GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Short Communication

Discovery of a large set of SNP and SSR genetic markers by high-throughput sequencing of pepper (*Capsicum annuum*)

M. Nicolaï[1], C. Pisani[1], J.-P. Bouchet[1], M. Vuylsteke[2,3] and A. Palloix[1]

[1] INRA, PACA, UR1052, Unité de Génétique et Amélioration des Fruits et Légumes, Domaine Saint Maurice, Montfavet Cedex, France
[2] Department of Plant Systems Biology, VIB, Ghent University, Ghent, Belgium
[3] Department of Plant Biotechnology and Genetics, Ghent University, Ghent, Belgium

Corresponding author: A. Palloix
E-mail: alain.palloix@avignon.inra.fr

Genet. Mol. Res. 11 (3): 2295-2300 (2012)
Received November 11, 2011
Accepted March 12, 2012
Published August 13, 2012
DOI http://dx.doi.org/10.4238/2012.August.13.3

ABSTRACT. Genetic markers based on single nucleotide polymorphisms (SNPs) are in increasing demand for genome mapping and fingerprinting of breeding populations in crop plants. Recent advances in high-throughput sequencing provide the opportunity for whole-genome resequencing and identification of allelic variants by mapping the reads to a reference genome. However, for many species, such as pepper (*Capsicum annuum*), a reference genome sequence is not yet available. To this end, we sequenced the *C. annuum* cv. "Yolo Wonder" transcriptome using Roche 454 pyrosequencing and assembled *de novo* 23,748 isotigs and 60,370 singletons. Mapping of 10,886,425 reads obtained by the Illumina GA II sequencing of *C. annuum* cv. "Criollo de Morelos 334" to the "Yolo Wonder" transcriptome allowed for SNP identification. By setting a threshold value that allows selecting reliable SNPs with minimal loss of information, 11,849 reliable SNPs spread across 5919 isotigs were identified. In addition, 853 single sequence repeats were obtained. This information has been made available online.

Key words: High-throughput sequencing; Transcriptome; Pepper; Single-nucleotide polymorphism; Single-sequence repeats

INTRODUCTION

Genetic association analysis and genomic selection require thousands of genetic markers that cover a genome. Recent advances in sequencing technology have provided the opportunity to generate numerous sequences. Mapping these sequences to a reference genome allows the identification of allelic variants. Such a strategy has been applied successfully in the transcriptome analyses of eucalyptus (*Eucalyptus grandis*) (Novaes et al., 2008) and maize (*Zea mays*) (Barbazuk et al., 2007) and the genomic analysis of soybean (*Glycine max*) (Wu et al., 2010). The best trade-off for massive SNP detection at a reasonable cost in organisms for which genomes have not yet been sequenced has proven to be the combination of Roche 454 GS-FLX and Illumina genome analyzer (GA) sequencing (Hyten et al., 2010). Given that the pepper (*Capsicum annuum* L.) genome has not yet been sequenced, and only 1 incomplete expressed sequence tag database is available (Kim et al., 2008), we used these strategies to sequence the pepper transcriptome and obtain single-nucleotide polymorphism (SNP) and single-sequence repeat (SSR) libraries.

Pepper is a major vegetable crop worldwide, particularly in the tropics. The main breeding objectives, like those in many other crop plants, include adaptation to abiotic stresses, improvement in fruit quality and yield, and resistance to pests and diseases. During the last 20 years, various molecular markers have been developed (Lee et al., 2004; Minamiyama et al., 2006; Wu et al., 2009). The objective of this study was to provide additional genetic markers that can be used for genome mapping, fingerprinting of breeding populations, and germplasm collection in pepper.

MATERIAL AND METHODS

Plants were grown in growth chambers under a 12-h/12-h light/dark period at 24°/20°C. RNA was extracted using the TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) from 7 tissues (apex from stressed and not stressed plants, youngest internodes, flowers, roots, fruit pericarps, and seeds) and combined in equal proportions to maximize the sample diversity of transcriptional units. First-strand complementary DNA (cDNA) synthesis was primed with a hexanucleotide randomized primer and Moloney murine leukemia virus reverse transcriptase (M-MLV RT). Adapters were ligated at the 5'- and 3'-ends and the cDNAs were amplified. cDNA was normalized with 1 cycle of denaturation and reassociation. Reassociated double-stranded cDNA was eliminated from the remaining single-stranded cDNA (normalized cDNA). High-density sequencing was carried out by the GATC Biotech Company (Tübingen, Germany).

The sequences of ‘Yolo Wonder’ were generated via Roche 454 pyrosequencing and assembled with the Newbler software option “cDNA assembly using the optimized algorithm” (Roche Applied Science). The short reads of ‘Criollo de Morelos 334’ were generated with the Illumina GA II and mapped to the "Yolo Wonder" reference transcriptome using Burrows-Wheeler transform-based alignment (Li and Durbin, 2009), which created a sequence alignment/map (SAM) (Li et al., 2009). A pileup file was created with SAMtools and used for SNP detection. For SNP validation, the presence of putative SNPs was analyzed in 74 sequences. Based on the "Yolo Wonder" Roche 454 transcriptome reference sequence, PCR primers were designed to the flanking sequences of SNPs, and the amplified fragments were sequenced with both forward and reverse primers. The SSR motifs were identified with the MIcroSAtellite (MISA) mode of the SSR Classification and Investigation by the Robert Kofler (SciRoKo) software (Kofler et al., 2007).

RESULTS AND DISCUSSION

In this study, we sequenced the transcriptome of 2 *C. annuum* inbred lines that have been used as parental lines of mapping populations (Barchi et al., 2007): 'Yolo Wonder', a homozygous line with a large and sweet fruit, and Criollo de Morelos 334, a homozygous line issued from a small, hot-fruited pepper landrace from Mexico. The Roche 454 sequencing of the 'Yolo Wonder' transcriptome generated 578,460 reads with an average of 308 bp spanning 175.7 Mb of sequence (Table 1). The transcriptome was assembled *de novo* with the GS Roche Assembler Newbler that, unlike other assemblers, creates isotigs from contigs that are consistently connected by a subset of reads and correspond to alternative transcripts (owing to splicing variants). Contigs or isotigs that shared reads were put into the same isogroup. A total of 494,850 reads (85.5%) was assembled into 84,118 reference sequences, including 23,748 isotig sequences with an average of 672 bp (grouped in 19,418 isogroups) and 60,370 singleton sequences (see Table 1). These data are available on the Sol Genomics Network (http://solgenomics.net) and are listed in Supplementary Dataset 1.

Table 1. Data summary of Yolo Wonder and Criollo de Morelos 334 transcriptome sequencing, and discovery of SNPs.

	Number
Roche 454 sequencing of Yolo wonder transcriptome	
Reads	578,460 corresponding to 175.7 Mb
Assembled reads	494,850 corresponding to 85.5% reads
Isotigs	23,748 with an average of 672 bp
Singletons	60,370
Illumina GA sequencing of Criollo de Morelos 334 transcriptome	
Reads	23,745,693 corresponding to 2513 Mb
Mapped reads	10,886,425 corresponding to 45.8% reads
SNPs discovery	
Putative SNPs in	
Isotigs	128,504
Singletons	90,185
Reliable SNPs in	
Isotigs	11,849 present in 5919 isotigs

The average length of the contigs (or isotigs in our case) is larger - namely 100 bp for maize (Barbazuk et al., 2007), 247 bp for eucalyptus (Novaes et al., 2008), and 500 bp for lodgepole pine (*Pinus contorta*) (Parchman et al., 2010) (mean = 672 bp) - than lengths obtained previously, probably owing to improvements in the Roche 454 instrumentation for longer reads (Parchman et al., 2010) and the enhanced performance of the assembler software. Assuming that the number and length of genes are similar in pepper and *Arabidopsis*, we estimated that the coverage of the Yolo Wonder transcriptome was 4.29X, which does not allow us to conclude that the whole transcriptome has been covered. Comparisons to the TAIR9 database and the tomato unigene library indicate that an important part of the 'Yolo Wonder' transcribed genes have been sequenced. From the Illumina GA sequencing of the 'Criollo de Morelos 334' transcriptome, 23,745,693 reads were obtained spanning 2513 Mb (see Table 1). In total, 10,886,425 reads (45.8%) were mapped to the 'Yolo Wonder' transcriptome sequences using the Burrows-Wheeler transform-based alignment.

This mapping yielded 218,689 putative SNPs: 128,504 SNPs in isotigs and 90,185 SNPs in singletons (see Table 1). The reliability of these putative SNPs was estimated with quality values calculated in the MAQ program (Li et al., 2008), which estimates the error probability for each read alignment. This quality index includes the frequency of base calls, which is critical for evaluating the reliability of the SNP in that position. To establish a threshold value, we checked for the presence of putative SNPs in 74 sequences homologous to genes involved in the cell cycle and growth of plants or fruit in various species. A total of 93 SNPs were found in 48 sequences. All of these SNPs had a quality value higher than 100, whereas none of the SNPs with values lower than 78 were valid. Therefore, we defined a quality value of 100 as a cutoff for reliable SNP identification, minimizing information loss. As a result, only 11,849 SNPs (9.2%) were classified as reliable from a total of 128,504 putative SNPs identified in isotig sequences. These SNPs occurred in 5919 isotigs (24.9%; see Table 1 and Supplementary Dataset 2, also available in the Sol Genomics Network), and they provide a high-density marker set for further genetic mapping or diversity studies within cultivated pepper species - the major genetic pool used in breeding programs - using the VeraCode or Illumina GoldenGate technologies on a BeadXpress platform, as carried out for common bean and pea (*Pisum sativum*) (Deulvot et al., 2010; Hyten et al., 2010).

In the present study, we also screened the 'Yolo Wonder' transcriptome for the presence of SSRs using the MISA mode in the SciRoKo software. The minimum repeat number considered was 7 for dinucleotides and 5 for tri-, tetra-, penta-, and hexanucleotides. Mononucleotide SSRs were excluded because of the frequent homopolymer errors found in the Roche 454 pyrosequencing data. A total of 853 SSRs, mainly trinucleotides (502), were obtained (Table 2). This result agrees with results reported earlier for SSRs in pepper (Yi et al., 2006; Portis et al., 2007), maize and rice (Chin et al., 1996; Temnykh et al., 2000), and *Homo sapiens* (Subramanian et al., 2003). SAMtools identified putative insertions or deletions (INDELs) of 2 to 5 bases in the 'Criollo de Morelos 334' Illumina GA sequences compared with the 'Yolo Wonder' transcriptome. A total of 3470 isotigs contained a putative INDEL of 2 to 5 bases. The identical sequence position of these putative INDELs and of the SSRs directly confirmed the presence of polymorphism for 27 SSRs: 13 dinucleotides, 12 trinucleotides, and 2 tetranucleotides. Because the size of detected INDELs was restricted to 5 bases, only single SSR motif polymorphism (and double for dinucleotide motifs) was confirmed.

In conclusion, 11,849 reliable SNPs were found in 5919 isotigs in addition to 853 putative SSRs. On average, SNPs are present in 1 of 4 transcribed sequences between *C. annuum* parents. This large set of transcribed SNPs, expected to be distributed across the entire genome and representing a large diversity of gene functionality, will be extraordinarily valuable in fine-mapping quantitative trait loci, genome-wide association analyses, and genomic selection in the cultivated genetic pool of pepper and related *Capsicum* species.

Table 2. Microsatellites detected in isotigs.

A			**B**		
Motif	No. of SSRs found	No. of SSRs confirmed	Isotig No.	SSR motif	INDELs
Dinucleotide	145	13	isotig02592	(AC)$_{7}$	+AC
Trinucleotide	502	12	isotig07787	(AT)$_{10}$	-AT
Tetranucleotide	139	2	isotig12165	(AT)$_{10}$	-AT
Pentanucleotide	32		isotig18730	(AT)$_{7}$	-AT
Hexanucleotide	35		isotig19218	(AT)$_{7}$	+AT
Total	853	26	isotig08903	(CT)$_{15}$	-CT
			isotig02154	(GA)$_{7}$	-AG
			isotig10700	(GA)$_{7}$	-GA
			isotig14361	(TA)$_{7}$	-TA
			isotig17254	(TA)$_{8}$	-TA
			isotig22695	(TA)$_{9}$	+TA
			isotig04064	(TG)$_{7}$	-TG
			isotig16420	(TG)$_{7}$	+TG
			isotig13286	(AGC)$_{5}$	-AGC
			isotig22913	(CAT)$_{6}$	+CAT
			isotig10782	(CTT)$_{6}$	+CTT
			isotig16001	(GAA)$_{5}$	-GAA
			isotig23459	(GTT)$_{5}$	-GTT
			isotig12681	(TAA)$_{5}$	-TAA
			isotig15734	(TAA)$_{8}$	+TAA
			isotig04382	(TAT)$_{8}$	+TAT
			isotig04383	(TAT)$_{8}$	+TAT
			isotig12716	(TAT)$_{6}$	+TAT
			isotig14363	(TCT)$_{8}$	-TCT
			isotig09681	(TTC)$_{11}$	+TC
			isotig01958	(ATGT)$_{6}$	-ATGT
			isotig01960	(ATGT)$_{6}$	-ATGT

A. Number of di-, tri-, tetra-, penta-, and hexanucleotide simple sequence repeats (SSRs) occurring in isotigs, and of validated SSRs by comparison with the putative INDELs. **B.** Isotig number, SSR motif found in Yolo Wonder sequences, INDELs ("-" deleted or "+" inserted motifs in Criollo de Morelos 334 sequences).

ACKNOWLEDGMENTS

We thank Christophe Meynier, Nastasja Trunk, and Ulrike Schöck of GATC Biotech Company for the Roche 454 and Illumina GA sequencing and assembly, Véronique Lefebvre for discussion and critical review of the manuscript, and Anne-Marie Sage-Palloix, Ghislaine Nemouchi, and Bruno Savio for the pepper germplasm and plant maintenance. Research supported by the Smart tools for Prediction and Improvement of Crop Yield (SPICY) European project (#KBBE-2008-211347).

REFERENCES

Barbazuk WB, Emrich SJ, Chen HD, Li L, et al. (2007). SNP discovery via 454 transcriptome sequencing. *Plant J.* 51: 910-918.

Barchi L, Bonnet J, Boudet C, Signoret P, et al. (2007). A high-resolution, intraspecific linkage map of pepper (*Capsicum annuum* L.) and selection of reduced recombinant inbred line subsets for fast mapping. *Genome* 50: 51-60.

Chin EC, Senior ML, Shu H and Smith JS (1996). Maize simple repetitive DNA sequences: abundance and allele variation. *Genome* 39: 866-873.

Deulvot C, Charrel H, Marty A, Jacquin F, et al. (2010). Highly-multiplexed SNP genotyping for genetic mapping and germplasm diversity studies in pea. *BMC Genomics* 11: 468.

Hyten DL, Song Q, Fickus EW, Quigley CV, et al. (2010). High-throughput SNP discovery and assay development in common bean. *BMC Genomics* 11: 475.

Kim HJ, Baek KH, Lee SW, Kim J, et al. (2008). Pepper EST database: comprehensive *in silico* tool for analyzing the chili pepper (*Capsicum annuum*) transcriptome. *BMC Plant Biol.* 8: 101.

Kofler R, Schlotterer C and Lelley T (2007). SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics* 23: 1683-1685.

Lee JM, Nahm SH, Kim YM and Kim BD (2004). Characterization and molecular genetic mapping of microsatellite loci in pepper. *Theor. Appl. Genet.* 108: 619-627.

Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

Li H, Ruan J and Durbin R (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18: 1851-1858.

Li H, Handsaker B, Wysoker A, Fennell T, et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.

Minamiyama Y, Tsuro M and Hirai M (2006). An SSR-based linkage map of *Capsicum annuum*. *Mol. Breed.* 18: 157-169.

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, et al. (2008). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9: 312.

Parchman TL, Geist KS, Grahnen JA, Benkman CW, et al. (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* 11: 180.

Portis E, Nagy I, Sasva Z, Stagelri A, et al. (2007). The design of *Capsicum* spp. SSR assays via analysis of *in silico* DNA sequence, and their potential utility for genetic mapping. *Plant Sci.* 172: 640-648.

Subramanian S, Madgula VM, George R, Kumar S, et al. (2003). SSRD: simple sequence repeats database of the human genome. *Comp Funct. Genomics* 4: 342-345.

Temnykh S, Park WD, Ayres N, Cartinhour S, et al. (2000). Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 100: 697-712.

Wu F, Eannetta NT, Xu Y, Durrett R, et al. (2009). 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.

Wu X, Ren C, Joshi T, Vuong T, et al. (2010). SNP discovery by high-throughput sequencing in soybean. *BMC Genomics* 11: 469-479.

Yi G, Lee JM, Lee S, Choi D, et al. (2006). Exploitation of pepper EST-SSRs and an SSR-based linkage map. *Theor. Appl. Genet.* 114: 113-130.