Single Nucleotide Polymorphisms and Indel Markers from the Transcriptome of Garlic

Michael J. Havey 1
U.S. Department of Agriculture, Agricultural Research Service and Department of Horticulture, University of Wisconsin, Madison, WI 53706

Yul-Kyun Ahn
National Institute of Horticultural and Herbal Science, Rural Development Administration, 100 Nongsaeng-ro, Wanju-gun, Jeollabuk-do 565-852, South Korea

ADDITIONAL INDEX WORDS. Allium sativum, complementary DNA, genetic variation, molecular markers

ABSTRACT. Garlic (Allium sativum) is cultivated worldwide and appreciated for its culinary uses. In spite of primarily being asexually propagated, garlic shows great morphological variation and adaptability to diverse production environments. Molecular markers and phenotypic characteristics have been used to assess the genetic diversity among garlics. In this study, we undertook transcriptome sequencing from a single garlic plant to identify molecular markers in expressed regions of the garlic genome. Garlic sequences were assembled and selected if they were similar to monomorphic sequences from a doubled haploid (DH) of onion (Allium cepa). Single nucleotide polymorphisms (SNPs) and insertion–deletion (indel) events were identified in 4355 independent garlic assemblies. A sample of the indels was verified using Sanger sequencing from the original complementary DNA (cDNA) library and genomics DNAs from diverse garlics, and segregations confirmed by sexual progenies of garlic. These molecular markers from the garlic transcriptome should be useful for estimates of genetic diversity, identification and removal of duplicate accessions from germplasm collections, and the development of a detailed genetic map of this important vegetable crop.

Garlic is cultivated worldwide and widely used for culinary purposes. As garlic cultivation spread from its origin in central Asia (Takagi, 1990) across the globe, this asexually propagated plant was selected for diverse production environments through the movement of genetically diverse clones that arose through mutation and/or occasional production of sexual progenies (Etoh and Simon, 2002). Viable true seed of garlic has been observed in the wild (Etoh, 1986) and can be produced from controlled crosses, opening the door for genetic improvement of this important plant (Jenderek and Hannan, 2004; Pooler and Simon, 1994; Simon and Jenderek, 2003).

Garlic is a member of the family Alliaceae in the monocot order Asparagales, a monophyletic group sister to the commelinid monocots [grasses, banana, palms, among others (Chase et al., 2000)]. The Asparagales is the second most economically important order in the monocots after the Poales (grasses) and includes food and ornamental plants such as the alliums [chive (Allium schoenoprasum), garlic, leek (Allium ampeloprasum), and onion], asparagus (Asparagus officinalis), agave (Agave sp.), and iris (Iris sp.). Plants in the Asparagales tend to have large nuclear genomes (Leitch et al., 2005); garlic is a diploid 1C nucleus (Bennett and Leitch, 2012). For with extremely large nuclear genomes, sequencing of expressed regions is an efficient approach to identify genes and avoid issues with repetitive DNAs (Gore et al., 2009; Kuhl et al., 2004; Newman et al., 1994). Transcriptome sequencing of garlic has been undertaken. Kim et al. (2009) used Sanger sequencing to produce 21,595 reads from four non-normalized cDNA libraries synthesized from leaf and stem tissues of garlic. Sun et al. (2012) completed Illumina sequencing of cDNAs from garlic buds to generate over 127,000 assembled sequences, of which about one-third showed significant similarities to sequences in databases. Kamenetsky et al. (2015) also used the Illumina platform for sequencing of non-normalized cDNA libraries from six to produce expression profiles for fertile garlic.

Transcriptome sequencing can be used to reveal polymorphisms in expressed regions of larger genomes useful for diversity analyses and development of genetic maps (Duangjit et al., 2013; Gore et al., 2009; Ipek et al., 2015; Kuhl et al., 2004; Martin et al., 2005). Molecular markers, such as amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), and SNPs, have been identified in garlic and sexual progenies used to produce the first rudimentary genetic maps of this important plant (Ipek et al., 2005, 2015; Ma et al., 2009; Zewdie et al., 2005; Zhao et al., 2011). We completed Sanger and 454 sequencing from a normalized cDNA library synthesized from a single garlic plant and identified SNPs and indel events in the garlic genome, to provide the first large set of sequence-characterized molecular markers for garlic.

Materials and Methods

CDNA SYNTHESIS, NORMALIZATION, AND SEQUENCING. Garlic accession W6 1861 (cultivar Dunganski) from the U.S. Department of Agriculture (USDA) PI station (Pullman, WA) was selected as a representative of purple hardneck garlic (Volk et al., 2004). Equal fresh weights of tissue from leaves, pseudostems, and roots from a single plant of ‘Dunganski’ were harvested,
combined, and frozen in liquid nitrogen. Total RNA was isolated from these combined tissues. Synthesis of cDNAs and normalization of the library were completed by CoreBioSystem (Seoul, South Korea) using proprietary approaches. After normalization, the cDNAs were directionally cloned into the pCNS.D-2 vector (Oh et al., 2004). Individual colonies were picked using blue-white selection, plasmids purified, and 2500 random inserts were Sanger sequenced from the 5' end as previously described (Kuhl et al., 2004). For next-generation sequencing, plasmids were purified through CsCl-EtBr gradients (Sambrook et al., 1989) and the cDNAs released from the vector by double digestions with EcoRI and NotI as recommended by the manufacturer (New England Biolabs, Gaithersburg, MD). The digested DNA was run through a 1.0% agarose gel overnight at 7 mA. The gel was briefly stained in EtBr, fragments visualized in ultraviolet light, and the gel region carrying fragments between 500 base pair (bp) and 2 kbp was harvested. DNA was extracted from the gel using a gel extraction kit (Qiagen, Valencia, CA). DNA was ethanol precipitated, washed with 76% ethanol, dried, and dissolved in sterile distilled water. cDNAs were sequenced using 1.5 plates of the Roche (Branford, CT) 454-FLX platform as recommended by the manufacturer.

**Bioinformatic analyses.** Sanger and 454 reads from the normalized cDNA library were initially assembled using SOAPdenovo-trans version 1.03 (Xie et al., 2014) with parameters max_rd_length = 1288 and map_len = 32 and the spliced transcripts alignment to a reference [STAR (Dobin et al., 2013)] software. SNPs and indels were identified after alignment using BWA 0.7.7 (Li and Durbin, 2009) with default parameters, removal of duplicates using Picard (2015) with default parameters, and realignment using with base quality score recalibration and default parameters. GATK 3.1-1 UnifiedGenotyper (DePristo et al., 2011) with default parameters was then used to identify SNP and indels among the assembled garlic sequences.

The 454 reads from a normalized cDNA library from onion DH 2150 (Baldwin et al., 2012) were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archives (SRA) srx031644 and srx031645 and assembled using the Roche gsAssembler 2.8 with default parameters. These assembled sequences and Illumina reads from SRA srx392929, srx393053, srx393054, srx393055, srx393056, and srx393057 (all from DH 2150) were aligned, duplicates removed, and realigned as described above. GATK 3.1-1 UnifiedGenotyper with default parameters was used to identify SNP and indels. All assembled cDNA sequences from DH 2150 with polymorphisms on greater than two independent reads with the variant state were considered as paralogs and removed.

To select against SNPs or indels among expressed paralogs in garlic, the garlic and onion assemblies were compared using NCBI-blast 2.2.28. Blast hits (<1e-2) were assumed to represent homologous sequences between garlic and onion. Garlic assemblies with a putative homolog from DH2150 carrying polymorphism(s) were discarded as potentially paralogous. SNPs or indels in the remaining garlic assemblies with quality scores above 30 were selected and 60 to 100 bp of sequence flanking the polymorphism were extracted.

Pairs of oligonucleotide primers flanking indels in 48 independent assemblies (Supplemental Table 2) were designed using Batch Primer 3 (You et al., 2008). Amplicons were produced by the polymerase chain reaction (PCR) using genomic DNAs from ‘Dunganski’ and USDA Pls 493098 (USDA, 1985b) as representatives of softneck garlic, as well as the original normalized cDNA library from ‘Dunganski’ (described above). PCR conditions were 1x PCR Buffer (Promega, Madison WI), 2.5 mM MgCl2, 0.25 mM dNTPs, 1 μM each primer, 1.25 units of Taq polymerase, and 100 ng garlic genomic DNA. PCR cycling was 95 °C for 5 min; 30 cycles of 95 °C for 30 s, 52 °C for 1 min, and 72 °C for 45 s; and one final cycle at 72 °C for 15 min. Amplicons were run on 9% polyacrylamide gels, silver stained (Sambrook et al., 1989), and visually scored for size differences. DNAs were isolated from individual sexually produced seed of garlic, originating from random pollination among a diverse collection of flowering garlic plants (gift of D. Brotslaw, Sensient Natural Flavorings, Turlock, CA) using the NucleoSpin Plant II kit (Macherey-Nagel, Düren, Germany) and concentrations established spectrophotometrically. The same PCR and electrophoresis conditions as described above were used for indel markers 83390 and 83548 (Supplemental Table 2) to establish segregation among sexual progenies.

**Results and Discussion.** **Garlic transcriptome sequencing.** We undertook Sanger and 454 sequencing of garlic cDNAs because of the longer reads produced by these technologies relative to the Illumina platform used by others (Kamenetsky et al., 2015; Sun et al., 2012). We produced 2422 Sanger reads from the 5' ends of random clones from the normalized cDNA library, of which 1134 were unique and deposited into GenBank (accessions JZ773807 to JZ774940). Six hundred and twenty (26%) of the random reads were highly similar to garlic virus D, an RNA virus that commonly infects garlic (Wylie et al., 2014) and known to occur in the northwestern United States (Gieck et al., 2009) where the ‘Dunganski’ bulb was produced. Kamenetsky et al. (2015) also reported sequences of garlic viruses in their transcriptome sequencing, but not specifically garlic virus D. To avoid 454 sequencing of this virus, the garlic cDNAs were released from the plasmid vector by restriction enzyme digests and size fractionated through an agarose gel. An obvious band was visible at ≈0.6 kbp, putatively the viral sequence, and cDNA fragments were purified from the gel avoiding this band and the vector. The 454 sequencing of these gel-purified cDNAs produced over 2.2 million reads and 1.1 gigabases of expressed sequence of garlic (SRA accessions SRST28573, SRX740874, and SRR1624682). The Sanger and 454 reads assembled into 35,936 contigs over 100 bp in size totaling 15,314,468 bp. A total of 965,358 of the 454 reads remained as singletons. The average read length of singletons (sequences remaining after assembly) from the Sanger and 454 platforms was 537 bp, longer than the read lengths reported by Sun et al. (2012) (90 bp) and Kamenetsky et al. (2015) (250 bp) who generated paired-end reads using the Illumina platform.

**Polymorphisms in expressed regions of the garlic genome.** To reduce the numbers of polymorphisms identified among expressed paralogs in garlic, we selected assemblies showing similarities to monomorphic cDNA sequences from a DH onion. Expressed sequences from DH 2150 (Baldwin et al., 2012) were assembled and any assembly with SNPs supported by at least two independent reads was rejected, assuming these to be from expressed paralogs. Garlic cDNA assemblies similar (<1e-2) to these monomorphic expressed regions from DH 2150 were selected. A total of 14,879 SNPs and indels were identified in these selected garlic ESTs, of
which 4,355 were on independent assemblies (Supplemental Table 1). We synthesized primers flanking 48 of the indels (Supplemental Table 2), of which 44 produced amplicons from the normalized cDNA library and genomic DNA from 'Dunganski' and genomic DNAs from USDA Plant Introductions 493098 and 497951. Water control (Water) contained no DNA. Expected amplicon sizes from the cDNA library were ±100 bp or slightly larger.

Fig. 1. Polyacrylamide gel of amplicons carrying indels identified in the garlic transcriptome. Sizes in base pairs of marker (SM) are shown on left. Numbers at bottom refer to markers listed in Supplemental Table 2. DNAs were the normalized cDNA library and genomic DNA from ‘Dunganski’ and genomic DNAs from USDA Plant Introductions 493098 and 497951. Water control (Water) contained no DNA. Expected amplicon sizes from the cDNA library were ±100 bp or slightly larger.

Fig. 2. Segregation among sexual progenies of amplicons carrying indels identified in the garlic transcriptome. Size marker (SM) is 100 bp. Numbers at bottom refer to indel markers listed in Supplemental Table 2.

Assessments of DNA diversity among garlics have primarily used anonymous markers such as RAPDs or AFLPs (Al-Zahim et al., 1997; Bradley et al., 1996; Ipek et al., 2003; Lampasona et al., 2003; Maas and Klaas, 1995; Volk et al., 2004), complicating direct comparisons among independent studies. Relatively few sequenced characterized DNA polymorphisms, such as SSRs and SNPs, have been reported (Ma et al., 2009; Zewdie et al., 2005). Zhao et al. (2011) used six SSRs to assess variation among 613 garlic accessions, and reported a total of 113 “alleles” with an average of 14 “alleles” per SSR. However, it was not clear if these polymorphisms are indeed alleles at one locus or amplicons from paralogous loci, the latter of which would complicate estimates of genetic diversity and relationships. Nevertheless, this study and those of Ma et al. (2009) and Zhao et al. (2011) reveal substantial DNA variation in garlic, consistent with the phenotypic diversity among garlic cultivars (Lallemand et al., 1997; Panthee et al., 2006; Pooler and Simon, 1993). The large number of sequence-characterized polymorphisms identified in this study from the garlic transcriptome should be useful to produce robust estimates of genetic diversity in garlic, aid in the confident identification and removal of duplicate accessions from germplasm collections, and contribute to the development of a detailed genetic map of this important vegetable crop.

Literature Cited

Al-Zahim, M., H.J. Newbury, and B.V. Ford-Lloyd. 1997. Classification of genetic variation in garlic (Allium sativum L.) revealed by RAPD. HortScience 32:1102–1104.

Baldwin, S., R. Revanna, S. Thomson, M. Pither-Joyce, K. Wright, R. Crowhurst, M. Fiers, L. Chen, R. MacKnight, and J. McCallum. 2012. A toolkit for bulk PCR-based marker design from next-generation sequence data: Application for development of a framework linkage map in bulb onion (Allium cepa L.). BMC Genomics 13:637.

Bennett, M.D. and I.J. Leitch. 2012. Plant DNA C-values database (release 6.0). 27 Oct. 2015. <http://www.kew.org/cvalues/>.

Bradley, K.F., M.A. Rieger, and G.G. Collins. 1996. Classification of Australian garlic cultivars by DNA fingerprinting. Austral. J. Exp. Agr. 36:613–618.

Chase, M.W., D.E. Soltis, P.S. Soltis, P.J. Rudall, M.F. Fay, W.H. Hahn, S. Sullivan, J. Joseph, M. Molvray, P.J. Kores, T.J. Givnish, K.J. Sytsma, and J.C. Pires. 2000. Higher-level systematics of the monocotyledons: An assessment of current knowledge and a new classification, p. 3–16. In: K.L. Wilson and S. Morrison (eds.). Monocots: Systematics and evolution 1. CSIRO, Melbourne, Australia.

DePristo, M., E. Banks, R. Poplin, K. Garimella, J. Maguire, C. Hartl, A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, A. McKenna, T. Fennell, A. Kernshtky, A. Sivachenko, K. Cibulskis, S. Gabriel, detected among the garlic genomic DNAs (Fig. 1) and polymorphisms segregated among sexual progenies (Fig. 2).
Lampasona, G.S., L. Martinez, and J.L. Burba. 2003. Genetic diversity among selected Argentinean garlic clones (Allium sativum L.) using AFLP (amplified fragment length polymorphism). Euphytica 132:115–119.

Maass, H.I. and M. Klaas. 1995. Intraspecific differentiation of garlic (Allium sativum L.) by isozyme and RAPD markers. Theor. Appl. Genet. 91:89–97.

Martin, W., J. McCallum, M. Shigyo, J. Jakse, J.C. Kuhl, N. Yamane, M.P. Joyce, A.F. Gokce, K.C. Sink, C.D. Town, and M.J. Havey. 2005. Genetic mapping of expressed sequences in onion and in silico comparisons with rice show scant collinearity. Mol. Genet. Genomics 274:197–204.

Newman, T., F.J. de Bruijn, P. Green, K. Keggestra, H. Kende, L. McIntosh, J. Ohlrogge, N. Raikhel, S. Somerville, M. Thomashow, E. Retzel, and C. Somerville. 1994. Genes galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. Plant Physiol. 106:1241–1255.

Oh, J.H., H.Y. Sohn, J.M. Kim, Y.S. Kim, and N.S. Kim. 2004. Construction of multi-purpose vectors, pCNS and pCNS-D2, are suitable for collection and functional study of large-scale cDNAs. Plasmid 51:217–226.

Panthee, D.R., R.B. KC, H.N. Regmi, P.P. Subedi, S. Bhattarai, and J. Dhakal. 2006. Diversity analysis of garlic (Allium sativum L.) germplasms available in Nepal based on morphological characters. Genet. Resources Crop Evol. 53:205–212.

Picard, 2015. A set of tools for working with next generation sequencing data in the BAM format, version 1.4. 25 Oct. 2015. <http://broadinstitute.github.io/picard/>.

Pooler, M.R. and P.W. Simon. 1994. Characterization and classification of isozyme and morphological variation in a diverse collection of garlic clones. Euphytica 68:121–130.

Pooler, M.R. and P.W. Simon. 1994. True seed production in garlic. Sex. Plant Reprod. 7:282–286.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. Molecular cloning: A laboratory manual. Cold Spring Harbor Press, New York, NY.

Simon, P.W. and M.M. Jenderek. 2003. Flowering, seed production, and the genesis of garlic breeding. Plant Breeding Rev. 23:211–244.

Sun, X., S. Zhou, F. Meng, and S. Liu. 2012. De novo assembly and characterization of the garlic (Allium sativum) bud transcriptome by Illumina sequencing. Plant Cell Rpt. 31:1823–1828.

Takagi, H. 1990. Garlic, Allium sativum L., p. 109–146. In: J.L. Brewster and H.D. Rabinowitch (eds.). Onions and allied crops. CRC Press, Boca Raton, FL.

U.S. Department of Agriculture (USDA). 1985a. PI 493098. 27 Oct. 2015. <http://www.ars-grin.gov/cgi-bin/npgs/acc/display.pl?1388034>.

U.S. Department of Agriculture (USDA). 1985b. PI 497951. 27 Oct. 2015. <http://www.ars-grin.gov/cgi-bin/npgs/acc/display.pl?1392887>.

Volk, G.M., A.D. Henk, and C.M. Richards. 2004. Genetic diversity among U.S. garlic clones as detected using AFLP methods. J. Amer. Soc. Hort. Sci. 129:559–569.

Wylie, S.J., H. Li, M. Saqib, and M. Jones. 2014. The global trade in fresh produce and the vagility of plant viruses: A case study in garlic. PLoS One 9(8):e105044.

Xie, Y., G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T. Lam, Y. Li, X. Xu, G. Wong, and J. Wang. 2014. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666.

You, F.M., N. Huo, Y. Gu, M. Luo, D. Ma, H. Lano, G.R. Lazo, J. Dvorak, and O.D. Anderson. 2008. Batch Primer3: A high throughput web application for PCR and sequencing primer design. BMC Bioinformatics 9:253.

Zewdie, Y., M.J. Havey, J.P. Prince, and M.M. Jenderek. 2005. The first genetic linkages among expressed regions of the garlic (Allium sativum L.) genome. J. Amer. Soc. Hort. Sci. 130:569–574.

Zhao, W.G., J.W. Chung, G.A. Lee, K.H. Ma, H.H. Kim, K.T. Kim, I.M. Chung, J.K. Lee, N.S. Kim, S.M. Kim, and Y.J. Park. 2011. Molecular genetic diversity and population structure of a selected core set in garlic and its relatives using novel SSR markers. Plant Breed. 130:46–54.