De novo Sequencing, Characterization, and Comparison of Inflorescence Transcriptomes of *Cornus canadensis* and *C. florida* (Cornaceae)

Jian Zhang1,2, Robert G. Franks2, Xiang Liu2, Ming Kang3, Jonathan E. M. Keebler4, Jennifer E. Schaff4, Hong-Wen Huang3*, Qiu-Yun (Jenny) Xiang2*

1 State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, P.R. China, 2 Department of Plant and Microbial Biology, North Carolina State University, Raleigh, North Carolina, United States of America, 3 CAS Key Laboratory of Plant Resource Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, P.R. China, 4 Bioinformatics Analyst and Consultant Genomic Sciences Laboratory, North Carolina State University, Raleigh, North Carolina, United States of America

**Abstract**

**Background:** Transcriptome sequencing analysis is a powerful tool in molecular genetics and evolutionary biology. Here we report the results of de novo 454 sequencing, characterization, and comparison of inflorescence transcriptomes of two closely related dogwood species, *Cornus canadensis* and *C. florida* (Cornaceae). Our goals were to build a preliminary source of genome sequence data, and to identify genes potentially expressed differentially between the inflorescence transcriptomes for these important horticultural species.

**Results:** The sequencing of cDNAs from inflorescence buds of *C. canadensis* (cc) and *C. florida* (cf), and normalized cDNAs from leaves of *C. canadensis* resulted in 251799 (ccBud), 96245 (ccLeaf) and 114648 (cfBud) raw reads, respectively. The de novo assembly of the high quality (HQ) reads resulted in 36088, 17802 and 21210 unigenes for ccBud, ccLeaf and cfBud. A reference transcriptome for *C. canadensis* was built and assembled using HQ reads of ccBud and ccLeaf, containing 40884 unigenes. Reference mapping and comparative analyses found 10926 sequences were putatively specific to ccBud, and 6979 putatively specific to cfBud. Putative differentially expressed genes between ccBud and cfBud that are related to flower development and/or stress response were identified among 7718 shared sequences by ccBud and cfBud. Bi-directional BLAST found 87 (41.83% of 208) of Arabidopsis genes related to inflorescence development had putative orthologs in the dogwood transcriptomes. Comparisons of the shared sequences by ccBud and cfBud yielded 65931 high quality SNPs between two species. The twenty unigenes with the most SNPs are listed as potential genetic markers for evolutionary studies.

**Conclusions:** The data provide an important, although preliminary, information platform for functional genomics and evolutionary developmental biology in *Cornus*. The study identified putative candidates potentially involved in the genetic regulation of inflorescence evolution and/or disease resistance in dogwoods for future analyses. Results of the study also provide markers useful for dogwood phylogenomic studies.

Citation: Zhang J, Franks RG, Liu X, Kang M, Keebler JEM, et al. (2013) De novo Sequencing, Characterization, and Comparison of Inflorescence Transcriptomes of *Cornus canadensis* and *C. florida* (Cornaceae). PLoS ONE 8(12): e82674. doi:10.1371/journal.pone.0082674

Editor: Ting Wang, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China

Received August 25, 2013; Accepted October 25, 2013; Published December 27, 2013

Copyright: © 2013 Zhang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by a grant from National Nature Science Foundation of the United States grant (IOS-1024629) and a CAS/SAFEA International Partnership Program for Creative Research Teams from China. This study was also benefited from a National Nature Science Foundation of China (grant no. 31100171). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: jenny_xiang@ncsu.edu (QX); huanghw@mail.scbg.ac.cn (HH)

Introduction

The dogwood genus (*Cornus* L.) belongs to the eu dicot family Cornaceae in the order Cornales of the Asterids clade. It consists of approximately 55 species that are mostly evergreen or deciduous trees or shrubs, and rarely rhizomatous herbs [1–4]. Many dogwood species are valued in horticulture because of their spectacular blooms in the spring and brightly colored fruits, le afes, or stems in the fall or winter. A short list of the well-known examples includes the flowering dogwood *C. florida* L., the kousa dogwood *C. kousa* Hance, the cornelian cherry *C. mas* L., the bloodtwig dogwood *C. sanguinea* L., the red-osier dogwood *C. sericea* L. (syn. *C. stolonifera* Michx.), and the bunchberry *C. canadensis* L. f. One of the cornelian species *C. officinalis* Siebold & Zucc. is also highly valued in Chinese medicine and cultivated as a crop in China. Dogwoods are widely distributed in the north temperate regions extending to the tropical and subtropical areas of Asia, America, and Africa, with several species isolated in small areas of these continents [3,5]. One striking feature of the dogwood genus is the considerable variation in morphology of inflorescences, bracts, and fruits among species [1,3–7]. A number of studies on the taxonomy, phylogeny, biogeography, and morphological evolution of this genus have been recently conducted to better understand the biodiversity and evolutionary history of this
Comparative Transcriptome Sequencing in *Cornus* L.

Economically important group [1,3–15,39]. For example, phylogenetic studies using molecular data have shown that the modern dogwood species represent descendants of four closely related evolutionary lineages that have diverged in inflorescence morphology: (1) the blue- or white-fruited dogwoods bearing large, elongated inflorescences with rudimentary bracts on the inflorescence branches, e.g., *C. sericea*, *C. sanguinea*, *C. macrophylla* Wall.; (2) the cornelian cherries with umbel-like inflorescences that are subtended by four non-petaloid involucral bracts (bracts at the base of inflorescences), e.g., *C. mas*, *C. officinalis*; (3) the big-bracted dogwoods with head-like inflorescences subtended by four or six large, petaloid, involucral bracts, e.g., *C. florida*, *C. kousa*, *C. nuttallii* Aud. and; (4) the dwarf dogwoods, rhizomatous herbs with small, condensed, dichasia subtended by four large, petaloid, involucral bracts, e.g., *C. canadensis*, *C. suecica* L. [1,10–12]. This variation among the four closely related lineages of dogwoods provides a very useful system to study inflorescence evolution and the underlying molecular mechanisms. In particular, the genus is an excellent system for studying the origins of umbel-like and head-like inflorescences and petaloid bracts that have evolved many times during angiosperm diversification, but are poorly understood. Comparative analyses of inflorescence developmental morphology based on the dogwood phylogeny suggested that umbel-like and head-like inflorescences in *Cornus* evolved independently from elongated forms through an umbellate dicaulism ancestor, and the petaloid bracts in *C. canadensis* and *C. florida* also evolved independently via different developmental mechanisms [14–15]. These previous phylogenetic and developmental studies provide the necessary framework for investigating the genetic basis of inflorescence evolution in the genus. However, the genus lacks sequence data at genomic scale to facilitate such investigation.

In addition to the striking evolutionary divergence in inflorescence morphology, the four major dogwood lineages also display variation in fruit type (simple vs. compound), fruit color (white, blue, black, red, purple red), growth habit (trees, shrubs, herbs), chromosome number (x = 11, 10, 9), pollination mechanism, freezing tolerance, wood anatomy, phytochemistry, as well as disease resistance [4,16–17]. Of particular interest is the susceptibility of this genus to the fungal disease - dogwood anthracnose that affects some North American species and has caused serious damage and decrease of natural populations of *C. florida* throughout its range [17–19]. *Cornus florida* is an important ecological element of the temperate forests in eastern North America and the state flower of North Carolina and the state tree of Virginia. This species is the most vulnerable victim of this disease, while other species have shown to be resistant to the dogwood anthracnose to various degrees [17–19]. At present, the genetic basis and molecular mechanism for the resistance and susceptibility of dogwoods to the fungal pathogen are still unknown. It is well recognized that genome and transcriptome sequences are fundamental to genetic research of morphology. To our knowledge, there are few genomic or transcriptomic resources of any dogwood species available to the public.

Transcriptome analysis provides valuable insights into genes and gene activities responsible for differences of organ morphology during the developmental processes. As demonstrated in recent studies, transcriptome analyses using next generation sequencing tools have become increasingly common to unravel genetic network regulating the development and morphology [20–22]. Here we applied this approach to characterize two inflorescence transcriptomes and one leaf transcriptome of *Cornus* L. to provide the first source of transcriptomic data for the genus. Comparisons between the two inflorescence transcriptomes permitted identification of a pool of potentially differential expressed genes (DEGs) between the two inflorescence types, in addition to a number of genes containing SNP markers useful for evolutionary studies. Specifically, we employed Roche 454 GS-FLX, next generation sequencing (NGS) method, for de novo sequencing of inflorescence transcriptomes of *C. canadensis* (referred to as ccBud hereafter) and *C. florida* (referred to as cfBud hereafter) that are from two sister lineages, the dwarf dogwoods and the big-bracted dogwoods, respectively [14] (Figure 1). The leaf transcriptome of *C. canadensis* (referred to as ccLeaf hereafter) was also sequenced using this method to maximize the coverage for transcripts of the species and to provide a background reference in identifying only the genes involved in inflorescence development. The transcriptomes were assembled de novo using the pool of sequencing data, and the corresponding contigs and singletons were annotated. According to gene ontology (GO) terms, homologs of genes potentially involved in flower development (including those related to inflorescence development) and stress response (including those related to disease response) were identified for each species transcriptome. Putative DEGs from each of these categories between *C. canadensis* and *C. florida* inflorescence transcriptomes were also identified based on Reads per Kilobase per Million mapped Reads (RPKM) [23]. Furthermore, we used bi-directional (or reciprocal) BLAST to determine if orthology of *Arabidopsis* genes known to regulate inflorescence architecture and/or express in inflorescence meristem were present in the ccBud and cfBud transcriptomes. The putative DEGs are to serve as the candidates of future quantitative Real-Time PCR (qRT-PCR) in situ hybridization analyses in order to identify genes contributing to the changes of inflorescence architectures in *Cornus*. Finally, genes containing high quality [HQ] single nucleotide polymorphisms (SNPs) between these two species were also identified to provide genetic markers for future phylogenomic and evolutionary ecological genomic studies that are fundamental to the conservation of dogwood biodiversity.

**Results**

**Sequencing and assembly.** The 454 sequencing of cDNA libraries of ccBud, ccLeaf and cfBud yielded 251799, 96245, and 114648 raw reads with average lengths of 395 bp, 317 bp and 392 bp, respectively and the most frequent length of 450–550 bp (Table 1; Figure S1). After the sequence trimming, 251679, 85703 and 114605 HQ reads of ccBud, ccLeaf and cfBud were obtained, corresponding to 99.95%, 89.05% and 99.96% of the original raw reads (Table 1). The average length of HQ reads was 391 bp for ccBud, 322 bp for ccLeaf and 388 bp for cfBud. The most frequent length of HQ reads was 395 bp, 317 bp and 392 bp, respectively and the most frequent length of 450–550 bp (Table 1; Figure S1).

**Figure 1. Inflorescence morphology of *Cornus canadensis* (A) and *C. florida* (B).**

doi:10.1371/journal.pone.0082674.g001
reads in all samples was also in the 450–550 bp range (Table 1; Figure S1). Through de novo assembly, 226675, 73775 and 98096 HQ reads of ccBud, ccLeaf and cfBud were assembled into 14656, 7574 and 7366 contigs, respectively (Table 1). The mean of average coverage of assembled contigs was assessed as five for ccBud, five for ccLeaf and four for cfBud (Table 1). And 21432, 10228 and 13844 HQ reads of ccBud, ccLeaf and cfBud were retained as singletons (Table 1). The average length of contigs and singletons was 699 bp and 350 bp for ccBud, 607 bp and 345 bp for ccLeaf, 680 bp and 369 bp for cfBud, respectively (Table 1). In total, the contigs and singletons resulted in 36008, 17802 and 21210 unigenes for ccBud, ccLeaf and cfBud, with the average length of 496 bp, 436 bp and 477 bp, respectively (Table 1). The distributions of contig, singleton, and unigene lengths and average coverage of assembled contigs for ccBud and cfBud are shown in Figure 2 and those for ccLeaf are shown in Figure S2.

After the HQ reads from ccBud and ccLeaf were combined together for de novo assembly of transcriptome for C. canadensis (referred to as ccTranscriptome hereafter), a total of 190494 contigs (average coverage of five), 21835 singletons, and 40804 unigenes were obtained, with the average length of 714 bp, 358 bp and 524 bp, respectively (Table 2). The distributions of contig length, singleton length, average coverage of assembled contigs, and unigene length for ccTranscriptome are shown in Figure 3.

Functional annotation

After BLAST search, a total of 22231 (61.60%) unigenes from ccBud, 12434 (69.85%) from ccLeaf and 13723 (64.70%) from cfBud had significant BLAST matches (Table 1). For both C. canadensis and C. florida, the top three species of BLAST hit were Vitis vinifera, Populus trichocarpa and Ricinus communis (Figure 4). And 17229 (47.74%) unigenes from ccBud, 9728 (54.65%) from ccLeaf and 10694 (50.42%) from cfBud had at least one GO term assigned (some genes have more than one GO term) (Table 1). The distribution of GO terms was very similar between C. canadensis and C. florida inflorescence transcriptomes (ccTranscriptome and cfBud) (Figure 5). At the GO level 2, the annotation configuration found 32965 sequences assigned to biological process, 25111 to molecular function, and 24486 to cellular component. The corresponding numbers in cfBud were 20163, 14127 and 13645, respectively. In the biological process category, the most abundant sequences in both transcriptomes were classified to “cellular process” (GO: 0005623) (53.85%) was highly represented term “cell” (GO: 0005623) (53.85%) was highly represented (Figure 6).

Comparative analysis

After reference mapping and BLAST search, there were 22057 consensus aligned sequences with BLAST hits for ccBud, 11206 for ccLeaf, and 8275 for cfBud. The de novo assembly of the unaligned HQ reads from cfBud yielded 14325 unigenes as the putative cfBud specific transcriptome (referred to as cfBud specific hereafter), with an average length of 408 bp (Table S1). The distributions of contig length, singleton length, average coverage of assembled contigs, and unigene length for cfBud specific transcriptome are shown in Figure S3. Among these unaligned unigenes of cfBud, 6979 (48.72%) sequences had BLAST matches (Table S1). Comparing these results from BLAST search between ccBud and cfBud, we found that 7718 sequences with BLAST hits were shared by ccBud and cfBud, and 10926 only in ccBud and 6979 only in cfBud.

Furthermore, among these putative species-specific sequences, 7984 for ccBud and 5103 for cfBud could be assigned to GO terms. The annotation results showed that 180 ccBud-specific sequences were assigned to “flower development” (GO: 0009906) and 633 “response to stress” (GO: 0006950), while in cfBud-specific sequences, the numbers were 67 sequences related to “flower development” and 405 “response to stress”. Among these, the contigs with at least five mapped reads are shown in Table S2 for ccBud and Table S3 for cfBud. For example, the amino acid sequences of Ccanadensis_transcriptome_contig6848 (23 mapped reads and 215.16 RPKM value) and cfBud_transcriptome_contig966 (15 mapped reads and 464.94 RPKM value) in the “flower development” category were annotated as a member of basic helix-loop-helix (bHLH) DNA-binding superfamily proteins and an Eos-bah-pha domain-containing protein, respectively (Table S2 and S3). The amino acid sequences of Ccanadensis_transcriptome_contig14719 (38 mapped reads and 465.54 RPKM value) and cnBud_transcriptome_contig1202 (77 mapped reads and 1705.30 RPKM value) in the “response to stress” category were annotated as adefensin protein and a major allergen, respectively (Table S2 and S3).

Among a total of 7718 sequences shared by ccBud and cfBud, putative DEGs between ccBud and cfBud were identified based on the RPKM values that are calculated from the read counts mapped onto the reference transcriptome (ccTranscriptome) and the corresponding unigene length. In these shared sequences, 6350 were annotated, including 945 with at least two-fold increase of RPKM value in ccBud than in cfBud, and 2513 with at least two-fold increase of RPKM value in cfBud than in ccBud. In “flower development” and “response to stress” categories, the putative DEGs with at least five mapped reads in ccBud or cfBud are listed in Table S4 and Table S5. To provide an example, in the “flower development” category, there were 66 ccBud reads mapped onto the Ccanadensis_transcriptome_contig3484 with 225.04 RPKM value while only two cfBud reads aligned onto this contig with 17.33 RPKM value (Table S4), which probably reflects a higher level of expression of the gene corresponding to contig3484 in ccBud. The putative protein of this contig was annotated to belong to the Enol-CoA hydratase/isomerase family (Table S4). In another example, in the “flower development” category, 13 cfBud reads were aligned onto the Ccanadensis_transcriptome_contig10032 with 261.48 RPKM value but only one ccBud read mapped onto the same contig with 7.92 RPKM value (Table S5). The putative protein of contig10032 was annotated to be a HUA enhancer 2 (Table S5).
Bi-directional BLAST searches between *Arabidopsis* and two dogwood transcriptomes showed that 87 (41.83% of 208) of *Arabidopsis* genes that regulate inflorescence architecture (27) and/or are expressed in inflorescence meristem (60) had putative orthologs in *C. canadensis* transcriptome and/or *C. florida* inflorescence specific transcriptome (Table S6). And some putative orthologs had greater RPKM values in ccBud than in cfBud (Table S6). Moreover, among the 87 genes with bi-directional best
hits, 35 (40.23%) genes only had putative orthologs in *C. canadensis* inflorescence transcriptome, and 13 (14.94%) only had putative orthologs in *C. florida* inflorescence specific transcriptome (Table S6).

**SNP characterization**

Mapping cfBud reads onto the ccTranscriptome identified 65931 high quality SNPs (excluding all gaps) from 2542 unigenes, including 2305 contigs and 237 singletons, with an average of 26 SNPs per unigene, between *C. canadensis* and *C. florida*. These predicted SNPs included 38385 transitions and 27546 transversions, at approximately a 1.4 : 1 ratio (Table 3). In addition, the frequencies between A/G and C/T transitions and frequencies among the four transversion types (A/T, G/T, C/G, A/C) were similar (Table 3). Among the unigenes containing SNPs, 2053 had annotation information, including 1983 contigs and 70 singletons. The information for the top 20 contigs containing relatively more SNPs is provided in Table S7. Taken the ccTranscriptome_contig4195 as an example, there was nearly equal read number from ccBud (33) and cfBud (34) mapped onto this contig, and 115 SNPs were found between the *C. canadensis* sequence and *C. florida* one, a proportion of 4.96% given the length of 2318 bp of the contig.

|                         | *Cornus canadensis* | *Cornus florida* |
|-------------------------|---------------------|------------------|
| Raw reads               | 251799              | 96245            |
| Average length of raw reads (bp) | 395                | 317              |
| HQ reads (% of total raw reads) | 251679 (99.95%) | 85703 (89.05%) |
| Average length of HQ reads (bp) | 391                | 322              |
| Assembled contigs       | 14656               | 7574             |
| Mean of average coverage | 5                  | 5                |
| Average length of contigs (bp) | 699                | 607              |
| Singletons              | 21432               | 10228            |
| Average length of singletons (bp) | 358               | 345              |
| Unigenes                | 36088               | 17802            |
| Average length of unigenes (bp) | 496               | 456              |
| Number of unigenes with BLAST matches (% of total unigenes) | 22231 (61.60%) | 12434 (69.85%) |
| Number of unigenes with GO annotation (% of total unigenes) | 17229 (47.74%) | 9728 (54.65%)|

![Figure 3](https://example.com/f3.png)

**Figure 3.** Assembly characteristics of *Cornus canadensis* transcriptome. (A) Length frequency distribution of assembled contigs. (B) Length frequency distribution of singletons. (C) Average coverage frequency distribution of assembled contigs. (D) Length frequency distribution of unigenes.
The annotation data showed that the putative protein of this contig was a homolog of a chloroplast heat shock protein (Table S7).

**Discussion**

Our preliminary comparative sequencing of inflorescence transcriptomes in two *Cornus* species, *C. canadensis* and *C. florida*, using the Roche 454 GS-FLX method and de novo assembly generated abundant useful data. The sequencing produced HQ reads mostly 350–550 bp in length, and a great proportion of these (>85%) could be assembled into contigs that were longer than 500 bp (Figure S1 and Table 1). Although the average coverage of assembled contigs (mostly 5) was relatively low due to single sequencing run [24–26], the assembled ccTranscriptome of *C. canadensis* still provides the first reference transcriptome for *Cornus* species, offering a platform to perform comparative analysis to identify putative DEGs and interspecies SNPs for future evolutionary developmental biology and phylogenomic studies.

A large proportion of the unigenes had significant BLAST hits (61.60%, 69.85% and 64.70% of ccBud, ccLeaf and cfBud, respectively, Table 2).

![Top-hit species distribution of BLAST matches of two *Cornus* inflorescence transcriptome sequences.](https://plosin.org/articles/82674.g004)
respectively), and most of the best BLAST hits were plant proteins (Table 1 and Figure 4). However, a fraction of unigenes from our study had no significant matches to NCBI-NR database and TAIR database at the E-value threshold of $10^{-6}$. This phenomenon has also been reported in many other plants and the proportion of presumably unique sequences without BLAST hits were considered to be affected by species, sequencing depth, read length, and BLAST parameters [21,27–29]. And “non-BLASTable” genes may include those rapidly evolving genes having homologs in other species but too divergent in Cornus to be recognized during BLAST. In addition, taxon-specific genes in the Cornus species that are missing from other databases may also contribute to the “non-BLASTable” category [21,27–29].

Identification of putative differentially expressed genes

Due to reasonably high cost of 454 sequencing, biological replicates were not included in the study, which prevented proper statistical testing on identification of DEGs. Nonetheless, annotated contigs abundant in one sample, rare or absent in the other still provide a valuable data source for selection of putative DEGs for analyses using qRT-PCR, in situ hybridization, and genetic transformation analyses (currently in progress in our lab) to evaluate their potential contributions to the inflorescence differences in Cornus (Table S2 and S3). For example, the putative ccBud specific contig, Ccanadensis_transcriptome_contig3886, is a homolog of Arabidopsis ER (ERECTA) gene (AT2G26330) (Table S2). A previous study in Arabidopsis suggested that loss of function in ER gene confers a corymb-like inflorescence due to a reduction in the length of stem internodes and pedicels [30–31]. Besides promoting inflorescence elongation, this gene was also found to regulate multiple developmental processes as well as environmental and biotic responses [32–33]. Therefore, we speculated that the absence of expression or the defect of function for these orthologous genes might contribute to the development of head-like and umbel-like compact inflorescence architecture in the big- bracted dogwoods (such as C. florida) and cornelian cherries (such as C. officinalis). Furthermore, among the 7718 unigenes shared by ccBud and cfBud, a number of these showed evident differences in RPKM values (Table S4 and S5). For instance, Ccanadensis_transcriptome_contig3484 had almost 13-fold increase of RPKM value in ccBud than in cfBud, and was annotated to be homologous to AIM1 (ABNORMAL INFLORESCENCE MERISTEM 1) gene (AT4G29010) in Arabidopsis (Table S4). Mutation of AIM1 gene was reported to affect inflorescence and floral development in Arabidopsis [34]. Another example is Ccanadensis_transcriptome_contig18209, which had nearly 19-fold increase of RPKM value in cfBud than in ccBud, and was identified to be a homolog of ACL5 (ACAULIS 5) in Arabidopsis (Table S5). The ACL5 gene is required for internodal elongation after flowering and its mutant exhibit a severe dwarf phenotype, with dramatically shortened inflorescence internodes and premature arrest of the inflorescence meristem [35–36], which is similar to the phenotype of C. canadensis. In addition, Ccanadensis_transcriptome_contig16216 was annotated to be a homolog of WOX9 (WUSCHEL-related homeobox 9) with ten mapped-reads in ccBud and zero in cfBud (Table S2). In Arabidopsis, WOX9 is required for meristem growth and maintenance [37]. In tomato, the homolog of WOX9 gene (COMPOUND INFLORESCENCE, S) has been reported to...
determine inflorescence branching [38]. In the sympodial tomato, a mutant exhibited highly branched inflorescence [38], while in the monopodial Arabidopsis, WOX9 RNAi constructs driven by a floral specific promoter also resulted in branching of floral meristems (Katie Liberatore, unpublished data). These data suggest a possible network of genes in regulating the development of inflorescences in C. canadensis and C. florida.

However, it must be noted that the total number of sequences identified to be specific to ccBud or cBud via mapping onto ccTranscriptome reference might be overestimated in this study, due to the following reasons. First, the relatively low coverage might miss some low expressed genes in one of the two species, resulting in false identification of species-specific genes/transcripts.

Second, in the cases where non-overlapping regions of the homologous gene sequences were recovered in cBud and ccTranscriptomes, the alignment in reference mapping would have been failed, resulting in false identification of species-specific transcripts in cBud. We investigated this by performing local BLAST search of randomly selected species-specific reads and found that some did match the same homologous gene sequences but aligned to the different regions of the genes. For example, the putative protein of cBud_specific_singleton7973 that could not be aligned onto ccTranscriptome, matched the same protein sequence of Arabidopsis as that of ccTranscriptome_contig1425 did, indicating that the corresponding gene was not cBud specifically expressed (Figure S4). This fact suggested that the

**Figure 6. Gene ontology categories of unigenes in Cornus canadensis transcriptome.** The proportion of annotated unigenes from ccTranscriptome classified into three gene ontology categories: (A) biological process, (B) molecular function and (C) cellular component. doi:10.1371/journal.pone.0082674.g006
non-overlapping regions of the homologous gene sequences recovered in ccTranscriptome and cfBud transcriptome can lead to the failure of the reference mapping for some cfBud sequences, therefore, resulting in false identification of some cfBud specifically expressed genes. Third, the potentially “high” divergence of homologous gene sequences between C. canadensis and C. florida due to phylogenetic divergence could also have led to false identification of cfBud specific transcripts. The two species have diverged >40 million years ago [12,39]. However, if a sequence from cfBud of C. florida could be aligned to the homologous sequence from Arabidopsis, it would not fail to align to the homologous sequence from ccTranscriptome of C. canadensis, a congeneric sister lineage to C. florida, unless the sequences in each species are from the non-overlapping regions of the homologous gene sequences. Therefore, we rechecked the putative species-specific genes of interest, e.g., those listed in Table S2 and S3, to confirm that they were not falsely identified. The putative DEGs (including the species-specific ones and those shared by ccBud and cfBud, but with significantly different RPKM values in the two species) will serve as our best initial choices of candidate genes for analyses using qRT-PCR and in situ hybridization to characterize, in detail, their expression patterns in different dogwood species to evaluate their potential roles. Those displaying differences in expression pattern among species would then serve as the candidates for functional analyses through in vivo gene transformation using the systems established in C. canadensis [40,41] and Arabidopsis [42].

Orthologs of Arabidopsis inflorescence architecture related genes in dogwood inflorescence transcriptomes

We detected putative orthologs of most (27/41) of the reported inflorescence architecture related genes of Arabidopsis from the inflorescence transcriptomes of C. canadensis and C. florida (Table S6). These included some well-known regulators in flowering and inflorescence development, such as SOC1, FUL, KNAT1, KNAT6, LFY, etc (Table S6). This evidence suggests that the fundamental programs in flowering and inflorescence development may be conserved between Arabidopsis and dogwoods. In Arabidopsis, there were two major molecular programs controlling the inflorescence architecture. One regulates inflorescence internode elongation that includes the ER, ACL5, and ACA10 genes among others. Orthologs of these genes regulating inflorescence internode elongation were also found in the inflorescence transcriptomes of both C. canadensis and C. florida, with some exhibiting differences in expression levels (Table S6). These genes might play a role in the divergence of inflorescence architecture between the dogwoods species, which can be tested by further investigation. The other program regulates inflorescence and floral meristem identity, which includes the well-known inflorescence architecture regulators, LFY (LEAFY), TFL1 (TERMINAL FLOWER 1), and API (APETALA1). LFY and API were found to primarily promote floral meristem identity, while TFL1 specify shoot identity and repress the floral meristem identity ([43] for a review). The lfy and api mutants caused delay of flowering and partial conversion of flowers into shoots or shoot-like structures, whereas tfl1 mutant generated short inflorescences that terminate with a flower [44–51]. The orthologs of these genes and their inflorescence related functions have been reported in other plants, including tomato and petunia from the Asterids clade [38,52–54]. We found LFY orthologous in our transcriptome data (Table S6), and API and TFL1 orthologs in the inflorescence cDNAs of both species by gene cloning (unpublished data). These data similarly suggest that the key regulators on inflorescence architecture among the dogwood lineages remain to be studied. Our recent investigation of LFY orthology in Cornus (Cornus) [56]. It suggested that accumulation of transcripts of CorLFY might not contribute to the evolutionary changes of inflorescence architectures in the dogwood genus.

Identification of SNPs for evolutionary study

We identified 65931 high quality SNPs between C. canadensis and C. florida sequences that were distributed among 2542 unigenes, with 26 SNPs per unigene. The level of sequence divergence between C. canadensis and C. florida is not particularly high, given that the two species have started to diverge in the Paleogene [4,12]. Some examples of the SNP-containing genes found in this study are listed in Table S7. It is noteworthy that the SNP-containing genes identified in this study included some that have already been used as phylogenetic markers in other plant groups, such as the sucrose synthase gene (homologous to Ccanadensis_transcriptome_contig3878) which has been used in phylogenetic study of Leguminosae, and the eukaryotic translation elongation factor (homologous to Ccanadensis_transcriptome_-contig3553) which has been used in phylogenetic study of eukaryotes [57–58]. The large number of SNP-containing genes identified in this study provides markers at the genomic scale to resolve phylogenetic relationships among dogwood species and their close relatives, e.g., for phylogenetic study of Cornaceae and Cornales, and also offers a source of candidate genes to perform ecological and evolutionary genetic study of adaptive traits in dogwood species, as done in other species [59]. Although the major clades within Cornales have been identified using conventional molecular markers (e.g., several chloroplast DNA genes and nuclear ribosomal genes) in previous studies, the relationships among the clades are still uncertain and/or controversial among different markers [1,12,39]. Therefore, it is expected that more nuclear gene data will be helpful to clarify the phylogenetic relationships among different clades.

| SNP type | Variations | Count number/Percentage |
|----------|------------|-------------------------|
| Transitions | A<->G | 19041/28.9% |
| C<->T | 19344/29.3% |
| Total | 38385/58.2% |
| Transversions | A<->C | 7697/11.7% |
| G<->T | 6840/10.4% |
| C<->G | 6130/9.3% |
| A<->C | 6879/10.4% |
| Total | 27546/41.8% |

doi:10.1371/journal.pone.0082674.t003
platforms to generate markers for the analyses. However, it should be noted that the choice of a putative marker in this study requires orthogonal comparison and variation assessment among species, which were not evaluated in our analyses.

Conclusion

The data from comparative 454 sequencing of transcriptomes in C. canadensis and C. florida provide the first transcriptome information platform for functional genomics and evolutionary developmental biology studies in Cornus. The study identified meaningful candidate genes for future analyses to understand the genetic mechanisms underlying dogwood inflorescence evolution. Furthermore, the study generated a wealth of potential genetic markers useful for genetic mapping, phylogenetic and population genetic studies.

Materials and Methods

RNA extraction and cDNA synthesis

The inflorescence buds of C. canadensis (ccBud) and C. florida (cfBud), as well as young and mature leaves of C. canadensis (ccLeaf), were collected from living plants and stored in RNAlater (Ambion of Applied Biosystems, Foster City, CA, USA) immediately after removing from plants. Materials were stored at −20°C until total RNA extraction. Total RNAs were extracted from tissues pooled from two or more plants of the same species using the modified CTAB RNA isolation method [60]. The unmaturated cDNAs of ccBud and cfBud were synthesized and processed into sequencing libraries according to Roche’s cDNA Rapid Library Preparation Method and standard Rapid Library kit (cat# 05 608 228 001). The cDNAs of ccLeaf were first synthesized using Evrogen’s (Moscow, Russia) MINT-Universal cDNA synthesis kit (cat# SK005), normalized using the Evrogen Trimmer kit (cat# NK003), and then processed into a sequencing library using Roche’s standard Rapid Library kit (all procedures following manufacturer’s recommendations). Concentration and quality of the libraries were assayed using Agilent’s (Santa Clara, CA, USA) Bioanalyzer.

454 sequencing and assembly

The rapid library adapters A and B were ligated on the cDNA samples, and then the three cDNA libraries (ccBud, ccLeaf and cfBud) were subjected to clonal amplification by emulsion PCR. The clonally amplified beads of the three samples were enriched and sequenced by a half-plate run on the 454 GS-FLX Titanium platform following manufacturer’s protocol (Roche Diagnostics, USA). Roche’s onboard base caller was used to generate files that were then processed into a sequencing library using CLC Genomics Workbench 4.6.1 with the long reads mapping parameters, mismatch cost of 2, insertion cost of 3, deletion cost of 3, length fraction of 0.5, and similarity of 0.8 [61–62]. The portion of HQ reads from cfBud that could not be mapped onto the reference transcriptome (ccTranscriptome) (referred to as “cfBud specific”) were assembled de novo using CLC Genomics Workbench 4.6.1 with default parameters described above.

BLAST and annotation

The contigs and singletons from all de novo assembled transcriptomes were compared to NCBI non-redundant (nr) protein database (NCBI-NR database) and Arabidopsis Information Resource (TAIR) database (TAIR10) using BLASTx algorithm with the E-value threshold of 10−8, in order to find the corresponding homologous genes from other species [63]. The Gene Ontology (GO) annotation was performed using Blast2GO program with the annotation cut-off of 10−8 [64–65]. The best five protein hits for each query were parsed out to create annotated tables, containing available information such as taxonomy, protein function, accession number, etc. And the best blast hits were used to retrieve associated GO terms describing biological process, molecular function, and cellular component [64]. In order to get the information on general functional categories, the final GO graph was generated, which summarized the distribution of the GO level 2 terms for three main categories (biological process, molecular function and cellular component) [26].

Comparative analysis of transcriptomes

The HQ reads from ccBud, cfBud, and ccLeaf transcriptomes were mapped onto the reference transcriptome (ccTranscriptome). After the reference mapping, the alignment was visualized by Integrative Genomics Viewer (IGV) [66]. The number of reads mapped onto each unigene and the length of each unigene were exported into excel table for expression comparison. The relative expression levels for unigenes were assessed by Reads per Kilobase per Million mapped Reads (RPKM), using the formula, $R(G)=\frac{(10^{9}C)}{NL}$, C = number of reads mapped to unigene, N = number of reads mapped in each library, L = length of unigene G [67]. A RPKM value with at least two-fold difference between the two samples was used as criteria to determine putative differentially expressed genes (DEGs) [68–69]. The contigs and singletons of ccTranscriptome that only had mapped reads from ccBud were treated as putative specific expressed genes of ccBud. The de novo assembled contigs and singletons from unmapped reads of cfBud (cfBud specific) were treated as putative specific expressed genes for cfBud. Furthermore, to search whether inflorescence regulators found in model plants are present in our transcriptome data, we selected 41 inflorescence architecture related genes reported in Arabidopsis and 167 inflorescence meristem expressed genes in Arabidopsis for local BLASTn (blast-2.2.28+) search against the assembly data from ccTranscriptome and cfBud specific (as local BLAST databases) with the E-value threshold of 10−8, and then the best BLAST hits from our local databases were used as queries to do BLASTx against TAIR10 with the E-value threshold of 10−8. Through this bidirectional BLAST [70–74], the evolutionary orthologs between Arabidopsis and two dogwood species were found (Table S6).
SNP discovery

In order to detect interspecies single nucleotide polymorphisms (SNPs) between *C. canadensis* and *C. florida*, we used cctranscriptome assembled from ccBud and ccLeaf HQ reads as a reference, and mapped the HQ reads from eBud onto the reference transcriptome through CLC Genomics Workbench 4.6.1. A SNP was called when all *C. florida* reads produced a consensus base that was different from that in the *C. canadensis* reference. Following recent publications [75–76], we called a SNP when it was supported by at least three reads and all the reads agreed on the same consensus base calls. The SNP sites between ccBud and eBud could be visualized by IGV.

Supporting Information

Figure S1 Length distribution of raw reads and high quality (HQ) reads. (A) ccBud, (B) eBud and (C) ccLeaf.

Figure S2 Assembly characteristics of *Cornus canadensis* leaf transcriptome (ccLeaf). (A) Length frequency distribution of assembled contigs. (B) Length frequency distribution of singletons. (C) Average coverage frequency distribution of assembled contigs. (D) Length frequency distribution of unigenes.

Figure S3 Assembly characteristics of *Cornus florida* inflorescence specific transcriptome (cfBud specific). (A) Length frequency distribution of assembled contigs. (B) Length frequency distribution of singletons. (C) Average coverage frequency distribution of assembled contigs. (D) Length frequency distribution of unigenes.

Figure S4 An example of protein alignment for local BLAST results. The protein sequences in the alignment were translated from Ccanadensis_transcriptome_contig1425, cfBud_transcriptome_singleton7973 and LFY of Arabidopsis, a query gene.

References

1. Xiang QY, Thomas DT, Zhang WH, Manchester SR, Murrell ZE. (2006) Species level phylogeny of the genus *Cornus* (Cornaceae) based on molecular and morphological evidence - implications for taxonomy and Tertiary intercontinental migration. Taxon 55(1): 9–36.
2. Murrell ZE. (1994) Dwarf dogwoods: intermediary and the morphological landscape. Syst Bot 19(4): 539–556.
3. Xiang QY, Boufford DE. (2005) Cornaceae, Mastiniaceae, Toricelliacae, Helwingiaceae, Asclepiadaceae. Wu ZY, Raven PH, Hong DY, Flora of China (Apiaceae through Ericaceae): Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis. 14: 206–234.
4. Xiang QY, Thomas DT. (2008) Tracking character evolution and biogeographic history through time in *Cornaceae* - does choice of methods matter? J Syst Evol 46(3): 349–374.
5. Eyde RH. (1988) Comprehending *Cornus*: puzzles and progress in the systematics of the dogwoods. Bot Rev 54: 233–351.
6. Murrell ZE. (1993) Phylogenetic Relationships in *Cornus* (Cornaceae). Syst Bot 18(3): 469–495.
7. Xiang QY, Brunsfeld SJ, Solis DE, Solis PS. (1996) Phylogenetic relationships in *Cornus* based on chloroplast DNA restriction sites: implications for biogeography and character evolution. Syst Bot 21(4): 515–534.
8. Xiang QY. (1990) Taxonomy of *Cornus shindleri* complex based on quantitative analysis of some characters. Bulletin of Botanical Research 9(1): 125–138.
9. Murrell ZE. (1996) A new section of *Cornus* in South and Central America. Syst Bot 21(5): 273–288.
10. Fan CZ, Xiang QY. (2001) Phylogenetic relationships within *Cornus* (Cornaceae) based on 26S rDNA sequences. Am J Bot 88(8): 1131–1136.
11. Fan CZ, Xiang QY. (2003) Phylogenetic analyses of Cornales based on 26S rRNA and combined 26S rDNA-matK-5.8S sequence. Am J Bot 90(9): 1157–1372.
12. Xiang QY, Thorne JL, Seo T, Zhang W, Thomas DT, et al. (2008) Rates of nucleotide substitution in *Cornaceae* (Cornales) - Pattern of variation and underlying causal factors. Mol Phylogenet Evol 49(1): 327–342.
13. Zhang WH, Xiang QY, Thomas DT, Wiegmann BM, Frohlich MW, et al. (2008) Molecular evolution of *PENTATATHY*-like genes in the dogwood genus *Cornus* (Cornaceae). Mol Phylogenet Evol 47(1): 175–195.
14. Feng CM, Xiang QY, Franks RG. (2011) Phylogeny-based developmental analyses illuminate evolution of inflorescence architectures in dogwoods (*C. s. l.*, *Cornaceae*). New Phytol 196(2): 631–643.
15. Karlson DT, Xiang QY, Stirm VE, Shirazi AM, Ashworth EN. (2004) Phylogenetic analyses in *Cornus* substantiate ancestry of xylem supercooling freezing behavior and reveal lineage of desicciation related proteins. Plant Physiol 135(3): 1654–1665.
16. Brown DA, Windham MT, Trigiano RN. (1996) Resistance to dogwood anthracnose among *Cornus* species. J Arboriculture 22(2): 83-85.
17. Sherard JL, Stidham TM, Hatfield JM, Hoekhke JE. (1996) Progression of the dogwood anthracnose epidemic and the status of flowering dogwood in Catactin Mountain Park. Plant Disease 80: 310–312.
18. Eric H. (2006) Ecology of flowering dogwood (*Cornus florida* L.) in response to anthracnose and fire in Great Smoky Mountain National Park, United States of America (North Carolina, Tennessee). PhD thesis. University of Florida.
19. Gao S, Zheng Y, Joung JG, Liu S, Zhang Z, et al. (2010) Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. BMC Genomics 11: 304.
20. Logacheva MB, Kasianov AS, Vinogradov DV, Samipilllin TH, Gelend M, et al. (2013) De novo sequencing and characterization of floral transcriptome in two species of backshade (*Fagopyrum*). BMC Genomics 12: 30.
40. Feng CM, Qu RD, Zhou LL, Xie DY, Xiang QY (2009) Shoot regeneration of
39. Xiang QY, Thomas DT, Xiang QP (2011) Resolving and dating the phylogeny
37. Wu X, Dabi T, Weigel D (2005) Requirement of homeobox gene
34. Richmond TA, Bleecker AB (1999) A defect in beta-oxidation causes abnormal
33. van Zanten M, Snoek LB, Proveniers MC, Peeters AJ (2009) The many
29. Liu C, Ma N, Wang PY, Fu N, Shen HL (2013) Transcriptome Sequencing and
28. Zhang XM, Zhao L, Larson-Rabin Z, Li DZ, Guo ZH (2012)
27. Parchman TL, Geist KS, Grahnen JA, Benkman GW, Buerkle CA (2010)
26. Zhang YQ, Wang X, Huang J, Xiong Z, Wang X, et al. (2011) Transcriptome sequencing of de novo analysis for Yello...usum L.). PLoS One 7(6): e16202.
25. Hale MC, Jackson JR, Dewoody JA (2011) Discovery and evaluation of candidate sex-determining genes and xenobiotics...goiter. PLoS One 8(5): e65141.
24. de Vries K, Shaw A, Li Y, Voskuil J, et al. (2012) De novo transcriptome annotation, and marker discovery. BMC Genomics 11: 462.
23. Mortazavi A, Williams BA, Mccue K, Schaferl L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth 5(7): 621–628.
22. Ness RW, Siol M, Barrett SC (2011) A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18): 3674–3676.
21. Tahvildar-Zadeh H, Robinson JT, Mesarov´ J (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14(2): 178–192.
20. Mortazavi A, Williams BA, Mccue K, Schaferl L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Meth 5(7): 621–628.
19. Waters AJ, Makarevitch I, Echtenk RA, Swanson-Wagner RA, Yeh C-T, et al. (2011) Parent-of-origin effects on gene expression and DNA methylation in the maize endosperm. Plant Cell 23(12): 4221–4233.
18. Xu T, Guo X, Wang H, Du X, Gao X, et al. (2013) De novo transcriptome assembly and differential gene expression profiling of three Cape hiruncus skin types during amagen of the hair growth cycle. Int J Genomics doi: 10.1155/2013/269191.
17. Huslen T, Huynen MA, de Vlieg J, Groenen FM (2006) Benchmarking ortholog identification methods using functional genomics data. Genome Biol 7(4): R31.
16. Price MN, Dehal PS, Arkin AP (2007) Orthologous transcription factors in bacteria have different functions and regulate different genes. PLoS Comput Biol 3(9): 1739–1750.
15. Costa GG, Cardoso KC, Del Bem LE, Lima AC, Cunha MA, et al. (2010) Transcriptome analysis of the oil-rich seed of the bioenergy crop Jatropha curcas L. BMC Genomics 11: 462.
14. Martínez-Barnetche J, Gómez-Barroto RE, Ovilla-Muñoz M, Téllez-Sosa J, López DE, et al. (2012) Transcriptional of the anahil female malaria mosquito vector Anopheles albimanus. BMC Genomics 13: 207.
13. Terabayashi Y, Shimizu M, Iizumi T, Sato M, Fujii T, et al. (2012) Conserved and specific responses to hypoxia in Aspergillus oryzae and Aspergillus niger determined by comparative transcriptomics. Appl Microbiol Biotechnol 93(3): 305–317.
12. Lu P, Han X, Qi J, Yang J, Wijermajte AJ, et al. (2012) Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. Genome Res 22(3): 308–318.
11. Ramos AM, Crooijmans RP, Alfara NA, Amaraal AJ, Archbald AL, et al. (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One 4(6): e6024.