PLOS ONE

RESEARCH ARTICLE

SpinachDB: A Well-Characterized Genomic Database for Gene Family Classification and SNP Information of Spinach

Xue-Dong Yang[1☯], Hua-Wei Tan[2☯], Wei-Min Zhu[1]*

1 The Protected Horticulture Institute, Shanghai Academy of Agricultural Sciences, Shanghai, China,
2 College of Horticulture, Nanjing Agricultural University, Nanjing, Jiangsu, China

☯ These authors contributed equally to this work.
* yy17@saas.sh.cn

Citation: Yang X-D, Tan H-W, Zhu W-M (2016) SpinachDB: A Well-Characterized Genomic Database for Gene Family Classification and SNP Information of Spinach. PLoS ONE 11(5): e0152706. doi:10.1371/journal.pone.0152706

Editor: Xiang Jia Min, Youngstown State University, UNITED STATES

Received: January 5, 2016

Accepted: March 17, 2016

Published: May 5, 2016

Data Availability Statement: Data created during this research are openly available through SpinachDB (http://222.73.98.124/spinachdb/) and provided as supplementary information accompanying this paper.

Funding: This work was supported by Natural Science Foundation of Shanghai (15ZR1436700) and Shanghai Science and Technology Talents Project (14XD1425100) and the Scientific Research Project in Public Agricultural Industry (201403032). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Spinach (*Spinacia oleracea* L.), which originated in central and western Asia, belongs to the family Amaranthaceae. Spinach is one of most important leafy vegetables with a high nutritional value as well as being a perfect research material for plant sex chromosome models. As the completion of genome assembly and gene prediction of spinach, we developed SpinachDB (http://222.73.98.124/spinachdb) to store, annotate, mine and analyze genomics and genetics datasets efficiently. In this study, all of 21702 spinach genes were annotated. A total of 15741 spinach genes were catalogued into 4351 families, including identification of a substantial number of transcription factors. To construct a high-density genetic map, a total of 131592 SSRs and 1125743 potential SNPs located in 548801 loci of spinach genome were identified in 11 cultivated and wild spinach cultivars. The expression profiles were also performed with RNA-seq data using the FPKM method, which could be used to compare the genes. Paralogs in spinach and the orthologous genes in *Arabidopsis*, grape, sugar beet and rice were identified for comparative genome analysis. Finally, the SpinachDB website contains seven main sections, including the homepage; the GBrowse map that integrates genome, genes, SSR and SNP marker information; the Blast alignment service; the gene family classification search tool; the orthologous and paralogous gene pairs search tool; and the download and useful contact information. SpinachDB will be continually expanded to include newly generated robust genomics and genetics data sets along with the associated data mining and analysis tools.

Introduction

Spinach (*Spinacia oleracea* L.), which originated in central and western Asia, belongs to the family Amaranthaceae. Spinach is particularly resistant to cold, has wide adaptability, and is cultivated worldwide, mostly in temperate regions. As one of most important leafy vegetables, spinach has a high nutritional value, and is rich in carotene, vitamin C, amino acids and iron, among others [1]. Recent studies show that spinach may help to protect people against

Competing Interests: The authors have declared that no competing interests exist.

inflammatory problems, oxidative stress-related problems, cardiovascular problems, and cancers [2–4]. In 2013, the total harvested area of spinach reached 910833 Ha and gross production value (USD current) reached $11.995 billion worldwide, according to FAOSTAT statistics (http://faostat3.fao.org).

Because spinach is a particularly desirable vegetable, germplasm resources and breeding are important. At present, spinach cultivars in the USA, Europe and Japan, are mostly F1-hybids. Molecular breeding for complex traits in crop plants requires understanding and manipulation of many genes. Utilization of genome-wide molecular markers is an effective tool for plant breeding, which drives researchers to focus on constructing detailed genetic maps with high density markers [5]. A genetic map of spinach was constructed using 101 AFLP and 9 microsatellite markers, and the result showed a small chromosomal region co-segregating with sex determination in the species [6]. In the recent years, 10 AFLP and 2 male-specific markers were identified in close vicinity to the X/Y locus and a single major gene responsible for the monoecious condition was found [7]. Spinach is a model for genetic and physiological studies on sex determination and expression, and the molecular basis of sex determination needs further study [7]. Study on the spinach C class floral identity genes indicated that gene SpAGAMOUS was differentially expressed prior to reproductive organ development and involved in the sexual dimorphism of spinach [8].

In recent years, with the advent of high-throughput sequencing technologies, genomic nucleotide sequence data of many crops has accumulated, and *de novo* assembly of genomic sequence data can provide whole-genome sequences, which are valuable resources for investigating the genetic characteristics; identification of the candidate genes especially associated with agronomic traits; and understanding of important evolutionary processes [9]. With the accumulation of the large amount of plant genome sequences, various genomics databases have been constructed for such species, including genome database for carrot [10], rice [11], radish [12] and gene family database like SALAD database [13], GreenPhylDB [14] and Phytozome [15].

Spinach possesses a small genome, thus is suitable for basic genomic studies, and many physiologically important genes have been cloned from the species [16]. Recently, the spinach genome has been sequenced and *de novo* assembled [17,18]. The genome sequence and gene predictions are available at The *Beta vulgaris* Resource website (http://bvseq.molgen.mpg.de) [18]. Although researchers can browse genes and make a blast query in that website, useful tools for deeply mining spinach genes are still lacking. Thus, it is urgent to construct a central database for spinach to efficiently store, annotate, mine and analyze the genomics and genetics datasets. Therefore, we developed SpinachDB, to store assembled and annotated EST transcripts, predicted metabolic pathways, and EST-derived simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers. A set of tools and user-friendly query interfaces have also been developed in the database to help researchers in identifying and deciphering biologically important information from the datasets. SpinachDB will be continually expanded to include newly generated robust genomics and genetics data sets and the associated data mining and analysis tools.

Materials and Methods

The workflow was, essentially, gene annotation, SSR and SNP detection, and expression profiling of genes using spinach genomic and transcriptomic data. This was followed by pooling the data into a database and providing the researcher with a user-friendly website interface (Fig 1).

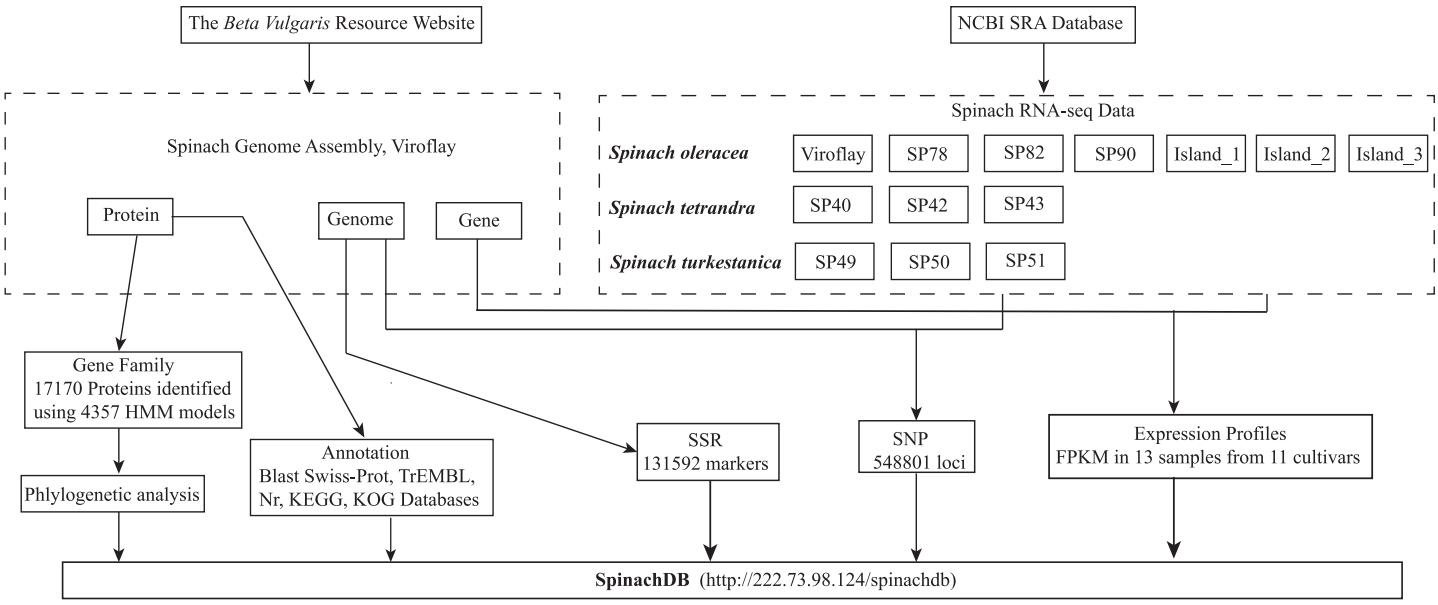

Fig 1. The workflow for data mining and construction of SpinachDB.

doi:10.1371/journal.pone.0152706.g001

Collection and annotation of genome resources

Since the spinach genomic data was available in The *Beta vulgaris* Resource website (http://bvseq.molgen.mpg.de/Genome/Download/Spinach/), we downloaded the Spinach genome assembly version 1.0.1 and annotation version SpiSet-1 (August 2014) for further analysis. RNA-seq raw data were also downloaded in the NCBI SRA (Sequence Read Archive) database (http://www.ncbi.nlm.nih.gov/sra) [19]. All of the spinach proteins were annotated by search, using NR (Non-reduntant, NCBI), Swiss-Prot and TrEMBL (www.uniprot.org) [20], KEGG (http://www.genome.jp/kegg) [21], and KOG (http://www.ncbi.nlm.nih.gov/COG) [22] databases using BLASTP with a cut-off E-value of 10^{-5}. SSRs were identified using the MIcroSAtellite (MISA) identification tool (http://pgrc.ipk-gatersleben.de/misa/) based all the spinach genome sequences. For mono-, di-, tri-, tetra-, penta- or hexanucleotide SSRs, the minimum number of units were set to be 10, 6, 5, 5, 5 and 5 repeats, respectively. The software Primer3 (http://bioinfo.ut.ee/primer3/) was used to design potential primers for each SSR, with default parameters following the previous rules [23].

Classification of gene families

The HMM (Hidden Markov Models) models of gene families were downloaded from the Pfam database (http://pfam.xfam.org) on September 26, 2015 [24]. The proteins of another 12 representative species in the plant kingdom including algae, moss, fern, spruce, *Arabidopsis*, rice, grape, poplar, tomato, apple, soybean and sugar beet were also downloaded. The proteins of *Arabidopsis* and sugar beet were downloaded from TAIR 10 (http://www.arabidopsis.org) and The *Beta vulgaris* Resource website respectively; the others were downloaded from the Phytozome portal (http://phytozome.jgi.doe.gov, v10) [15]. For those genes that had several corresponding protein products due to splicing models, only the primary ones were retained for further analysis. The hmmsearch program (http://www.hmmer.org, v3.1b1) [25] was used to identify gene families with the same E-value of e^{-5}. The output of all identified gene families was classified according to the species. The software orthoMCL (http://orthomcl.org/orthomcl,

v2.0.3) [26] was performed to identify orthologous and paralogous gene pairs in *Arabidopsis*, grape, rice, sugar beet and spinach.

Expression profiles of genes and SNP mining

Thirteen SRA files were used to characterize the gene expression profiles and for SNP mining of spinach, including 7 from cultivated *S. oleracea* (Viroflay cultivar, SRR1542623; SP78 cultivar, SRR1766310; SP82 cultivar, SRR1766311; SP90 cultivar, SRR1766312; Island cultivar, SRR1763297, SRR1763298, SRR1763299), 3 from *S. tetrandra* (SP40 cultivar, SRR1766329; SP42 cultivar, SRR1766329; SP43 cultivar, SRR1766329) and 3 from *S. turkestania* (SP49 cultivar, SRR1766332; SP50 cultivar, SRR1766333; SP51 cultivar, SRR1766334). The fastq-dump program in the SRA Toolkit was used to release fastq format files from SRA files; then NGS QC Toolkit (http://www.nipgr.res.in/ngsqctoolkit.html, v2.3.3) was used to obtain high quality reads which contained at least 90% bases beyond Phred-Qual 20 [27]. The program Tophat (https://ccb.jhu.edu/software/tophat/index.shtml, v2.1.0) [28] was used to map the reads to the spinach genome; then the expression profile of all spinach genes was obtained with FPKM (Fragments Per Kilobase of exon per million fragments Mapped) value using software Cufflinks (http://cole-trapnell-lab.github.io/cufflinks, v2.2.1) that under the guidance of annotated gene models with a GFF file [29,30]. Meanwhile, the mapped profiles were used to identify the different bases while being compared with reference spinach genome sequence. The Genome Analysis Toolkit (GATK) software package (http://www.broadinstitute.org/gatk/, v3.5) [31] was used for SNP calling using HaplotpeCaller with default parameter before hard filters were applied to the call sets. Another python script in software Seqgene (http://sourceforge.net/projects/seqgene/, v2.5) [32] were used to identify SNPs in consideration of the quality of bases, the minimum coverage and percentage of SNP allele calling.

Server and website construction

Linux system CentOS6.6 (http://www.centos.org) was installed on an HP blade server machine with 80 threads and 700 GB memory. The MySQL database server (http://www.mysql.com, v5.1.73) was installed for storing all genomic and website data. The Apache web server with PHP (v5.3.3) was used for powerful website service, and the software Joomla (http://www.joomla.org, v2.5) was used for user-friendly website interface construction. Some frequently used bioinformatics programs like GBrowse (http://gmod.org, v1.70) and wwwblast (ftp://ftp.ncbi.nlm.nih.gov/blast/executables, v2.2.26) were installed to provide user-friendly search capability for genes of interest. Some Perl and CGI (Common Gateway Interface) scripts were built within to provide additional search services.

Results and Discussion

Overview of the database contents and functions

The genomic and transcriptomic data of spinach are publicly available in SpinachDB (http://222.73.98.124/spinachdb). The Fig 2 shows that the website contains seven main sections, including the homepage; the GBrowse map that integrates genome, genes, SSR and SNP markers information; the Blast alignment service; the gene family classification search tool; the orthologous and paralogous gene pairs search tool; and the download and useful contact information.

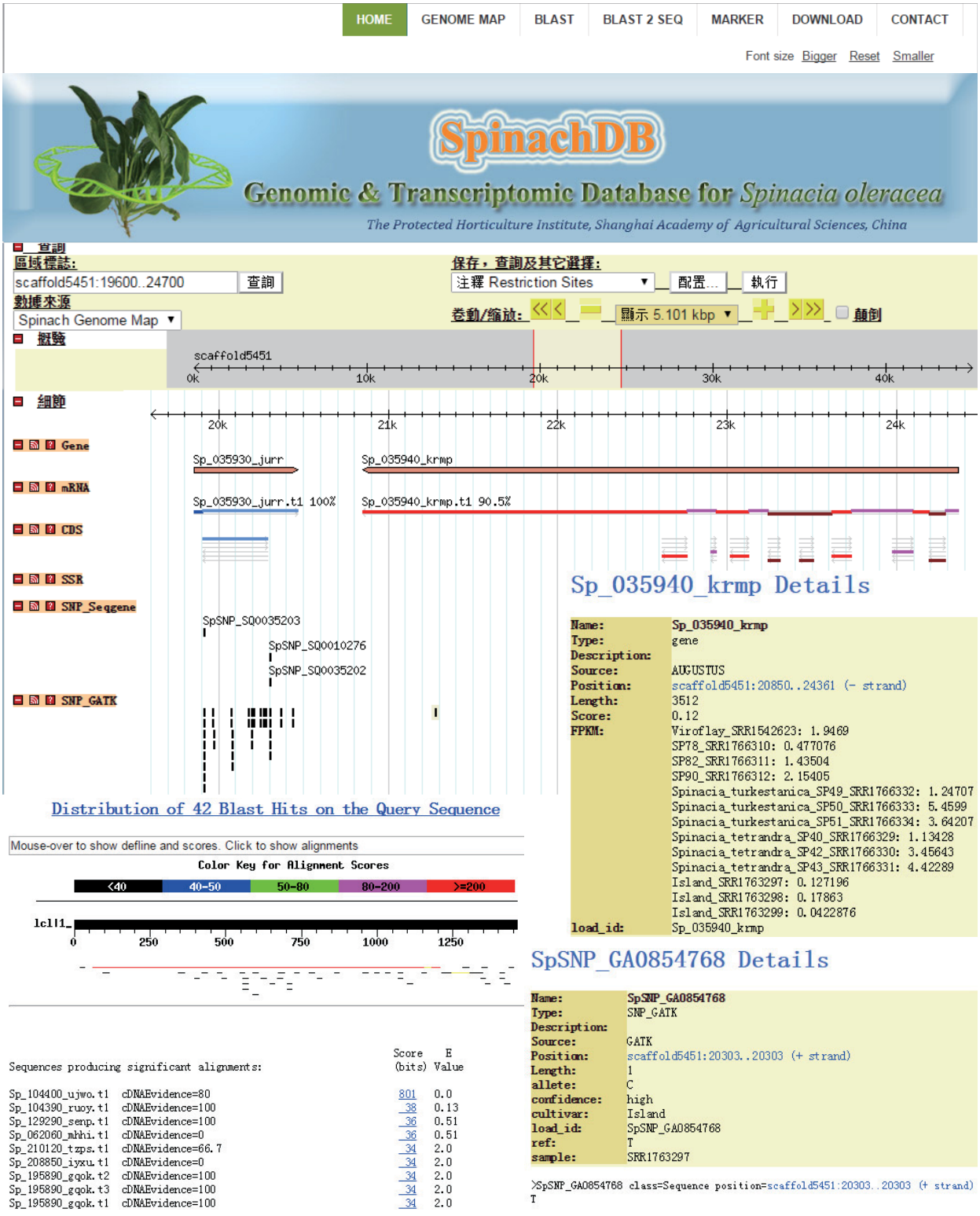

Fig 2. The organizational structure of the SpinachDB website.

doi:10.1371/journal.pone.0152706.g002

The genome of spinach and SSR identification

The span of the updated genome assembly of spinach was 489 Mb in total, with 21702 predicted genes with high confidence found in the *Beta vulgaris* Resource website. Based on the information in that website, the software Augustus (http://augustus.gobics.de/) predicted 40309 genes, among which only 21702 were high-evident ones that were collected for the dataset used for final gene assembly. To functionally categorize these genes, there were 15550, 18864, 18659, 7956, 18634 genes assigned annotation after aligning with Swiss-Prot, TrEMBL, KEGG, KOG, NR and NT databases, respectively (Table A in S1 File). These annotations give insight into spinach genes, based on molecular research on corresponding genes of other organisms.

A total of 131592 SSRs were found in the whole-genome sequence of spinach cultivar Viroflay, including 94741, 16749, 15379, 1748, 1754 and 1221 for mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats, respectively. The number of genomic SSRs in spinach was almost 2.7 times that in carrot (48398 SSRs) while the genome sizes of these two species were similar. A collection of 13229 SSRs that fit the definition of a compound SSR were found in this study. Among all of the SSRs, the A_{10} and T_{10} type were the most abundant ones. In addition, to develop SSR markers useful for experimental validation, 57519 pairs of primers specific for SSRs were designed using Primer3 software. The SSRs and corresponding primers were integrated in the GBrowse tool, so as to provide useful information for genetic map construction.

Gene expression in different samples

Recently, biology has become a data intensive science because of huge data sets produced by high throughput molecular biological experiments in the field of genomics, transcriptomics, proteomics, and metabolomics. These data is useful for constructing molecular biological networks [33]. To characterize the expression profiles of genes in different cultivars, the RNA-seq data of four spinach cultivars that had been sequenced in two stand-alone projects were collected from the NCBI SRA database. To reduce false positivity and bias, the low-quality bases were filtered using the software NGS QC Toolkit (http://www.nipgr.res.in/ngsqctoolkit.html) [31]. After that, 10.57 million, 1.04 million, 0.88 million, 1.15 million, 0.87 million, 1.48 million, 1.60 million, 1.03 million, 0.97 million, 0.85 million, 46.08 million, 43.61 million and 47.06 million sequencing bases, for cultivars Viroflay, SP40, SP42, SP50, SP51, SP78, SP82, SP90 and three samples of cultivar Island respectively, were retained for further analysis. The assembly of transcriptomic sequences from these spinach cultivars was performed using Tophat and Cufflinks with a guided GFF (General Feature Format) file, yielding expression profiles for each gene and the corresponding isoforms. There were 19715 (90.8%), 19639 (90.5%), 19326 (89.1%), 19559 (90.1%), 19508 (89.9%), 19653 (90.6%) and 20894 (96.3%) genes expressed in cultivated cultivars Island 1, Island 2, Island 3, SP78, SP82, SP90 and Viroflay, respectively (Table B in S1 File). For the wild spinach, there were 19646, 16599, 17117, 17927, 19984 and 20014 genes expressed in sample SP40, SP42, SP43, SP49, SP50 and SP51 (Table C in S1 File).

SNP detection in different cultivars

SNP detection is usually performed by mapping the DNA sequencing reads to the referent genome sequences with pipelines and software, such as GATK (https://www.broadinstitute.org/gatk/) and SOAPsnp (http://soap.genomics.org.cn/). Although DNA-oriented sequencing reads were regularly used for SNP prediction for better results, RNA-oriented sequencing reads such as RNA-seq have also been performed occasionally, but increasingly, in many studies. In some studies, the RNA-seq data and EST sequences were also used to predict potential SNPs.

The subsequent SNP validation experiment supported the use of RNA-seq and also EST data for contribution to SNP identification. The validation of selected high-confident SNPs in tea plant [34], longan [35], and pineapple [36] were carried out with nearly half of them having been validated.

Although there were several studies on the SNP and comparative analysis using transcriptomic data, they were conducted by mapping reads to assembled expressed transcripts rather than whole genome sequencing [37,38]. In this study, a total of 1125743 potential SNPs located in 548801 loci of spinach genome were identified. Initially, we attempted to use the software packages SOAP and SOAPsnp to detect SNPs, but too much false positive results were obtained. The professional software GATK called 1074400 potential SNPs in 535439 loci. Among them, 169915 (15.8%) were defined as high confident SNPs that passed hard filters, whereas 834549 were intermediate confident SNPs and 69936 Low quality SNPs. The cultivar SP42 and SP43 of *Spinacia tetrandra* contained more SNPs than other cultivated or wild spinach, which indicated its relationship with other cultivars were more far away. In addition, we used the software Seqgene with lower criteria, resulting in a total of 51343 SNPs in 33631 loci.

The differentiation of gene expression and SNPs among these cultivars was a useful resource for molecular biology as well as for breeding. Some of these differentiated genes could be related to resistant/tolerant genes or to those for high nutritional value. The SNPs of interest in gene regions could be easily selected for validation according to gene annotation. At the pre-breeding stage, the validated SNPs were useful in selecting good breeding materials and good F1 hybrid lines, which could potentially facilitate the breeding of cultivars that are fleshy, high yielding, highly tolerant to abiotic stresses and highly resistant to pests.

Orthologous and paralogous genes

The concepts of orthology and paralogy originated in 1970s, and have been used for functional annotation and classifications on large scale whole-genome comparisons [39]. To give insight into the crossing-referencing and classification of genes from multiple species among the flowing plants, the well-studied plants *Arabidopsis*, rice, and grape from different clades; sugar beet in the same family; and spinach were selected to detect orthologs and paralogs. Among these species, the percentage of classified genes was fairly high, i.e. between 66.4% (rice) and 83.9% (*Arabidopsis*). All five of these species were flowering plants and the orthology they shared consisted of 64184 orthologous gene pairs in 8441 groups. The four eudicots shared more orthologous genes, while rice, the monocotyledonous species, shared the fewest orthologs with the others. The distribution of orthologous gene pairs was totally consistent with the genetic relationship. The species in the same family shared the largest number of orthologous gene pairs, and then with other eudicots, and the least with the monocot (Fig 3, Table C in S1 File).

To investigate the orthologous genes in spinach, a Venn diagram was drawn to show the number of clustered groups and genes that spinach shared with the other four plants (Fig 4). A total of 17376 genes were identified in 13421 groups in this study. Among them, 740 spinach genes in 254 groups were identified to be paralogous genes, without corresponding orthologous genes in the other four species. A total of 11197 spinach genes were found to be orthologs that were shared with all other four species. Another 1090 spinach genes were orthologous with sugar beet, *Arabidopsis* and grape simultaneously, in 980 groups. There were 2190, 355, 47 and 24 spinach genes that were found only to have orthologous genes with sugar beet, *Arabidopsis*, grape and rice, respectively. These orthologous and paralogous gene pairs can be usefully employed to facilitate both evolutionary and functional analyses of spinach.

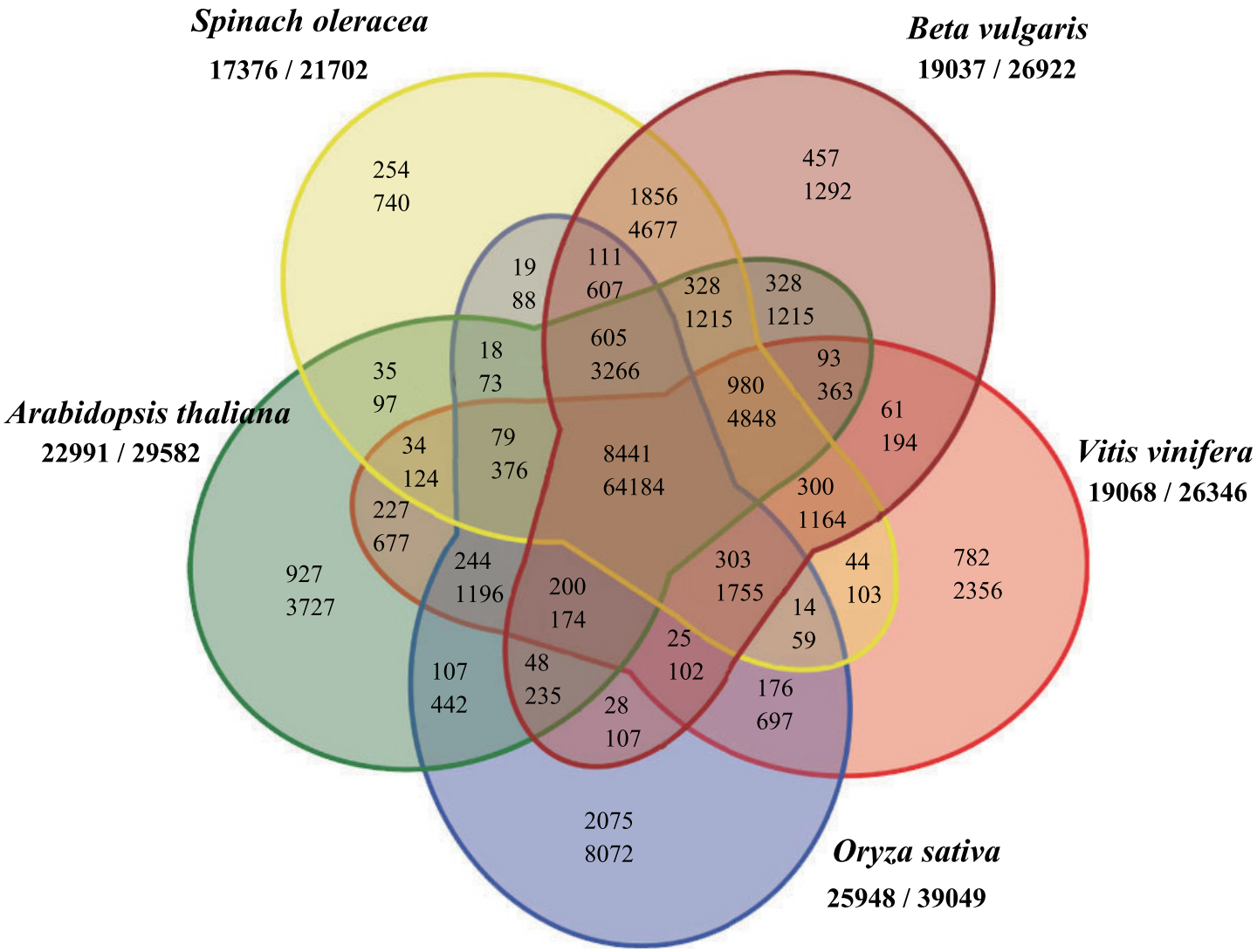

Fig 3. The Venn diagram shows shared and distinct cluster classes from an orthoMCL analysis of proteins from Arabidopsis thaliana, Oryza sativa spp. indica, Vitis vinifera, Beta vulgaris and Spinach oleracea. Numbers below each species name indicate numbers of clustered genes and of all genes in the corresponding genome. For each area in the Venn diagram, the top line and the bottom line represent the number of orthoMCL clusters and the number of all accumulated shared genes in all organisms.

doi:10.1371/journal.pone.0152706.g003

Gene families

A total of 16230 HMM models from Pfam database were used to search proteins in 13 organisms. A total of 301184 proteins were catalogued to 5699 HMM models. Among all 21702 proteins of spinach, 15741 proteins were catalogued into 4351 HMM models (Table D in S1 File). Among these 13 organisms, the spinach proteins had the largest number of proteins in 272 HMM model groups, such as the gene families FeoB_N (PF02421), MS_channel (PF00924), MTS (PF05175), PD40 (PF07676) and Whirly (PF08536). Users, especially molecular researchers without a robust programming background, can easily find the name and sequence of proteins in a particular gene family and perform a genome-wide analysis of the gene family in a spinach study.

In order to explain a possible usage for these classified gene families, the genome-wide identification and analysis of BES1/BZR1 in spinach were performed. The BES1 gene family is a

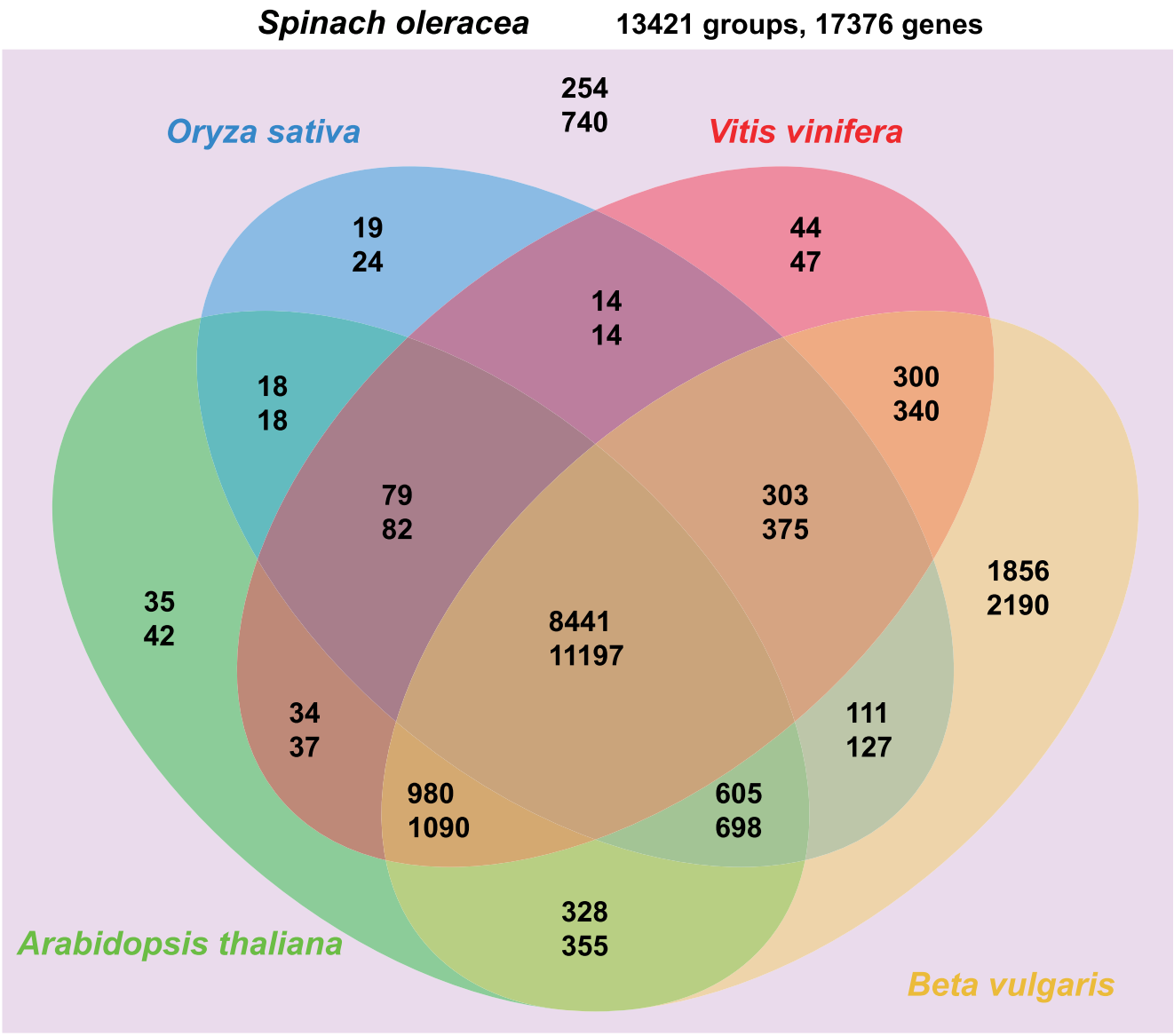

Fig 4. The Venn diagram shows the number of clustered groups and genes that spinach shared with other four plants. For example, the numbers 328 and 355 at the bottom of this figure indicate that for all clustered groups of Arabidopsis thaliana, Beta vulgaris and Spinach oleracea, the number of these groups is 328 and there are 355 spinach genes in these groups.

doi:10.1371/journal.pone.0152706.g004

class of plant-specific transcription factors that play a key role in the Brassinosteroid signaling pathway. The BES1 gene interacts with the basic helix-loop-helix protein BIM1 to synergistically bind to E box (CANNTG) sequences present in many BR-induced promoters [40]. In this study, it was straightforward to obtain the names and sequences of 118 proteins from 12 species using the same criteria. The program MEGA 6.6 (http://www.megasoftware.net/) [41] was used to align and construct an N-J tree with a bootstrap of 500. In the phylogenetic analysis, there was no BES1 protein found in algae but 6 were found in moss, indicating that BES1 may occur first in land plants. All of the proteins were classified into 5 groups according to the similarity of domain sequences. The BES1 genes were relatively conserved in the Amaranthaceae family; there are 6 and 7 BES1 proteins in spinach and sugar beet respectively, and 6 from each

of them have counterparts. There was no correspondent for protein Bv4_085230 in spinach, which may have resulted from actual gene loss in spinach or imperfect annotation of the spinach genome. Interestingly, in groups A and C, the BES1 proteins from moss and spruce shared more similarity with one another than with fern, which was not consistent with plant classification rules (Fig 5).

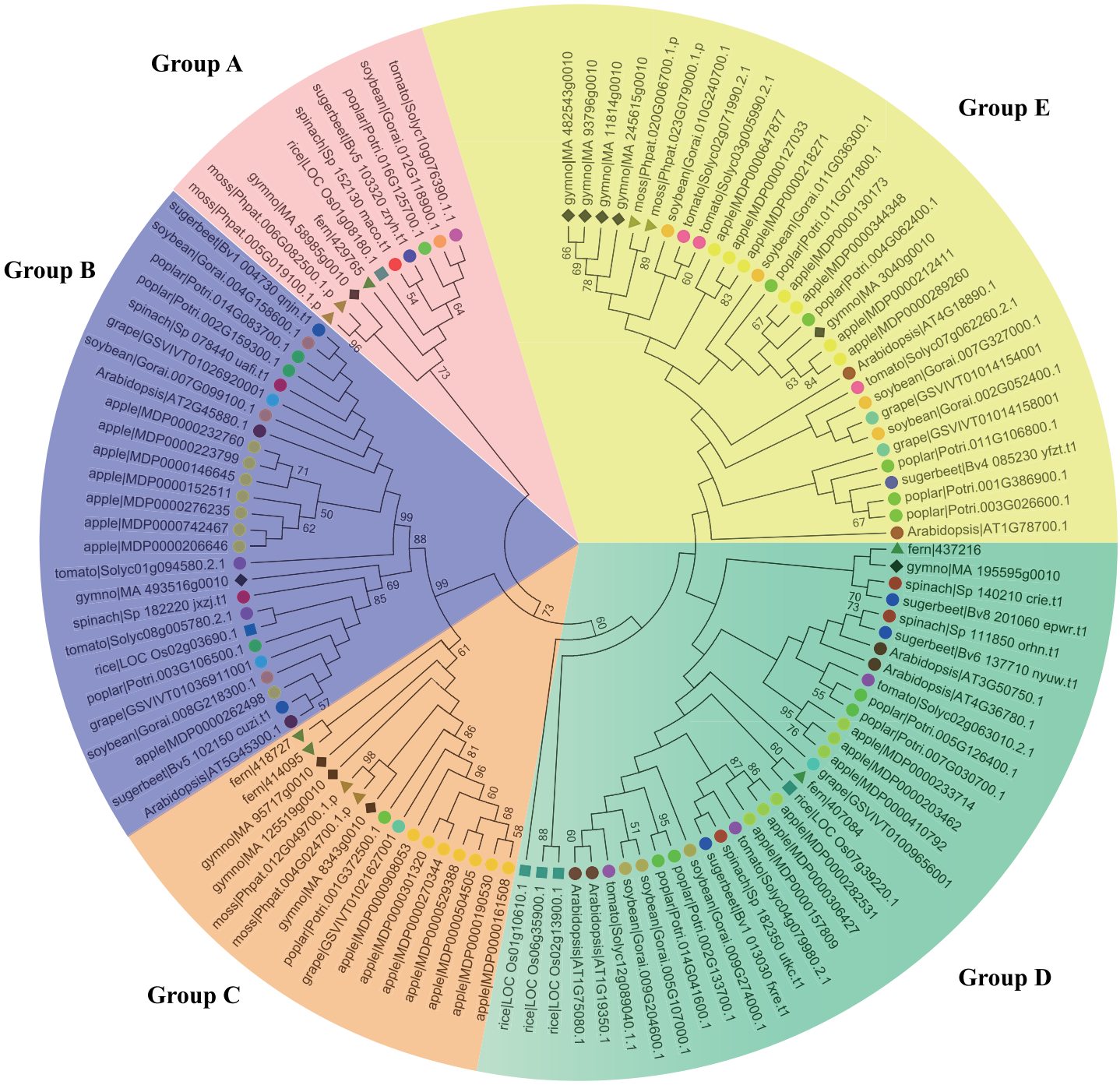

Fig 5. Phylogenetic tree of twelve plants constructed based on protein sequences of BES1/ domains using the maximum-parsimony method. The different species can be distinguished by different shapes and colors.

doi:10.1371/journal.pone.0152706.g005

Useful tools for genome and genes

GBrowse, developed by the Generic Model Organism Database Project, is the most widely used genomic browser among organism databases, because the browser is both universal and customizable. It is possible to quickly and accurately locate regions of interest in spinach genes, and to display sequences and annotations visually. In SpinachDB, the GBrowse tool is highly configurable and portable; users can view eight data tracks including Gene, mRNA, CDS (Coding Sequence), SSR, SNP, 6-frame translation, the DNA/GC Content, and restriction sites of the spinach genome. The annotation and expression of genes that were measured using the FPKM method were integrated inside GBrowse, to provide users a way to comparably study gene function and expression in different samples. Furthermore, all of the useful SNP data are also available in GBrowse, allowing users to view the SNP information from cultivars of interest and referent species. This is robust tool for breeders to identify markers that may relate to traits of interest.

All sequences and annotations of spinach genome, genes, proteins and ESTs were deposited in SpinachDB. SpinachDB provides a BLAST search interface which allows users to perform sequence similarity searches against the spinach assembled genome, genes, cDNA, CDS, EST and proteins. Since the annotations using NR, KEGG, KOG, Swiss-Prot and TrEMBL databases were deposited in SpinachDB, users can also perform keyword searches in SpinachDB, which will obtain a list of annotated genes with the specific keyword. Similar searches could also be performed to search orthologs, paralogs and gene families.

Supporting Information

S1 File. Annotation, classification, and expression information of spinach genes in tables.
(XLSX)

Acknowledgments

This work was supported by Natural Science Foundation of Shanghai (15ZR1436700) and Shanghai Science and Technology Talents Project (14XD1425100) and the Scientific Research Project in Public Agricultural Industry (201403032). Thanks to Dr. Sue Mischke in USDA/ARS/Beltsville Research Center for her carefully check on our manuscript. Special thanks to the Editor and Reviewers for their suggestions and comments that helped to improve our manuscript.

Author Contributions

Conceived and designed the experiments: WMZ HWT. Performed the experiments: HWT XDY. Analyzed the data: HWT XDY. Contributed reagents/materials/analysis tools: HWT XDY. Wrote the paper: XDY HWT.

References

1. Koh E, Charoenprasert S, Mitchell AE (2012) Effect of organic and conventional cropping systems on ascorbic acid, vitamin C, flavonoids, nitrate, and oxalate in 27 varieties of spinach (Spinacia oleracea L.). J Agric Food Chem 60: 3144–3150. doi: 10.1021/jf300051f PMID: 22393895
2. Longnecker MP, Newcomb PA, Mittendorf R, Greenberg ER, Willett WC (1997) Intake of carrots, spinach, and supplements containing vitamin A in relation to risk of breast cancer. Cancer Epidemiol Biomarkers Prev 6: 887–892. PMID: 9367061
3. Yang Y, Marczak ED, Yokoo M, Usui H, Yoshikawa M (2003) Isolation and antihypertensive effect of angiotensin I-converting enzyme (ACE) inhibitory peptides from spinach Rubisco. J Agric Food Chem 51: 4897–4902. PMID: 12903942

4. Song W, Derito CM, Liu MK, He X, Dong M, Liu RH (2010) Cellular antioxidant activity of common vegetables. J Agric Food Chem 58: 6621–6629. doi: 10.1021/jf9035832 PMID: 20462192
5. Xu Y, Lu Y, Xie C, Gao S, Wan J, Prasanna BM (2012) Whole-genome strategies for marker-assisted plant breeding. Molecular breeding 29: 833–854.
6. Khattak JZK, Torp AM, Andersen SB (2006) A genetic linkage map of Spinacia oleracea and localization of a sex determination locus. Euphytica 148: 311–318.
7. Onodera Y, Yonaha I, Masumo H, Tanaka A, Niikura S, Yamazaki S, et al. (2011) Mapping of the genes for dioecism and monoecism in Spinacia oleracea L.: evidence that both genes are closely linked. Plant Cell Rep 30: 965–971. doi: 10.1007/s00299-010-0998-2 PMID: 21301852
8. Sather DN, York A, Pobursky KJ, Golenberg EM (2005) Sequence evolution and sex-specific expression patterns of the C class floral identity gene, SpAGAMOUS, in dioecious Spinacia oleracea L. Planta 222: 284–292. PMID: 15940462
9. Bevan MW, Uauy C (2013) Genomics reveals new landscapes for crop improvement. Genome Biol 14: 206. doi: 10.1186/gb-2013-14-6-206 PMID: 23796126
10. Xu ZS, Tan HW, Wang F, Hou XL, Xiong AS (2014) CarrotDB: a genomic and transcriptomic database for carrot. Database (Oxford) 2014.
11. Nagamura Y, Antonio BA, Sato Y, Miyao A, Namiki N, Yonemaru J, et al. (2011) Rice TOGO Browser: A platform to retrieve integrated information on rice functional and applied genomics. Plant Cell Physiol 52: 230–237. doi: 10.1093/pcp/pcq197 PMID: 21216747
12. Shen D, Sun H, Huang M, Zheng Y, Li X, Fei Z (2013) RadishBase: a database for genomics and genetics of radish. Plant Cell Physiol 54: e3. doi: 10.1093/pcp/pcs176 PMID: 23239846
13. Mihara M, Itoh T, Izawa T (2010) SALAD database: a motif-based database of protein annotations for plant comparative genomics. Nucleic Acids Res 38: D835–842. doi: 10.1093/nar/gkp831 PMID: 19854933
14. Rouard M, Guignon V, Aluome C, Laporte MA, Droc G, Walde C, et al. (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res 39: D1095–1102. doi: 10.1093/nar/gkq811 PMID: 20864446
15. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40: D1178–1186. doi: 10.1093/nar/gkr944 PMID: 22110026
16. Sather DN, Golenberg EM (2009) Duplication of AP1 within the Spinacia oleracea L. AP1/FUL clade is followed by rapid amino acid and regulatory evolution. Planta 229: 507–521. doi: 10.1007/s00425-008-0851-9 PMID: 19005675
17. Minoche AE, Dohm JC, Schneider J, Holtgrawe D, Viehover P, Montfort M, et al. (2015) Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol 16: 184. doi: 10.1186/s13059-015-0729-7 PMID: 26328666
18. Dohm JC, Minoche AE, Holtgrawe D, Capella-Gutierrez S, Zakrzewski F, Tafer H, et al. (2014) The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature 505: 546–549. doi: 10.1038/nature12817 PMID: 24352233
19. Coordinators NR (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44: D7–D19. doi: 10.1093/nar/gkv1290 PMID: 26615191
20. Jungo F, Bougueleret L, Xenarios I, Poux S (2012) The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data. Toxicon 60: 551–557. doi: 10.1016/j.toxicon.2012.03.010 PMID: 22465017
21. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research 42: D199–D205. doi: 10.1093/nar/gkt1076 PMID: 24214961
22. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5: R7. PMID: 14759257
23. Xu ZS, Tan HW, Wang F, Hou XL, Xiong AS (2014) CarrotDB: a genomic and transcriptomic database for carrot. Database-the Journal of Biological Databases and Curation.
24. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. (2014) Pfam: the protein families database. Nucleic Acids Research 42: D222–D230. doi: 10.1093/nar/gkt1223 PMID: 24288371
25. Eddy SR (2011) Accelerated Profile HMM Searches. PLoS Comput Biol 7: e1002195. doi: 10.1371/journal.pcbi.1002195 PMID: 22039361

26. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189. PMID: 12952885

27. Patel RK, Jain M (2012) NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. Plos One 7.

28. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14: R36. doi: 10.1186/gb-2013-14-4-r36 PMID: 23618408

29. Ghosh S, Chan CK (2016) Analysis of RNA-Seq Data Using TopHat and Cufflinks. Methods Mol Biol 1374: 339–361. doi: 10.1007/978-1-4939-3167-5_18 PMID: 26519415

30. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28: 511–U174. doi: 10.1038/nbt.1621 PMID: 20436464

31. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303. doi: 10.1101/gr.107524.110 PMID: 20644199

32. Deng X (2011) SeqGene: a comprehensive software solution for mining exome- and transcriptome-sequencing data. BMC Bioinformatics 12: 267. doi: 10.1186/1471-2105-12-267 PMID: 21714929

33. Altaf-Ul-Amin M, Katsuragi T, Sato T, Kanaya S (2015) A Glimpse to Background and Characteristics of Major Molecular Biological Networks. Biomed Res Int 2015: 540297. doi: 10.1155/2015/540297 PMID: 26491677

34. Fang WP, Meinhardt LW, Tan HW, Zhou L, Mischke S, Zhang D (2014) Varietal identification of tea (Camellia sinensis) using nanofluidic array of single nucleotide polymorphism (SNP) markers. Hortic Res 1: 14035. doi: 10.1038/hortres.2014.35 PMID: 26504544

35. Wang B, Tan HW, Fang W, Meinhardt LW, Mischke S, Matsumoto T, et al. (2015) Developing single nucleotide polymorphism (SNP) markers from transcriptome sequences for identification of longan (Dimocarpus longan) germplasm. Hortic Res 2: 14065. doi: 10.1038/hortres.2014.65 PMID: 26504559

36. Zhou L, Matsumoto T, Tan HW, Meinhardt LW, Mischke S, Wang B, et al. (2015) Developing single nucleotide polymorphism markers for the identification of pineapple (Ananas comosus) germplasm. Hortic Res 2: 15056. doi: 10.1038/hortres.2015.56 PMID: 26640697

37. Xu C, Jiao C, Zheng Y, Sun H, Liu W, Cai X, et al. (2015) De novo and comparative transcriptome analysis of cultivated and wild spinach. Sci Rep 5: 17706. doi: 10.1038/srep17706 PMID: 26635144

38. Yan J, Yu L, Xuan J, Lu Y, Lu S, Zhu W (2016) De novo transcriptome sequencing and gene expression profiling of spinach (Spinacia oleracea L.) leaves under heat stress. Sci Rep 6: 19473. doi: 10.1038/srep19473 PMID: 26857466

39. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, et al. (2000) Comparative genomics of the eukaryotes. Science 287: 2204–2215. PMID: 10731134

40. Fujioka S, Yokota T (2003) Biosynthesis and metabolism of brassinosteroids. Annu Rev Plant Biol 54: 137–164. PMID: 14502988

41. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30: 2725–2729. doi: 10.1093/molbev/mst197 PMID: 24132122