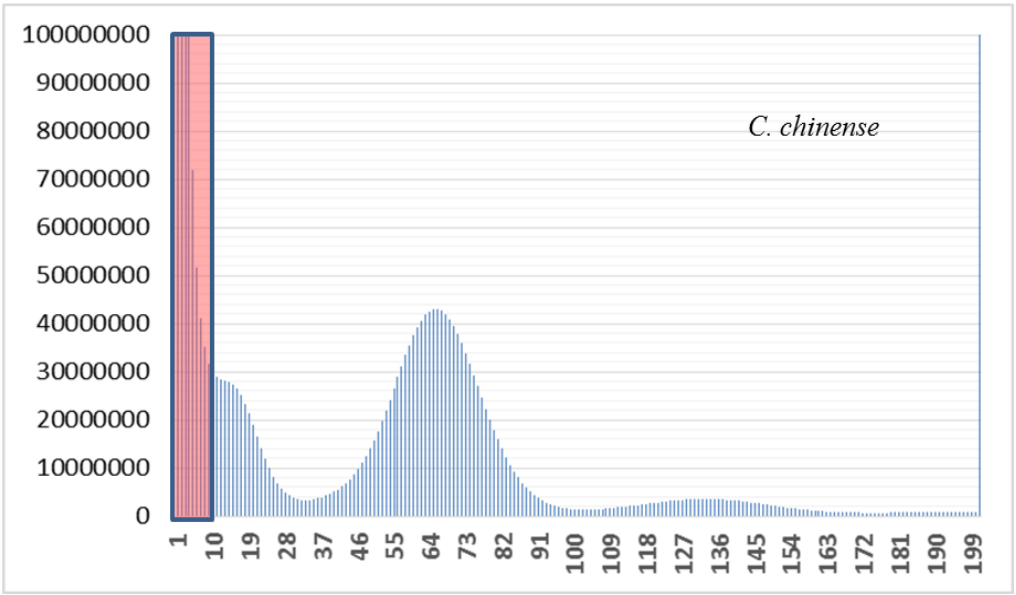


Figure 2. The 19-mer distribution of the sequenced pepper genomes. The X-axis means frequency of 19-mer and the y-axis indicates amount of the 19-mer. The low frequency sequences in red boxes are masked considering putative error sequences.

Table 3. Comparison of the pepper genome assemblies

	C. annuum		*C. baccatum*		*C. chinense*	
	Contig	Scaffold	Contig	Scaffold	Contig	Scaffold
N10	112,542bp (1,927th)	7,426,011bp (31th)	129,849(1,762th)	6,083,938(40th)	182,821(1,174th)	10,886,548(22th)
N20	76,075bp (5,182th)	4,993,279bp (82th)	89,601(4,691th)	4,421,161(102th)	120,769(3,247th)	7,868,644(56th)
N30	55,068bp (9,787^{h})	3,957,205bp (151th)	66,915(8,733th)	3,451,198(186th)	87,499(6,213th)	6,260,044(99th)
N40	40,567bp (16,087th)	3,068,199bp (240th)	50,928(14,069th)	2,702,886(292th)	65,597(10,208th)	4,637,893(157th)
N50	29,995bp (24,618th)	2,472,394bp (352th)	38,859(21,062th)	2,085,930(427th)	49,067(15,542th)	3,643,404(232th)
N60	21,778bp (36,237th)	1,947,990bp (491th)	29,012(30,328th)	1,681,667(599th)	35,954(22,731th)	2,719,709(328th)
N70	15,269bp (52,522th)	1,469,508bp (672th)	20,841(42,928th)	1,248,822(821th)	25,090(32,772th)	1,978,036(460th)
N80	9,796bp (76,680th)	1,083,843bp (915th)	13,598(61,264th)	851,922(1,135th)	15,871(47,826th)	1,344,266(647th)
N90	4,797bp (119,071th)	628,252bp (1,276th)	6,798(92,910th)	447,255(1,645th)	7,047(75,632th)	605,366(968th)
Max	442,125bp	18,549,843bp	494,009bp	16,429,935bp	872,291bp	20,331,050bp
Amount of initial contig		3,572,686,812 bp		4,180,852,308 bp		3,220,229,745 bp
Total Length / Number	2.96Gb / 337,329ea	3.06Gb / 37,989ea	3.1Gb / 251,555ea	3.2Gb / 25,206ea	3.02Gb / 273,397ea	3.07Gb /89,332ea

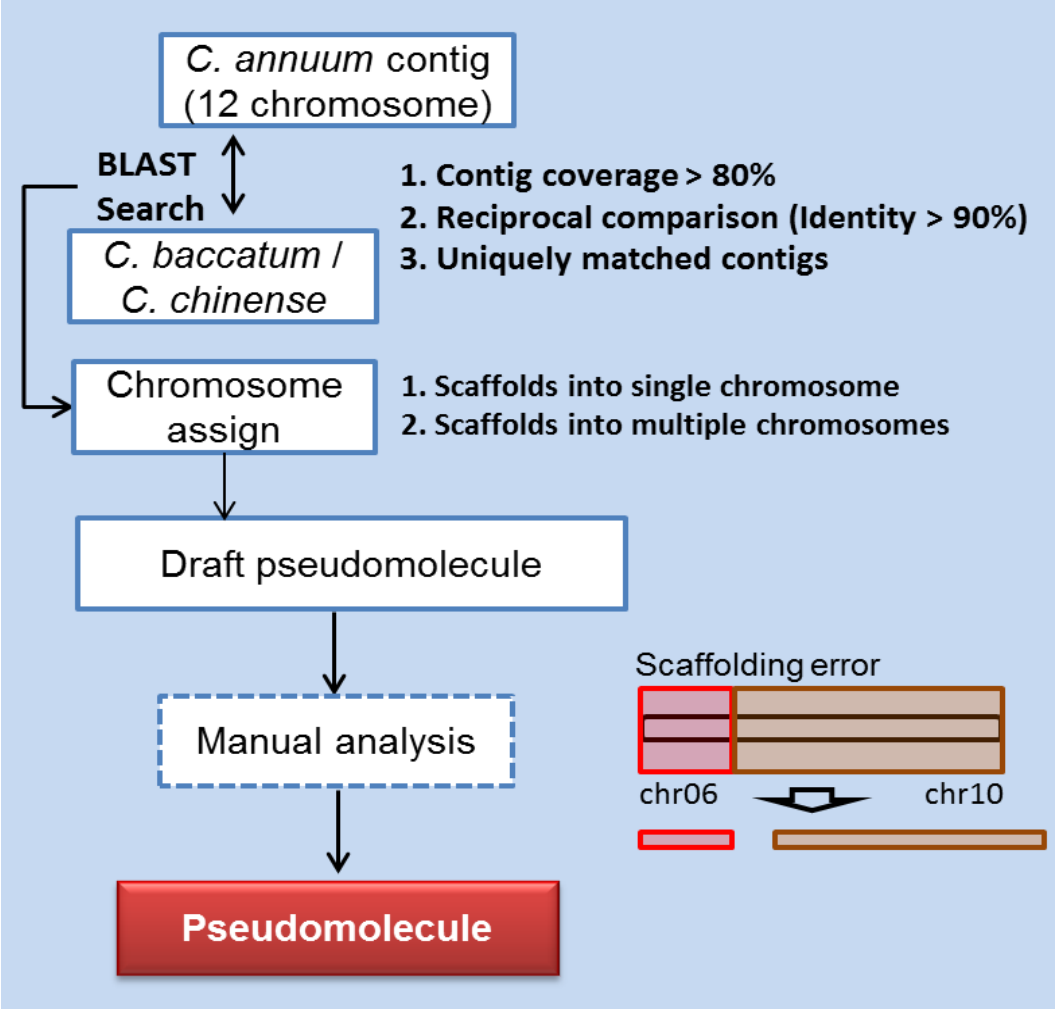


Figure 3. Strategy for pseudomolecule construction of the *C. baccatum* and *C. chinense*. The flowchart summarizes steps for pseudomolecule construction of *C. baccatum* and *C. chinense* genome using the final contigs of *C. annuum*.

the assembled *C. annuum* genome (Table 4).

To predict gene structure, gene annotation was performed for *C. chinense* and *C. baccatum* genome assemblies using in-house annotation pipe line (see method). As a result, 35,284 genes were predicted with average CDS length of 1,090bp in *C. chinense* (Table 5). A total of 34,640 protein coding genes were also annotated (average CDS length : 1,108bp) in *C. baccatum*. The predicted number of genes of *C. chinense* and *C. baccatum* were similar to the gene numbers in *C. annuum* genome, but genes of *C. chinense* and *C. baccatum* were slightly longer than genes in *C.annuum* (Table 5). To assign biological function of the annotated genes in *C. chinense* and *C. baccatum*, functional annotation was performed using INTERPROSCAN version 5.3-46, then gene functions were assigned and used for further analysis.

Evolution of gene families in the genus *Capsicum*

To define the gene family clusters from the pepper and the other plant genomes listed in Table 6, gene clustering analysis was performed using the OrthoMCL version 2.0.2[35]. A total of 207,410 proteins from *C. annuum*, *C. baccatum*, *C. chinense*, tomato, potato, *Arabidopsis*, grapevine and rice were clustered into 29,949 families (except singletons) using OrthoMCL. A total of 133, 472 genes in 12,194 families were shared among all eight species and 4,284 gene families were Solanaceae specific with 10,818 genes. The 6,005 families

Table 4. Statistics of twelve pseudomolecule chromosomes of the pepper genomes

Chr	Chromosome length (Mb) / number of anchored scaffolds		
	C. annuum	*C. baccatum*	*C. chinense*
1	273 / 149	276 / 221	272 / 175
2	171 / 106	174 / 138	163 / 123
3	258 / 164	265 / 198	263 / 168
4	223 / 109	212 / 158	217 / 115
5	233 / 128	239 / 179	246 / 106
6	237 / 130	257 / 169	235 / 121
7	232 / 116	258 / 196	222 / 133
8	145 / 45	159 / 116	143 / 77
9	253 / 136	234 / 159	243 / 129
10	234 / 144	231 / 191	209 / 138
11	260 / 146	266 / 185	252 / 151
12	236 / 129	220 / 177	217 / 120
Total	2,755 / 1,502	2,791 / 2,087	2,682 / 1,556

Table 5. Comparison of annotated gene models of the pepper genomes

Species	Version	Protein coding Loci	Total CDS Length (bp)	Ave CDS Length (bp)
C. annuum	1.55	34,898	35,254,530	1,010
C. baccatum	0.9	34,640	38,388,616	1,108
C. chinense	0.9	35,284	38,469,398	1,090

Table 6. Protein sets used for gene family analysis

Group	Species	Protein coding loci	Total CDS length (aa)	Ave CDS length (aa)
Different group	Arabidopsis	27,416	11,109,813	405.2
	Rice	39,049	13,811,501	353.7
	Grape	26,346	9,986,113	379.0
Solanum group	Tomato	33,907	11,757,019	346.7
	Potato	38,446	11,902,405	309.6
Capsicum group	Pepper (*C. annuum*)	34,898	11,747,802	336.6
	Pepper (*C. baccatum*)	34,640	12,796,179	369.4
	Pepper (*C. chinense*)	35,284	12,822,898	363.4

unique to *Capsicum* species contained 18,704 genes (Figure 4a). Among the *Capsicum* species specific gene families, 284, 341, and 400 families containing 801, 3,776, and 1,210 genes were unique to *C. annuum*, *C. baccatum*, and *C. chinense*, respectively (Figure 4b). All families were assigned biological function via INTERPROSCAN.

To calculate divergence time of the eight species and construct species tree, 1, 485 single copy genes that present in all tested 8 genomes were extracted and used. The estimation of divergence time was performed using Bayesian Evolutionary Analysis Sampling Trees[41] (BEAST) version 1.8.2 considering phylogenetic relationships among the eight genomes. The results showed that genus *Capsicum* was diverged from Solanaceae family at 19.6 million years ago (MYA) approximately coincided to the previous result[42]. In addition, the estimated speciation times among the plant genomes also corresponded to the previous results[43-46]. Considering consensus between the estimation and previous results, divergences among the three peppers occurred firstly between *C. baccatum* and the other peppers at 1.7 million years ago (MYA) and subsequently between *C. annuum* and *C. chinense* at 1.1 MYA (Figure 5).

A great change of gene repertories is one of the results derived from genome evolution. To survey variations of gene families among the pepper genomes, expanded or species-specific gene families in each pepper genome were classified using Chi-square and Fisher's exact tests (p-value $\leq$ 0.05). Regarding gene families

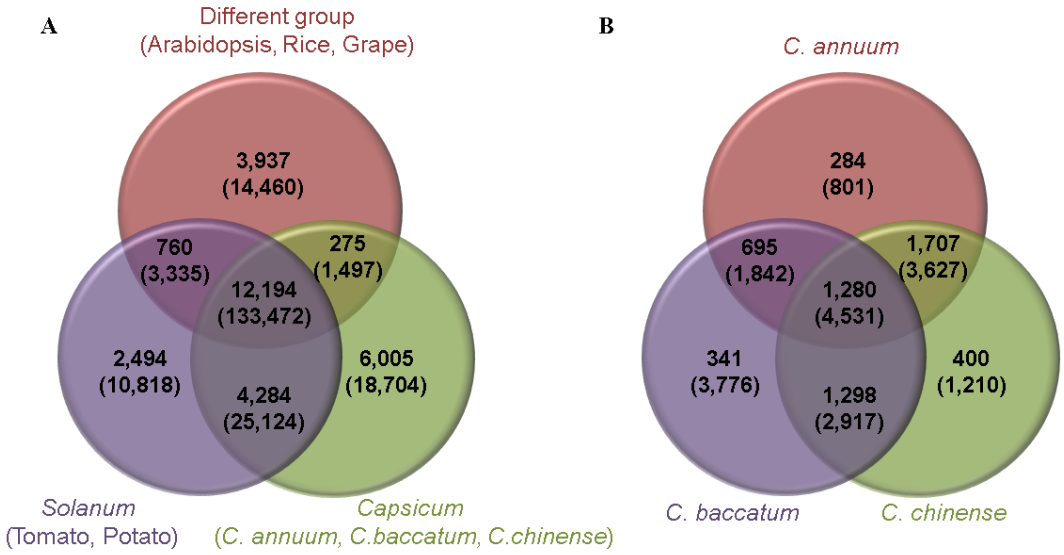


Figure 4. Comparison of the orthologous gene families. (A) Number of orthologous or specific gene families (genes) among the plant genomes. (B) Distribution of orthologous or species-specific gene families (genes) in the pepper genomes.

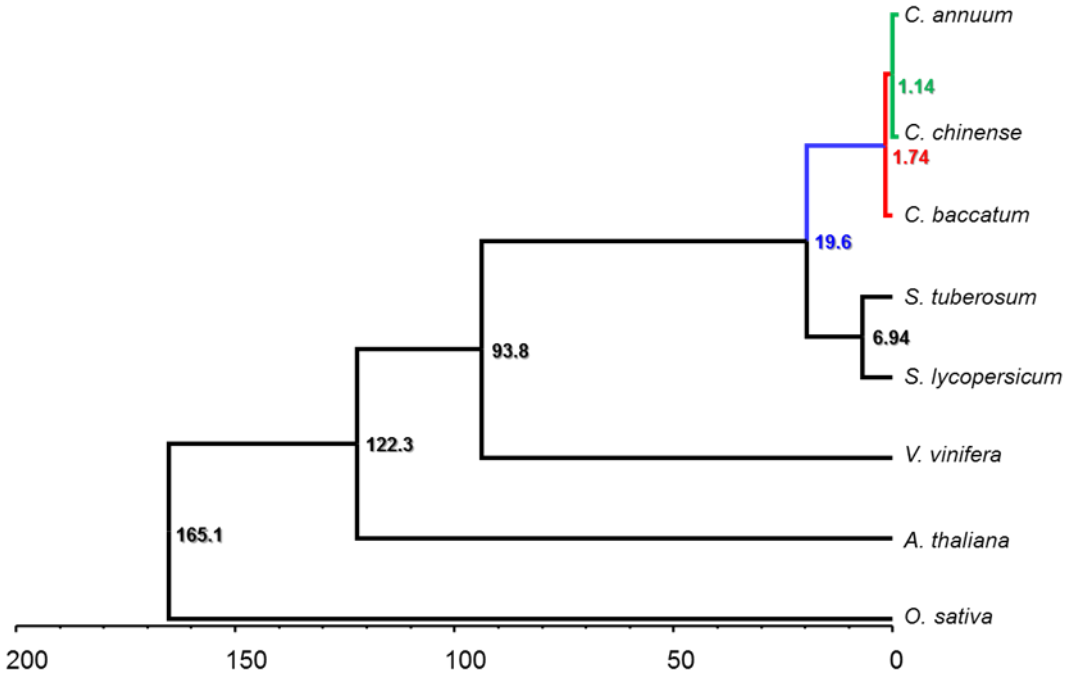


Figure 5. Estimated divergence time of the sequenced plant genomes. The phylogenetic tree shows the speciation times of the plants. The blue line indicates divergence of the *Capsicum* species and the red line represents speciation between *C. baccatum* and the other peppers. The green line means speciation between *C. annuum* and *C. chinense*.

of *C. baccatum* and the other pepper genomes, 17 gene families containing 618 genes were expanded in *C. baccatum* compared to 105 and 64 genes in *C. annuum* and *C. chinense*, respectively (Table 7). The expanded gene families included various biological functions containing endonuclease family, NB-ARC domain, zinc knuckle, and Ulp1 protease family (Figure 6). Twelve families (376 genes) including Ulp1 protease family, zinc-finger domain and aspartyl protease were expanded in *C. chinense* compared to 34 genes in *C. annuum*. In *C. annuum*, only 4 gene families were expanded including NB-ARC, leucine rich repeat, and ring finger domain related genes (Figure 6).

Gene duplication to adapt various environmental conditions is the main source of functional diversity for eukaryotes[47]. To estimate creation time of the expanded gene families, duplication time of the expanded gene families in each genome was calculated following the previously described methods[48-50]. As results, most of gene families (13 out of 18 families) have been actively duplicated during speciation time between *C. baccatum* and the other peppers (Figure 7a). Furthermore, the expanded gene families of *C. annuum* and *C. chinense* respectively have also been experienced duplication during speciation time (Figure 7b-c). Even though dramatic gene duplication event has not occurred in the pepper genomes around the speciation period, these results in this study provide insights that the recent gene duplication contributed change of gene repertories and is ongoing.

Table 7. Number of expanded gene families in the three pepper genomes

Type	*C. annuum*	*C. baccatum*	*C. chinense*
CA+[a]	4 (281)	4 (293)	4 (150)
CC+[b]	12 (34)	12 (70)	12 (376)
CB+[c]	18 (105)	18 (618)	18 (64)

[a] Expanded families (genes) in *C. annuum* compared to *C. chinense*

[b] Expanded families (genes) in *C. chinense* compared to *C. annuum*

[c] Expanded families (genes) in *C. baccatum* compared to the other peppers

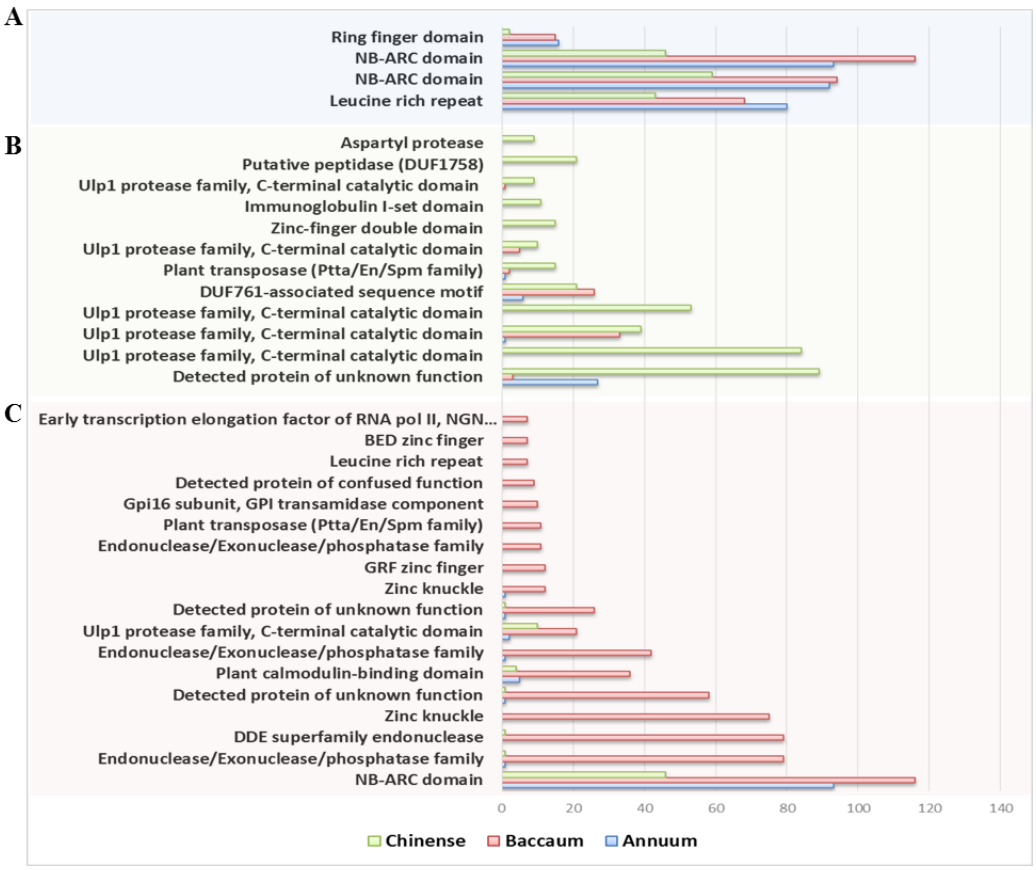
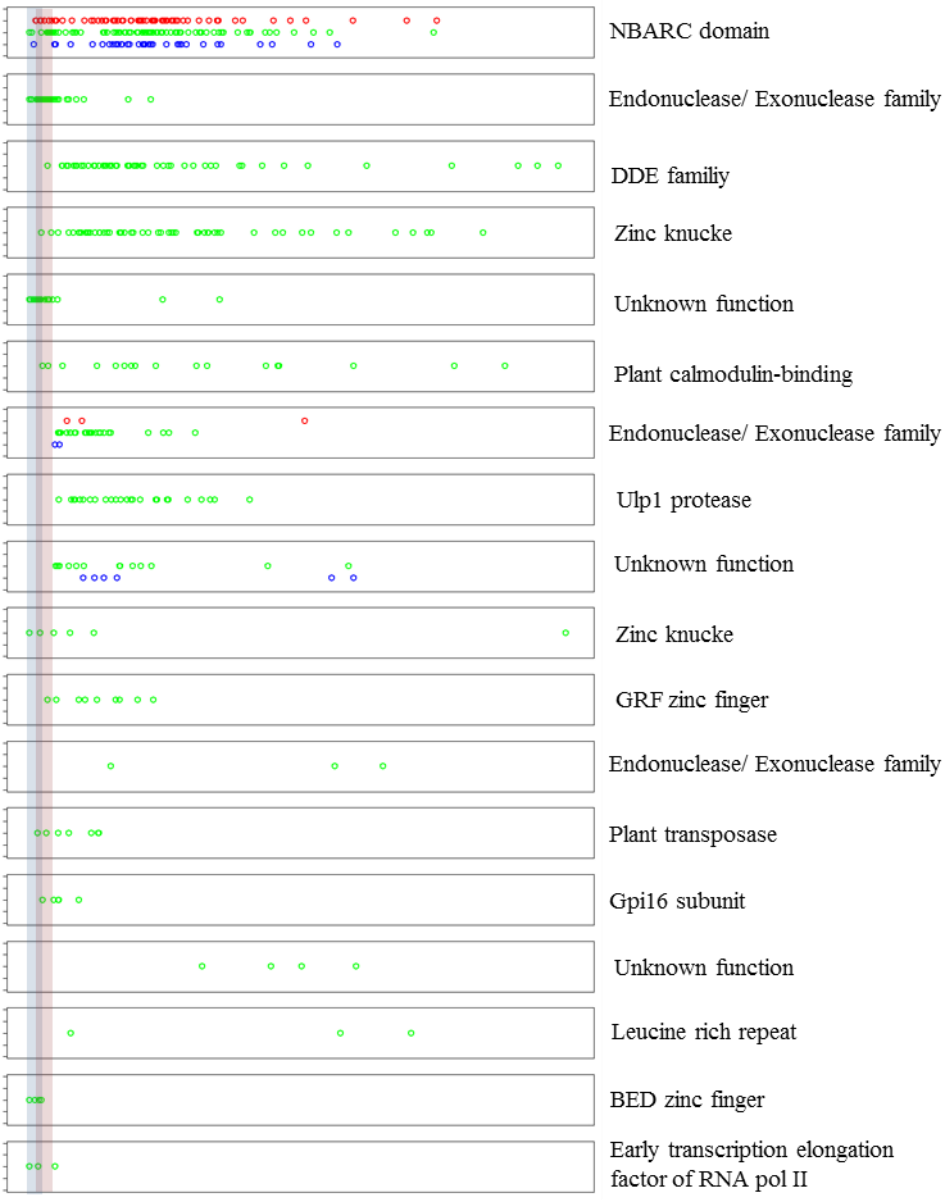


Figure 6. Biological function of the expanded gene families in the pepper genomes. The graphs show the number of expanded genes and their functional descriptions for (A) *C. chinense*, (B) *C. annuum*, and (C) *C. baccatum*, respectively. The pale green, red, and blue bars indicate number of genes for *C. chinense*, *C. baccatum*, and *C. annuum*.

A


NBARC domain
Endonuclease/ Exonuclease family
DDE familiy
Zinc knucke
Unknown function
Plant calmodulin-binding
Endonuclease/ Exonuclease family
Ulp1 protease
Unknown function
Zinc knucke
GRF zinc finger
Endonuclease/ Exonuclease family
Plant transposase
Gpi16 subunit
Unknown function
Leucine rich repeat
BED zinc finger
Early transcription elongation factor of RNA pol II
KS value

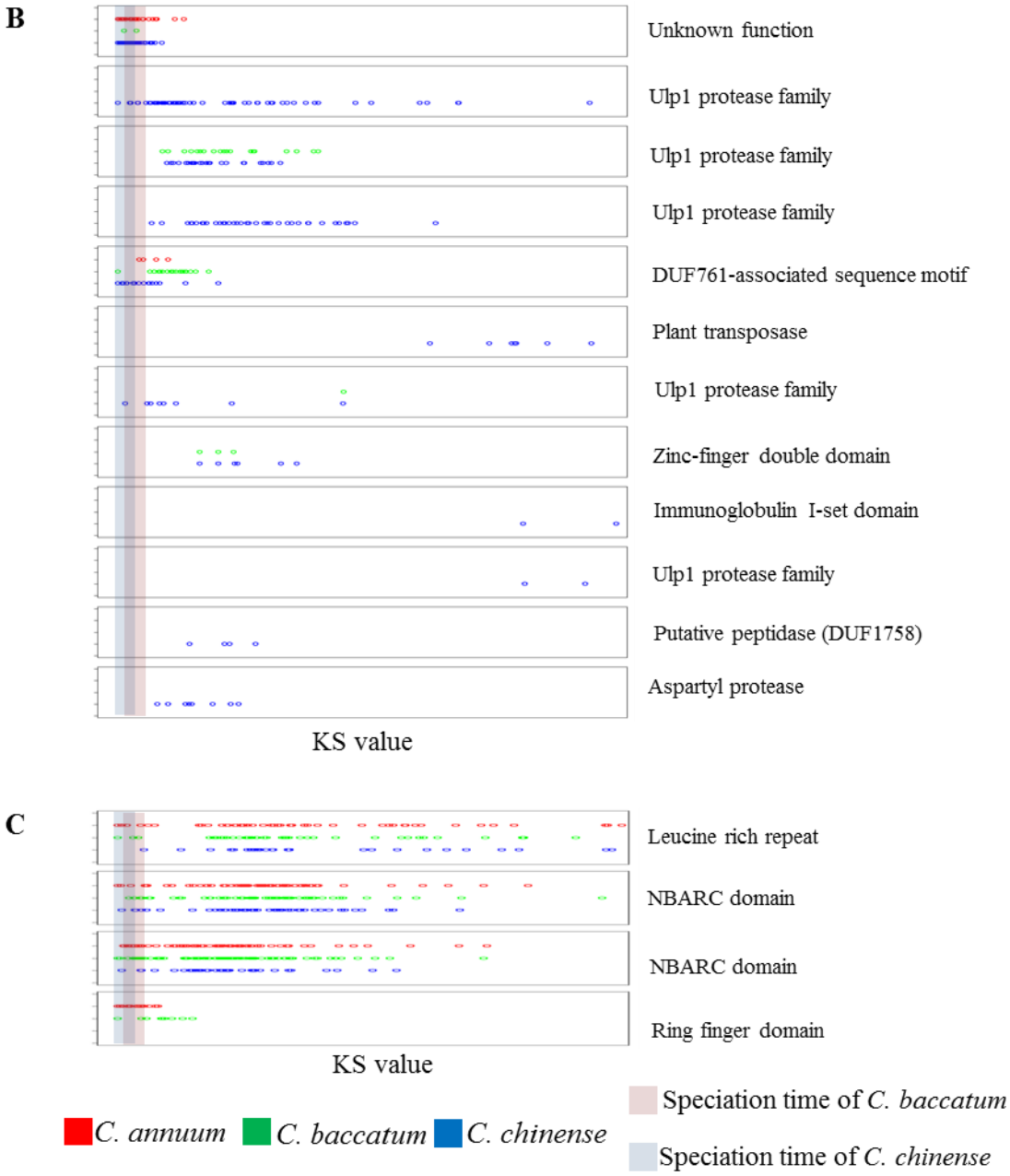


Figure 7. Age of the expanded gene families in the pepper genomes. Distributions of gene duplication time for the expanded gene families in (A) *C. baccatum*, (B) *C. chinense*, and (C) *C. annuum* are showed, respectively. The x-axis indicates synonymous substitution values ranging from 0 to 0.5.

Repeat annotation and genome size variation in the *Capsicum* spp

Repeat sequences in the assembled *C. chinense* and *C. baccatum* genomes were annotated by REPEATMASKER using custom repeat library generated from the published and newly assembled pepper genomes[19,51]. In addition, repeat sequences of the *C. annuum* genome[19] was newly annotated using the same library. However, the assembled genomes could not cover the whole genomes for *C. annuum*, *C. baccatum*, and *C. chinense*. Therefore, initial contigs covering 100 % of genome size (3.57 Gb, 3.22 Gb, and 4.18 Gb) for *C. annuum, C. baccatum* and *C. chinense* were used to predict whole repeats and the repeat sequences were predicted using the repeat library (Table 3).

A total of 3.01 Gb, 3.49 Gb, and 2.72 Gb (84.3%, 83.6%, and 84.5%) was classified as repeat sequences in the whole genome of *C. annuum, C. baccatum* and *C. chinense*, respectively (Table 8). In these genomes, proportions of repeat were positively correlated with the genome size. Furthermore, differences of genome size among the pepper species were mostly correlated with different fraction of repeats indicating that genome size variation was caused by repeat sequences in the pepper genomes (Table 8). In the repeat sequences, transposable elements (TEs) occupied most of proportion in the three pepper genomes (Table 8). Among the TEs, LTR-retrotransposons were mostly abundant and *Gypsy* family of the LTR elements was major type similar to other plants representing approximately 2.11 Gb, 2.47 Gb, and 1.91 Gb in *C. annuum, C. chinense* and *C. baccatum*, respectively (Table 8).

Table 8. Amount of repeat sequences in whole genome of the peppers

Type of TE	*C. annuum*		*C. baccatum*		*C. chinense*	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA elements	382,235,654	10.7%	440,064,743	10.5%	340,237,147	10.6%
LINE elements	77,911,930	2.2%	81,189,335	1.9%	73,411,515	2.3%
SINE elements	25,122,592	0.7%	32,099,007	0.8%	23,921,919	0.7%
LTR/Others	1,100,097	0.0%	757,320	0.0%	967,113	0.0%
LTR/Gypsy	2,106,918,172	59.0%	2,465,425,579	59.0%	1,910,090,915	59.3%
LTR/Copia	213,764,896	6.0%	223,175,961	5.3%	195,445,515	6.1%
LTR/Caulimoviridae	128,678,293	3.6%	115,881,420	2.8%	108,068,343	3.4%
rRNA	4,262,558	0.1%	50,792,057	1.2%	5,247,809	0.2%
Simple repeat	54,331,613	1.5%	64,261,670	1.5%	46,182,071	1.4%
Others	17,637,748	0.5%	19,588,521	0.5%	17,889,634	0.6%
Total	3,011,963,553	84.3%	3,493,235,613	83.6%	2,721,461,981	84.5%

In the pepper genomes, most of TE-related repeats in each species were directly proportional to genome size. Of the TEs, *Gypsy* family of LTR-retrotransposons covered approximately 60 % of pepper genomes (Table 8). Although the proportion of *Gypsy* family was similar, subgroups of *Gypsy* family were significantly different among the three genomes. Among the *Gypsy* families, del elements occupied the largest fraction representing 1.48 Gb, 1.45 Gb, and 1.34 Gb (41.5%, 34.7%, and 41.7%) in *C. annuum*, *C. baccatum*, and *C. chinense*, respectively (Table 9). Considering genome size, the proportion of del in *C. baccatum* was relatively lower than that in the other pepper genomes, whereas athila was exceptionally abundant compared to the other genomes (Table 9). This result indicates that genome expansion in *C. baccatum* was mainly resulted by accumulation of athila subgroup. For other LTR-retrotransposons, the proportion of *Copia* and *Caulimoviridae* families were slightly higher in *C.annuum* and *C. chinense* than in *C. baccatum*. However, the differences among the genomes were not significant (Table 10 and Table 11).

Except TE-related repeats, it is known that copy number of ribosomal RNA (rRNA) is associated with genome size of eukaryotes[52]. A large number of rRNA (up to 10-fold) was discovered in *C. baccatum* compared to other pepper genomes (Table 8). To confirm the genome-wide distribution of rRNA in the pepper genomes, fluorescence *in situ* hybridization (FISH) analysis was performed as described[53] and the result revealed that rRNA is frequently distributed and more abundant in *C. baccatum* compared to the other genomes (Figure 8).

Table 9. Amount of *Gypsy* subgroups in the pepper genomes

Type of TE	*C. annuum*		*C. baccatum*		*C. chinense*	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
athila	303,649,406	8.5%	637,602,219	15.3%	278,467,087	8.6%
crm	54,760,423	1.5%	78,091,056	1.9%	49,718,804	1.5%
del	1,481,780,930	41.5%	1,451,197,733	34.7%	1,342,931,960	41.7%
galadriel	1,693,369	0.0%	2,268,114	0.1%	1,462,859	0.0%
reina	2,842,382	0.1%	3,142,177	0.1%	3,160,385	0.1%
tat	186,670,596	5.2%	214,661,605	5.1%	158,618,578	4.9%
others	75,521,066	2.1%	78,462,675	1.9%	75,521,066	2.3%
Total	2,106,918,172	59.0%	2,465,425,579	59.0%	1,909,880,739	59.3%

Table 10. Total volume of *Copia* subgroups in the pepper genomes

Type of TE	*C. annuum*		*C. baccatum*		*C. chinense*	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
oryco	5,954,551	0.2%	6,273,671	0.2%	5,453,683	0.2%
retrofit	24,577,062	0.7%	28,427,014	0.7%	25,348,731	0.8%
sire	57,692,847	1.6%	52,951,543	1.3%	55,694,012	1.7%
tork	107,204,817	3.0%	114,548,981	2.7%	90,048,010	2.8%
others	18,335,619	0.5%	20,974,752	0.5%	18,901,079	0.6%
Total	213,764,896	6.0%	223,175,961	5.3%	195,445,515	6.1%

Table 11. Statistics of *Caulimoviridae* subgroups in the pepper genomes

Type of TE	*C. annuum*		*C. baccatum*		*C. chinense*	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
badnavirus	13,822,944	0.4%	13,030,161	0.3%	10,943,628	0.3%
caulimovirus	13,681,510	0.4%	16,784,654	0.4%	10,840,822	0.3%
cavemovirus	94,834,388	2.7%	79,568,694	1.9%	79,697,436	2.5%
Total	122,338,842	3.4%	109,383,509	2.6%	101,481,886	3.2%

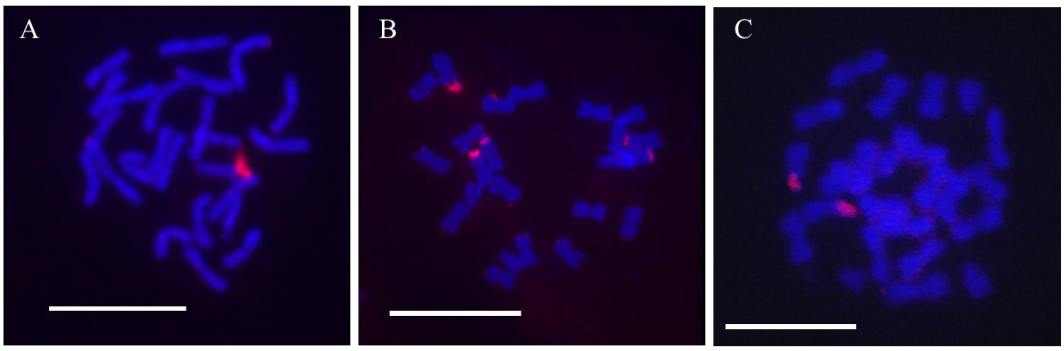


Figure 8. FISH with 25S (red) ribosomal DNA probes to somatic metaphase chromosomes of *Capsicum* species. **.** A: *C. annuum* cv. Chilsungcho, B: *C. chinense* habanero, and C: *C. baccatum* PBC81. The scale bar is 15 μm.

Taken together, genome size variation between the *C. baccatum* and the other pepper species was mainly caused by differences in distinct accumulation of LTR-retrotransposon and rRNA. Additionally, difference of genome size between *C. annuum* and *C. chinense* was mainly generated from a quantitative difference of TEs in each genome.

Evolutionary history of LTR-retrotransposons in the *Capsicum* species

To study evolution of LTR-retrotransposons as a major type for genome size variation among the pepper genomes, intact LTR elements were predicted and extracted. To confirm that the intact LTR elements were evenly extracted without bias, ratio of subgroups of the intact LTR-retrotransposons was compared to that of whole LTR-retrotransposons and had similar distribution (Figure 9). The intact LTR elements were used to calculate insertion time of LTR-retrotransposons in the pepper genomes. The insertion times of LTR elements were estimated according to the method described by SanMiguel *et al*.[54]

Given that the speciation times of '*C. annuum* and *C. chinense*' and '*C. baccatum* and other peppers' were 1.1 and 1.7 MYA with the substitution rate of 0.01 and 0.02, respectively, accumulation of LTR-retrotransposon firstly peaked at the substitution value of 0.1 like other pepper genomes and then, exponentially occurred in *C. baccatum* shortly before speciation at the value of 0.03 (Figure 10a). In addition, although the composition of LTR-retrotransposons was similar between *C. annuum*

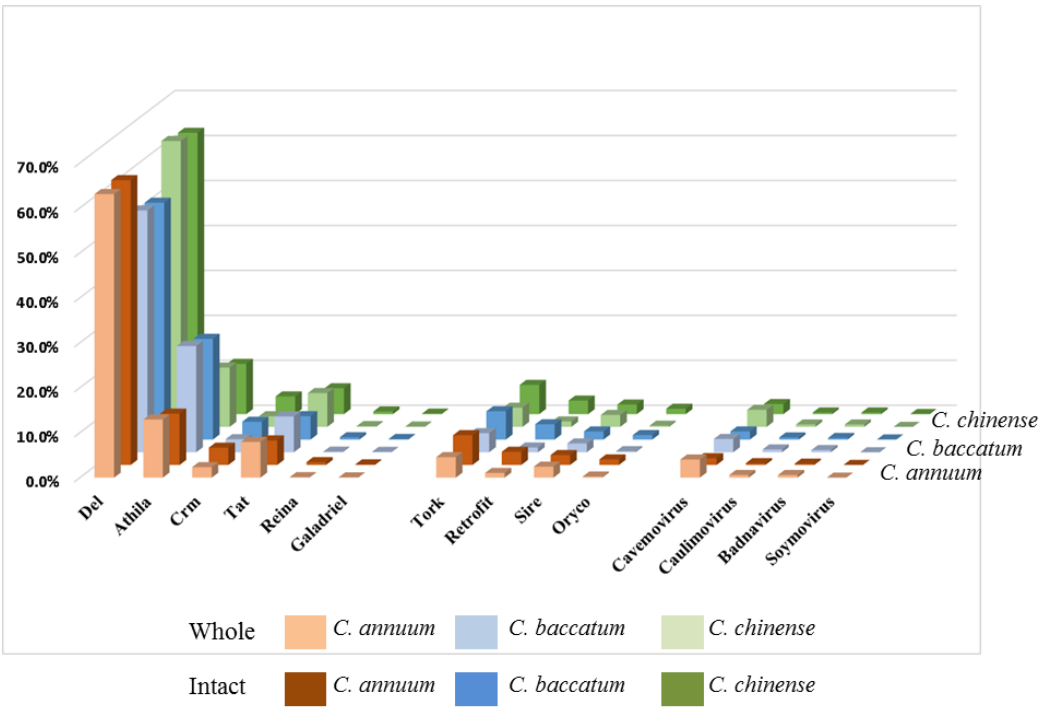
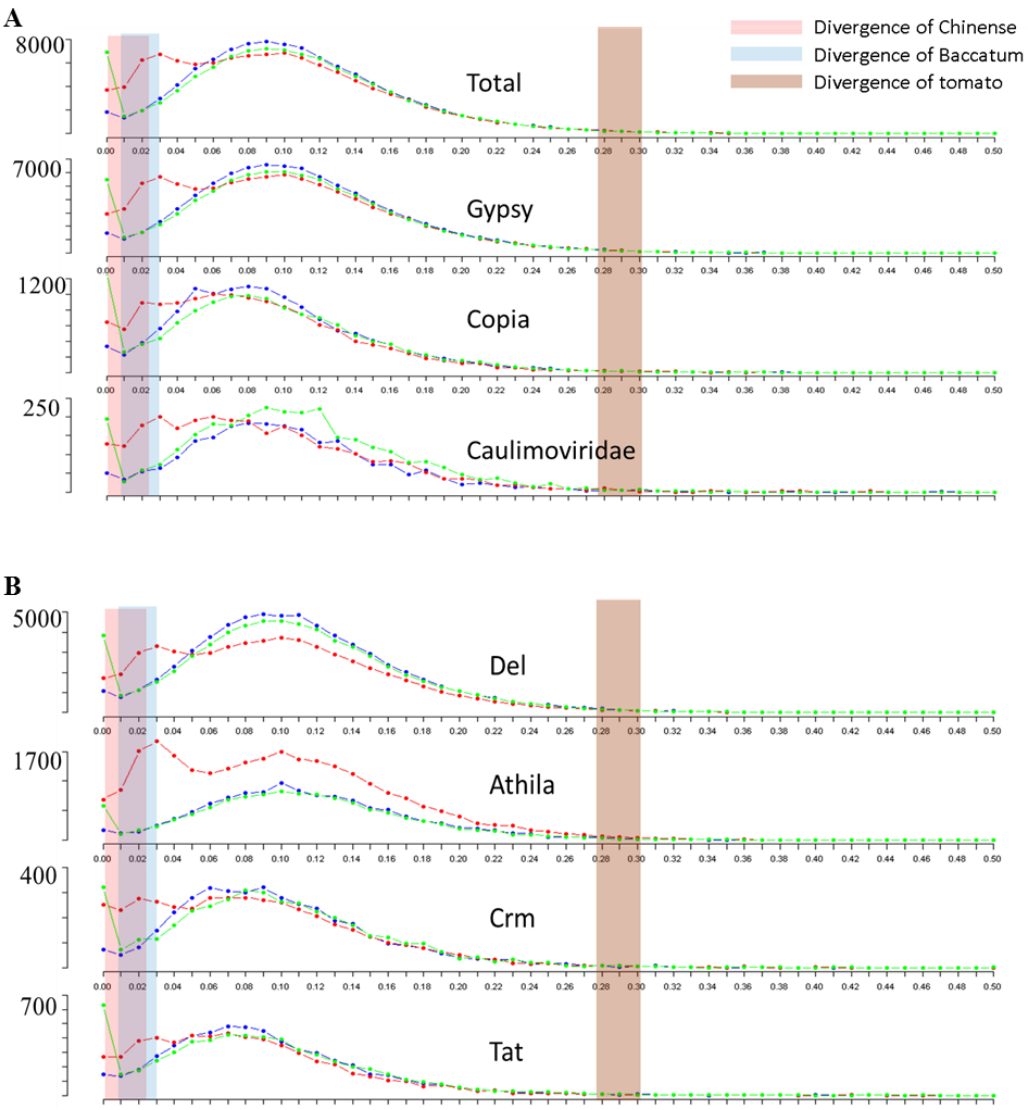


Figure 9. Comparison for ratio of subgroups between intact and whole LTR-retrotransposons. The x-axis indicates type of LTR subgroups and the y-axis represents the ratio of LTR subgroups for total LTR elements.


A
Divergence of Chinense
Divergence of Baccatum
Divergence of tomato
8000
Total
7000
Gypsy
1200
Copia
250
Caulimoviridae
B
5000
Del
1700
Athila
400
Crm
700
Tat

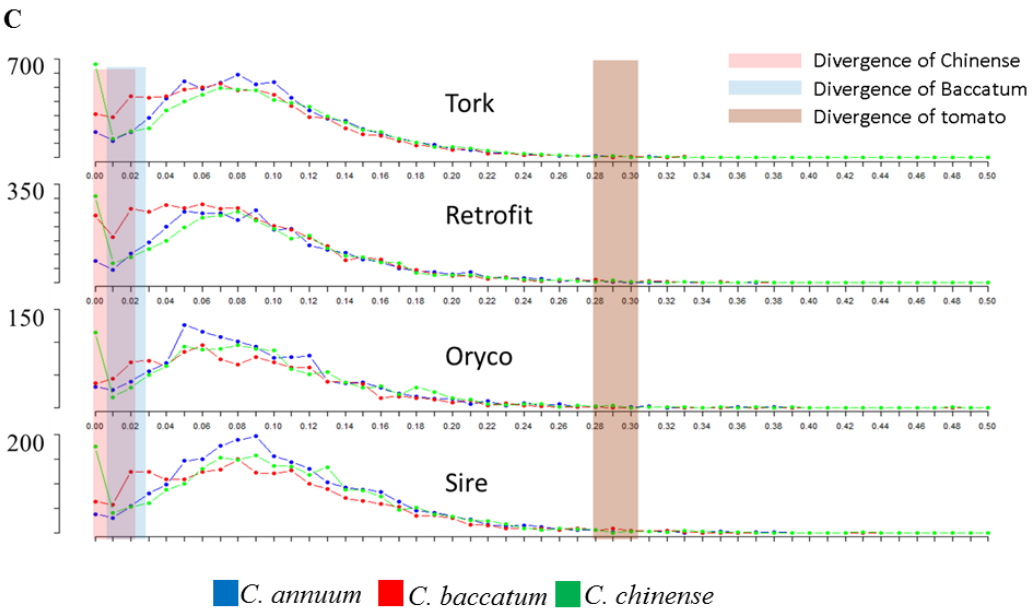


Figure 10. Age of LTR-retrotransposons in the pepper genomes. (A) Distribution of insertion time for Total, *Gypsy*, *Copia*, and *Caulimoviridae* elements of LTR-retrotransposons. (B) Accumulation patterns of subgroups for *Gypsy* family. (C) Insertion patterns of *Copia* subgroups. The x-axis indicates DNA substitution values ranging from 0 to 0.5 and the y-axis represents number of LTR-elements.

and *C. chinense*, insertion patterns of the LTR-retrotransposons were clearly different in the two genomes. Distribution of LTR-retrotransposon in *C. chinense* showed similar insertion pattern with that of *C. annuum* until speciation time between the two genomes at the substitution value of 0.01. However, during the speciation period, LTR-retrotransposons were explosively accumulated in the *C. chinense* genome (Figure 10a).

Considering subgroup of LTR-retrotransposons, species-specific insertion patterns were discovered in each pepper genome. In the *C. baccatum*, athila was also explosively accumulated and peaked before speciation time between *C. baccatum* and the other peppers (Figure 10b). In addition, proliferation of del elements occurred during the same time, even though a slightly low number of del elements were accumulated before the substitution value of 0.05 (Figure 10b). For *Copia* family, all subgroups were actively accumulated shortly before speciation between *C. baccatum* and the other peppers. Consequently, these results represent that distinct proliferation of athila as well as the other *Gypsy* subgroups of *C. baccatum* at shortly before its divergence time may have driven speciation. The significantly increased subgroups of *Copia* around speciation time may have also affected species divergence for *C. baccatum*.

Furthermore, insertion pattern of subgroups in *C. annuum* and *C. chinense* was distinctly differed near speciation time of the two species. Mostly, del elements of *C. chinense* were explosively piled up at its divergence time and also accumulation

patterns of all subgroups were also similar to that of del elements (Figure 10). Therefore, although composition of LTR-retrotransposons is more abundant in *C. annuum* than in *C. chinense*, these results indicate that recent proliferation of subgroups of LTR-retrotransposons may have led divergence of *C. annuum* and *C. chinense*.

DISCUSSION

In plant, reference genomes have actively led development of the plant genomic and genetic research providing novel information for whole genome and genes. To date, the rapidly accumulated genome sequences have represented entire genomic characteristics of the species as well as the genus. In this study, *de novo* genome sequencing of the two pepper genomes (*C. chinense* and *C. baccatum*) was performed to construct high-quality multiple reference genomes for the genus *Capsicum*. Comparing to quality of the reference pepper genome, the newly constructed genomes have high-quality considering N50 value of genome assembly and amount of anchored scaffolds into 12 chromosomes (Table 3 and Table 4). Furthermore, simplified but more comprehensive annotation strategy efficiently performed gene annotation and reasonably predicted protein-coding genes (Figure 1 and Table 5). In addition, repeat annotation was performed using initial contigs to represent whole genome and the amount of whole repeat sequences in the pepper genomes were accurately predicted as resources for further analysis (Table 9-11).

Comparative genome analysis accounts for large fraction of genomic changes derived from genome evolution of both within- and between-species. Speciation time was estimated and species tree was constructed through comparative gene analysis of the pepper genomes with the sequenced plant genomes (Table 6). Compared to the previous reports[42-46], the speciation times were accurately

calculated, and based on the species tree, *C. chinense* and *C. baccatum* were diverged from the *Capsicum* species at 1.1 and 1.7 MYA, respectively. Considering the speciation time, these results coincide that *C. annuum* and *C. chinense* are closely related species and contained in the same complex, whereas *C. baccatum* is included in a distinct complex (Figure 5). The other results caused by genome evolution, change of gene repertories among the pepper genomes was occurred and several gene families were expanded in the respective genomes. The expanded gene families were actively duplicated around the speciation time of the each pepper species (Figure 7).

The previous study[19] reported that the pepper genome was expanded as four-fold of the tomato genome mainly by del subgroup of *Gypsy* family. In the peppers, 1Gb of sequences was more represented in *C. baccatum* compared to *C. chinense*. In addition, The *C. baccatum* genome was also larger than the *C. annuum* genome as around 600 Mb (Table 3). In the *C. baccatum*, athila subgroup of *Gypsy* family has been and caused genome size variation between the *C. baccatum* and the other peppers (Table 9). The insertion pattern of subgroups of LTR-retrotransposons showed that athila subgroup has been explosively inserted in *C. baccatum* during the speciation time compared to the other pepper genomes (Table 8 and Figure 10b). These results may indicate that the distinct accumulation of athila family during the speciation mainly regulated genome expansion of *C. baccatum* and played a role for divergence of *C. baccatum* from the *Capsicum* species.

Slight size difference between *C. annuum* and *C. chinense* genomes was discovered and this may be due to similar composition of TEs in both genomes (Table 8). However, the accumulation pattern of LTR elements in the both genomes was significantly different. Interestingly, all subgroups of LTR elements in the *C. chinense* were explosively inserted during speciation time between *C. annuum* and *C. chinense*. Although, the total amount of LTR elements was larger in *C. annuum* than in *C. chinense*, more recent accumulation of the LTR elements was actively occurred in *C. chinense*. These results represent that recent proliferation of all LTR-retrotransposons in the *C. chinense* genome was associated with the divergence of *C. annuum* and *C. chinense*.

Even though these results provide that accumulation of LTR-retrotransposons was correlated with speciation of the pepper species, they could not explain how the accumulation of LTR-retrotransposons have led speciation of the peppers. Thus further studies will be needed to identify how LTR-retrotransposons have affected divergence of the pepper species including insertion effects of LTR retrotransposons for genic regions which change gene expression patterns or carry gene fragments during speciation time of the pepper species. In conclusion, the newly constructed genome sequences of *C. baccatum* and *C. chinense* will play a major role as reference genomes of the genus *Capsicum* combined with the pre-existing pepper genome for comparative, population genomic, and evolutionary studies of pepper.

REFERENCES

1. Perry, L. *et al.* Starch fossils and the domestication and dispersal of chili peppers (Capsicum spp. L.) in the Americas. *Science* **315**, 986-988 (2007).
2. Onus, A.N. & Pickersgill, B. Unilateral incompatibility in Capsicum (Solanaceae): Occurrence and taxonomic distribution. *Annals of Botany* **94**, 289-295 (2004).
3. Rodriguez, J.M., Berke, T., Engle, L. & Nienhuis, J. Variation among and within Capsicum species revealed by RAPD markers. *Theor. Appl. Genet.* **99**, 147-156 (1999).
4. Pickersgill, B. The Genus Capsicum - a Multidisciplinary Approach to the Taxonomy of Cultivated and Wild Plants. *Biologisches Zentralblatt* **107**, 381-389 (1988).
5. Smith, P.G. & Heiser, C.B. Taxonomic and Genetic Studies on the Cultivated Peppers, Capsicum-Annuum L and C-Frutescens L. *Am. J. Bot.* **38**, 362-368 (1951).
6. Guerra, M. Patterns of heterochromatin distribution in plant chromosomes. *Genet. Mol. Biol.* **23**, 1029-1041 (2000).
7. Pickersgill, B. Genetic resources and breeding of Capsicum spp. *Euphytica* **96**, 129-133 (1997).
8. Ammiraju, J.S.S. *et al.* Dynamic Evolution of Oryza Genomes Is Revealed by Comparative Genomic Analysis of a Genus-Wide Vertical Data Set. *Plant Cell* **20**, 3191-3209 (2008).
9. Paterson, A.H., Freeling, M., Tang, H. & Wang, X. Insights from the comparison of plant genome sequences. *Annu. Rev. Plant Biol.* **61**, 349-372 (2010).
10. Bennetzen, J.L. Patterns in grass genome evolution. *Curr. Opin. Plant Biol.* **10**, 176-181 (2007).
11. Haudry, A. *et al.* An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* **45**, 891-898 (2013).
12. Hu, T.T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476-481 (2011).

13. Gan, X. *et al.* Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477**, 419-423 (2011).
14. Chen, J. *et al.* Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. *Nat. Commun.* **4**, 1595 (2013).
15. Feschotte, C., Jiang, N. & Wessler, S.R. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329-341 (2002).
16. Carroll, S.B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25-36 (2008).
17. Seehausen, O. *et al.* Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176-192 (2014).
18. Ibiza, V.P., Blanca, J., Canizares, J. & Nuez, F. Taxonomy and genetic diversity of domesticated Capsicum species in the Andean region. *Genet. Resour. Crop Evol.* **59**, 1077-1088 (2012).
19. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nat. Genet.* **46**, 270-278 (2014).
20. Boiteux, L.S. & Deavila, A.C. Inheritance of a Resistance Specific to Tomato Spotted Wilt Tospovirus in Capsicum-Chinense Pi-159236. *Euphytica* **75**, 139-142 (1994).
21. Mahasuk, P., Taylor, P.W.J. & Mongkolporn, O. Identification of Two New Genes Conferring Resistance to Colletotrichum acutatum in Capsicum baccatum. *Phytopathology* **99**, 1100-1104 (2009).
22. Kim, S. *et al.* Integrative structural annotation of de novo RNA-Seq provides an accurate reference gene set of the enormous genome of the onion (Allium cepa L.). *DNA Res.***22**, 19-27 (2015).
23. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).
24. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
25. Kelley, D.R., Schatz, M.C. & Salzberg, S.L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **11**, R116 (2010).

26. Schatz, M.C., Witkowski, J. & McCombie, W.R. Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.* **13**, 243 (2012).
27. Zhong, S. *et al.* High-throughput illumina strand-specific RNA sequencing library preparation. *Cold Spring Harb. Protoc.* **2011**, 940-9 (2011).
28. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
29. Arabidopsis Genome, I. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).
30. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641 (2012).
31. Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130-D135 (2012).
32. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435-W439 (2006).
33. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
34. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).
35. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178-2189 (2003).
36. The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-95 (2011).
37. Huala, E. *et al.* The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* **29**, 102-105 (2001).
38. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463-467 (2007).
39. Goff, S.A. *et al.* A draft sequence of the rice genome (Oryza sativa L. ssp. japonica).

Science **296**, 92-100 (2002).

40. Loytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155-170 (2014).
41. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969-1973 (2012).
42. Wu, F. *et al.* A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus Capsicum. *Theor. Appl. Genet.* **118**, 1279-1293 (2009).
43. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641 (2012).
44. Kritsas, K. *et al.* Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. *Genome Res.* **22**, 2455-2466 (2012).
45. Wu, F. & Tanksley, S.D. Chromosomal evolution in the plant family Solanaceae. *BMC Genomics* **11**, 182 (2010).
46. Krom, N. & Ramakrishna, W. Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, Arabidopsis, and populus. *Plant Physiol.* **147**, 1763-1773 (2008).
47. Kondrashov, F.A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. Biol. Sci.* **279**, 5048-5057 (2012).
48. Sanzol, J. Dating and functional characterization of duplicated genes in the apple (Malus domestica Borkh.) by analyzing EST data. *BMC Plant Biol.* **10**, 87 (2010).
49. Yang, T.J. *et al.* Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of Brassica rapa. *Plant Cell* **18**, 1339-1347 (2006).
50. Cui, L. *et al.* Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738-749 (2006).
51. Qin, C. *et al.* Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proc. Natl. Acad. Sci. USA* **111**, 5135-5140 (2014).
52. Prokopowich, C.D., Gregory, T.R. & Crease, T.J. The correlation between rDNA

copy number and genome size in eukaryotes. *Genome* **46**, 48-50 (2003).

53. Kwon, Jin-Kyung, and Byung-Dong Kim. Localization of 5S and 25S rRNA genes on somatic and meiotic chromosomes in Capsicum species of chili pepper. *Mol. cells* **27**, 205-209 (2009).

54. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43-45 (1998).

ABSTRACT IN KOREAN

고추 (*Capsicum* spp.)는 널리 재배되는 매운맛을 함유한 작물이며 전세계 요리의 주된 재료이다. 본 연구에서는, 유전체 어셈블리, 어노테이션 그리고 가상염색체를 포함한 매운 고추 (*Capsicum annuum* CM334)의 고품질 표준 유전체가 구축되었다. 전체 3.48 Gb 중 3.06 Gb의 유전체가 어셈블이 되었고 이중, 2.63Gb가 12개의 염색체로 통합되었다. 완성된 BAC 서열과 비교하여, 어셈블리된 유전체 서열이 검증되었고 그 검증 결과는 어셈블리된 유전체와 완성된 BAC 서열간의 유사도가 99%이상임을 보여주었다. 전체 34,903개의 유전자와 그 유전자들의 생물학적 기능이 예측되었다. 고추의 다중 표준유전체 구축을 위하여 작물화된 2개의 고추 (*Capsicum chinense* and *Capsicum baccatum*)의 고품질 유전체 서열을 구축하였고 이미 완성된 고추 유전체와의 비교 유전체 분석을 수행하였다. 계통발생 분석을 통하여 고추속 식물의 종 분화가 약 100만년에서 200만년 사이에 발생하였다는 것을 밝혀내었다. *C. baccatum* 과 다른 고추들의 유전체 크기의 차이는 극도로 축적된 Athila 반복서열 그룹에 의하여 발생하였다. LTR-retrotransposon 반복서열의 축적패턴을 통하여 살펴봤을 때 *C.*

*chinense*와 *C. baccatum*의 LTR 반복서열의 축적은 명확하게 종 분화 시점에 증가를 하는 것을 확인하였다. 유전자 군의 비교 및 진화분석 결과, 각 고추 종에서 확장된 유전자 군은 최근 유전자 중복에 의해 종 분화 시점에 활발히 그 수가 증가하였음이 밝혀졌다. 결론적으로, 다중 고추 유전체는 비교 유전체 및 진화 연구를 위한 중요한 자원으로서의 역할을 할 것이며 더 나아가 고추 속 식물의 집단 유전체학 및 육종 연구를 위한 필수적인 원천을 제공할 것이다.