DNA Research, 2018, 25(6), 587–596
doi: 10.1093/dnares/dsy027
Advance Access Publication Date: 6 August 2018
Full Paper

OXFORD

Full Paper

Transcriptome-referenced association study of clove shape traits in garlic

Xiaojun Chen[1,†], Xia Liu[2,†], Siyuan Zhu[1], Shouwei Tang[1], Shiyong Mei[1,*], Jing Chen[2], Shan Li[2], Mengdi Liu[2], Yuejiao Gu[2], Qiuzhong Dai[1], and Touming Liu[1,*]

[1]Institute of Bast Fiber Crops and Center of Southern Economic Crops, Chinese Academy of Agricultural Sciences, Changsha 410205, China, and [2]Novogene Bioinformatics Institute, Beijing 100012, China

*To whom correspondence should be addressed. Tel. 86 731 88998555. Fax. 86 731 88998528. Email: liutouming@caas.cn (T.L.); Tel. 86 731 88998578. Fax. 86 731 88998528; hbvegbt@163.com (S.M.)

†These authors contributed equally to this work.

Edited by Prof. Kazuhiro Sato

Received 4 February 2018; Editorial decision 17 July 2018; Accepted 25 July 2018

Abstract

Genome-wide association studies are a powerful approach for identifying genes related to complex traits in organisms, but are limited by the requirement for a reference genome sequence of the species under study. To circumvent this problem, we propose a transcriptome-referenced association study (TRAS) that utilizes a transcriptome generated by single-molecule long-read sequencing as a reference sequence to score population variation at both transcript sequence and expression levels. Candidate transcripts are identified when both scores are associated with a trait and their potential interactions are ascertained by expression quantitative trait loci analysis. Applying this method to characterize garlic clove shape traits in 102 landraces, we identified 22 candidate transcripts, most of which showed extensive interactions. Eight transcripts were long non-coding RNAs (lncRNAs), and the others were proteins involved mainly in carbohydrate metabolism, protein degradation, etc. TRAS, as an efficient tool for association study independent of a reference genome, extends the applicability of association studies to a broad range of species.

Key words: garlic, associative transcriptomics, clove shape

1. Introduction

Garlic (*Allium sativum* L.) is a diploid ($2n = 2x = 16$) species that has been cultivated globally for more than 5,000 yrs as a vegetable, spice, and medicinal plant. The bulb, which is the most widely consumed plant part, consists of several cloves that are abnormal axillary buds. Clove shape traits (CSTs) are economically important quantitative traits. However, garlic has a giant genome[1] and the cultivars are generally sterile, both features hindering research on this crop species.

Plant quantitative traits are typically controlled by several major and minor effect genes that constitute complex regulatory networks.[2] Characterization of quantitative traits is time-consuming and labor-intensive when using traditional quantitative trait loci (QTL) mapping methods that involve the identification and cloning of dozens of trait-control genes. Alternatively, a genome-wide association study (GWAS) provides a powerful tool for the identification of genes underlying complex traits.[3–6] However, to identify candidate trait genes, a reference genome or linkage map of the species under study

is essential for GWAS, which markedly restricts the application of this technology in species with uncharacterized genomes.

Unlike genome analysis, transcriptome analysis by next-generation sequencing is rapid, inexpensive, and unconstrained by genomic complexity.[7] Recently, transcriptomes have been used in association analysis of traits by associative transcriptomics, which extended the genetic association studies to a broad range of species, especially in complex polyploid species.[8,9] However, associative transcriptomics still requires the genome sequence of a progenitor or related species.

Here, we propose a novel method termed 'transcriptome-referenced association study' ('TRAS') that not only identifies trait-associated transcripts in species for which a reference genome sequence is lacking, but also detects potential interactions among these transcripts. In this approach, a transcriptome obtained by single-molecule long-read sequencing is used as a reference sequence for scoring the variation in both sequence (single-nucleotide polymorphisms, SNPs) and expression level (GE) in the population. Both the SNPs and GE are used in an association analysis of a trait to determine candidate transcripts involved in the trait. In addition, SNPs and GE data are used as genotype and phenotype, respectively, in expression quantitative trait loci (eQTL) analysis to ascertain the potential interactions among the trait-control transcripts. To assess the feasibility of TRAS, we applied this method to the characterization of three traits of garlic clove shape.

2. Materials and methods

2.1. Plant material, experimental design, and phenotypic measurements

We collected 92 garlic landraces from China and 10 from other countries (Supplementary Table S1). Thus, in total, 102 landraces were planted in the experimental farm (28°11′49″N, 112°58′42″E) of the Institute of Bast Fiber Crops (IBFC), Chinese Academy of Agricultural Sciences (CAAS), Changsha, China, on 18 September 2014. Field experiments were carried out following a randomized complete block design with two replicates. Thirty-six cloves of garlic for each landrace were planted into a three-row plot, with a distance of 10 cm between the plants and 20 cm between the rows, in each replicate. Thirty plants in the middle of the three rows of each plot were harvested individually in 2015 when their bulbs were ripe. After air drying, the surface of the bulbs, clove length (CL), clove width (CW), and clove thickness (CT) of each bulb were measured using a Vernier caliper.

2.2. Tissue sampling and RNA isolation

Twelve plants of each landrace were grown in one row for tissue sampling. The experimental farm, sowing date, and planting density were the same as those used for phenotypic measurements. Starting from February 2015, the growth of the bulbs was observed in five plants from each row (from the second to the sixth plant) by carefully removing the soil covering the bulbs; the soil around the roots was not disturbed. When all the observed bulbs in one row began to expand, the bulbs of another five plants (from the seventh to the eleventh plant) were collected and pooled as a sample of the corresponding landrace, immediately frozen in liquid nitrogen, and stored at −80 °C until use. Total RNA of each sample was extracted using an E.Z.N.A. Plant RNA Kit (OMEGA Bio-Tek, Norcross, GA, USA) according to the manufacturer's protocol.

2.3. PacBio SMRTbell library construction and single-molecule sequencing

The landrace from Yangxi (China) was selected among the 102 garlic landraces for single-molecule sequencing. Approximately 1 μg of total RNA of developing bulbs was used for reverse transcription (RT) using a Clontech SMARTer cDNA synthesis kit (Clontech Laboratories, Mountain View, CA, USA) and oligo dT primer to generate full-length cDNA. Three RT reactions were run in parallel, and primer IIA from the kit was used for all PCR reactions following RT. The optimal amplification cycle number for generating dsDNA was determined. PCR products were purified with AMPure PB beads (Pacific Biosciences, Menlo Park, CA, USA). The obtained dsDNA was subjected to size selection using the BluePippin System and subsequent re-amplification. Finally, the products were purified and subjected to Iso-Seq SMRTBell library preparation (https://pacbio.secure.force.com/SamplePrep) to yield three libraries (1–2, 2–3, and 3–6 kb), which were sequenced on the PacBio RSII platform using P6-C4 chemistry.

2.4. PacBio sequence data analysis

Sequence data were processed using the SMRT analysis software (www.pacb.com/products-and-services/analytical-software/devnet/). Circular consensus sequence (CCS) reads were generated from the sub-read files using the following parameters: min_length, 300; max_drop_fraction, 0.8; min_passes, 1; min_predicted_accuracy, 0.8. The generated CCS reads were classified into full-length and non-full-length sequences using the parameters ignorepolyA false and minSeqLength 300. The non-full-length and full-length CCS reads were subjected together to isoform-level clustering under default parameter settings. Sequences that were non-redundant and unextended on either end were defined as transcripts. Because PacBio reads have a higher frequency of nucleotide errors than the shorter reads generated by the second-generation sequencing technologies, the software *proovread*[10] was used to correct those errors based on Illumina RNA-sequencing data of the Yangxi landrace of garlic. Redundant sequences were removed with CD-HIT.[11]

2.5. Annotation of the transcriptome

We detected the protein-coding potential of transcripts using Coding Potential Calculator, which is based on the detection of quality, completeness, and sequence similarity of the open reading frame to proteins in current protein databases, with default parameters.[12] Only the transcripts that did not pass the protein-coding-score test were classified as long non-coding RNAs (lncRNAs), and the others were subjected to functional annotation by searching against seven public databases, including the National Center for Biotechnology Information (NCBI) non-redundant (NR) protein sequences, NCBI nucleotide sequences (NT), eukaryotic ortholog groups (KOG), Kyoto Encyclopedia of Genes and Genomes ortholog (KO), Swiss-Prot protein, Gene Ontology (GO), and protein family (PFAM) databases, as described previously.[13] The coding sequence (CDS) of each transcript was predicted by BLAST search against the NCBI non-redundant protein sequence database (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and the SwissProt protein database (http://www.uniprot.org/uniprot/), as well as by using the ESTscan program.[14]

2.6. Illumina RNA-sequencing library construction and sequencing

To characterize the population variation in the SNPs and GE, all 102 landraces were subjected to Illumina RNA sequencing individually. Total RNA from each sample was used to construct cDNA libraries with a fragment length of 250 bp (±25 bp) using a NEBNext® UltraTM RNA Library Prep Kit for Illumina® (New England BioLabs, Ipswich, MA, USA) following the manufacturer's instructions. Paired-end sequencing was performed for each library using a HiSeq PE Cluster Kit v4 cBot (Illumina, San Diego, CA, USA) in conjunction with the Illumina sequencing platform (HiSeq™ 2500) according to the manufacturer's instructions. Adapter sequences were trimmed from the raw sequence reads, and the reads were then subjected to a stringent filtering process to yield clean reads; reads with adaptor contamination, reads with more than 20% of $Q < 20$ bases, and low-quality reads with more than 5% of ambiguous bases (N) were discarded.

2.7. Expression analysis

To quantify the expression of each transcript in all 102 landraces, all clean Illumina reads of each landrace were mapped to the reference transcriptome generated by PacBio sequencing by using Bowtie 2[15] with default parameters. The expression level of each transcript in each sample was analysed by estimating the expected number of fragments per kilobase of transcript sequence per million base pairs sequenced (FPKM) using RSEM.[16] A co-expression analysis of the transcripts in the 102 genotypes was conducted using weighted gene co-expression network analysis (WGCNA)[17] implemented in the R package, and the transcripts that showed co-expression in the population were assigned to a module according to the description of Zhang and Horvath.[18] Briefly, co-expression similarity was determined based on the expression levels of all transcripts in the population. According to the similarity value, the co-expressed transcripts were then clustered as a module with the following parameters: the minimum module size was set to 30 transcripts, and modules were merged if they shared at least 25% similarity. Thereafter, the expression value of each transcript in one co-expressed module was used as the eigenvector to estimate the eigenvalue of this module. The eigenvalue of each module was estimated and was used to analyse the modules' correlation with traits based on Pearson's correlation coefficients, and to identify the trait-related co-expressed module.

2.8. Population SNP detection and phylogenetic analysis

To identify the SNPs from the 102 landraces, all Illumina reads of each sample were aligned to the reference transcriptome using Samtools in Picard.[19] After removing the reads without unique location, SNP calling from the population was performed using the GATK software;[20] this process required an SNP quality ≥ 40. To exclude SNP calling errors caused by incorrect mapping, only high-quality SNPs (coverage depth ≥ 2, minor allele frequency ≥ 0.05, and missing genotype rate ≤ 0.5) were retained. An individual-based neighbour-joining tree was constructed using TreeBest (http://treesoft.sourceforge.net/treebest.shtml) to clarify the phylogenetic relationships of the 102 landraces in the population, and the tree was visualized in Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

2.9. Identifying candidate transcripts for traits

To mine the candidate transcripts for CL, CW, and CT, we performed a transcriptome-wide analysis to detect the suggestive loci associated with the traits. In our association panel containing 102 samples, a total of 19,912 high-quality SNPs was used in the association analysis for the three CSTs. We first analysed the population structure, which is helpful for minimizing false positives and increases statistical power. The mixed linear model program GEMMA v.0.94.1[21] was then used for the association analysis. The first three principal component analysis values (eigenvectors of kinship matrix) derived from all 19,912 SNPs were used as fixed effects in the mixed model to correct for stratification.[22] The random effect was estimated from groups clustered based on kinship among all accessions, which was derived from all SNPs. The *P*-value threshold for suggestive association loci was set to 5×10^{-5}; this was calculated by Bonferroni correction based on the effective number of independent markers.[6,23] To validate the association between the suggestive loci and the trait, Pearson's correlation between the expression of the transcript harbouring the suggestive locus and the phenotype of the trait was calculated; significant correlation was assumed at $P < 0.05$. A transcript associated with a trait at the level of both sequence and expression was defined as a candidate transcript for the trait.

2.10. Potential interaction detected by eQTL analysis

eQTLs that link the variations in gene expression levels to genotypes have been proven to detect gene interactions.[24] To determine the relationship among candidate transcripts, eQTL analysis was performed by TRAS of 102 landraces using a mixed linear model.[23] In our eQTL analysis, the 19,912 SNPs were defined as markers, and the expression of candidate transcripts were considered as phenotypes. The *P*-value threshold was set to 2.5×10^{-6} so as to 5% significance, which was calculated by Bonferroni correction based on the effective number of independent markers. If one transcript was located in the eQTL of another transcript, we assumed that these two transcripts potentially interacted.

2.11. Prediction of lncRNA function

The co-location and co-expression methods have been proposed for predicting lncRNA function.[25] However, because a full garlic genome sequence is not available, we used only the latter method to predict lncRNA functions. In brief, the transcripts co-expressed with lncRNA were identified by WGCNA, and GO enrichment analysis was carried out to predict the function of the lncRNA. Adjusted *P*-values were calculated using the false discovery rate to determine the significance of the enrichment analysis.[26] Co-expression networks were constructed with Cytoscape.[27]

3. Results

3.1. Single-molecule long-read survey of the transcriptome of developing bulbs

To obtain a high-quality reference transcriptome, we sequenced the transcriptome of developing bulbs of a garlic landrace from Yangxi (China) using a single-molecule long-read sequencer from Pacific Biosciences. In total, 8.07 million circular consensus reads representing 10.96 billion bases were generated from three libraries of different fragment sizes, which resulted in a transcriptome that consisted of 36,321 transcripts, accounting for 54.48 million bases in total

(Supplementary Table S2). Transcript length ranged from 120 to 4,803 bp, and the average length was 1,500 bp (Fig. 1 and Supplementary Table S2). To evaluate the completeness of the obtained transcripts, we analysed the ends of the transcript sequences and found that 71.8% of the transcripts had either PolyA sequences in the 3′ ends or PolyT sequences in the 5′ ends. Because the presence of PolyA or PolyT depends on the orientation of the template, the results indicate that more than 70% of the transcripts were intact at the 3′ ends. We also examined the protein-coding potential of these 36,321 transcripts and identified 4,909 (13.5%) lncRNAs (Supplementary Tables S2 and S3). Of the remaining coding transcripts, 31,125 transcripts were functionally annotated and the function of only 287 transcripts remained unknown (Supplementary Tables S2 and S3).

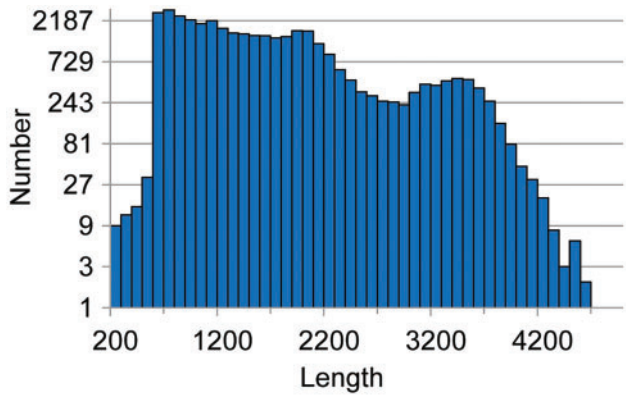

Figure 1. Length distribution of garlic transcripts generated by single-molecule long-read sequencing. A logarithmic scale with a base of three was used in the *Y*-axis.

3.2. Population variation in traits, and in transcript sequences and expression

We investigated three CSTs in the 102 landraces of garlic collected from China and other countries (Supplementary Table S1), namely CL, CW, and CT (Fig. 2a). The three CSTs varied extensively among the 102 landraces, with 2.3-, 3.7-, and 3.6-fold differences in CL, CW, and CT, respectively (Fig. 2b and Supplementary Table S4). We observed a significant correlation among the three traits (corrected *P*-value < 0.01; Supplementary Table S4), suggesting they were mutually dependent.

To characterize the genotypes of the 102 landraces in both sequence and expression, we sequenced the transcriptomes of developing bulbs in the population and obtained approximately 6.05 billion clean reads, with an average number of 59.3 million reads for each landrace (Supplementary Tables S1 and S5) The read sequences were aligned with the reference transcriptome generated by single-molecule long-read sequencing; the mapping rate ranged from 51.3% to 80.2% (Supplementary Tables S1 and S5). These mapped reads were further used for scoring the variation in both sequence (SNPs) and GE of transcripts. In total, 55,012 SNPs were identified from 8,245 transcripts, which suggested that the sequences of approximately 77% of the transcripts were conserved in the 102 landraces. After filtering, 19,912 high-quality SNPs were obtained from the 5,408 transcripts, half of which were within the coding sequence region, and the remaining SNPs were located within the 3′ and 5′ untranslated regions (Fig. 2c). Phylogenetic analysis based on the SNP genotypes resolved the 102 garlic landraces into three distinct groups (Fig. 2d). Interestingly, the landraces from China in the three

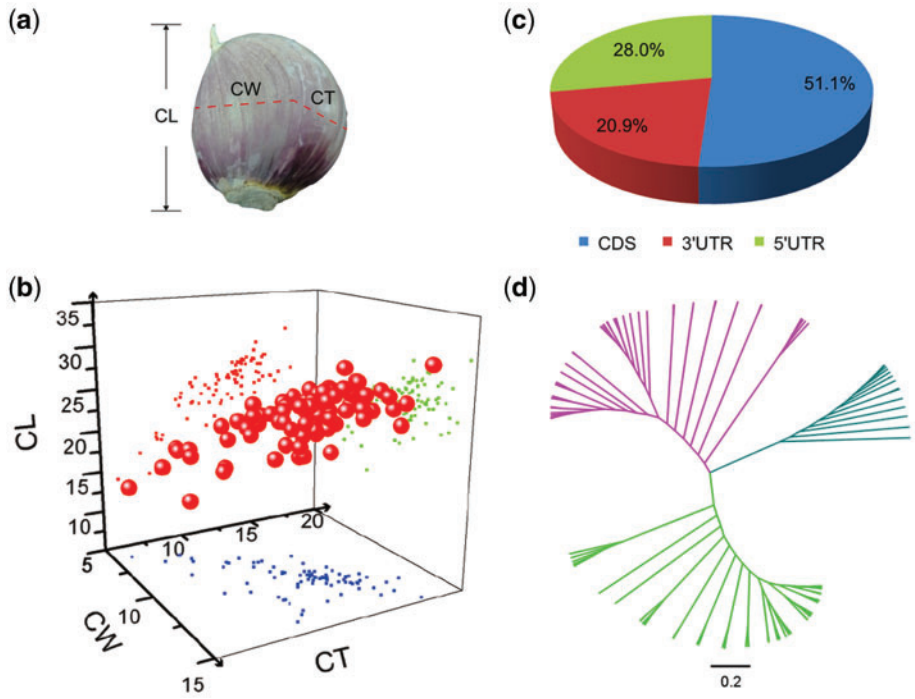

Figure 2. Characterization of the garlic population. (a) Tridimensional presentation of garlic clove illustrating CL, CW, and CT. The direction from the bottom to the top represents the direction of the bulb axis. (b) Phenotype variation of CL, CW, and CT in the population. The unit of scale in *X*-, *Y*-, and *Z*-axes is millimeter (mm). (c) Distribution of SNPs identified in the population. CDS: coding sequence; UTR: untranslated region. (d) Phylogenetic tree of the 102 landraces based on the 19,912 SNPs identified from the population. The unrooted tree was generated using the TreeBest program with the neighbour-joining method. The 102 landraces were clustered into three distinct groups, and cyan, green, and purple clusters represent groups I, II, and III, respectively.

groups showed a clear geographic distribution (Supplementary Fig. S1). We also quantified GE and characterized the genome-wide expression profile of each landrace in the bulb-expansion stage. Based on WGCNA, 36,321 transcripts were involved in 46 co-expression modules (CEMs) in which the transcript number ranged from 37 to 7,132, and these 46 CEMs were named from M1 to M46 (Supplementary Table S3 and Figs S2 and S3).

3.3. Candidate transcripts involved in the determination of CT, CW, and CL

We performed the association analysis by involving the first three principal components in the analysed model as fixed effects, because the first three components explained significantly large variations in the population and we found the presence of three subpopulations in the phylogenetic tree (Fig. 2d). Finally, we identified 42 SNPs from 27 transcripts, 27 SNPs from 18 transcripts, and 10 SNPs from 8 transcripts that were associated with CL, CW, and CT, respectively ($P < 5 \times 10^{-5}$; Fig. 3a–c and Supplementary Table S6). In total, 36 transcripts were associated with the CSTs. Among them, two were associated with all three CSTs, and 13 were associated with two of the three traits (Fig. 3d).

To ascertain that the candidate transcripts were involved in the CSTs, we performed a correlation analysis of the GE of the 36 transcripts and the corresponding trait phenotype in the population, which revealed that the expression of 14, 11, and 1 transcript was significantly associated with CL, CW, and CT, respectively

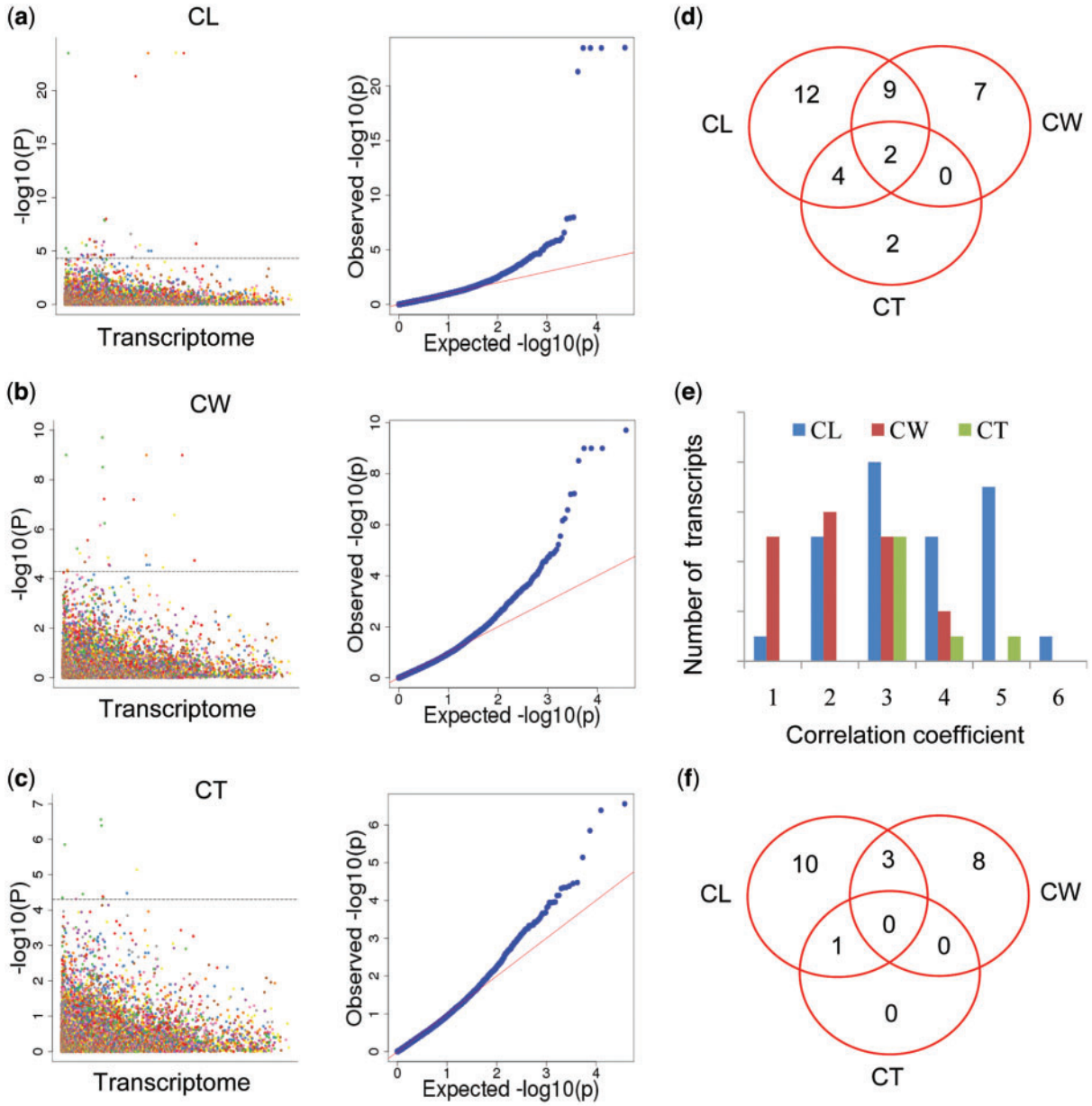

Figure 3. Identification of candidate transcripts involved in clove thickness (CT), weight (CW), and length (CL). (a–c) Manhattan and quantile-quantile plots result from the transcriptome-based association study (TRAS) data for CT, CW, and CL, respectively. The dashed horizontal line indicates the significance threshold ($P < 5 \times 10^{-5}$). (d) Overlap of the transcripts associated with CT, CW, and CL. (e) The distribution of correlation coefficient between the expression of associated transcripts and the traits. (f) Venn diagram of the candidate transcripts related to CT, CW, and CL.

Table 1. Candidate transcripts related to clove shape

Traits	Candidate transcript	CEM[a]	*P*-value[b]		RD[c]	Annotation
			Genotype	Expression		
CL	*ASTG1209*	M16	3.3×10^{-24}	2.3×10^{-5}	−	Alpha-1, 4 glucan phosphorylase L-1 isozyme
	ASTG1548	M3	9.9×10^{-6}	1.8×10^{-2}	+	lncRNA
	ASTG248	M43	4.3×10^{-5}	2.2×10^{-2}	−	DNA replication
	ASTG25643	M9	3.1×10^{-24}	1.0×10^{-2}	+	U3 small nucleolar RNA-associated protein
	ASTG299	M43	1.2×10^{-8}	1.8×10^{-4}	−	lncRNA
	ASTG3024	M2	2.8×10^{-5}	1.1×10^{-2}	+	Chromosomal replication initiator DnaA
	ASTG32719	M9	3.3×10^{-24}	3.8×10^{-3}	+	F-box protein
	ASTG32909	M3	1.4×10^{-6}	2.8×10^{-4}	+	Uncharacterized protein
	ASTG33530	M3	1.4×10^{-5}	8.5×10^{-5}	+	Uncharacterized protein
	ASTG3410	M3	2.8×10^{-6}	3.0×10^{-3}	+	E3 ubiquitin-protein ligase
	ASTG35109	M43	1.8×10^{-6}	7.8×10^{-3}	−	lncRNA
	ASTG35115	M16	3.8×10^{-5}	4.3×10^{-4}	−	lncRNA
	ASTG35782	M3	1.4×10^{-6}	4.0×10^{-2}	+	Uncharacterized protein
	ASTG8464	M9	5.8×10^{-6}	1.1×10^{-2}	−	Protein-tyrosine-phosphatase
CW	*ASTG1209*	M16	1.0×10^{-9}	8.5×10^{-4}	−	Alpha-1, 4 glucan phosphorylase L-1 isozyme
	ASTG1416	M43	1.1×10^{-5}	6.9×10^{-8}	−	Transcriptional regulation activity, sequence-specific DNA binding
	ASTG180	M43	2.8×10^{-6}	5.6×10^{-4}	−	lncRNA
	ASTG299	M43	2.0×10^{-10}	5.4×10^{-5}	−	lncRNA
	ASTG32200	M43	6.1×10^{-6}	$1.0 \times 10^{--6}$	−	lncRNA
	ASTG3501	M43	3.5×10^{-5}	$2.2 \times 1^{--6}$	−	ATP-dependent DNA helicase
	ASTG35109	M43	9.3×10^{-6}	1.0×10^{-7}	−	lncRNA
	ASTG35427	M3	6.9×10^{-7}	3.1×10^{-2}	−	lncRNA
	ASTG35908	M15	2.7×10^{-5}	4.0×10^{-2}	−	lncRNA
	ASTG4729	M43	$6.4 \times 1^{--8}$	1.7×10^{-2}	−	Probable folate-biopterin transporter
	ASTG628	M43	1.0×10^{-9}	1.3×10^{-4}	−	Uncharacterized protein
CT	*ASTG3410*	M3	7.2×10^{-6}	4.6×10^{-2}	+	E3 ubiquitin-protein ligase

[a]CEM, co-expression modules.

[b]The *P*-value of genotype indicates the significance of the association between SNPs of candidate transcript and traits, and the *P*-value of expression represents the significance of correlation between candidate transcript expression and traits.

[c]RD is an abbreviation of regulation direction, and '+' and '−' indicate that the candidate transcript positively and negatively regulates the trait, respectively.

($P < 0.05$; Fig. 3e and Supplementary Table S7). Among those, 6 and 11 transcripts showed a significant negative correlation with CL and CW, respectively, suggesting that the growth of garlic cloves is negatively controlled by these transcripts; whereas 8 and 1 transcript was found to positively regulate CL and CT, respectively. Because both the sequence and GE level were associated with the traits, we concluded that the 14, 11, and 1 transcript identified above were candidate transcripts for CT, CW, and CL, respectively (Table 1). Among them, four were pleiotropic, reducing the total number of candidate transcripts for CSTs to 22.

3.4. Characterization of the 22 candidate transcripts

We analysed the CEM harbouring the candidate transcripts and found that the 22 candidate transcripts were assigned into six CEMs, including nine in M43 and six in M3, respectively (Table 1). Interestingly, five of the six candidate transcripts involved in M3 were found to positively regulate CL, whereas eight of the candidate transcripts involved in M43 negatively controlled the CW trait (Table 1). Correlation analysis of the expression of these two CEMs and the traits revealed that the expression of M3 was significantly positively correlated with the CL trait ($r = 0.39$, $P = 4 \times 10^{-5}$), and the expression of M43 was negatively correlated with CW ($r = -0.4$, $P = 2 \times 10^{-5}$) (Supplementary Table S8 and Fig. S4). Since eight transcripts in M43 and five in M3 were candidate transcripts for CW and CL, respectively, and M43 and M3 expression was significantly correlated with CW and CL, respectively, we concluded that M3 and M43 are CEMs that regulate CL and CT, respectively.

Among the 22 candidate transcripts, 14 were annotated as protein-encoding RNAs that are involved in carbohydrate metabolism (*ASTG1209*), transcriptional regulation (*ASTG1416*), U3 small nucleolar RNA function (*ASTG25643*), DNA replication and helicase activity (*ASTG248*, *ASTG3501*, and *ASTG3024*), and protein modification (*ASTG8464*), transport (*ASTG4729*) and degradation (*ASTG3410* and *ASTG32719*), and four encoded uncharacterized proteins (Table 1). In addition, eight candidate transcripts were predicted to be lncRNAs. The co-expression method was used to predict the potential function of the eight lncRNAs, and the results revealed that a variable number of transcripts, ranging from 4 to 6,301, were co-expressed with the lncRNAs (Fig. 4 and Supplementary Table S9) and they were enriched for various GO terms, including protein ubiquitin activity involved in protein degradation (Supplementary Table S10). Because the candidate transcripts *ASTG3410* and *ASTG32719* were annotated to function in protein degradation and the transcripts co-expressed with the lncRNA *ASTG1548* were enriched for GO terms associated with protein degradation, we deduced that protein degradation potentially has a role in the regulation of CSTs.

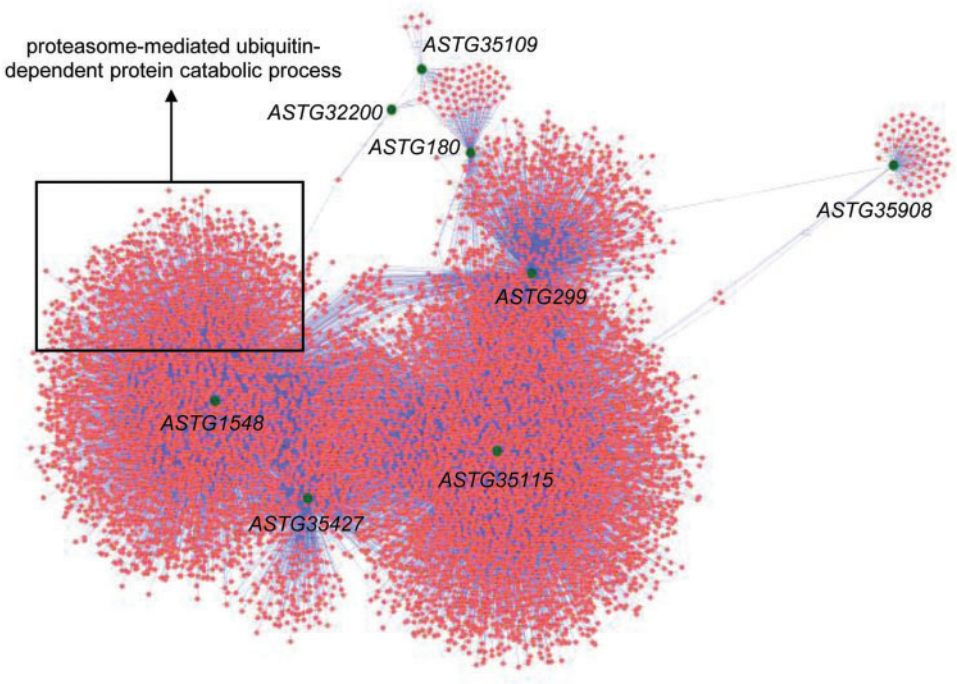

Figure 4. Visualization of a network consisting of eight lncRNAs related to CSTs and their co-expressed transcripts. Green and red dots represent lncRNAs identified as candidate transcripts and their co-expressed transcripts, respectively. Transcripts co-expressed with *ASTG1548* were significantly enriched in the GO term 'proteasome-mediated ubiquitin-dependent protein catabolic process.'.

3.5. Potential interactions among clove shape-associated candidate transcripts

The eQTL analysis is an efficient tool for identifying genetic variants that affect the gene expression, and thus, it has been frequently applied in detecting the interactions and determining the regulation relationship between genes.[24,28] Herein, to detect potential interactions among the trait-associated transcripts, we carried out eQTL analysis for the 22 trait-associated transcripts. The results showed that a variable number of SNPs were associated with the expression of the 22 transcripts (Supplementary Table S11 and Figs S5–S26). No SNPs associated with the expression of *ASTG42793* were identified, suggesting that the expression of this transcript is rarely regulated by other transcripts and that they possibly function upstream from the pathway involved in trait regulation. The expression of *ASTG628*, *ASTG1416*, and *ASTG25643* was associated with at least 30 SNPs (in 39, 38, and 149 transcripts, respectively), indicating that these three candidate transcripts are modulated by the expression of many transcripts and that they probably function downstream of the pathway for trait regulation.

Thereafter, the transcript in which expression-associated SNPs (eQTLs) were located was analysed. We assumed that if a transcript and its eQTL-located transcript were associated with the same trait in sequence, then the two transcripts interacted to control the trait. We next identified the eQTL-located transcript for each candidate trait-associated transcript. The correlation between the expression level of the transcript and that of the corresponding eQTL-located-transcript was used to confirm their interaction. Among the 22 candidate transcripts, 13 had potential interactions. Of the 14 CL candidate transcripts, five showed potential interaction (Supplementary Table S12) and constituted two regulatory pathways: one, in which *ASTG32719* influences the expression of *ASTG25643*, and another, in which *ASTG3410* controls the expression of *ASTG299*, thereby modulating the expression of *ASTG1209* (Fig. 5). In addition, 10 of the 11 CW candidate transcripts were found to form a potential interaction network (Fig. 5) in which *ASTG35908* was a key transcript that determined CW, not only through the regulation of *ASTG1209*, but also by integrating a set of at least eight transcripts in a complex pathway (Fig. 5). Among these eight transcripts, *ASTG35908* regulates the expression of *ASTG1416*, which then regulates *ASTG3501*; this in turn controls the expression of *ASTG35427* and *ASTG180*, whose feedback affects the expression of *ASTG3501* in the regulatory loop *ASTG3501-ASTG35427-ASTG180*. *ASTG35427* and *ASTG180* also modulate the expression of *ASTG1416* to precisely control the expression of *ASTG3501*. Finally, stable expression of the *ASTG3501-ASTG35427-ASTG180* regulatory loop controls the expression of *ASTG35109* to determine CW (Fig. 5). Thus, *ASTG1209* and *ASTG35109* are two key transcripts in the pathway downstream of the regulatory network of CL and CW that function simultaneously. These results provide insight into the interaction network that controls CSTs of garlic.

4. Discussion

The present study developed a TRAS approach based on association mapping and eQTL analysis that is independent of a reference genome of the studied species, and thus, theoretically can be applied to nearly all species. TRAS offers three additional advantages in comparison with the GWAS approach. First, TRAS can directly detect candidate transcripts for a trait by integrating sequence data with expression data. In contrast, GWAS identifies only a genome region in which markers are in linkage disequilibrium for the loci controlling the trait. Unlike GWAS, after identifying a genome region based on the sequence variation, TRAS uses the information on transcript

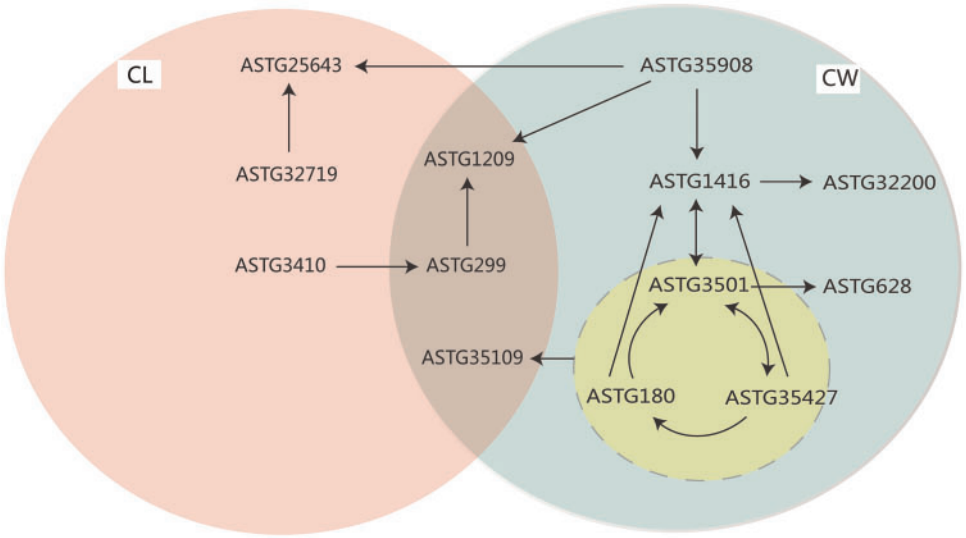

Figure 5. Visualization of the interaction network of 13 candidate transcripts for CL and CW.

expression in the identified region to determine whether or not the corresponding transcript is associated with a given trait. Second, TRAS can detect potential interaction of transcripts by eQTL analysis. Although it remains questionable whether the interaction identified by TRAS is due to direct or indirect regulation, the potential relationship among the transcripts is helpful for further validation of these interactions.

One main shortcoming of TRAS is that it cannot identify SNPs in non-expressed genes or regulatory regions of genes. However, this disadvantage can be addressed by experimental design. For example, depending on target traits, selection of a tissue at a proper growth stage of the organ can ensure that most of the genes involved in trait regulation of this organ are included in the TRAS analysis. In addition, if sequence variation in a regulatory region causes changes in the expression of the corresponding gene, this gene, whether or not involved in the trait, can be determined by association between expression of this gene and the trait.

To test the feasibility of the TRAS approach, a reference transcriptome of garlic was generated and used to characterize three CSTs in this study. Several previous studies conducted *de novo* assembly of the garlic transcriptome, producing more than 120,000 transcripts with an average length of less than 600 bp, of which only 35–42% were functionally annotated.[13,29,30] These data indicated that a large number of transcripts in these transcriptomes were incomplete and could not qualify for the reference sequence in TRAS. Therefore, this study used the single-molecule long-read sequencing technology for RNA sequencing, which significantly improved the transcriptome quality—the mean length of the transcripts was 1,500 bp; more than 70% of the transcripts had a complete 3′ end; and only less than 1% of the transcripts remained functionally unannotated. Based on this transcriptome, substantial amounts of SNPs were identified in the population. When association analysis for the CSTs was performed, principal component analysis was used to decrease the rate of potential false-positive association caused by population stratification. Finally, 42, 27, and 10 SNPs (involved in 36 transcripts in total) were found to be significantly associated with CL, CW, and CT, respectively, and their observed *P*-values were lower than the expected values, suggesting that the associations between these genomic regions where 36 transcripts were located and the CSTs are reliable. Irrespective of the association in terms of sequence variation, the gene expression levels of 22 of the 36 transcripts showed significant correlation with the CSTs, suggesting that the association of these 22 transcripts were authentic, and these 22 transcripts were candidate transcripts involved in regulating CSTs. However, the remaining 14 transcripts were associated with the CSTs only in terms of the sequences, but not in terms of expression, which was probably caused by the following: association study based on sequence variation identifies a genome region in which the markers are in linkage disequilibrium for the gene involved in regulating the trait; this linkage disequilibrium region probably harbours several genes, and only one of them is the gene that regulates the trait; thus, if the associated markers-located gene is not the trait-conferring candidate, then the expression of this gene would not be correlated with the trait. In summary, there were 22 candidate transcripts identified to be involved in regulating the CSTs, suggesting that TRAS is suitable for genetic dissection of complex traits in species lacking reference genomes.

Garlic cloves are abnormal, expanding axillary buds that are rarely found among vascular plants. The genetic mechanism of their development is poorly understood. This study identified 22 candidate transcripts involved in regulating the CSTs, eight of which (36.4%) were lncRNAs. This ratio was far higher than that in the transcriptome (13.5%), indicating that non-coding RNAs play important roles in the determination of clove morphology. Interestingly, we found that the lncRNA *ASTG1548* and two candidate transcripts—*ASTG3410* and *ASTG32719*—were potentially involved in protein degradation. Protein degradation possibly plays important roles in the regulation of the CSTs. Analogously in rice, at least three genes that function in protein degradation (*GW2*, *qSW5*/*GW5*, *HGW*) are involved in regulating grain shape.[31–34] Because both rice grain and garlic clove obtain a tridimensional appearance during their development, it is possible that the pathway of protein degradation plays a similarly important role in the development of rice grain and garlic clove.

It is notable that carbohydrates are the main component of garlic bulbs.[35] We identified a transcript—*ASTG1209*—encoding alpha-1, 4 glucan phosphorylase that is associated with both CW and CL, suggesting that carbohydrate metabolism contributes to clove development. One candidate transcript, *ASTG1416*, encodes a sequence-

specific DNA-binding protein with transcriptional regulatory activity. The bulb of garlic undergoes expansion, which requires cell proliferation that necessitates DNA replication. In the present study, three CST-related transcripts were found to be involved in DNA replication. In addition, one transcript associated with protein modification, one encoding a protein transporter, and two related to protein degradation were linked to the regulation of the CSTs, suggesting that post-translational modification plays a role in the control of the CSTs. However, some candidate protein-coding transcripts were uncharacterized; clarifying their role in the regulation of the CSTs can provide new insights into the genetic basis of garlic bulb formation. In conclusion, 22 CST-related candidate transcripts were identified in garlic, a species without a reference genome, and their potential interactions were ascertained. Our results demonstrate that TRAS is a useful approach for association studies, and its independence from a reference genome will extend the applicability of association studies to a broad range of species. Additionally, the CST-related candidate transcripts identified herein can provide a basis for improving the CSTs in garlic breeding.

5. Data Availability

The CCS reads from the single-molecule sequencing effort are available in the NCBI SRA database under the accession number SRX3145702. The sequence of the reference transcriptome can be downloaded from the NCBI GenBank database under the accession number GFYZ00000000. The transcript expression data of the 102 landraces have been deposited in the NCBI GEO database under the accession number GSE102157.

Funding

This work was supported by grants from the Agricultural Science and Technology Innovation Program of China (CAAS-ASTIP-IBFC).

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Ohri, D. and Pistrick, K. 2001, Phenology and genome size variation in Allium L. - a tight correlation? *Plant Biol.*, **3**, 654–60.
2. Zuo, J. and Li, J. 2014, Molecular dissection of complex agronomic traits of rice: a team effort by Chinese scientists in recent years, *Natl. Sci. Rev.*, **1**, 253–76.
3. Zhao, K., Tung, C., Eizenga, G., et al. 2011, Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*, *Nat. Commun.* , **2**, 467.
4. Huang, X., Sang, T., Zhao, Q., et al. 2010, Genome-wide association studies of 14 agronomic traits in rice landraces, *Nat. Genet.* , **42**, 961–7.
5. Huang, X., Zhao, Y., Li, C., et al. 2011, Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm, *Nat. Genet.* , **44**, 32–9.
6. Yang, W., Guo, Z., Huang, C., et al. 2014, Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice, *Nat. Commun.* , **5**, 5087.
7. Mutz, K., Heilkenbrinker, A., Lonne, M., Walter, J. and Stahl, F. 2013, Transcriptome analysis using next-generation sequencing, *Curr. Opin. Biotechnol.* , **24**, 22–30.
8. Harper, A., Trick, M., Higgins, J., et al. 2012, Associative transcriptomics of traits in the polyploid crop species *Brassica napus*, *Nat. Biotechnol.*, **30**, 798–802.
9. Lu, G., Harper, A., Trick, M., et al. 2014, Associative transcriptomics study dissects the genetic architecture of seed glucosinolate content in *Brassica napus*, *DNA Res.*, **21**, 613–25.
10. Hackl, T., Hedrich, R., Schultz, J. and Forster, F. 2014, proovread: large-scale high-accuracy PacBio correction through iterative short read consensus, *Bioinformatics*, **30**, 3004–11.
11. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. 2012, CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics* , **28**, 3150–2.
12. Kong, L., Zhang, Y., Ye, Z., et al. 2007, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, *Nucleic Acids Res.*, **35**, W345–9.
13. Zhu, S., Tang, S., Tan, Z., Yu, Y., Dai, Q. and Liu, T. 2017, Comparative transcriptomics provide insight into the morphogenesis and evolution of fistular leaves in *Allium*, *BMC Genomics*, **18**, 60.
14. Iseli, C., Jongeneel, C. and Bucher, P. 1999, ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 138–48.
15. Langmead, B. and Salzberg, S. 2012, Fast gapped-read alignment with Bowtie 2, *Nat. Methods.* , **9**, 357–9.
16. Li, B. and Dewey, C. 2011, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC Bioinformatics*, **12**, 323.
17. Langfelder, P. and Horvath, S. 2008, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics*, **9**, 559–72.
18. Zhang, B. and Horvath, S. 2005, A general framework for weighted gene co-expression network analysis, *Stat. Appl. Genet. Mol. Biol.*, **4**, Article 175.
19. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The sequence alignment/-map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
20. McKenna, A., Hanna, M., Banks, E., et al. 2010, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Res.*, **20**, 1297–303.
21. Zhou, X. and Stephens, M. 2012, Genome-wide efficient mixed-model analysis for association studies, *Nat. Genet.*, **44**, 821–4.
22. Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N. and Reich, D. 2006, Principal components analysis corrects for stratification in genome-wide association studies, *Nat. Genet.*, **38**, 904–9.
23. Duggal, P., Gillanders, E. M., Holmes, T. N. and Bailey-Wilson, J. E. 2008, Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies, *BMC Genomics*, **9**, 516.
24. Gilad, Y., Rifkin, S. and Pritchard, J. 2008, Revealing the architecture of gene regulation: the promise of eQTL studies, *Trends Genet.*, **24**, 408–15.
25. Liao, Q., Liu, C., Yuan, X., et al. 2011, Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network, *Nucleic Acids Res.*, **39**, 3864–78.
26. Boyle, E., Weng, S., Gollub, J., et al. 2004, GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics*, **20**, 3710–5.
27. Shannon, P., Markiel, A., Ozier, O., et al. 2003, Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.*, **13**, 2498–504.
28. Keurentjes, J., Fu, J., Terpstra, I., et al. 2007, Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci, *PNAS*, **104**, 1708–13.
29. Sun, X., Zhou, S., Meng, F. and Liu, S. 2012, *De novo* assembly and characterization of the garlic (*Allium sativum*) bud transcriptome by Illumina sequencing, *Plant Cell Rep.*, **31**, 1823–8.

30. Liu, T., Zeng, L., Zhu, S., et al. 2015, Large-scale development of expressed sequence tag-derived simple sequence repeat markers by deep transcriptome sequencing in garlic (*Allium sativum* L.), *Mol. Breeding*, **35**, 20.
31. Li, J., Chu, H., Zhang, Y., et al. 2012, The rice *HGW* gene encodes a ubiquitin-associated (UBA) domain protein that regulates heading date and grain weight, *PLoS One*, **7**, e34231.
32. Shomura, A., Izawa, T., Ebana, K., et al. 2008, Deletion in a gene associated with grain size increased yields during rice domestication, *Nat. Genet.*, **40**, 1023–8.
33. Song, X., Huang, W., Shi, M., Zhu, M. and Lin, H. 2007, A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase, *Nat. Genet.*, **39**, 623–30.
34. Weng, J., Gu, S., Wan, X., et al. 2008, Isolation and initial characterization of *GW5*, a major QTL associated with rice grain width and weight, *Cell Res.*, **18**, 1199–209.
35. Arguello, J., Ledesma, A., Nunez, S., Rodriguez, H. and Goldfarb, M. 2006, Vermicompost effects on bulbing dynamics, nonstructural carbohydrate content, yield, and quality of 'Rosado Paraguayo' garlic bulbs, *Hortscience*, **41**, 589–92.