MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 662–674 doi: 10.1111/j.1755-0998.2011.02990.x

De novo transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques

F. ANGELONI,* C. A. M. WAGEMAKER,* M. S. M. JETTEN,† H. J. M. OP DEN CAMP,† E. M. JANSSEN-MEGENS,‡ K.-J. FRANCOIJS,‡ H. G. STUNNENBERG‡ and N. J. OUBORG*
**Radboud University Nijmegen, Institute for Water and Wetland Research, Department of Molecular Ecology, Heyendaalseweg 135, 6525 AJ Nijmegen, the Netherlands, †Radboud University Nijmegen, IWWR, Department of Microbiology, Heyendaalseweg 135, 6525 AJ, Nijmegen, the Netherlands, ‡Radboud University Nijmegen, Department of Molecular Biology, Nijmegen Centre for Molecular Life Sciences, Geert Grooteplein-Zuid 26, 6525 GA, Nijmegen, the Netherlands*

Abstract

Next-generation sequencing (NGS) technologies are increasingly applied in many organisms, including nonmodel organisms that are important for ecological and conservation purposes. Illumina and 454 sequencing are among the most used NGS technologies and have been shown to produce optimal results at reasonable costs when used together. Here, we describe the combined application of these two NGS technologies to characterize the transcriptome of a plant species of ecological and conservation relevance for which no genomic resource is available, *Scabiosa columbaria*. We obtained 528 557 reads from a 454 GS-FLX run and a total of 28 993 627 reads from two lanes of an Illumina GAII single run. After read trimming, the *de novo* assembly of both types of reads produced 109 630 contigs. Both the contigs and the >75 bp remaining singletons were blasted against the Uniprot/Swissprot database, resulting in 29 676 and 10 515 significant hits, respectively. Based on sequence similarity with known gene products, these sequences represent at least 12 516 unique genes, most of which are well covered by contig sequences. In addition, we identified 4320 microsatellite loci, of which 856 had flanking sequences suitable for PCR primer design. We also identified 75 054 putative SNPs. This annotated sequence collection and the relative molecular markers represent a main genomic resource for *S. columbaria* which should contribute to future research in conservation and population biology studies. Our results demonstrate the utility of NGS technologies as starting point for the development of genomic tools in nonmodel but ecologically important species.

Keywords: genetic resources, microsatellites, next-generation sequencing, *Scabiosa columbaria*, SNPs, transcriptome characterization

Received 5 November 2010; revision received 20 December 2010; accepted 24 December 2010

Introduction

The study of genomes using next-generation sequencing (NGS) technologies is an efficient way to generate a large amount of valuable genomic information of the organism involved (Morozova & Marra 2008; Shendure & Ji 2008; Metzker 2010). This is especially important for nonmodel organisms of ecological importance, for which genomic resources are scarce or lacking (Wheat 2010). In particular, there is a growing interest in using NGS technologies in the field of conservation genetics (Kohn *et al.* 2006; Ouborg *et al.* 2010a,b), because these technologies have proven useful for many purposes, including gene discovery and annotation (Emrich *et al.* 2007), comparative genomics (Vera *et al.* 2008), molecular marker development (Novaes *et al.* 2008; Zeng *et al.* 2010) and for studies of genetic variation associated with adaptive traits (Namroud *et al.* 2008). 454 pyrosequencing is a fast and efficient NGS technology commonly used to define the transcriptome of specific individuals (Margulies *et al.* 2005), and it has been successfully applied to several nonmodel species (Garcia-Reyero *et al.* 2008; Novaes *et al.* 2008; Quinn *et al.* 2008; Vera *et al.* 2008). Illumina sequencing is another NGS technology which, when compared to 454 pyrosequencing, has a much higher throughput sequencing capacity (thus providing a higher coverage), although producing shorter reads (Strausberg *et al.* 2008; Zhang *et al.* 2010). While 454 pyrosequencing is commonly used for organisms for which little or

Correspondence: F. Angeloni, Fax: +31-24-36 52409;
E-mail: f.angeloni@science.ru.nl

no genomic information exists (Garcia-Reyero *et al.* 2008; Novaes *et al.* 2008; Vera *et al.* 2008), Illumina sequencing technology is commonly used to re-sequence genomes to study genetic variation and develop genetic markers (Ossowski *et al.* 2008; Yamamoto *et al.* 2010) and for gene expression studies (Marioni *et al.* 2008; Szittya *et al.* 2008). Because of their short length, only recently the reads obtained by Illumina sequencing proved to be sufficient for *de novo* assemblies of transcriptomes of eukaryotic organisms (Li *et al.* 2009). The limitations of these two sequencing methods are, therefore, substantially different, and their combined use should result in an improved transcriptome characterization (Aury *et al.* 2008; Wall *et al.* 2009). This has been shown in prokaryotes (Reinhardt *et al.* 2009; Nowrousian *et al.* 2010), fungi (Diguistini *et al.* 2009), animals (Bai *et al.* 2010) and nonmodel plant species (Buggs *et al.* 2010).

In the last decades, there has been an increasing tendency to adopt genomics tool to study the interaction between the phenotype and the environment at the molecular level (i.e. eco-genomics), in an effort to study the variation of functional genes in an ecological context (e.g. Feder & Mitchell-Olds 2003, Van Straalen & Roelofs 2006; Ouborg & Vriezen 2007). The major challenge of this new research field is to expand genomic research from few well-characterized model organisms to nonmodel but ecologically important organisms. The application of NGS technologies has proven to be an effective way to study functional gene variation in nonmodel species (Novaes *et al.* 2008; Vera *et al.* 2008; Hale *et al.* 2009; Meyer *et al.* 2009).

Scabiosa columbaria is a protandrous, entomophilous, short-lived perennial plant species of dry calcareous grasslands (Van Treuren *et al.* 1993) and is of great ecological and conservation importance. During the last decades, *S. columbaria* has been thoroughly studied in relation to inbreeding and inbreeding depression, habitat fragmentation and genetic erosion (Van Treuren *et al.* 1991, 1993; Waldmann & Andersson 2000; Andersson & Waldmann 2002; Waldmann 2002; Picó *et al.* 2004; Pluess & Stöcklin 2004; Reisch & Poschlod 2009) becoming an increasingly important model for conservation genetics. The availability of genomic resources for *S. columbaria* is valuable in a conservation context. In fact, this genomic information could be used as reference for mapping transcripts to study the variation in gene activity as a function of habitat fragmentation and environmental changes. In addition, conservation genetics can benefit from this sequence information, which can increase the molecular insight of fundamental processes relevant for conservation genetics (e.g. inbreeding). The use of NGS technologies provides a wealth of non-neutral molecular markers (e.g. SNPs and SSRs) which can be used to identify areas in the genome that are under selection (Ouborg *et al.* 2010b).

In this study, we present the transcriptome characterization of *S. columbaria* using the combination of 454 GS-FLX and Illumina GAII sequencing technologies. Along with other recent studies (e.g. Novaes *et al.* 2008; Vera *et al.* 2008; Kristiansson *et al.* 2009; Parchman *et al.* 2010), our results demonstrate the utility and the potential of NGS technologies when applied to nonmodel but ecologically relevant species such as S. columbaria.

Methods

454 sequencing

To maximize the representation of the *Scabiosa columbaria* transcriptome, promoting extensive gene discovery and a comprehensive survey of allelic variation in the transcriptome, RNA was extracted from untreated plant material and plant material subjected to different treatments, from a total of 46 inbred and outbred individuals (Table 1). Because cold stress is known to trigger gene expression in plants and to affect many plant processes (e.g. Kreps *et al.* 2002), inbred and outbred seedlings of *S. columbaria* were placed for 21 days in a 4 °C environment and then germinated at 20 °C. Another subset of inbred and outbred seedlings were germinated under a 1-ppm ethylene flow (flow rate: 500 mL/min) for 25 days, as ethylene can also influence plant gene expression (e.g. Chapman & Estelle 2009). Control seedlings were germinated at 20 °C. RNA was also extracted from roots, rosette leaves and flowerbuds of inbred and outbred adult plants growing in different soil conditions (Table 1) using the Aurum Total RNA Mini Kit (Bio-Rad, Hercule, CA). An amount of 2.5 μg of RNA of each sample was mixed to create the RNA pool for the cDNA synthesis (Table 1). The poly(A)+ RNA was prepared from this total RNA mix. At Eurofins MWG (Germany), first-strand cDNA synthesis was primed with a N6 randomized primer. Then, 454 adapters A and B were ligated to the 5′ and 3′ ends of the cDNA. The cDNA was finally amplified with PCR using a proof-reading enzyme (19 cycles). Normalization was carried out by one cycle of denaturation and reassociation of the cDNA. Reassociated ds-cDNA was separated from the remaining ss-cDNA (normalized cDNA) by passing the mixture over a hydroxylapatite column. After hydroxylapatite chromatography, the ss-cDNA was amplified with eight PCR cycles. For 454 sequencing, the cDNA in the size of 400–600 bp was eluted from a preparative agarose gel (1.5%). The resulting cDNA fragments were used for sequencing, according to Margulies *et al.* (2005). Files containing sequences and associated quality scores were deposited at NCBI Short Read Archive [SRA:023452] and are freely

Table 1 Description of the plant material used to create normalized cDNA sequenced in the full GS-FLX run. Inbred and outbred individuals were created by selfing and crossing individuals collected from different natural populations of *Scabiosa columbaria* occurring throughout the Netherlands

RNA pool #	RNA extracted from	Treatment
RNA pool 1	5 inbred and 5 outbred seedlings	No treatment
RNA pool 2	5 inbred and 5 outbred seedlings	Ethylene treatment
RNA pool 3	5 inbred and 5 outbred seedlings	Cold treatment
RNA pool 4	Leaves of 2 inbred individuals	Grown in 1:1 sand:potting soil
RNA pool 5	Leaves of 2 inbred individuals	Grown in 1:3 sand:potting soil
RNA pool 6	Leaves of 2 outbred individuals	Grown in 1:1 sand:potting soil
RNA pool 7	Leaves of 2 outbred individuals	Grown in 1:3 sand:potting soil
RNA pool 8	Roots of 2 inbred individuals	Grown in 1:3 sand:potting soil
RNA pool 9	Roots of 2 outbred individuals	Grown in 1:3 sand:potting soil
RNA pool 10	Roots of 2 outbred individuals	Grown in 1:3 sand:potting soil
RNA pool 11	Flowerbuds of 1 inbred and 1 outbred individual	Grown in 1:3 sand:potting soil

available. All sequence data processing was carried out using a desktop computer with two Intel Xeon X5550 processors (2.67 GHz each) and 48 Gb of RAM. The reads were trimmed before entering the assembly using the software 'CLC Genomics Workbench' (CLCbio hereafter) for quality score, presence of repeated sequences >50 bp, poly(A) stretches, retrotransposons, primers and adaptors (http://www.clcbio.com).

Illumina sequencing

As a first test for future transcriptional profiling experiments of inbreeding depression, leaf samples from one inbred and one outbred individual of *S. columbaria* were prepared for Illumina sequencing. Total RNA was isolated using the Aurum Total RNA mini kit (Bio-Rad, Hercule, CA, USA). All the RNA was used for poly(A) enrichment by PolyATract mRNA Isolation system III (Promega). The total product (250 μL) was concentrated by a SpeedVac Rotor Vapor and dissolved in 12 μL of distilled water. cDNA was created out of the whole poly(A)+ RNA sample using M-MulV RT in accordance with Ribolock RNAse inhibitor. Double-strand cDNA was created by second-strand cDNA synthesis using DNA polymerase I *Escherichia coli* and RNAse-H *E. coli*. The ds-cDNA was further purified by phenol/chloroform extraction and ethanol precipitation. DNA concentration analysis on a Qbit™ fluorometer (Invitrogen) showed a total yield of 8.7 μg and 4.6 μg ds-cDNA for the inbred and outbred individual, respectively. Further processing was performed according to manufacturer's protocol (Illumina GAII analyzer). Files containing the sequences and the quality scores of the reads were deposited at NCBI Short Read Archive [SRA:023452] and are freely available. Reads were trimmed for quality score using the modified Mott-trimming algorithm present in CLCbio (Ewing & Green 1998; Ewing *et al.* 1998). Reads were also trimmed for the presence of repeated sequences >50 bp, poly(A), retrotransposons, primers and adaptors.

De novo assembly and annotation

Both 454 and Illumina-trimmed sequences were *de novo* assembled using CLCbio by setting minimum 95% identity, minimum 30% overlap for 454 reads and 40% for Illumina reads, and 100 bp as minimum contig length (Table 2). The quality of the assembly was assessed with a local BLASTn alignment of all the contigs against themselves and the singletons (*e*-value < 10^{-6}). Further assessment of the quality of the *de novo* assembly was carried out as follows. We compared the depth and the length of contig coverage with reference to orthologous genes in *Helianthus annuus* and *Arabidopsis thaliana*, by plotting the ratio of contig length to *H. annuus* and to *A. thaliana* orthologue coding region length against coverage depth. Orthologous genes were retrieved performing a local BLASTx alignment (*e*-value < 10^{-6}) with the TAIR9 *A. thaliana* database (Swarbreck *et al.* 2008) and *H. annuus* predicted proteins (Uniprot database, Boeckmann *et al.* 2003). To further assess the coverage and the quality of the assembly, we used BLASTx to align the contigs to the manually curated protein database Uniprot/Swissprot (Boeckmann *et al.* 2003) using BLAST2GO (Conesa *et al.* 2005). BLAST2GO is an automated tool for

Table 2 Results of the *de novo* assembly

Sequencing and *de novo* assembly results	Value
Total number of reads (before trimming)	29 522 184
Total base pair (before trimming)	2 287 088 896
Total number of reads (after trimming)	29 333 721
Total base pair (after trimming)	1 817 551 895
Transcriptome coverage (after trimming)	36.35×
Total number of contigs	109 630
Total number of singletons	7 726 134 (21 611 >75 bp)
Average contig length (bp)	26 978
Maximum contig length (bp)	3564
Minimum contig length (bp)	100

the assignment of Gene ontology (GO) terms to BLAST hits, and it has been designed for use with novel sequence data (Conesa *et al.* 2005). Assignment of GO terms to contigs with significant BLASTx match with swissprot was also performed using BLAST2GO. In addition, we generated GO assignments for *A. thaliana* annotated proteins to compare the distribution of functional annotation in *S. columbaria* to that of a plant species with a well-characterized transcriptome.

Marker identification and characterization

We used the software SciRoKo (Kofler *et al.* 2007) to identify SSRs in the contigs and in the >75-bp singletons. SciRoKo is a fast and suitable tool for whole-genomic identification of microsatellites, and, compared to other already existing tools, it allows the analysis of compound microsatellites (Kofler *et al.* 2007). We located di-, tri-, tetra-, penta- and hexa-nucleotide SSR using default settings. Newly identified microsatellite loci are in general useful only if it is possible to design primers in the non-repeated flanking regions and be successfully used for PCR amplification. We therefore used the software iQDD (Meglécz *et al.* 2010) to screen sequences with microsatellite loci for flanking regions with high-quality PCR-priming sites. These loci can be referred to as 'potentially amplified loci' or PALs (Castoe *et al.* 2010). iQDD is a user-friendly program capable to design primers to amplify microsatellite regions and is especially suitable for large sequencing projects (Meglécz *et al.* 2010).

Potential SNPs were detected with CLCbio using standard settings. However, we set maximum coverage at 40× because read coverage can be high in repetitive regions where the alignment is not very trustworthy. Setting the maximum coverage threshold a little higher than the average coverage (in our case 38.48) allows for some variation, therefore being helpful in ruling out false positives from such regions (e.g. Amaral *et al.* 2009).

Results

454 and Illumina sequencing

The full 454 GS-FLX run produced a total of 528 557 raw reads, averaging 212 bp length. CLCbio entirely eliminates low-quality reads and shortens reads, retaining only their high-quality section without primers, adaptors and Poly(A) stretches. Searching for 15 well-characterized retrotransposons in all the 454 GS-FLX reads revealed only three reads (0.0005%) containing retrotransposon sequences. We, therefore, assumed that this type of sequences did not affect the quality of the assembly. We found, and removed, a very low number of reads containing repeats >50 bp (N = 151, 0.03% of the total read number), which could cause problems during the *de novo* assembly. A total of 27 450 202 nucleotides (24.4% of the total) were removed during trimming.

mRNA from leaves of an inbred and outbred individual of *Scabiosa columbaria* was used to prepare two samples of rRNA free, non-normalized cDNA. An Illumina GAII single run of the two leaf samples produced a total of 28 993 627 reads, all of 75 bp length. We again used CLCbio to remove low-quality reads, primers, adaptors and poly(A) sequences. We did not find any of the 15 retrotransposon sequences in the Illumina reads. We found, and removed, a very low number of reads containing repeats >50 bp (N = 102 442; 0.35% of the total read number). In total, 442 086 799 nucleotides (20.3% of the total) were removed during trimming. Including both 454 and Illumina sequencing results, we obtained a total of 29 333 721 trimmed reads (1 817 599 195 bp) that were used for the *de novo* assembly.

Assuming that a comparable number of genes occur in *S. columbaria* as in *Arabidopsis thaliana* (25 000) and a similar average gene length of 2000 bp (Bevan & Walsh 2005), the average transcriptome coverage of *S. columbaria* based on the remaining high-quality 454 and Illumina reads was estimated as 36.35×.

De novo assembly and quality control

The *de novo* assembly of 454 and Illumina high-quality reads resulted in 109 630 contigs (Table 2), while 7 726 134 reads remained as singletons, of which 21 773 were 454 reads (length >75 bp). Contigs ranged from 100 to 3564 bp in size with an average of 270 bp and a median of 202 bp (Fig. 1a). Figure 1b shows the distribution of contig depth coverage of the assembly, with an average value of 38.48. Ideally, an accurate assembly of the reads would create an assembled sequence with at most a few short alignments to other contigs. Using a local BLASTn to align contigs against themselves and singletons against contigs, we found that every contig had a significant blast hit with itself, and no singleton sequence had significant hit with any of the contigs. In addition, we manually checked the BLASTn results of a random subset of contigs (N = 1096), and we found that 356 (32.5%) had at least 1 BLASTn hit (e-value < 10^{-6}) with other contigs, but in no case did these alignments extend over the entire length of the query or subject. Of these 356, as many as 193 contigs (54.2%) had a BLASTn hit against the *A. thaliana* TAIR9 protein data set, yet only 5 (1.4%) of those contigs had BLASTx hits to the same *A. thaliana* protein, and only 48 (13.5%) also had best BLASTx to the same protein in Uniprot/Swissprot. This suggests that these regions represent conserved motifs in different genes or variants of the same gene created by

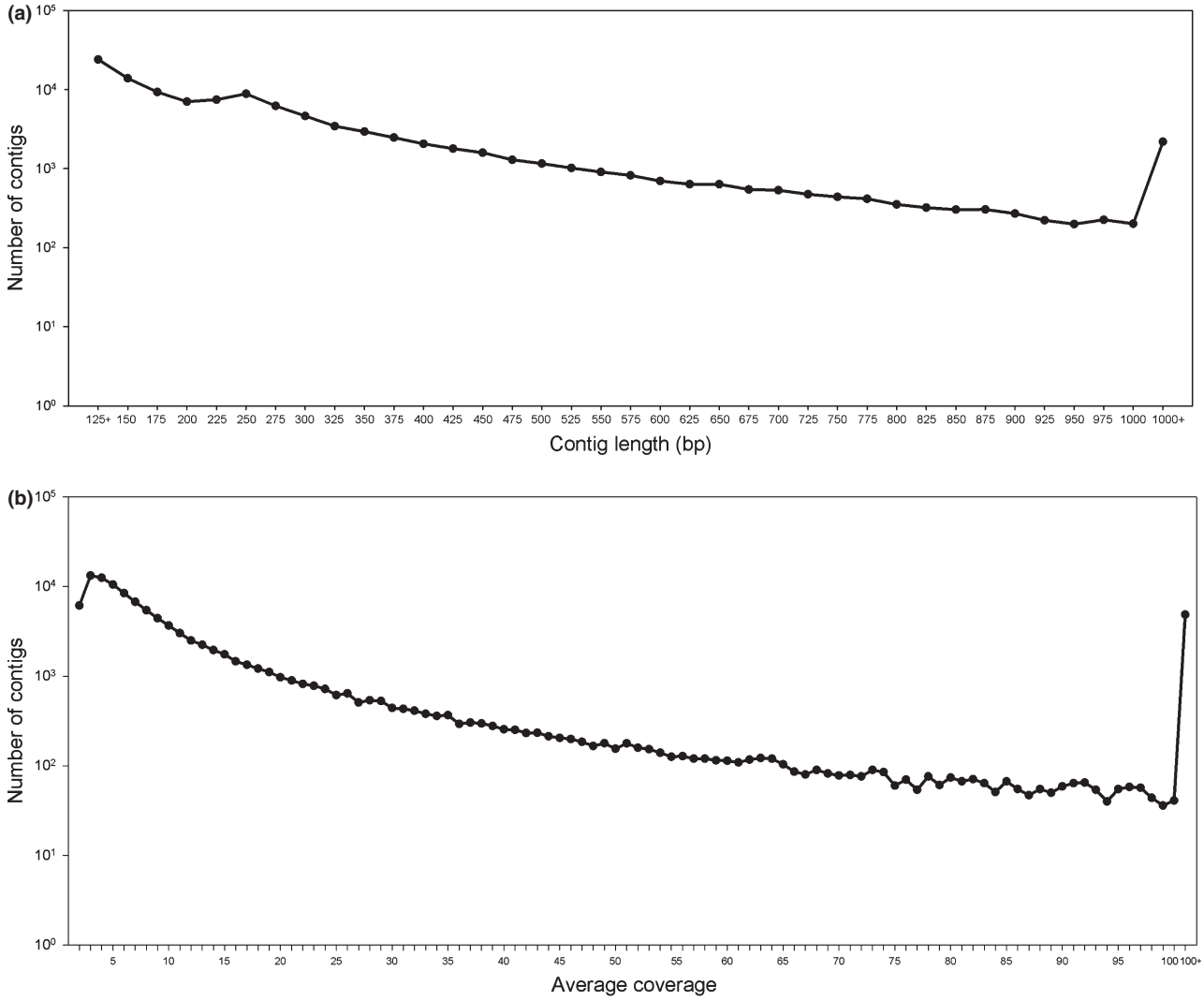

Fig. 1 (a) Contig length distribution of the *de novo* assembly (average: 270 bp, median: 202 bp) and (b) distribution of average contig coverage depth (average: 38.48, median: 6.56 bp). The contig coverage depth is defined as the average number of reads included to create a contig, averaged per nucleotide.

alternative splicing. In the latter case, our assembly seems to have correctly separated the splicing variants into different contigs. Finally, we found that 32 contigs (2.9% of our random contig subset) showed the presence of at least one SSR.

As many as 48 740 *S. columbaria* contigs had a significant BLASTx hit with *A. thaliana* and 5277 with *H. annuus*. The effect of the transcriptome coverage (the number of reads that, on average, cover a given base pair of a certain contig) on the quality of the assembly is shown in Fig. 2a,b. Curves represent sigmoid fit ($P < 0.0001$) for all the genes (solid curve), for genes of 800–1800 bp, and for genes >1800 bp in length (dashed curves) of *H. annuus* (Fig. 2a) and *A. thaliana* (Fig. 2b). For both *H. annuus* and *A. thaliana*, the ratio between contig length and its orthologous gene length increased with the depth of coverage of the contigs and reached an asymptote near unity at average coverage >200 for *H. annuus* and between 10 and 100 for *A. thaliana*. Values greater than 1 in both figures are possibly because of noncoding regions (e.g. UTRs) present in our contigs, but also indicate considerable coding region coverage for many individual contigs.

Annotation

We further characterized the contigs data set using BLASTx to align the contigs to the curated protein database of Uniprot/Swissprot (Boeckmann *et al.* 2003), considering any significant hit with an *e*-value $<10^{-6}$. Of

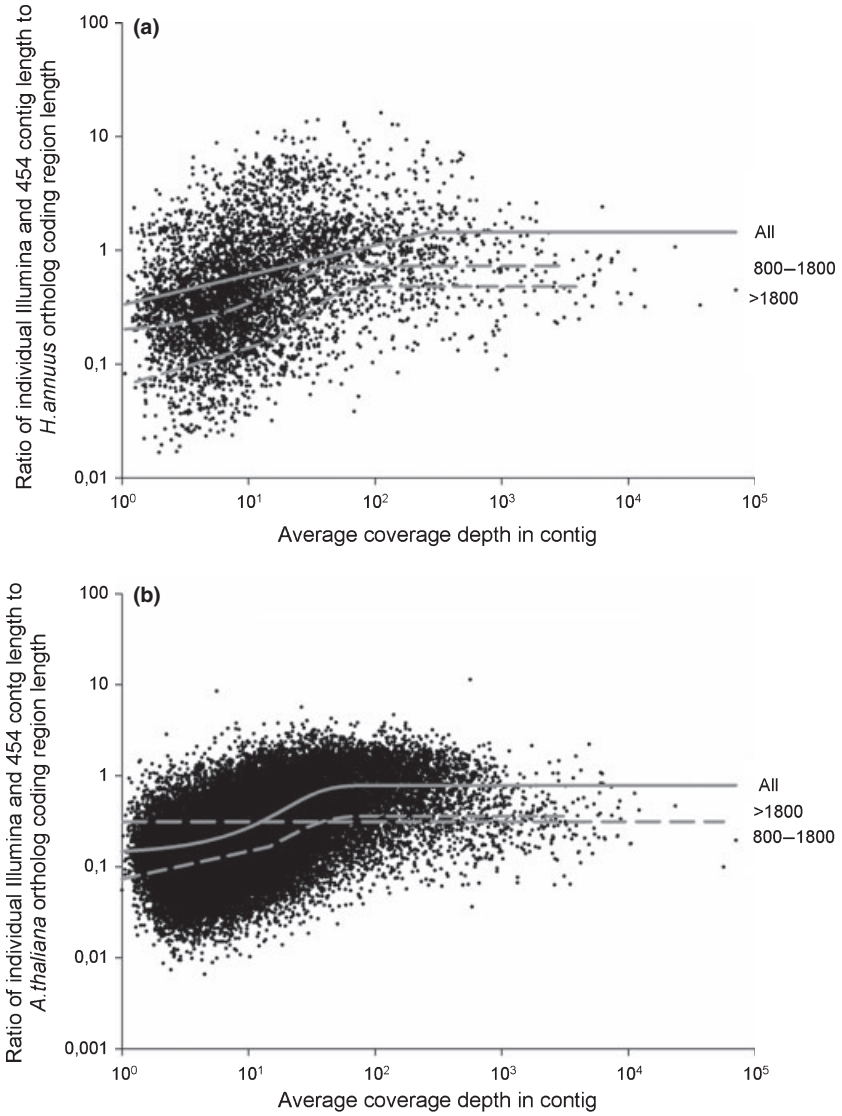

Fig. 2 Ratio of assembled contigs to *H. annuus* (a) and *Arabidopsis thaliana* (b) orthologue length in relation to the average contig coverage depth. Orthologues were identified by BLASTx (*e*-value < 10^{-6}). Total number of significant BLASTx hits: (a) = 5277, (b) = 48740. Total number of sequences whose ratio is above 1: (a) = 1352, (b) = 3201.

109 630 contigs, 29 676 (27.1% of the total) had a significant BLASTx hit to proteins present in Uniprot/Swissprot and matched 11 699 unique protein accessions (Table 3). The full list of annotated contigs is available in supplementary material S1. BLASTx in the order of 30% of the sequences is common when using NGS technologies to characterize the transcriptome of nonmodel species (Novaes *et al.* 2008; Vera *et al.* 2008; Parchman *et al.* 2010). Singletons >75 bp (21 611), thus 454 pyrosequencing singletons, were also analysed by BLASTx, resulting in 10 515 (48.6% of the total) significant blast in the Uniprot/Swissprot database, and matched 2177 unique protein accessions (Table 3). Including both contigs and >75-bp singletons, we identified a total of 12 516 unique protein accessions (Table 3), the majority of which corresponded to known plant proteins (62.8%) or to other eukaryote species (19.5%).

Using BLAST2GO, we assigned GO classes to 11 240 (89.8%) of the 12 516 unique genes with BLAST matches to known proteins in Uniprot/Swissprot. We retrieved a total of 94 184 GO terms associated to these unique genes. Assignments to the biological process ontology were the majority (*N* = 45 341; 48.1%) followed by molecular function (*N* = 24 824; 26.4%) and cellular component (*N* = 24 019; 25.5%). We also compared the distribution of the GO slim annotations in our *S. columbaria* data set to that of *A. thaliana* (Fig. 3). GO slim terms are a cut-down version of the GO ontologies containing a subset of the terms in the whole GO, providing a broad overview for genome–genome comparison (Lomax 2005). All the GO slim categories were represented by the annotated *S. columbaria* sequences (Fig. 3). The distribution of annotated *S. columbaria* sequences over the various GO categories was similar to the distribution in *A. thaliana*, which

Table 3 Summary and taxonomic source of BLASTx matches to the contigs and singletons resulted from the *de novo* assembly. Number and percentages of unique best BLASTx matches (*e*-value < 10^{-6}) of 454/Illumina contigs and >75-bp singletons to Uniprot/Swissprot database grouped by taxonomic category

Taxonomic category	Contigs (*N* = 11 699)	>75-bp singletons (*N* = 2177)	Combined set (*N* = 12 516)
Arabidopsis thaliana	4596 (39.29%)	977 (44.88%)	4812 (38.45%)
Other plants	2766 (23.65%)	743 (34.13%)	3006 (24.02%)
Algae	39 (0.33%)	8 (0.37%)	44 (0.35%)
Fungi	605 (5.17%)	194 (8.91%)	781 (6.24%)
Other eukaryote	2352 (20.10%)	141 (6.48%)	2445 (19.54%)
Bacteria	776 (6.63%)	40 (1.84%)	814 (6.50%)
Others	565 (4.83%)	74 (3.39%)	614 (4.90%)

underlines that the obtained transcriptome of *S. columbaria* is representative of the entire transcriptome.

Marker identification

The discovery of SSRs and SNPs in *S. columbaria* is of great interest because it can be potentially applied in several different research fields, such as population and conservation genetics, comparative genomics and the genetic control of adaptive traits. Using the software SciRoKo, we found 3813 SSRs in contigs (of which 827 (21.7%) in annotated contigs) and 507 in singletons (of which 51 (10.1%) in annotated singletons) for a total of 4320 microsatellite loci (Tables 4 and 5). In both cases, the trinucleotide repeats had the greatest number of microsatellite loci (1383 and 160, respectively). Figure 4 shows the number of microsatellite loci per number of repeat units in the five different classes of repeats. We identified, using iQDD, 791 PAL in the contigs and 65 in the singletons, for a total of 856 PAL. A detailed description of the primers is given in supplementary material S2. Comparing across the five different classes of repeats, dinucleotide repeats had the greatest number of PAL and hexanucleotide the least (*N* = 484 and *N* = 3, respectively). Figure 5 shows the observed counts of identified PAL for the five different repeats sequence motifs for both contigs and singletons. AG was the most frequent repeat overall (*N* = 243), while several others motifs had only 1 PAL.

The assembled 454 and Illumina contigs also provide a rich data source for discovery of common SNPs. Using CLCbio, we identified a large number of SNPs in our most deeply covered contigs. Across all the contigs, we identified a total of 75 054 putative SNPs, with frequency of the alternative allele being at least 20% and a minimum fourfold coverage. Considering that the total number of bp of the contigs is 29 570 408 bp, SNP occurrence was on average 2.53 SNPs per thousand base pairs.

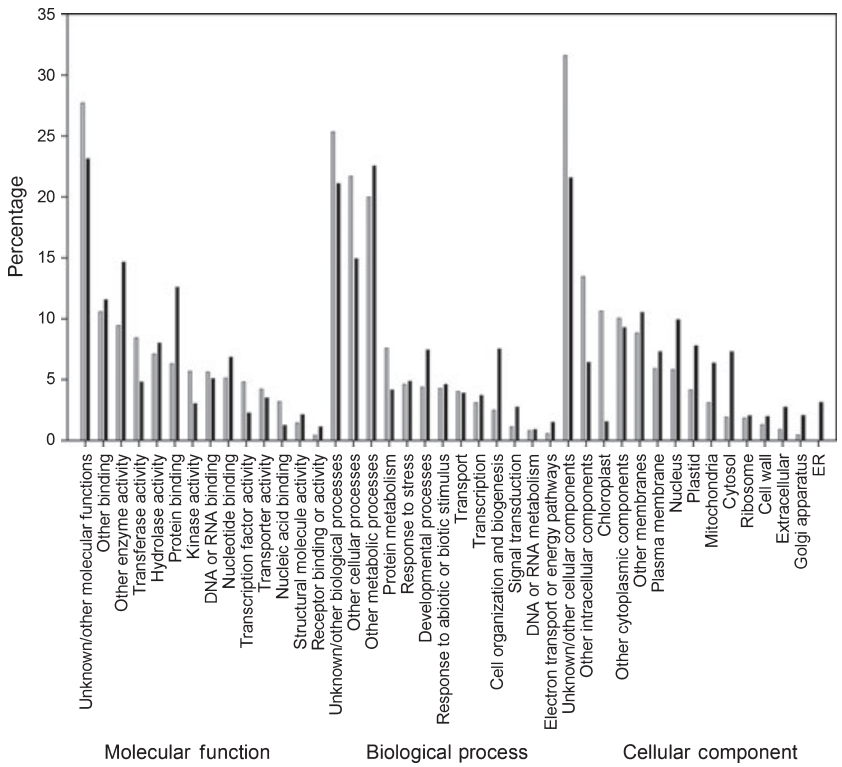

Fig. 3 Gene Ontology (GO slim) assignments for *Scabiosa columbaria* (black columns) and *Arabidopsis thaliana* (grey columns). The columns show the proportion of annotated contigs and singletons from *S. columbaria* combined assembly contigs and annotated *A. thaliana* proteins (from TAIR database) that matched a variety of GO categories.

Table 4 Number of single sequence repeats (SSR) occurring in contigs resulting from the *de novo* assembly

Motif	Counts	Average length (bp)	Counts/Mbp
Mononucleotide	475	19.36	16.06
Dinucleotide	429	19.29	14.51
Trinucleotide	1383	19.08	46.77
Tetranucleotide	479	18.56	16.20
Pentanucleotide	600	20.66	20.29
Hexanucleotide	447	22.84	15.12
All motifs	3813	19.76	21.49

Table 5 Number of single sequence repeats (SSR) present in the >75-bp singletons

Motif	Counts	Average length (bp)	Counts/Mbp
Mononucleotide	17	15.59	3.64
Dinucleotide	135	20.54	28.94
Trinucleotide	160	21.59	36.23
Tetranucleotide	60	20.12	12.86
Pentanucleotide	77	21.58	16.51
Hexanucleotide	49	24.69	10.50
All motifs	507	21.23	108.69

Visual inspection of 15 BLAST-annotated contigs (total bp = 43.425) with the largest number of reads revealed 18 SNPs with alternate alleles in at least 20% of the reads and at least a 8-fold coverage, on average 0.4 SNPs per thousand base pairs. Of 75 054 putative SNPs, 74 900 (99.79%) showed the presence of two alleles: 152 three alleles (0.20%) and 1 four alleles (0.01%). Within the biallelic SNPs group, we found 44 110 transitions and 30 790 transversions. The detailed description of the contig sequences and the relative SNPs is given in supplementary material S3.

All supplementary material files were deposited in the Dryad database (doi: 10.5061/dryad.8160) and are freely available (http://datadryad.org/handle/10255/dryad.8160).

Discussion

Wall *et al.* (2009) have predicted by means of a computer simulation that the combination of 454 pyrosequencing and Illumina technologies will achieve optimal performance at affordable cost. In addition, several empirical studies have successfully applied this combination to perform *de novo* genome assemblies and shown that this results in improved performance when compared to the individual use of the two NGS techniques (e.g. Aury *et al.* 2008; Reinhardt *et al.* 2009). Thus, the combined use of 454 and Illumina reads represents an effective way to assemble the genome or the transcriptome of any species of interest. Obtaining NGS technology sequence data is fairly straightforward. However, dealing with the enormous amount of sequence information produced can be challenging, especially when no reference genome is available. In this case, a *de novo* assembly should be performed, for which several software packages have been developed (reviewed by Miller *et al.* 2010). Because many assembly algorithms exist which may lead to different *de novo* assemblies, assessing the quality of the assembly is of fundamental importance for any genomic study. The contig coverage depth, the average number of reads included per nucleotide to create a contig, can be used as a first measure of the quality of the assembly. In fact, deep coverage both provides reliable SNP data and reduces the influence of sequencing error on *de novo* assemblies (Wheat 2010). In addition, if the number of reads for a given cDNA increases, contig assembly length should also increase (Wheat 2010). This control quality step is important, but not sufficient, to completely assess the quality of a *de novo* assembly. In fact, assembled contigs could correspond to repetitive or noncoding DNA (e.g. UTR) regions. A common solution to this problem is to calculate the ratio of contig length to the length of the coding region in the orthologous gene in a reference species and plot this ratio against coverage depth (Fig. 2) (Vera *et al.* 2008; Wheat 2010). This makes it possible to visualize whether increased coverage depth results in an increased coverage of orthologous genes. In this study, we used *Arabidopsis thaliana* and *H. annuus* as references species. *A. thaliana* is the model plant species for which a large amount of genomic resources are available and therefore represent one of the most widely used reference genomes for this type of analysis (Parchman *et al.* 2010). *H. annuus* is not a completely sequenced plant species but is phylogenetically the closest plant species to *Scabiosa columbaria* for which genomic resources are available (e.g. predicted proteins data set). In general, the completeness of orthologue coverage depended on *A. thaliana* and *H. annuus* ortholog length, decreasing for longer orthologue sequences (Fig. 2). This graphical representation is also useful to visualize how many contigs are likely to represent full-length assembly of different coding regions. In our assembly, the ratio of the length of individual contigs to the length of their *A. thaliana* orthologue coding region reached an asymptote (ratio = 1) when average coverage depth was above 50. A comparable trend was found when repeating the analysis using *H. annuus* as reference species. The results were somewhat less evident, probably due to the relatively low number of predicted protein sequences available for *H. annuus*, which were often not full length (N = 1215 for

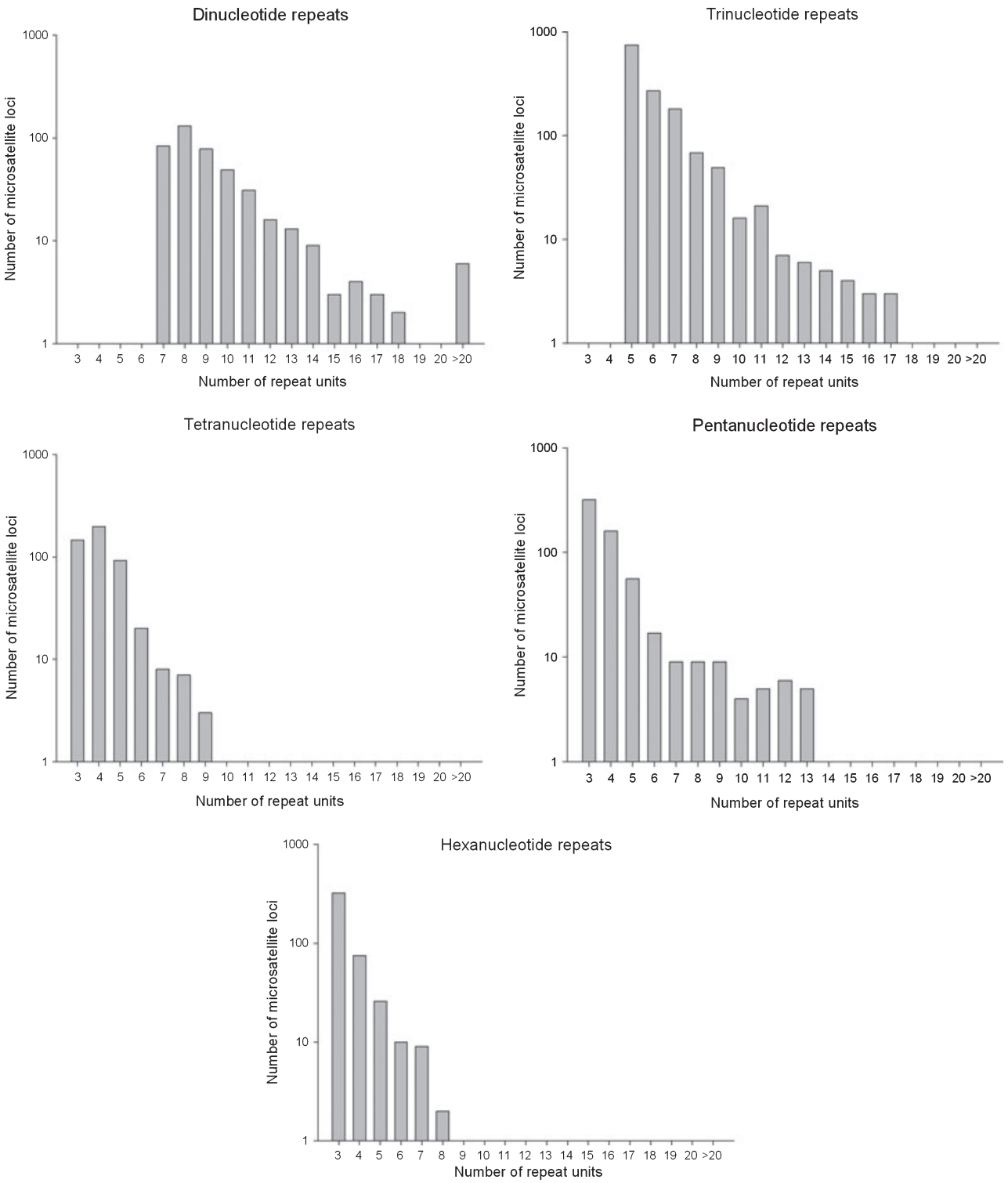

Fig. 4 Counts of the number of repeats units for the five different repeat unit classes.

H. annuus and $N = 33\,411$ for *A. thaliana).* Thus, our results demonstrate the utility of NGS technologies to develop genomic tools in nonmodel but ecologically important species.

Annotation and marker identification

We further characterized the de novo assembly with respect to the curated protein database Uniprot/

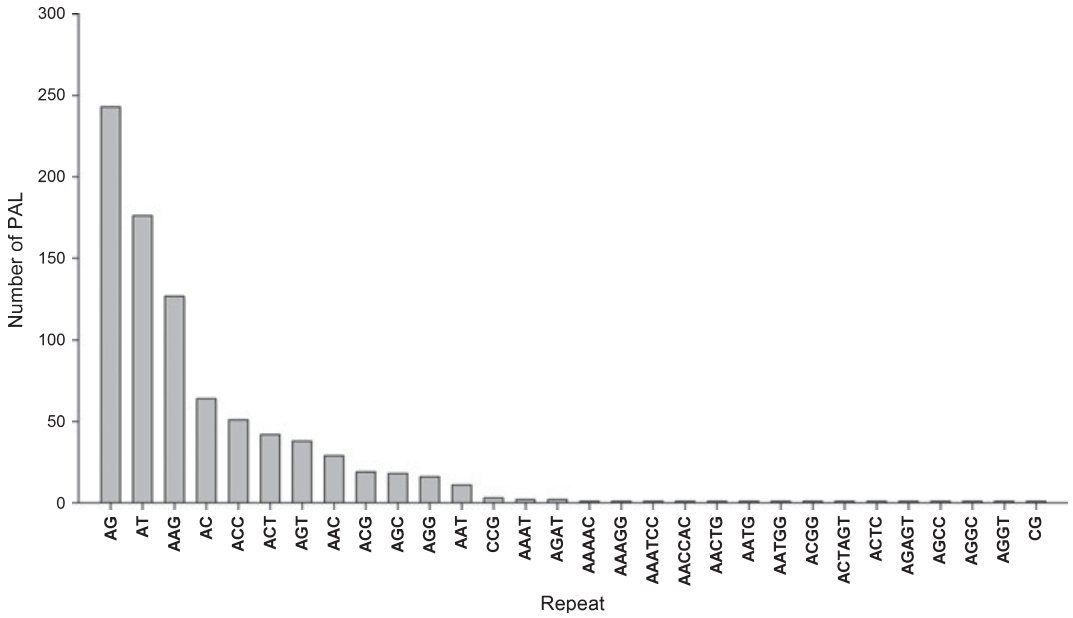

Fig. 5 Observed counts of microsatellite loci which contain PCR-primer sites (potentially amplifiable loci).

Swissprot (Boeckmann *et al.* 2003). Nearly 30% of the contigs and 50% of the singletons had significant BLASTx hits in Uniprot. In addition, many of the contigs and singletons without a significant BLASTx hit could still represent valuable information. In fact, they could correspond to additional genes not present in the database or genes that lacked a BLAST match because of their short length. Assuming that *S. columbaria* and *A. thaliana* have a similar number of genes, our annotated sequences are estimated to represent half of the genes of *S. columbaria*. Almost 90% of the unique genes were assigned to a wide range of GO categories. This is a remarkable result when considering the current state of functional annotation of the *A. thaliana* genome. For all the three GO ontologies (molecular function, biological process and cellular component), around 50% of *A. thaliana* genes are either not annotated or have an unknown function (TAIR database, http://www.arabidopsis.org). Moreover, the distribution of annotated genes over the various GO categories was similar for *S. columbaria* and *A. thaliana*, illustrating that the *S. columbaria* transcriptome assessed here was unbiased.

Our data set of *S. columbaria* annotated sequences represents a detailed foundation for future ecology and conservation biology studies. Studies examining the variation in gene activity of *S. columbaria* (e.g. as function of habitat fragmentation and environmental changes) will be able to use this database as reference for transcript mapping. This sequence information can also be valuable to study conservation genetics processes such as inbreeding depression, selection and adaptation. In particular, the study of inbreeding depression can take advantage of genomics techniques (Kristensen *et al.* 2010; Paige 2010), as they will help to improve our understanding of the molecular basis of inbreeding depression, including the search for candidate genes and functional classes of genes responsible for the decrease of fitness associated with inbreeding, the underlying genetic cause of inbreeding depression and the environmental influences on gene expression patterns (Kristensen *et al.* 2010; Ouborg *et al.* 2010a,b; Paige 2010). Our sequence data set also represents an extensive resource for the development of molecular markers, because of the massive amount of data in which molecular markers can be identified. In addition, the discovered markers are not neutral markers, but related to coding regions of the genome and therefore represent a valuable tool for the detection of functional variation and the effect of selection (Bouck & Vision 2008; Ouborg *et al.* 2010a,b). We located nearly 4500 potential SSRs in our contig data set of which 856 can potentially be used to amplify and score microsatellite alleles based on length variation. The deep coverage produced by the Illumina and 454 reads and the RNA sampling of 48 different individuals enabled us to discover more than 75 000 putative SNPs, a large fraction of which are in annotated sequences. The large number of SSRs and SNPs that we detected represents a large resource of molecular markers that can be potentially applied in several different research fields, such as population genetics, comparative genomics and the genetic control of adaptive traits (Luikart *et al.* 2003; Beaumont 2005; Avise 2010).

Conclusions

NGS is increasingly being used to obtain genome-wide sequence information of plant and animal species for which no genomic data are available. These species, including *Scabiosa columbaria*, are investigated from an ecological or conservation perspective, and obtaining sufficient genomic information of high quality allows the step towards an ecological genomics or conservation genomics research programme. Our study demonstrates that a combination of 454 and Illumina sequencing can be successfully applied in combination to characterize the transcriptome of a nonmodel species. NGS technologies are being developed at a stunning pace. At the moment, a single 454 pyrosequencing run on a Roche GS-FLX Titanium can produce more than 1 million reads, with average read length in the range of 400 base pairs, while the newest Illumina technology can currently provide about 100 bp per read (Zhang *et al*. 2010). In addition, paired-end sequencing (Schatz *et al*. 2010) (i.e. both ends of a given DNA fragments are sequenced) is now available for all existing NGS technologies (Deschamps & Campbell 2010) and can become useful in *de novo* assemblies. New systems, already tagged as 'third-generation' NGS technologies, are being developed, promising even longer reads and higher throughput per sample when compared to existing NGS technologies (Rusk 2009). Thus, it can be expected that in the near future, it will be possible to develop different and even more nuanced approaches to characterize the transcriptome of nonmodel species.

There is a rising interest to adopt ecological and evolutionary functional genomics approach in conservation genetics, often referred to as conservation genomics (e.g. Primmer 2009; Avise 2010; Ouborg *et al*. 2010a,b). Here, we demonstrated that it is possible to use the combination of 454 and Illumina technologies to rapidly characterize the transcriptome of a nonmodel plant species such as *S. columbaria*. We chose *S. columbaria* as a model because of its growing importance in conservation genetics and population biology (Van Treuren *et al*. 1991, 1993; Waldmann & Andersson 2000; Andersson & Waldmann 2002; Waldmann 2002; Picó *et al*. 2004; Pluess & Stöcklin 2004; Reisch & Poschlod 2009) and expect it to become a case system for this type of analysis. In fact, very little genomic information was available for *S. columbaria* before this study, which is the case for most species of ecological and conservation importance (Stinchcombe & Hoekstra 2008; Ouborg *et al*. 2010b). The genomic information that we obtained using this combination of NGS technologies can be effectively used in a conservation context in future transcriptional profiling studies of the effect of habitat fragmentation (e.g. inbreeding depression) and environmental changes on this or other plant species. In addition, thousands of SSR and SNP markers that we located should also contribute in facilitating future research in conservation genetics, population biology and functional genomics. This study thus represents a possible guideline for research groups oriented towards conservation genomics, trying to start up a sequencing project for a species that lacks genomic information.

Acknowledgements

We thank Dr. P. Vergeer for constructive comments on the manuscript. This study was financially supported by the Netherlands Organization for Scientific Research (NWO), project 817.01.001.

References

Amaral AJ, Megens HJ, Kerstens HHD *et al*. (2009) Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *BMC Genomics*, **10**, 374.

Andersson S, Waldmann P (2002) Inbreeding depression in a rare plant: *Scabiosa canescens* (Dipsacaceae). *Hereditas*, **136**, 207–211.

Aury JM, Cruaud C, Barbe V *et al*. (2008) High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics*, **9**, 603.

Avise JC (2010) Perspective: conservation genetics enters the genomic era. *Conservation genetics*, **11**, 665–669.

Bai X, Zhang W, Orantes L, Jun T-H, Mittapalli O *et al*. (2010) Combining Next-Generation Sequencing Strategies for Rapid Molecular Resource Development from an Invasive Aphid Species, *Aphis glycines*. *PLoS ONE*, **5**, e11370. DOI: 10.1371/journal.pone.0011370.

Beaumont MA (2005) Adaptation and speciation: what can FST tell us? *Trends in Ecology & Evolution*, **20**, 435–440.

Bevan M, Walsh S (2005) The Arabidopsis genome: a foundation for plant research. *Genome Research*, **15**, 1632–1642.

Boeckmann B, Bairoch A, Apweiler R *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL. *Nucleic Acids Research*, **31**, 365–370.

Bouck A, Vision T (2008) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology*, **16**, 907–928.

Buggs RJA, Chamala S, Wu W *et al*. (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology*, **19**, 132–146.

Castoe TA *et al*. (2010) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources*, **10**, 341–347.

Chapman EJ, Estelle M (2009) Mechanism of auxin-regulated gene expression in plants. *Annual Review of Genetics*, **43**, 265–285.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Deschamps S, Campbell MA (2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding*, **25**, 553–570.

Diguistini S, Liao NY, Platt D *et al*. (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology*, **10**, R94.

Emrich SJ, Barbazuk WB, Li L, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, **17**, 69–73.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186–194.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Feder ME, Mitchell-Olds TM (2003) Evolutionary and ecological functional genomics. *Nature reviews genetics*, **4**, 651–657.

Garcia-Reyero N, Griffitt RJ, Liu L *et al.* (2008) Construction of a robust microarray from a non-model species largemouth bass, *Micropterus salmoides* (Lacèpede), using pyrosequencing technology. *Journal of Fish Biology*, **72**, 2354–2376.

Hale MC, McCormick CR, Jackson JR, DeWoody JA (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploidy lake sturgeon *(Acipenser fulvescens)*: the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics*, **10**, 203.

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

Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006) Genomics and conservation genetics. *Trends in Ecology and Evolution*, **21**, 629–637.

Kreps JA, Wu Y, Chang H-S, Zhu T, Wang X, Harper JF (2002) Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiology*, **130**, 2129–2141.

Kristensen TN, Pedersen KS, Vermeulen CJ, Loeschcke V (2010) Research on inbreeding in the 'omics' era. *Trends in Ecology and Evolution*, **25**, 44–52.

Kristiansson E, Asker N, Forlin L, Larsson DGJ (2009) Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. *BMC Genomics*, **10**, 345.

Li R *et al.* (2009) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.

Lomax J (2005) Get ready to GO! A biologist's guide to the Gene Ontology. *Briefing in Bioinformatics*, **6**, 298–304.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Margulies M *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, **18**, 1509–1517.

Meglécz E, Costedoat C, Dubut V *et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics*, **26**, 403–404.

Metzker ML (2010) Sequencing technologies—the next generation. *Nature reviews*, **11**, 31–46.

Meyer E, Aglyamova GV, Wang S *et al.* (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GLFlx. *BMC Genomics*, **10**, 219.

Miller JR, Koren S, Sutton G (2010) Assembly algorithm for next-generation sequencing data. *Genomics*, **95**, 315–327.

Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255–264.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene SNPs involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599–3613.

Novaes ED, Drost R, Farmerie WG *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 312.

Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E *et al.* (2010) De novo Assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genetics*, **6**, e1000891. doi:10.1371/journal.pgen.1000891.

Ossowski S, Schneeberger K, Clark RM *et al.* (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Research*, **18**, 2024–2033.

Ouborg NJ, Vriezen WH (2007) An ecologist's guide to ecogenomics. *Journal of Ecology*, **95**, 8–16.

Ouborg NJ, Angeloni F, Vergeer P (2010a) An essay on the necessity and feasibility of conservation genomics. *Conservation genetics*, **11**, 643–653.

Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma R, Hedrick PW (2010b) Conservation genetics in transition to conservation genomics. *Trends in genetics*, **26**, 177–187.

Paige NK (2010) The functional genomics of inbreeding depression: a new approach to an old problem. *BioScience*, **60**, 267–277.

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

Picó FX, Ouborg NJ, van Groenendael JM (2004) Evaluation of the extent of among-family variation in inbreeding depression in the perennial herb *Scabiosa columbaria* (Dipsacaceae). *American Journal of Botany*, **91**, 1183–1189.

Pluess AR, Stöcklin J (2004) Genetic diversity and fitness in *Scabiosa columbaria* in the Swiss Jura in relation to population size. *Conservation Genetics*, **5**, 145–156.

Primmer CR (2009) From conservation genetics to conservation genomics. *Annals of the New York Academy of Science*, **1162**, 357–368.

Quinn NL, Levenkova N, Chow W *et al.* (2008) Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics*, **9**, 404.

Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD, Dangl JL (2009) De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Research*, **19**, 294–305.

Reisch C, Poschlod P (2009) Land use affects flowering time: seasonal and genetic differentiation in the grassland plant *Scabiosa columbaria*. *Evolutionary ecology*, **23**, 753–764.

Rusk N (2009) Cheap third-generation sequencing. *Nature Methods*, **6**, 244–245.

Schatz MC, Delcher AL, Saltzberg SL (2010) Assembly of large genomes using second generation sequencing. *Genome research*, in press. DOI: 10.1101/gr.101360.109.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature biotechnology*, **26**, 1135–1145.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Strausberg RL, Levy S, Rogers YH (2008) Emerging DNA sequencing technologies for human genomic medicine. *Drug Discovery Today*, **13**, 569–577.

Swarbreck D, Wilks C, Lamesch P *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, **36**, 1009–1014.

Szittya G, Moxon S, Santos DM *et al.* (2008) High-throughput sequencing of *Medicago truncatula* short RNAs identifies eight new miRNA families. *BMC Genomics*, **9**, 593.

Van Straalen NM, Roelofs D (2006) *An Introduction to Ecological Genomics*. Oxford University Press, Oxford, UK.

Van Treuren R, Bijlsma R, Van Delden W, Ouborg NJ (1991) The significance of genetic erosion in the process of extinction. 1. Genetic differentiation in *Salvia pratensis* and *Scabiosa columbaria* in relation to population size. *Heredity*, **66**, 181–189.

Van Treuren R, Bijlsma R, Ouborg NJ, van Delden W (1993) The significance of genetic erosion in the process of extinction. IV. Inbreeding depression and heterosis effects caused by selfing and outcrossing *Scabiosa columbaria*. *Evolution*, **47**, 1669–1680.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a non model organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Waldmann P (2002) Fluctuating asymmetry in *Scabiosa canescens* and *Scabiosa columbaria*. Association with genetic variation and population size. *International Journal of Plant Science*, **163**, 329–334.

Waldmann P, Andersson S (2000) Comparison of genetic (co)variance matrices within and between *Scabiosa canescens* and *S. columbaria*. *Journal of evolutionary Biology*, **13**, 826–835.

Wall PK, Leebens-Mack J, Chanderbali AS *et al.* (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, **10**, 347.

Wheat CW (2010) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica*, **138**, 433–451.

Yamamoto T, Nagasaki H, Ji Y *et al.* (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of singlenucleotide polymorphisms. *BMC Genomics*, **11**, 267.

Zeng S, Xiao G, Guo J *et al.* (2010) Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittattum* (Sieb. Et Zucc.) Maxim. *BMC genomics*, **11**, 94.

Zhang G, Guo G, Hu X *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Research*, **20**, 646–654.