Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa.

UNIVERSITÀ DEGLI STUDI DI TORINO

Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa.

Davide Scaglione[1], Sergio Lanteri[1], Alberto Acquadro[1]*, Zhao Lai[2], Steven J Knapp[3,a], Loren Rieseberg[2], Ezio Portis[1]

[1] *University of Torino, DIVAPRA, Grugliasco (TO) - Italy*

[2] *University of British Columbia, Department of Botany Vancouver, British Columbia, Canada*

[3]*University of Georgia, Department of Crop and Soil Sciences and Center for Applied Genetic Technologies, Athens, Georgia, USA*

[a] *current address: Monsanto Company, Woodland, California, USA*

***Corresponding author:**
Dr. Alberto Acquadro
DIVAPRA, Plant Genetics and Breeding - University of Torino
via L. da Vinci 44, 10095 Grugliasco (TO)
tel. +39 011 6708813; fax: +39 011 2368813; e-mail alberto.acquadro@unito.it

e-mail addresses of the authors:
davide.scaglione@unito.it
sergio.lanteri@unito.it
alberto.acquadro@unito.it
zlai@cgb.indiana.edu
steve.knapp@monsanto.com
loren.rieseberg@botany.ubc.ca
ezio.portis@unito.it

Running title: Transcriptome analysis & SNP discovery in globe artichoke

Key words: SNP, transcriptome, next generation sequencing, *C. cardunculus*, *de novo* assembly.

Sequence accessions: sequences have been submitted to NCBI and SRA archive numbers are going to be attributed.

Word count: 6880 words

SUMMARY

Cynara cardunculus (2n=2x=34) is a member of the *Asteraceae* family that contributes significantly to the agricultural economy of the Mediterranean basin. The species includes two cultivated varieties, globe artichoke and cardoon, which are grown mainly for comestible purposes. *C. cardunculus* is an orphan crop species whose genome/transcriptome has been relatively unexplored, especially in comparison to other *Asteraceae* crops. Hence, there is a significant need to improve its genomic resources through the identification of novel genes and sequence-based markers, in order to design new breeding schemes aimed at increasing quality and crop productivity.

We report the outcome of cDNA sequencing and assembly for eleven accessions of *C. cardunculus*. Sequencing of three mapping parental genotypes using Roche 454 Titanium technology generated 1.7M reads, which were assembled into 38,726 reference transcripts covering 32Mbp. Putative enzyme-encoding genes were annotated using the KEGG-database. Transcription factors and candidate resistance genes were surveyed as well. Paired-end sequencing was done for cDNA libraries of eight other representative *C. cardunculus* accessions on an Illumina Genome Analyzer (IGA) IIx, generating 46M reads. Alignment of the IGA and 454 reads to reference transcripts led to the identification of 195,400 SNPs with a Bayesian probability exceeding 95%; a validation rate of 90% was obtained by Sanger-sequencing of a subset of contigs. These results demonstrate that the integration of data from different NGS platforms enables large-scale transcriptome characterization, along with massive SNP discovery. This information will contribute to dissect key agricultural traits in *C. cardunculus* and facilitate the implementation of marker-assisted selection programs.

INRODUCTION

Cynara cardunculus L. belongs to the *Asteraceae* family (formerly *Compositae*) and includes three botanical *taxa*: the globe artichoke (var. *scolymus*), the cultivated cardoon (var. *altilis*), and the wild cardoon [var. *sylvestris* (Lamk) Fiori]. As an out-breeding diploid species (2n=2x=34) with proterandrous and asynchronous sexual maturity, *C. cardunculus* has a highly heterozygous genetic background (Basnitzki and Zohary, 1994). The wild cardoon is the common ancestor of both cultivated forms (Mauro et al., 2008; Lanteri et al., 2004), which evolved independently through directed selection for either the size of the immature inflorescence (globe artichoke) or the fleshiness of its leaves (cardoon). Globe artichoke contributes significantly to the agricultural economy of the Mediterranean basin, and in particular Italy, the world's leading producer with an annual yield of 486,600 tons (FAOSTAT, 2009). In addition to its culinary value, it also produces a number of nutraceutically and pharmaceutically active compounds (Gebhardt, 1997, 1998; Schutz et al., 2004; Shimoda et al., 2003; Lattanzio et al., 2009; Pandino et al., 2011). The cultivated cardoon has been identified as a potential source of both lignocellulosic biomass (Portis et. al., 2010, Ierna et al., 2010) that can be exploited for energy production by means of calorific power of up to 17 MJ/Kg (Foti et al., 1999) and of biodiesel from its seed (Maccarone et al., 1999; Encinar et al., 2002; Gonzalez et al., 2004; Lapuerta et al., 2005).

Unravelling the genetic basis of phenotypic traits is helpful for the design of breeding strategies directed at improved yield, biomass production, and end-use quality. The current *C. cardunculus* genetic map (Portis et. al 2009a) has been expanded significantly since the publication of the earliest AFLP-based version (Lanteri et al., 2006). A major

step forward was taken by the generation of a large set of microsatellite (SSR) assays mined from EST sequences (Scaglione et al., 2009), part of which were used to construct SSR-based consensus genetic maps (Portis et al., 2009b, Sonnante et al., 2011). The rapid development of sequencing technology has promoted the use of SNP (single nucleotide polymorphism) markers in large genotyping experiments, since they are virtually unlimited (Ganal et al., 2009) and a wide range of high-throughput analysis platforms is available (Ragoussis, 2009). Sequence-based molecular markers are essential for performing comparative analysis across related species and providing anchoring features for scaffold ordering in genome sequencing projects. Next generation sequencing (NGS) technology, along with the necessary bioinformatics support, is designed to rapidly acquire very large amounts of sequence data, and is thus well-suited for SNP discovery and detection (Metzker, 2010). Among the NGS platforms suitable for this purpose are the 454 FLX Titanium pyrosequencing, which generates relatively long reads that can facilitate accurate *de novo* assemblies, and the Illumina GAIIx platform, which is highly cost-effective and thus is particularly suitable for re-sequencing and SNP discovery (Delseny et al., 2010).

Sequencing the expressed portion of the genome is generally simpler than attempting whole-genome sequencing because of its lower complexity (Kaur et al., 2011). In orphan crop species, the cost of a whole-genome shotgun approach can be prohibitive, depending on its genome size and level of complexity (Parchman et al., 2010). Recently, a genome reduced complexity approach based on sequencing of RAD-tags (Scaglione et. al. 2012) has been conducted on the *C. cardunculus* genome to generate a large set of genomic SNPs. Indeed, a transcriptome deep sequencing together with an extensive gene

annotation represents a cost-effective and valuable resource for genetic, genomic, and proteomic investigations (Brautigam and Gowik, 2010).

Here we report the sequencing and assembly of the *C. cardunculus* transcriptome and its annotation, including the identification of transposable element (TE) signatures and putative miRNA targets. In addition, we have undertaken a broad-based program of SNP discovery based on sequencing of three mapping population parents and eight other representative accessions of the three *C. cardunculus* taxa.

RESULTS

Sequencing outcome

Roche 454-based cDNA sequencing was done of three mapping parents: var. *scolymus* "Romanesco C3", var. *altilis* "Altilis 41", and var. *sylvestris* "Creta 4" (Table 1, Figure 1). This produced 700 Mbp of raw sequence from ~1.7 M reads (435,375 from "Romanesco C3", 610,622 from "Altilis 41", and 696,300 from "Creta 4") (Table 2). Post-sequencing filtering reduced the total by only ~1%, resulting in 692.2 Mbp with a mean read length of 392 bp (Table 2). Quality trimming did not lead to significantly further data reduction (data not shown). cDNA libraries from the remaining eight accessions (Table 1) were sequenced using an Illumina GAIIx platform, producing 6.9 Gbp of raw data (46.4 M paired-end reads) with a mean of 5.8 M reads per accession. The data set was reduced to 6.7 Gbp following the removal of adaptor sequences and other contaminants, and it was further reduced to 6.2 Gbp after quality trimming. Although aggressive trimming of sequences led to a sizable reduction in bases at the 3'-ends, a common problem in Illumina-derived reads (Metzker, 2010), this measure substantially reduced the risk of false SNP calls. The relative representation and quality of the multiplexed samples in each sequencing lane were evenly distributed (Table 3).

454 de novo assembly

The first phase of the 454 data assembly approach generated 37,622 contigs for "Romanesco C3", 40,130 contigs for "Altilis 41", and 42,837 contigs for "Creta 4" with N50 contig lengths of 834 bp, 761 bp, and 772 bp, respectively, and mean coverage levels of 7.31X, 8.45X, and 9.17X, respectively. For the "Romanesco C3" assembly, a subset of

11,276 contigs resulted from the incorporation of a prior set of 28,641 Sanger ESTs (www.ncbi.nlm.nih.gov/dbEST). In the second phase, after contaminant removal by BLASTX analysis, the three datasets were merged into a set of 38,726 contigs. This "reference" assembly spanned 32.7 Mbp and had a GC content of 42.1%. The mean contig length was 844.3 bp (N50: 951bp) (Figure 2), which represents a 118% improvement in the coverage of the transcriptome from that described by Scaglione et al. (2009). As a result of the second assembly phase, 20,469 contigs were generated by merging at least two *taxon*-derived contigs from the first phase, consisting of a subset with a mean length of 1054 bp, while 5,375, 6,669, and 6,213 remained as single *taxon*-derived contigs of var. *scolymus*, var. *altilis*, and var. *sylvestris*, respectively (Dataset S1).

Sequence analysis

Open reading frame (ORF) and intron prediction. ORFs were associated with 38,567 sequences (99.6% of the total), defining 5.02 Mbp of 5'-UTR sequence, 21.17 Mbp of CDS and 6.42 Mbp of 3'-UTR sequence. BLASTX alignment succeeded in assigning ORFs to 587 contigs where *ab initio* prediction failed. The start codons were validated in 19,198 contigs via comparative analysis by BLASTX, while another 2,037 start codons were labeled as putative. Based on the predicted orthologs or out-paralogs from *Arabidopsis thaliana*, and assuming the general conservation of exon/intron structure (Fedorov et al., 2002; Rogozin et al., 2003), it was possible to infer the positions of 100,102 putative exon-exon junctions in 22,764 of the contigs.

Transcriptome representation

To estimate the transcriptome representation and its gene-level redundancy (e.g. splicing variants), different approaches were adopted. Basing on the *A. thaliana* gene content, the 454 sequencing output was predicted to be assembled in a total of 29.3 Mbp, scattered in 24,064 unigenes with an average length of 1,216 bp, and covering 96% of the transcriptome. Simulation analyses on each of the three independent *taxon*-derived assemblies (Der et al., 2011) resulted in similar asymptotic trends, with end-point values of 31,439, 32,766, and 31,869 for "Romanesco C3", "Altilis 41", and "Creta 4", respectively (Figure 3). To corroborate these results, we clustered the final contig set (38k) using the same criteria that were aimed at collapsing gene variants (*e.g.*, alternative splicing). This resulted in a set of 29,830 unigenes representing a *bona fide* estimation of the gene content of *C. cardunculus*, which suggested that 23% of splicing variants were present in our transcriptome assembly.

Functional annotation

The assignment of *C. cardunculus* genes to the chloroplast genome was based on similarity to those of lettuce and sunflower (Timme et al., 2007); this led to the categorization of 137 contigs. Of 84 annotated sunflower chloroplast genes, 80 were also present in *C. cardunculus*. Similarly, the grapevine (*Vitis vinifera*) mitochondrial genome (Goremykin et al., 2009) was used to identify putative transcripts of *C. cardunculus* mitochondria; out of the 74 grape genes, we identified 52 contigs with high sequence similarity. Detailed information regarding non-nuclear transcripts, their representation, and redundancy is provided in Dataset S2. The automated BLASTX analysis produced 711,220 hits against the NCBI nonredundant protein database, allowing for the retrieval of >1.3 million GO terms from several databases (80% from UniProtKB/TrEMBL). The

Blast2GO pipeline successfully annotated 32,408 *C. cardunculus* transcripts, with 4,399 falling below the threshold score and 1,919 remaining unassociated with any GO term. Finally, a total of 184,469 GO terms were assigned, with an average of five per sequence. Conserved domains were identified by InterProScan (IPS, www.ebi.ac.uk/interpro), resulting in codes for 25,485 sequences and an additional 14,720 GO terms for the final annotation. A comprehensive table showing the full annotation is given in Dataset S3. Transcription factor activity was assigned to 1,398 transcripts distributed among 67 families. Over half of the transcripts involved in the regulation of transcription belonged to one of the *bHLH*, *MYB*, *WRKY*, *C2H2*, *NAC*, *AP2-EREBP*, *bZIP*, *MYB*-related or *C3H* families (Table 4, Dataset S4).

Across a set of 12,449 transcripts, 16,419 enzyme codes were retrieved from the Blast2GO database and mapped onto KEGG's metabolic pathway encyclopedia (www.genome.jp/kegg/). The representative sample of *C. cardunculus* enzymes consisted of 1,133 unique enzyme codes. A summary and separate map images for each pathway are provided in Dataset S5. A subset of 71 enzymatic activities known to be involved in phenylpropanoid synthesis was identified among 921 sequences; 21 of these were annotated at varying levels of redundancy in the core phenylpropanoid pathway (KEGG's map: 00940), in which the synthesis of caffeoylquinic and di-caffeoylquinic acids (CQAs and dCQAs) takes place (Table 5). Pathogen resistance gene homologs were represented by 1,860 transcripts. After validation by a PFAM search, 316 were retained on the basis of 214 matches with leucine-rich repeats, 79 matches with TIR motifs, and 52 matches with NB-ARC motifs; 23 of the sequences carried both a TIR and an NB-ARC motif (Dataset S6).

Relict TE sequence in the transcriptome

When the *Viridiplantae* RepBase collection was interrogated, the search criteria identified 371 occurrences of TE relict sequence in the transcriptome (Dataset S7). There were discernible differences between the translated and non-translated regions with respect to both the identity of the TE family involved and the mean length of relict sequence present (Figure 4). LTR *copia*-like elements were the most frequent (103), with a mean sequence length of 151 bp. DNA/*Helitron* and RC/*Helitron* sequences had comparable mean lengths of 142 bp and 156 bp, respectively. DNA/*En-Spm* and LTR/*Gypsy*-like relicts were all relatively shorter with mean lengths of 95 bp and 108 bp, respectively. The *copia*-like elements tended to leave larger insertions in the 5'-UTR (mean 252bp) than in either the CDS (103bp) or the 3'-UTR (98bp). No such pattern of distribution was evident for either the DNA/*Em-Spm* or the *Gypsy*-like elements.

Presumptive miRNA targets

Each sequence was scanned for the presence of recognition sites for known plant miRNAs. In total, target annealing sites for 302 miRNAs were located in 1,043 transcripts (Table 6). *miR414* was removed from the dataset because of its low level of conservation in genomes other than *A. thaliana* and rice and its poor precursor hairpin structure (miRBase, release 16). A Fisher's exact test indicated some GO enrichment for miRNA targets (Table 7), particularly for the categories "immune system/defense response" and "programmed cell death/apoptosis", followed by "reproduction", "development of anatomical structure", "photosynthesis", "transmembrane receptor activity", and "transcription factor activity". A total of 40 sequences belonged to both the "programmed

cell death/apoptosis" and "immune system"-related paths of the DAG (directed acyclic graph); half of these were related to *miR2109*, a soybean miRNA (Wang et al., 2009) predicted to target 22 sequences across the whole *C. cardunculus* dataset. *miR2109* was the third most abundant target site after *mirR395* (26 transcripts) and *miR2275* (29 transcripts). A complete classification is given in Dataset S8.

Read mapping and SNP calling

About 1.5M of the 454-derived reads were aligned to the reference contig set (38,726 contigs). This number was reduced to ~1.0M by removing those that showed more than one unique alignment, thereby lowering the risk of false SNP calls due to misalignment of paralog-derived reads or to redundancy resulting from splicing variants. The same procedure was repeated for the Illumina-derived reads, producing an alignment of ~60M paired ends. Resolving paired ends reduced this to a set of ~21M reads. An assembly based on >35M sequences was generated by merging the two sequence datasets, resulting in a median genome coverage of 96X with 26,990 reference contigs containing at least 20 mapped reads.

Reliable SNPs (Bayesian probability >95%) were detected at 195,400 sites across the set of eleven accessions (Dataset S9). The average SNP frequency was calculated at 1 per 167 bp, with a mean of five per contig. Each SNP site was interrogated by scoring for the presence of at least one accession-specific sequence. Sequence information was available from an average of nine accessions per SNP site, and a core subset of 57,125 SNPs showed coverage from all the samples. Sanger sequencing was performed on 153 randomly chosen heterozygous SNP loci from the 454-derived "Romanesco C3" set, of

which 138 (90%) were confirmed (data not shown).

The merging of the Illumina-derived reads (eight accessions) with 454-generated reads substantially increased the number of parent-specific SNPs that were identified (Figure 5). A SNP-calling simulation was carried out using only the 454 reads or using the merged dataset. From the former set, ~46,600 SNPs were discovered, while ~81,700 SNPs were found in the latter set, largely because a much higher number of sequences available for SNP calling increases the likelihood that the imposed three-base threshold will be exceeded. In the 454-sequenced samples, the identification of exclusively homozygous mutations was increased by 74% in the merged data set compared to the set of 454-derived reads alone (23,393 versus 13,454), and the identification of mixed homozygous/heterozygous mutation sites showed 15% improvement (6,682 versus 5,819). The best results were obtained for SNPs with exclusive allelic variants in heterozygous states, with an 89% increase in SNP discovery in the merged data set (51,579 versus 27,360).

SNP categorization

The distribution of minor allele frequencies is shown in Figure 6: an even distribution was reported with a slightly higher proportion of low-frequency variants, which can be ascribed to the presence of wild accessions in the sample panel. *C. cardunculus* is a highly heterozygous species, so that mapping to date has been based on a pseudo-testcross strategy. The presence of intra-accession allelic variation is therefore of particular interest. As expected by their shallower coverage, the 454-derived sequences produced a somewhat lower frequency of SNPs with successful heterozygous SNP calling (Figure 7).

"Altilis 41" was relatively the least heterozygous of the accessions (17,570 loci), as has been observed previously (Portis et al., 2005a, 2009b), while "Romanesco Zorzi" was the most heterozygous (43,387), followed by "Violetto di Chioggia" (41,824). "Imperial Star" had the lowest ratio of heterozygous variants among globe artichoke genotypes (13.5%). Overall, SNPs were most frequent in 3'-UTR (one per 126bp), followed by the CDS (one per 169bp), and the 5'-UTR (one per 265bp); this same distribution of SNPs was also observed in *Glycine soja* (Kim et al., 2010). Of the SNPs located in the translated region, the balance between synonymous and non-synonymous was almost 1:1 (64,328 to 60,903). Premature stop codons were associated with 1,949 SNP sites, and the relative frequency of transitions to transversions was 63% to 37%. The allelic diversity within each accession was evaluated by considering "private" alleles (Table 8). The wild cardoon accessions carried the greatest number of allelic variants, of which 61% were synonymous and 39% non-synonymous. The cultivated cardoon accessions, with a lower amount of private SNPs, had a similar trend except for "Altilis 41", which showed an increased level of non-synonymous private SNPs, similar to those of "Romanesco C3" and "Romanesco Zorzi". With respect to private alleles in the homozygous state, the wild and cultivated cardoon accessions shared a similar frequency, but the globe artichokes were rather contrasting: while "Romanesco C3", "Violetto di Chioggia" and "Romanesco Zorzi" and each harboured a small proportion of homozygous private alleles (respectively 2.2%, 5.9% and 7.1%), "Imperial Star" harboured 43.3% (Table 8).

DISCUSSION

Rationale

NGS technology and the development of genomic resources in non-model plant species are significant from both an agricultural and an ecological point of view (Der et al., 2011; Kaur et al., 2011; Novaes et al., 2008; Trick et al., 2009). Much of the effort to date has focused on EST sequencing in an attempt to identify functional important genes, while large scale SNP characterization in crop species has concentrated mostly on that having a relevant economic scale. Globe artichoke plays a key role in the agricultural economy of the Mediterranean basin and its cultivation area, as well as its global economic significance, is increasing by progressive market expansion towards China, Africa and South America (FAOSTAT, 2009). However, unlike other species belonging to the Asteraeae family (like sunflower and lettuce) for which lot of efforts have been undertaken, it can be considered an orphan crop species being its genome/transcriptome relatively unexplored. The rapidly falling cost of DNA sequencing brought about by the development of NGS technology has now allowed us to fill this gap and, in this scenario, the generation of new comprehensive genomic/transcriptomic resources was expedited. The 454 Titanium (Roche) platform was applied to the genotypes of each of the three *C. cardunculus* taxa (globe artichoke, cultivated and wild cardoon) to produce a reference transcriptome, while the use of the Illumina GAIIx platform was intended to identify SNP variation by re-sequencing further five globe artichoke accessions, two cultivated cardoon and one wild cardoon genotypes (Table 1), demonstrating the complementary utility of the two platforms.

De novo assembly

Ideally, sequence assembly and mapping requires both a pre-existing scaffold of overlapping reads, and that each locus can be uniquely identified. In EST data, the latter requirement is disrupted by the presence of both splicing variants and the sequence similarity between distinct members of a single gene family. As a result, some fine-tuning of the assembly/mapping procedures becomes necessary to avoid misplacement of sequences and the false calling of SNPs. Here, a two-step approach was adopted; the first aiming to remove chimeric reads and spoilers (i.e. misassembled reads that interrupt the progression of overlap-layout-consensus algorithm) so that data loss in the second step (merging of the various datasets) could be minimized. The identification of a set of unigenes and the simulated assembly curves for each of the datasets suggested an even representation of transcripts across the three subspecies. As a result, it was possible to estimate that some 23% of the transcripts were splicing variants, which matches closely the estimated frequency of 21.5% in *A. thaliana* (TAIR10, www.arabidopsis.org). Further support for this estimate was provided by the observation that around 25% of the Illumina-derived reads aligned with multiple reference transcripts (data not shown). The normalization procedure adopted was effective in identifying the genes encoding a number of enzymes involved in phenylpropanoid synthesis. Only ~20,000 read ends (0.02% of the total sequence acquired) were mapping to the RuBisCO large subunit transcript sequence.

Transcriptome annotation

The 454-based assembly succeeded in identifying non-nuclear transcripts by exploiting the expected high level of sequence conservation reported between the chloroplast

genomes of the *Asteraceae* species lettuce and sunflower (Timme et al., 2007). In contrast, as expected, a reduced number of transcripts was identified using the sequence information from the grapevine mitochondrial genome, as its sequence complement is rather variable, even between closely taxa, as a result of gene loss over time (Adams and Palmer, 2003; Palmer et al., 2000). With respect to the nuclear component, a number of genes within the phenylpropanoid synthesis pathway were identified.

The categorization of transcription factors is particularly important for gaining an understanding of the control of gene expression. Their number in the *C. cardunculus* genome appears to be similar to that in the grapevine genome (PlnTFDB, http://plntfdb.bio.uni-potsdam.de), which lends further support to the capture of most of the transcriptome within the present sequence dataset. Many of the genes associated with pathogen resistance (*R* genes) are of the NBS-LRR type, which form a large gene family in plant genomes. They are well represented in the *C. cardunculus* transcriptome, allowing for the possibility of undertaking a candidate gene approach in a positional cloning strategy for genes determining pathogen resistance.

Transposable elements and miRNA targets

Certain LTR retrotransposons appear to target the gene space (Hirochika et al., 1996) and their contribution to gene evolution and expression has been widely explored (Bennetzen, 2000). In maize, the phenotypic consequences of TE activity have been widely investigated (Marillonnet and Wessler, 1997; Scott et al., 1996; Wessler et al., 1995) and several genes have shown to harbor relict TE sequence. About 100 instances of *copia*-like elements were noted in the *C. cardunculus* transcriptome; their mean length of 151bp was almost the same as that of the DNA/ and RC/*Helitrons* present (142bp and 156bp,

respectively), while the DNA/*En-Spm* and LTR/*Gypsy*-like sequences were somewhat shorter (95bp and 108bp, respectively). The mean insert size of *copia*-like elements present in the 5'-UTR fraction (252bp) was rather longer than their mean length in the CDS (103bp) or 3'-UTR (98bp), but this pattern did not extend to either the DNA/*Em-Spm* or the *Gypsy* elements. Overall, this suggests that Gypsy-like, copia-like and En/Spm-like TEs have had a measurable influence on the evolution of the host's gene space. The presence of such long *copia*-like element relicts in the 5'-UTR regions may imply that the constraints imposed on transposition are TE-dependent. Some useful information emerged from the analysis of conserved miRNA binding sites. Although requiring experimental validation, the sequence data suggested a marked contribution of *miR2109* which targets several TIR type NBS-LRR *R* genes (Wang et al., 2011).

SNP frequency and diversity

SNP frequency in the *C. cardunculus* transcriptome appears to be comparable to that found in the heterozygous grapevine whole genome sequence (Velasco et al., 2007) and among *Citrus* ssp. ESTs (Jiang et al., 2010). Within the UTRs, the frequency also matched that obtaining in tomato expressed sequence (Jimenez-Gomez and Maloof, 2009), while it was markedly higher to that present in the coding region (~2 per kb). This discrepancy may reflect either the greater tolerance by the heterozygous state of non-synonymous substitutions, or merely is an ascertainment bias due to the analysis a larger germplasm panel (which also included accession of a wild relative).

The merging of the 454- with the Illumina-derived data was a critical step towards SNP calling, in particular for rare heterozygous SNPs, which are needed to develop test-cross

markers (Figure 5). The highest improvement of exclusively heterozygous variants in 454 samples (89%) is expected to be symptomatic of a higher proportion of rescued SNPs which would not be identified by their lower reads count. The 90% validation rate of the SNPs called from the 454-derived sequence confirmed the reliability of the platform (Pavy et al., 2006). Despite the lower coverage of the 454 reads, the relative abundance of homozygous and heterozygous SNPs was maintained in the merged dataset. The globe artichoke cultivar "Imperial star" had a marked number of private alleles in the homozygous state, which likely reflects its development by directed breeding, possibly involving crosses with exotic material and the use of enforced self-fertilization. As opposed, farmer selection in clonally propagated varietal types has maintained a wide range of within genetic variation and heterozygosity (Portis et al., 2005a). The relatively high occurrence of private alleles among the wild cardoon accessions may be due to their high genetic diversity (Portis et al., 2005b), although the number of accessions analysed was just two. However, private allele richness remained consistent across 454- and Illumina-sequenced samples, suggesting a very low probability of false discovery events.

CONCLUSIONS

The combination of two NGS platforms has allowed for both an extensive characterization of the *C. cardunculus* transcriptome, and the rapid discovery of a large number of SNPs. The de novo assembly of the transcripts produced a catalogue of transcription factors, miRNA targets and putative *R* genes, which together represent an invaluable resource for upcoming genomic and genetic studies in this species. Actually, since many of the identified SNPs reside in homologs already mapped in lettuce and sunflower, they will facilitate comparative genetics across the Asteraceae family.

The adoption of stringent alignment and calling criteria has generated a robust EST-SNP database, the size of which is more than sufficient for any conceivable genotyping application. The identification of common and rare allelic variants will facilitate the design of diversity or association studies, while the inclusion of mapping parents in the germplasm panel has provided a ready-made source of informative assays for genetic mapping purposes. The availability of such a large number of sequence-based markers, in a format allowing for high throughput genotyping, offers many opportunities to conduct genetic analyses of key agricultural traits in a Mediterranean crop species which has end-uses both as a food, as a source of nutraceuticals and as a possible biomass producer. The outcomes of such analyses should facilitate the implementation of marker-assisted selection for the improvement of globe artichoke and cultivated cardoon.

EXPERIMENTAL PROCEDURES

Plant materials and RNA extraction

RNA was extracted from the eleven *C. cardunculus* accessions listed in Table 1. Leaf and root material was harvested from field-grown plants of "Romanesco C3", "Altilis 41" and "Creta 4", rinsed in sterile water, frozen in liquid nitrogen and stored at -80°C. Seeds of the other eight accessions were pot-grown in sand soaked in a hydroponic solution. After five weeks, the leaves and roots were collected, frozen in liquid nitrogen and stored at -80°C. Total RNA was extracted from 1g of each tissue using the TRIizol reagent (Invitrogen), following the manufacturer's instructions. The resulting RNA was quantified and controlled for purity using a Nanodrop 2000c spectrophotometer (Thermo Scientific). The RNA from the root and leaf of each accession was mixed in equimolar amounts and its integrity checked using BioAnalyzer® 2100 (Agilent).

Construction of normalized cDNA libraries and sequencing

In order to obtain the full length cDNAs for transcriptome analysis, we used SMART cDNA synthesis technology to generate cDNAs by following a protocol from the Clontech Creator SMART cDNA library construction kit (Clontech Laboratories, Mountain View, CA, USA). This requires an oligo-dT primer that anchors the polyA tail of mRNA to primer the cDNA synthesis process. However, for 454 pyrosequencing, mononucleotide runs reduce sequence quality and quantity due to excessive light production and crosstalk between neighboring cells (Margulies et al., 2005). To counteract this problem, we employed a "broken chain" short oligo-dT primer (primer sequence: 5'- AAGCAGTGGTATCAACGCAGAGTCGCAGTCGGTACTTTTTTCTTTTTTV

- 3', V=A, G or C) to prime the poly(A) tail of mRNA during first strand cDNA synthesis (Meyer et al., 2009). Approximately 1.5 µg of total RNA was reverse-transcribed to first-strand cDNA using this method. Following the first strand cDNA synthesis, double-stranded (ds) cDNA synthesis was performed using Phusion polymerase (New England Biolabs, Ipswich, MA, USA) with a hot start of 98°C for 30 sec, followed by 18 cycles of 98°C for 7 sec, 66°C for 20 sec, and 72°C for 4 min. The ds-cDNA PCR product was purified using a QIAquick PCR Purification column (Qiagen). To reduce the abundance of common transcripts, the cDNA library was normalized using a TRIMMER-DIRECT cDNA normalization kit (Evrogen, Moscow, Russia). Approximately 800ng of purified ds-cDNA was used as the starting material for normalization. A mixture of 0.25 µl and 0.5 µl DSN normalization tubes were used for the first and second amplifications. After normalization, cDNA was fragmented to 500 to 800 bp fragments by sonication and size-selected to remove small fragments using AMpure SPRI beads. Then the fragmented ends were polished and ligated with adaptors. The optimal ligation products were selectively amplified and subjected to two rounds of size selection including gel electrophoresis and AMpure SPRI bead purification (Lai et al., 2011). Each of the three libraries was sequenced using a half picotitre plate in a 454 FLX Titanium device, following the manufacturer's sequencing protocol. For 8 Illumina sample sequencing, we followed the SMART technology to generate the full length cDNA and normalized the cDNA based on TRIMMER cDNA normalization kit described above. The ds-cDNAs then were treated following the standard multiplexing genomic DNA shotgun library preparation kit to generate the genomic DNA library for sequencing. Sequencing was carried out in two Illumina GAIIx lanes (four multiplexed samples each), following the manufacturer's protocol.

Read pre-processing

The 454-derived reads were first screened to remove adaptor/primer leftovers by inspecting the dataset with the SeqClean perl script (compbio.dfci.harvard.edu/tgi/software/). Sequences were quality clipped using a sliding window analysis starting from the last 6bp on both the 5'- and the 3'- end; when average quality fell below Phred20, two external bp were discarded and the window shifted accordingly. Illumina reads were de-multiplexed using the Illumina Pipeline software, and the sequences were then screened for the presence of residual adaptor leftovers, using the ShortRead package (available at www.bioconductor.org). The subsequent quality clipping was performed as described for 454-derived sequence.

De novo assembly of 454 reads

The assembly of the 454-derived reads was achieved using a two-step procedure implemented in the MIRA assembler v3.2.0 (Chevreux et al., 2004). An overview of the analysis pipeline is provided in Figure 1. The reads from each of the three libraries were initially assembled independently. For the "Romanesco C3" library, a hybrid assembly was carried out by including 36,321 Sanger-ESTs (www.ncbi.nlm.nih.gov/dbEST) developed by the previous efforts of the Compositae Genome Project (CGP, http://compgenomics.ucdavis.edu). Reads which remained as singletons were discarded. Each contig was prefixed by "Scolymus_", "Altilis_" or "Sylvestris_". In the second step, the contigs were treated as Sanger reads and were merged. The different parameters applied in each of the two assembly phases are listed in Table 9. Both the post-merging contigs and those remaining as singletons were included in the transcript dataset. Contigs obtained by merging across subspecies were prefixed "C_cardunculus_joined_".

Potential contaminant-derived contigs were identified by Blastx against the *Viridiplantae* protein database (NCBI) and removed. A representative catalogue of unigenes was identified by additional clustering based on the CAP3 algorithm (Huang et al., 2009), applying a 50% criterion for maximum overhangslength and a 95% identity cutoff. Validation of the method was obtained through an independent cluster analysis based on CD-HIT-EST software (Huang et al., 2010, data not shown), applying a 95% identity cut-off, –r and –g options, and all other parameters left as default. The assembly performance was assessed by bootstrapped EST sampling (Der et al., 2011) applying 1,000 permutations.

Location of ORFs and introns

HMM-based ESTScan3 software (Iseli et al., 1999) was adopted for ORF identification (score threshold set with b=0.8), using a customized HMM matrix, which was created by means of scripts available in the prot4EST package (Wasmuth and Blaxter, 2004). To this regard, a simulated transcriptome was obtained by parsing the Blastx output against the TAIR9 dataset (http://www.arabidopsis.org/), setting as thresholds an E-value of e^{-15} and the number of alignment gaps as no more than six. Where ESTScan failed to generate a result, Blastx output parsing was used instead. Putative introns were inferred from an additional Blastx analysis against the TAIR 9 database, by parsing alignment offsets with genomic coordinates of exons. Intron positions were reported on the *C. cardunculus* contigs, using a perl script available at int-citrusgenomics.org/usa/ucr/files/intron_finder.zip. The default E-value cut-off of e^{-57} was considered to imply orthology.

Transcritome representation

The prediction of assembly performance was assessed using the simulation tool ESTcalc (fgp.huck.psu.edu/NG_Sims/ngsim.pl), considering the sequencing output obtained after the removal of adaptors. The permutation-based estimation of EST clustering was accomplished using a perl script developed by Der et al. (2011), combining MIRA assembly output with CAP3 clustering, and 1,000 permutations.

Transcriptome annotation

Non-nuclear transcripts were identified via Blastn analysis (E-value <e-^{35}, nucleotide identity >85%). The identification of chloroplast sequences within the *C. cardunculus* sequence dataset was based on the sunflower chloroplast sequence, and that of mitochondrial sequences from the grapevine sequence. The contigs were submitted to an annotation pipeline using a Blast2GO java interface (Conesa et al., 2005). First, all sequences were submitted to the NCBI nr protein database using the Blastx algorithm (E-value <e^{-3}, 20 best hits recorded). Gene names and GI (gene identifier) were recorded, while PIR code (Non-redundant Reference Protein Database) by reference to either UniProt, SwissProt, TrEMBL, RefSeq, GenPept or PDB. The GIs were used to search the GO database. The annotation of the transcripts was performed by applying the Blast2Go-embedded formula, setting a threshold score of 55. ANNEX augmentation was performed to retrieve implicit GO terms by certain "GO Molecular Functions". InterPro scan was carried out to collect additional GO annotations on the basis of conserved functional domains. Enzyme codes were retrieved from GO tables and mapped onto the KEGG pathway.

The identification and categorization of transcription factors were obtained on the basis of the GO terms “regulation of transcription”, “transcription factor activity”, “transcription activator activity”, “regulation of transcription”, “DNA-dependent”, “transcription repression activity”, “negative regulation of transcription”, “positive regulation of transcription” or “nucleic acid binding”, and these sequences were then used for a Blastx search against the Plant Transcription Factor Database (Perez-Rodriguez et al., 2010).

R gene candidates were identified by means of Blastp analysis against the Plant Resistance Genes database (PRGdb) (Sanseverino et al., 2010). Positive hits were validated via HMMER v3 (hmmer.janelia.org/software) software, searching against PFAM hidden Markov models for NB-ARC (PF00931), TIR (0PF1582) and several leucine-rich repeat (PF00560, PF01462, PF01463, PF12799, PF07723, PF08191, PF08263, PF07725, PF01816, PF12534) motifs (Finn et al., 2010); “gathering scores” method was applied to retrieve positive hits.

The presence of TE relict sequence was inferred by a RepeatMasker scan, using a local rmBlastn analysis against the *Viridiplantae* RepBase v16.01. Only alignments with an identity >75%, a minimum length of 60bp and a rmBlastn score >225 were retained. Transcribed sequences were subjected to psRNATarget analysis against miRBase 16 (Griffiths-Jones et al., 2008). A maximum expectation of 2.5 was adopted, allowing a maximum energy to unpair the target site of 25, and considering 17bp upstream and 13bp downstream of the TE sequence. Inhibition of translation was considered for mismatches in the 9th to 11th mature miRNA nucleotides. Any enrichment of GO terms was identified

by comparing the putative miRNA targets against the whole transcript dataset by means of the Gossip package implemented in the Blast2Go suite: a Fisher's exact test was applied collecting terms with *p* values $<e^{-4}$ and false discovery rate <0.01.

Read mapping and SNP calling

Reads produced by both sequencing platforms were aligned against the reference contig set. The alignment software package Mosaik (bioinformatics.bc.edu/marthlab/Mosaik) was used to process the 454- and Illumina-derived reads separately. The parameters for the 454 reads were: "m=unique, hs=13, act=36, bw=41, mmp=.07, mmal, minp=.90", while Illumina paired-end reads were aligned with "hs=11, act=25, bw=25, m=all, mmp=.07, mmal, minp=.90, ls=150" and paired-ends were sorted with "afl, ci=.9985, sa, rmm". Alignments were merged and assembled with MosaikMerge and MosaikAssembler modules, respectively. Automated SNP calling on aligned reads was carried out by gigaBayes (bioinformatics.bc.edu/marthlab/GigaBayes) software. SNPs were identified on the basis of the following search parameter set: "ploidy=diploid, QRL=30, CAL=3, PSL=0.95". A customized perl script (provided by JM Jimenez-Gomez (UC Davis, USA)) was used to locate the placement of the SNPs into the CDS, the 5'-UTR or the 3'-UTR, to assess whether the SNP was synonymous, to predict ORFs and to position the SNPs. The full SNP data set has been organized into a relational database, which is available upon request.

ACKNOWLEDGEMENTS

This research was supported by: (i) U.S. National Science Foundation grants (DBI-0820451) for LHR, SJK, and ZL, (ii) and by MIPAAF (Ministero delle Politiche Agricole, Alimentari e Forestali - Italy) through the CYNERGIA ("Costituzione e valutazione dell'adattabilita' di genotipi di *Cynara cardunculus* per la produzione di biomassa e biodiesel in ambiente mediterraneo") project and CARVARVI ("Valorizzazione di germoplasma di carciofo attraverso la costituzione varietale ed il risanamento da virus") project. We wish to thank Joan Wong for the manuscript revision.

REFERENCES

Adams, K. and Palmer, J. (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution* **29**, 380-395.

Basnitzki, J. and Zohary, D. (1994) Breeding of seed planted artichoke. *Plant Breeding Reviews* **12**, 253-269.

Bennetzen, J. (2000) Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology*, 251-269.

Brautigam, A. and Gowik, U. (2010) What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biology*, 831-841.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A., Muller, W., Wetter, T. and Suhai, S. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, 1147-1159.

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

Delseny, M., Han, B. and Hsing, Y. (2010) High throughput DNA sequencing: The new sequencing revolution. *Plant Science***, 179**, 407-422.

Der, J., Barker, M., Wickett, N., dePamphilis, C. and Wolf, P. (2011) De novo characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. *BMC Genomics*, **12**: 99.

Encinar, J., Gonzalez, J., Rodriguez, J. and Tejedor, A. (2002) Biodiesel fuels from vegetable oils: Transesterification of *Cynara cardunculus* L. oils with ethanol. *Energy & Fuels*, 443-450.

Fedorov, A., Merican, A. and Gilbert, W. (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proceedings of the National Academy of Sciences of the United States of America*, 16128-16133.

Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J., Gavin, O., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E., Eddy, S. and Bateman, A. (2010) The Pfam protein families database. *Nucleic Acids Research* **38**, D211-D222.

Foti, S., Mauromicale, G., Raccuia, S., Fallico, B., Fanella, F. and Maccarone, E. (1999) Possible alternative utilization of *Cynara* spp. I. Biomass, grain yield and chemical composition of grain. *Industrial Crops and Products*, 219-228.

Ganal, M., Altmann, T. and Roder, M. (2009) SNP identification in crop plants. *Current Opinion in Plant Biology*, 211-217.

Gebhardt, R. (1997) Antioxidative and protective properties of extracts from leaves of the artichoke (*Cynara scolymus* L) against hydroperoxide-induced oxidative stress in cultured rat hepatocytes. *Toxicology And Applied Pharmacology* **144**, 279-286.

Gebhardt, R. (1998) Inhibition of cholesterol biosynthesis in primary cultured rat hepatocytes by artichoke (*Cynara scolymus* L.) extracts. *Journal Of Pharmacology And Experimental Therapeutics* **286**, 1122-1128.

Geraldes, A., Pang, J., Thiessen, N., Cezard, T., Moore, R., Zhao, Y., Tam, A., Wang, S., Friedmann, M., Birol, I., Jones, S., Cronk, Q. and Douglas, C. (2011) SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology Resources* **11**, 81-92.

Gonzalez, J., Gonzalez-Garcia, C., Ramiro, A., Gonzalez, J., Sabio, E., Ganan, J. and Rodriguez, M. (2004) Combustion optimisation of biomass residue pellets for domestic heating with a mural boiler. *Biomass & Bioenergy*, 145-154.

Goremykin, V., Salamini, F., Velasco R. and Viola R. (2009) Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. *Molecular Biology and Evolution*, **26**, 99-110.

Griffiths-Jones, S., Saini, H., van Dongen, S. and Enright, A. (2008) MiRBase: tools for microRNA genomics. *Nucleic Acids Research*, D154-D158.

Hirochika, H., Otsuki, H., Yoshikawa, M., Otsuki, Y., Sugimoto, K. and Takeda, S. (1996) Autonomous transposition of the tobacco retrotransposon Tto1 in rice. *Plant Cell*, 725-734.

Huang, X. and Madan, A. (1999) CAP3 a DNA sequence assembly program. *Genome Research*, **9**, 868-877.

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**: 680-682.

Ierna, A. and Mauromicale, G. (2010) *Cynara cardunculus* L. genotypes as a crop for energy purposes in a Mediterranean environment. *Biomass & Bioenergy*, 754-760.

Iseli, C., Jongeneel, C.V. and Bucher, P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*, 138-148.

Jackson, S., Rounsley, S. and Purugganan, M. (2006) Comparative sequencing of plant genomes: Choices to make. *Plant Cell*, **18**, 1100-1104.

Jiang, D., Ye, Q., Wang, F. and Cao, L. (2010) The mining of *Citrus* EST-SNP and its application in cultivar discrimination. *Agricultural Sciences in China*, 179-190.

Jimenez-Gomez, J. and Maloof, J. (2009) Sequence diversity in three tomato species: SNPs, markers, and molecular evolution. *BMC Plant Biology*, **9**, 85.

Kaur, S., Cogan, N., Pembleton, L., Shinozuka, M., Savin, K., Materne, M. and Forster, J. (2011) Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genomics* **12**, 265.

Kim, M., Lee, S., Van, K., Kim, T., Jeong, S., Choi, I., Kim, D., Lee, Y., Park, D., Ma, J., Kim, W., Kim, B., Park, S., Lee, K., Kim, D., Kim, K., Shin, J., Jang, Y., Do Kim, K., Liu, W., Chaisan, T., Kang, Y., Lee, Y., Moon, J., Schmutz, J., Jackson, S., Bhak, J. and Lee, S. (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proceedings of the National Academy of Sciences of the United States of America*, 22032-22037.

Lai, Z., Kane, N.C., Kozik, A., Hodgins, K.A., Dlugosch, K.M., Barker, M.S., Matvienko, M., Yu, Q., Turner, K.G., Pearl, S.A., Bell, G.D.M., Zou, Y., Grassa, C., Guggisberg, A., Adams, K.L., Anderson, J.V., Horvath, D.P., Kesseli, R.V., Burke, J.M., Michelmore, R.W. and Rieseberg L.H. (2012) Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression. *American Journal of Botany*, **99**(2), 1–10.

Lanteri, S., Saba, E., Cadinu, M., Mallica, G.M., Baghino, L. and Portis E. (2004) Amplified fragment length polymorphism for genetic diversity assessment in globe artichoke. *Theoretical and Applied Genetics*, **108**, 1534-1544.

Lanteri, S., Acquadro, A., Comino, C., Mauro, R., Mauromicale, G. and Portis, E. (2006) A first linkage map of globe artichoke (*Cynara cardunculus* var. *scolymus* L.)

based on AFLP, S-SAP, M-AFLP and microsatellite markers. *Theoretical And Applied Genetics*, **112**, 1532-1542.

Lanteri, S., Portis, E., Acquadro, A., Mauro, R.P. and Mauromicale, G. (2011) Morphology and SSR fingerprinting of newly developed *Cynara cardunculus* genotypes exploitable as ornamentals. *Euphytica*, Online first, DOI 10.1007/s10681-011-0509-8.

Lattanzio, V., Kroon, P. A., Linsalata, V. and Cardinali, A., (2009) Globe artichoke: a functional food and source of nutraceutical ingredients. *Journal of Functional Foods*, **1**, 131-144.

Lapuerta, M., Armas, O., Ballesteros, R. and Fernandez, J. (2005) Diesel emissions from biofuels derived from Spanish potential vegetable oils. *Fuel*, 773-780.

Maccarone, E., Fallico, B., Fanella, F., Mauromicale, G., Raccuia, S. and Foti, S. (1999) Possible alternative utilization of *Cynara* spp. II. Chemical characterization of their grain oil. *Industrial Crops and Products*, 229-237.

Marillonnet, S. and Wessler, S. (1997) Retrotransposon insertion into the maize waxy gene results in tissue-specific RNA processing. *Plant Cell*, 967-978.

Mauro, R., Portis, E., Acquadro, A., Lombardo, S., Mauromicale, G. and Lanteri, S. (2009) Genetic diversity of globe artichoke landraces from Sicilian small-holdings: implications for evolution and domestication of the species. *Conservation Genetics*, **10,** 431-440.

McNally, K., Childs, K., Bohnert, R., Davidson, R., Zhao, K., Ulat, V., Zeller, G., Clark, R., Hoen, D., Bureau, T., Stokowski, R., Ballinger, D., Frazer, K., Cox, D., Padhukasahasram, B., Bustamante, C., Weigel, D., Mackill, D., Bruskiewich, R., Ratsch, G., Buell, C., Leung, H. and Leach, J. (2009) Genome wide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 12273-12278.

Metzker, M. (2010) Next Generation Technologies: Basics and Applications. *Environmental and Molecular Mutagenesis*, 691-691.

Meyer, E., Aglyamova, G., Wang, S., Buchanan-Carter, J., Abrego, D., Colbourne, J., Willis, B. and Matz, M. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, **10**, 219.

Myles, S., Chia, J., Hurwitz, B., Simon, C., Zhong, G., Buckler, E. and Ware, D. (2010) Rapid genomic characterization of the genus *Vitis*. *Plos One*, **5**(1): e8219.

Novaes, E., Drost, D., Farmerie, W., Pappas, G., Grattapaglia, D., Sederoff, R. and Kirst, M. (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 312.

Palmer, J., Adams, K., Cho, Y., Parkinson, C., Qiu, Y. and Song, K. (2000) Dynamic evolution of plant mitochondrial genomes: Mobile genes and introns and highly variable mutation rates. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 6960-6966.

Pandino, G., Lombardo, S. and Mauromicale, G., (2011) Chemical and Morphological Characteristics of New Clones and Commercial Varieties of Globe Artichoke (*Cynara cardunculus* var. *scolymus*). *Plant Foods for Human Nutrition*, **66**, 291-297.

Parchman, T., Geist, K., Grahnen, J., Benkman, C. and Buerkle, C. (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* **11**, 180.

Pavy, N., Parsons, L., Paule, C., MacKay, J. and Bousquet, J. (2006) Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. *BMC Genomics*, **7**, 174.

Perez-Rodriguez, P., Riano-Pachon, D., Correa, L., Rensing, S., Kersten, B. and Mueller-Roeber, B. (2010) PInTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Research*, **38**, D822-D827.

Portis, E., Acquadro, A., Longo, A., Mauro, R., Mauromicale, G. and Lanteri, S. (2010) Potentiality of *Cynara cardunculus* L. as energy crop. *Journal of Biotechnology*, S165-S166.

Portis, E., Mauromicale, G., Mauro, R., Acquadro, A., Scaglione, D. and Lanteri, S. (2009a) Construction of a reference molecular linkage map of globe artichoke (*Cynara cardunculus* var. *scolymus*). *Theoretical and Applied Genetics*, 59-70.

Portis, E., Acquadro, A., Scaglione, D., Mauromicale, G., Mauro, R., Taylor, C.A., Knapp, S.J. and Lanteri, S. (2009b) Construction of an SSR-based linkage map for *Cynara cardunculus*. *8th Plant Genomics European Meeting*.

Portis, E., Barchi, L., Acquadro, A., Macua, J. and Lanteri, S. (2005a) Genetic diversity assessment in cultivated cardoon by AFLP (amplified fragment length polymorphism) and microsatellite markers. *Plant Breeding*, **124**, 299-304.

Portis E., Acquadro A., Comino C., Mauromicale G., Saba E., Lanteri S. (2005b) Genetic structure of island populations of wild cardoon [*Cynara cardunculus* L. var. *sylvestris* (Lamk) Fiori] detected by AFLPs and SSRs. *Plant Science*, **169**, 199-210.

Ragoussis, J. (2009) Genotyping technologies for genetic research. *Annual Review of Genomics and Human Genetics*, 117-133.

Rogozin, I., Wolf, Y., Sorokin, A., Mirkin, B. and Koonin, E. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology*, 1512-1517.

Sanseverino, W., Roma, G., De Simone, M., Faino, L., Melito, S., Stupka, E., Frusciante, L. and Ercolano, M. (2010) PRGdb: a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Research* **38**, D814-D821.

Scaglione, D., Acquadro, A., Portis, E., Tirone, M., Knapp, S.J. and Lanteri, S. (2012) RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L. *BMC Genomics*, **13**: 3.

Scaglione, D., Acquadro, A., Portis, E., Taylor, C., Lanteri, S. and Knapp, S. (2009) Ontology and diversity of transcript-associated microsatellites mined from a globe artichoke EST database. *BMC Genomics*, **10**: 454.

Schutz, K., Kammerer, D., Carle, R. and Schieber, A. (2004) Identification and quantification of caffeoylquinic acids and flavonolds from artichoke (*Cynara scolymus* L.) heads, juice, and pomace by HPLC-DAD-ESI/MSn. *Journal Of Agricultural And Food Chemistry*, **52**, 4090-4096.

Schwab, R., Ossowski, S., Riester, M., Warthmann, N. and Weigel, D. (2006) Highly specific gene silencing by artificial microRNAs in *Arabidopsis*. *Plant Cell*, **18**, 1121-1133.

Scott, L., LaFoe, D. and Weil, C. (1996) Adjacent sequences influence DNA repair accompanying transposon excision in maize. *Genetics*, 237-246.

Shimoda, H., Ninomiya, K., Nishida, N., Yoshino, T., Morikawa, T., Matsuda, H. and

Yoshikawa, M. (2003) Anti-hyperlipidemic Sesquiterpenes and new sesquiterpene glycosides from the leaves of artichoke (*Cynara scolymus* L.): Structure requirement and mode of action. *Bioorganic & Medicinal Chemistry Letters*, 223-228.

Sonnante, G., Gatto, A., Morgese, A., Montemurro, F., Sarli, G., Blanco, E. and Pignone D. (2011) Genetic map of artichoke x wild cardoon: toward a consensus map for *Cynara cardunculus*. *Theoretical and Applied Genetics*, **123**, 1215-1229.

Timme, R., Kuehl, J., Boore, J. and Jansen, R. (2007) A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. *American Journal of Botany,* **94**, 302-312.

Trick, M., Long, Y., Meng, J. and Bancroft, I. (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnology Journal*, **7**, 334-346.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D., Cestaro, A., Pruss, D., Pindo, M., FitzGerald, L., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J., Perazzolli, M., Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Dematte, L., Mraz, A., Battilana, J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J., Sterck, L., Vandepoele, K., Grando, S., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S., Fontana, P., Gutin, A., Van de Peer, Y., Salamini, F. and Viola, R. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *Plos One*, **12**, e1326.

Wang, W., Barnaby, J., Tada, Y., Li, H., Tor, M., Caldelari, D., Lee, D., Fu, X. and Dong, X. (2011) Timing of plant immune responses by a central circadian regulator. *Nature,* **470**, 110-126.

Wang, Y., Li, P., Cao, X., Wang, X., Zhang, A. and Li, X. (2009) Identification and expression analysis of miRNAs from nitrogen-fixing soybean nodules. *Biochemical and Biophysical Research Communications*, **378**, 799-803.

Wasmuth, J.D. and Blaxter, M.L. (2004) Prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, **5**, 187.

Wessler, S., Bureau, T. and White, S. (1995) LTR-retrotransposons and MITES - important players in the evolution of plant genomes. *Current Opinion in Genetics & Development*, 814-821.

TABLES

#	Genotype	*C. cardunculus taxon*	Source*	Sequencing platform
1	“Romanesco C3”	*scolymus*	F	454 FLX
2	“Altilis 41”	*altilis*	F	454 FLX
3	“Creta 4”	*sylvestris*	F	454 FLX
4	“Romanesco Zorzi”	*scolymus*	C	GAIIx
5	“Violetto di Chioggia”	*scolymus*	C	GAIIx
6	“Violetto Pugliese”	*scolymus*	C	GAIIx
7	“Spinoso Sardo”	*scolymus*	G	GAIIx
8	“Imperial Star”	*scolymus*	C	GAIIx
9	“Blanco de Peralta”	*altilis*	G	GAIIx
10	“Gobbo di Nizza”	*altilis*	G	GAIIx
11	“Sylvestris_LOT23”	*sylvestris*	G	GAIIx

Table 1. The *C. cardunculus* germplasm panel used for sequencing. *F: mapping parent; C: obtained from seed retailers/developers; G: germplasm held at DIVAPRA, Plant genetics and breeding sector.

	"Romanesco C3"	"Altilis 41"	"Creta 4"	Total
Sequencing results				
Total Mb	184 Mb	246 Mb	263Mb	693Mb
Reads #	435,375	610,622	696,300	1,742,297
Average length	421.5	402.8	377.0	392.0
Mode	502	481	490	491
Phred20 bases	167 Mb (90%)	216 Mb (88%)	227Mb (87%)	610(88%)
GC%	43.00%	41.30%	43.6%	42.1%
Ns/10K	47.48	56.72	62.02	56.53
Assembly results				
# of contigs	37,622	40,130	42,837	38,726*
N50	834	761	772	951bp*
Mean length	723.8	699.9	688.5	844.3bp*

Table 2. 454-derived sequencing and assembly. The output statistics were calculated following the removal of contaminating and adaptor sequences. "Romanesco C3" is a globe artichoke (var. *scolymus*), "Altilis 41" a cultivated cardoon (var. *altilis*) and "Creta 4" a wild cardoon (var. *sylvestris*). *Results obtained by merging the three independent assemblies (see Figure 1).

		Before quality clipping		After quality clipping		
	Num. of raw reads	First mates (Mbp)	Paired mates (Mbp)	First mates (Mbp)	Paired mates (Mbp)	Avg. data loss
"Romanesco Zorzi"	6.6M	496	444	458	408	7.86%
"Violetto di Chioggia"	6.6M	498	446	470	420	5.75%
"Violetto Pugliese"	5.2M	394	357	367	331	6.92%
"Spinoso Sardo"	6.7M	508	456	474	424	6.87%
"Imperial Star"	6.4M	489	442	459	415	6.09%
"Blanco de Peralta"	4.8M	360	325	340	305	5.91%
"Gobbo di Nizza"	5.6M	418	373	380	341	8.86%
"Sylvestris_LOT23"	4.6M	342	307	322	287	6.22%
Total	46.4M	3,505	3,151	3,271	2,931	
PE sum		**6,657**		**6,202**		**6.82%**

Table 3. GAIIx (Illumina)-derived sequencing. A total of 46.6M raw reads were generated in two GAIIx lanes and 6.7Gbp were retained after adaptor and contaminating sequence was removed. The windowed quality clipping routine produced a final dataset of 6.2Gbp. A higher number of bases was obtained for single ends, because 84 sequencing cycles were used instead of the 76 used for the paired ends.

TF family	count	avg. identity	TF family	count	avg. identity
bHLH	110	56.0%	C2C2-GATA	15	59.7%
MYB	109	65.9%	Jumonji	15	66.3%
WRKY	88	59.6%	SET	14	55.7%
HB	73	66.4%	SBP	14	58.1%
C2H2	71	66.6%	TCP	13	60.3%
NAC	65	62.5%	FAR1	12	66.9%
AP2-EREBP	64	53.8%	TUB	11	51.7%
bZIP	58	56.4%	ABI3VP1	11	69.3%
Orphans	58	68.0%	HMG	11	73.5%
MYB-related	56	58.9%	Sigma70-like	10	56.6%
C3H	56	59.7%	SNF2	10	63.0%
MADS	40	63.3%	ARR-B	9	71.3%
G2-like	37	62.6%	GeBP	8	52.9%
PHD	37	67.4%	Pseudo ARR-B	8	60.3%
AUX/IAA	30	62.6%	TRAF	8	57.1%
CCAAT	30	72.1%	Alfin-like	7	70.4%
Trihelix	28	60.6%	C2C2-YABBY	7	64.7%
HSF	28	64.5%	RWP-RK	7	49.3%
ARF	27	64.9%	SOH1	7	86.0%
GRAS	26	65.7%	BES1	6	58.8%
C2C2-Dof	23	55.0%	Others	63	
CAMTA	18	64.3%	**TOTAL**	**1398**	

Table 4. Categorization of transcription factors (TF). These activities were identified using a set of relevant GO-terms, and confirmed by a Blastx search of the Plant Transcriptional Factors Database (PlnTFDB). The mean identity was calculated on the basis of the best hits from PlnTFDB.

Automated annotation pipeline

Enzyme name	Enzyme code	Ortology code	Best hit	Reference species	Reference Locus	# of annotated contigs
peroxidase	1.11.1.7	ko:K00430	Cynara_c_joined_rep_c6889	*V. vinifera*	LOC100241814	104
beta-glucosidase	3.2.1.21	ko:K01188	Cynara_c_joined_rep_c13098	*V. vinifera*	LOC100232900	45
caffeoyl-CoA O-methyltransferase	2.1.1.104	ko:K00588	Cynara_c_joined_rep_c8703	*V. vinifera*	LOC100243978	20
4-coumarate:CoA ligase	6.2.1.12	ko:K01904	Cynara_c_joined_c20822	*V. vinifera*	LOC100245991	13
caffeic acid 3-O-methyltransferase	2.1.1.68	ko:K13066	Cynara_c_joined_rep_c11395	*R. communis*	RCOM_1053300	13
cinnamyl-alcohol dehydrogenase	1.1.1.195	ko:K00083	Cynara_c_joined_rep_c11798	*V. vinifera*	LOC100245522	12
cinnamoyl-CoA reductase	1.2.1.44	ko:K09753	Cynara_c_joined_rep_c3845	*V. vinifera*	LOC100251623	11
coniferin beta-glucosidase	3.2.1.126	-	Cynara_c_joined_rep_c24689	*R. communis*	RCOM_0905360	4
trans-cinnamate 4-monooxygenase	1.14.13.11	ko:K00487	Cynara_c_joinedrep_1926	*A. thaliana*	AT2G30490	3
coniferyl-aldehyde dehydrogenase	1.2.1.68	ko:K12355	Cynara_c_joined_rep_c11395	*R. communis*	RCOM_1053300	3
shikimate O-hydroxycinnamoyltransferase	2.3.1.133 (HCT)	ko:K13065	Cynara_c_joined_rep_c15600	*V. vinifera*	LOC100265530	3
oniferyl-alcohol glucosyltransferase	2.4.1.111	ko:K12356	Cynara_c_joined_c11359	*V. vinifera*	LOC100265092	2
quinate O-hydroxycinnamoyltransferase	2.3.1.99 (HQT)	-	Sylvestris_c19732	*P. trichocarpa*	XP_002332068.1	2
sinapate 1-glucosyltransferase	2.4.1.120	ko:K13068	Sylvestris_rep_c16035	*A. thaliana*	AT3G21560	1
2-coumarate O-beta-glucosyltransferase	2.4.1.114	-	Cynara_c_joined_rep_c15950	*A. thaliana*	AT1G15950	1

Putative orthologs

Enzyme name	Enzyme code	Ortology code	Best hit	Reference species	Reference Locus	E-value	Similarity
p-coumarate 3-hydroxylase	1.14.13.- (C3H)	ko:K03184	Sylvestris_c19616	*V. vinifera*	LOC100263633	1E-131	94%
ferulate-5-hydroxylase	1.14.-.- (F5H)	ko:K11785	Cynara_c_joined_rep_c47	*V. vinifera*	LOC100267863	1E-97	60%
p-coumarate 3-hydroxylase	1.14.13.36 (C3H)	Ko:K09754	Sylvestris_c19616	*A. thaliana*	AT2G40890	1E-126	92%
sinapoylglucose-malate O-sinapoyltransferase	2.3.1.92	ko:K09757	Cynara_c_joined_rep_c3990	*A. thaliana*	AT2G22990	3E-79	63%
phenylalanine ammonia-lyase	4.3.1.24	ko:K10775	Cynara_c_joined_rep_c12350	*V. vinifera*	LOC100233012	0	92%
phenylalanine/tyrosine ammonia-lyase	4.3.1.25	ko:K13064	Cynara_c_joined_rep_c8636	*O. sativa*	4330034	1E-165	81%

Table 5. List of the enzymes involved in the phenylpropanoid synthesis pathway upstream of the CQAs and dCQAs. The upper section shows enzymes identified by automatic annotation, and the lower section lists the best candidate on the basis of homology searches.

miRNA family	# of targets	miRNA family	# of targets
miR2275	29	miR2638	8
miR395	26	miR838	8
miR2109	22	miR858	8
miR2673	21	miR865	8
miR482	20	miR1514	7
miR156	16	miR1861	7
miR169	16	miR2643	7
miR396	16	miR4221	7
miR821	14	miR837	7
miR172	13	miR1134	6
miR834	13	miR1530	6
miR854	13	miR1866	6
miR164	12	miR2595	6
miR1863	12	miR2665	6
miR399	12	miR2676	6
miR529	12	miR397	6
miR774	12	miR815	6
miR1520	11	miR1312	5
miR159	11	miR1428	5
miR1886	11	miR1507	5
miR2633	11	miR1854	5
miR444	11	miR1871	5
miR773	11	miR2087	5
miR171	10	miR2642	5
miR2118	10	miR2655	5
miR2629	10	miR2661	5
miR783	10	miR415	5
miR166	9	miR4350	5
miR393	9	miR4379	5
miR413	9	miR475	5
miR530	9	miR835	5
miR1439	8	miR847	5
miR1510	8	Others	408
miR167	8		

Table 6. Abundance of putative miRNA annealing sites in the *C. cardunculus* transcriptome. miRNA families occurring fewer than five times were incorporated into the category labelled "other". A complete report is provided in Dataset S8.

GO term	name	FDR	p-Value	#Test	#Ref
GO:0012501	programmed cell death	3.65301E-8	1.61503E-11	40	241
GO:0006915	apoptosis	3.65301E-8	4.19075E-11	39	159
GO:0004888	transmembrane receptor activity	3.65301E-8	4.5962E-11	33	247
GO:0008219	cell death	3.65301E-8	6.3646E-11	42	310
GO:0016265	death	3.65301E-8	6.3646E-11	42	310
GO:0015979	photosynthesis	1.53229E-7	5.6573E-10	22	133
GO:0005788	endoplasmic reticulum lumen	1.53229E-7	6.71023E-10	13	36
GO:0045087	innate immune response	5.93924E-6	2.966E-8	32	341
GO:0043178	alcohol binding	6.91042E-6	3.94759E-8	7	7
GO:0000822	inositol hexakisphosphate binding	6.91042E-6	3.94759E-8	7	7
GO:0006955	immune response	9.26856E-6	6.98081E-8	32	355
GO:0002376	immune system process	2.2297E-5	1.81835E-7	34	411
GO:0006952	defense response	3.16498E-5	2.73057E-7	59	974
GO:0000304	response to singlet oxygen	4.29656E-4	3.07325E-6	6	10
GO:0008097	5S rRNA binding	0.00204601	1.8459E-5	6	15
GO:0048856	anatomical structure development	0.00247686	2.13454E-5	43	730
GO:0003700	transcription factor activity	0.00333179	2.507E-5	59	1137
GO:0022414	reproductive process	0.00527285	4.53476E-5	65	1321

Table 7. The over-representation of GO-terms among the putative miRNA target transcripts. Fischer's exact test (FDR <0.01) was used to assess statistical significance.

	Total	In CDS		Synonymous		Non synonymous		In UTR		Homozygous fixed	
Private SNPs (per accession)											
var. *scolymus:*											
"Romanesco C3"	628	*557*	88,7%	*273*	49,0%	*284*	51,0%	*71*	11,3%	*14*	2,2%
"Romanesco Zorzi"	424	*322*	75,9%	*148*	46,0%	*174*	54,0%	*102*	24,1%	*30*	7,1%
"Violetto di Chioggia"	544	*414*	76,1%	*237*	57,2%	*177*	42,8%	*130*	23,9%	*32*	5,9%
"Violetto Pugliese"	496	*385*	77,6%	*216*	56,1%	*169*	43,9%	*111*	22,4%	*76*	15,3%
"Spinoso Sardo"	759	*542*	71,4%	*314*	57,9%	*228*	42,1%	*217*	28,6%	*158*	20,8%
"Imperial Star"	490	*385*	78,6%	*213*	55,3%	*172*	44,7%	*105*	21,4%	*212*	43,3%
var. *altilis:*											
"Altilis 41"	723	*638*	88,2%	*314*	49,2%	*324*	50,8%	*85*	11,8%	*118*	16,3%
"Gobbo di Nizza"	729	*608*	83,4%	*338*	55,6%	*270*	44,4%	*121*	16,6%	*78*	10,7%
"Blanco de Peralta"	719	*549*	76,4%	*323*	58,8%	*226*	41,2%	*170*	23,6%	*138*	19,2%
var. *sylvestris:*											
"Creta 4"	3572	*3054*	85,5%	*1838*	60,2%	*1216*	39,8%	*518*	14,5%	*394*	11,0%
"Lot 23"	4092	*3033*	74,1%	*1859*	61,3%	*1174*	38,7%	*1059*	25,9%	*759*	18,5%
Private SNPs (per varietas)											
var. *scolymus*	869	*681*	78,4%	*393*	57,7%	*288*	42,3%	*188*	21,63%	*78*	8,98%
var. *altilis*	463	*356*	76,9%	*218*	61,2%	*138*	38,8%	*107*	23,11%	*69*	14,90%
var. *sylvestris*	3123	*2459*	78,7%	*1559*	63,4%	*900*	36,6%	*664*	21,26%	*336*	10,76%

Table 8. Categorization of private SNPs. Private SNPs per *taxa* were considered for shared allelic variants within one *taxon* and missing in any other. Homozygous fixed SNPs consider both coding/non-coding regions.

First assembly phase	Second assembly phase
COMMON_SETTINGS	COMMON_SETTINGS
-GE:NOT=8	-GE:NOT=8
-AS:sep=yes:ugpf=no	-AS:sep=yes:ugpf=no:sd=yes
-SK:not=8:pr=50:mnr=no	-SK:not=8:pr=50:mnr=no
-CO:mr=yes:asir=yes	-CO:mr=yes:asir=yes
-OUT:ora=yes:org=no	-OUT:ora=yes:org=no
-SB:lsd=yes	-SB:lsd=yes
-CL:ascdc=yes	-CL:pec=yes
SANGER_SETTINGS*	SANGER_SETTINGS
-LR:wqf=no	-CL:cpat=no:pvlc=yes:emlc=yes:emrc=yes
-AS:epoq=no:bdq=20	-OUT:sssip=yes
-CL:cpat=no	-AL:mo=30:ms=30:mrs=90:egp=no
-OUT:sssip=yes	-CO:fnicpst=yes
-AL:mo=40:ms=40:mrs=93:egp=no	-ED:ace=no
-CO:fnicpst=yes	-DP:ure=yes.
-ED:ace=no	
454_SETTINGS	
-OUT:sssip=yes	
-AL:mo=40:ms=40:mrs=93:egp=no	
-CL:cpat=no:qc=yes:qcmq=15:qcwl=20	
-CO:fnicpst=yes:rodirs=10	
-DP:ure=yes	
-ED:ace=no.	

Table 9. Parameters configuration of the two assembly phases. For missing parameters, we employed the default values in MIRA v. 3.2.0. *SANGER_SETTINGS in the first assembly phase have been only applied to the hybrid assembly of "Romanesco C3" 454 reads with Sanger-ESTs from CGP database.

FIGURE CAPTIONS

Figure 1. Sequence analysis pipeline. The *de novo* assembly, read mapping, ORF prediction, intron location, SNP calls and other ancillary data have been assembled in a relational database (available upon request).

Figure 2. The distribution of contig length, following the two step MIRA assembly process.

Figure 3. Simulated assembly curves for each of the three subspecies-specific 454-derived datasets. Solid lines represent the permutation-based growth curve; dotted lines report the projection of the best fitting logarithmic functions.

Figure 4. Relict TE sequence in the transcriptome. The bar chart shows the number of inferred TE signatures sorted by the frequency of their occurrence, and the pie chart depicts the overall distribution of relict sequence length in the 5'-UTR, the CDS and the 3'-UTR.

Figure 5. Combined calling of SNPs. The number of calls based solely on the 454-derived reads is shown in blue, and the combined SNP discovery based on both the 454- and the Illumina-based sequence in red. "Exclusively homozygous" and "exclusively heterozygous" refer to allelic variants present in only one of the three 454-sequenced libraries.

Figure 6. The distribution of minor allele frequencies across the eleven accession germplasm panel.

Figure 7. The allelic state at SNP loci. Bars indicate the total number of SNP loci in the homozygous or heterozygous state (or missing) for each accession. Each bar's colour identifies the *C. cardunculus* varietas (green = *sylvestris*, orange = *altilis*, yellow = *scolymus*). White dots identify the three accessions sequenced using 454 technology.

ELECTRONIC SUPPLEMENTARY MATERIAL

- Dataset S1
- File format: .fasta (.gz archive)
- Title of data: Transcript contig set

- Dataset S2
- File format: .xls
- Title of data: Mitochondrial and chloroplast genes categorization

- Dataset S3
- File format: .xls
- Title of data: GO-based annotation of the transcript set

- Dataset S4
- File format: .txt (tab-delimited)
- Title of data: Classification of transcription factors

- Dataset S5
- File format: .tar archive (.xls + .gif images)
- Title of data: Enzymatic activities mapped on KEGG's methabolic pathway.

- Dataset S6
- File format: .xls
- Title of data: Identification of putative R-genes

- Dataset S7
- File format: .xls
- Title of data: Identification of transposable elements relicts.

- Dataset S8
- File format: .xls
- Title of data: Prediction of putative miRNA target

- Dataset S9
- File format: .gz archive (.txt, tab-delimited)
- Title of data: SNP information table

Figure 1

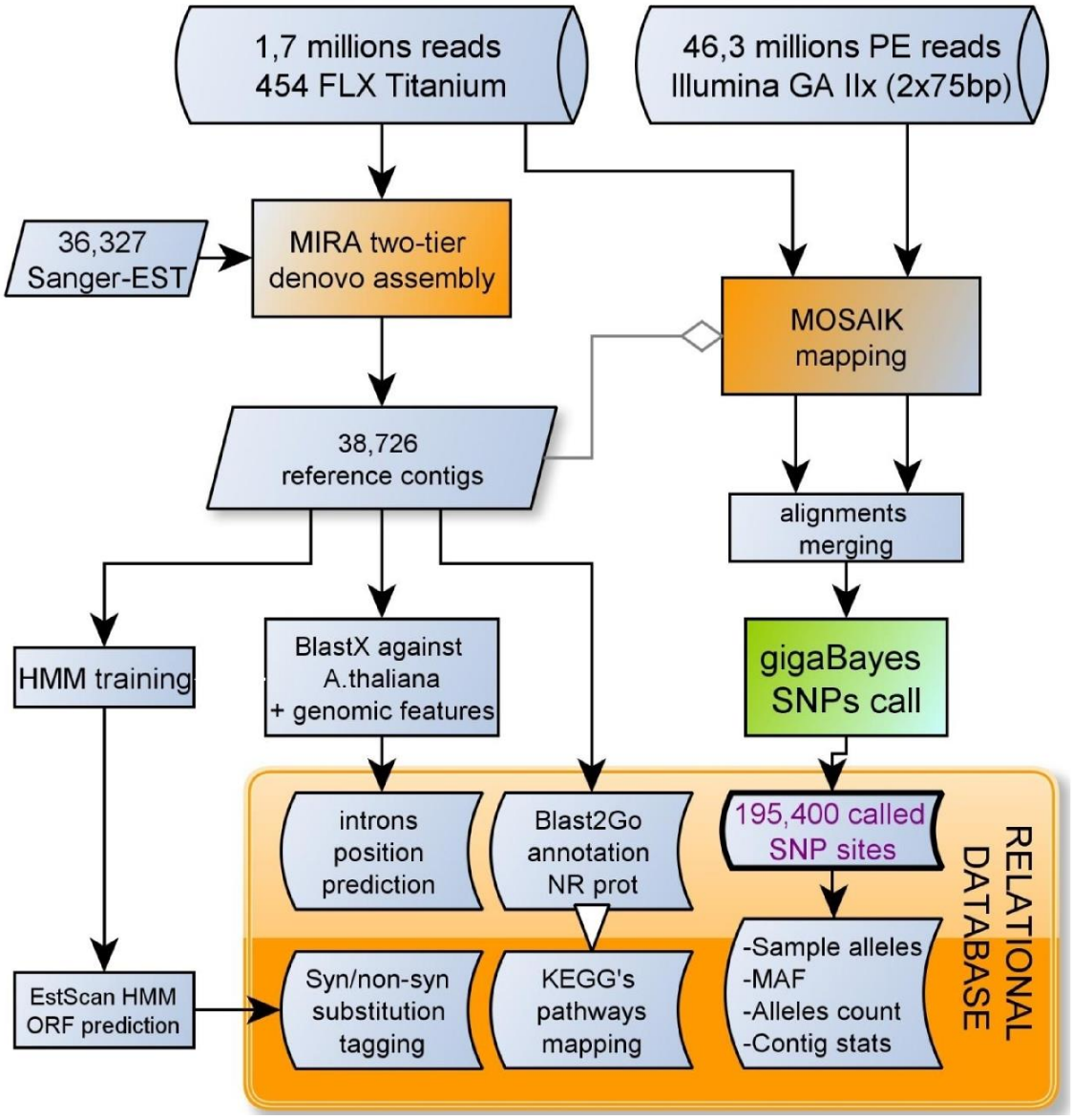

1,7 millions reads
454 FLX Titanium
46,3 millions PE reads
Illumina GA IIx (2x75bp)
36,327
Sanger-EST
MIRA two-tier
denovo assembly
MOSAIK
mapping
38,726
reference contigs
alignments
merging
HMM training
BlastX against
A.thaliana
+ genomic features
gigaBayes
SNPs call
introns
position
prediction
Blast2Go
annotation
NR prot
195,400 called
SNP sites
RELATIONAL
DATABASE
EstScan HMM
ORF prediction
Syn/non-syn
substitution
tagging
KEGG's
pathways
mapping
-Sample alleles
-MAF
-Alleles count
-Contig stats

Figure 2

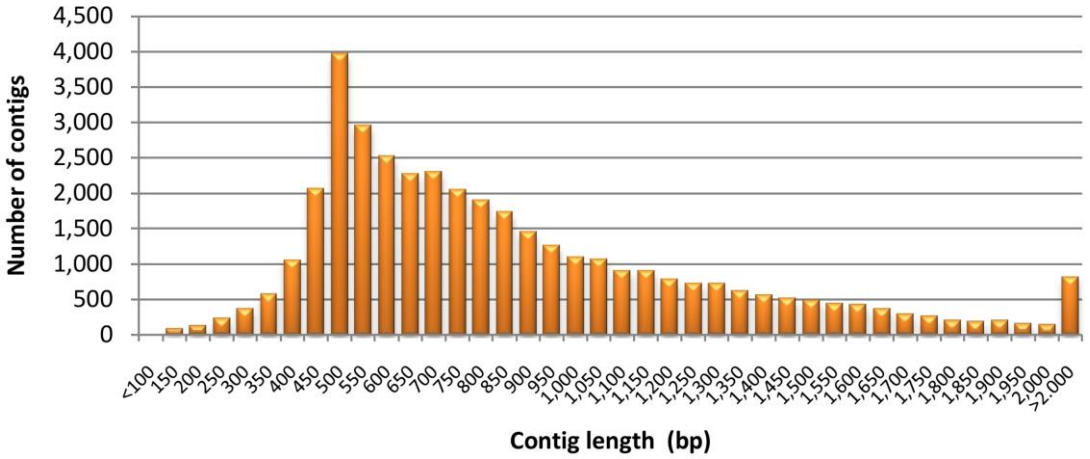

Number of contigs
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
<100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1,000
1,050
1,100
1,150
1,200
1,250
1,300
1,350
1,400
1,450
1,500
1,550
1,600
1,650
1,700
1,750
1,800
1,850
1,900
1,950
2,000
>2,000
Contig length (bp)

Figure 3

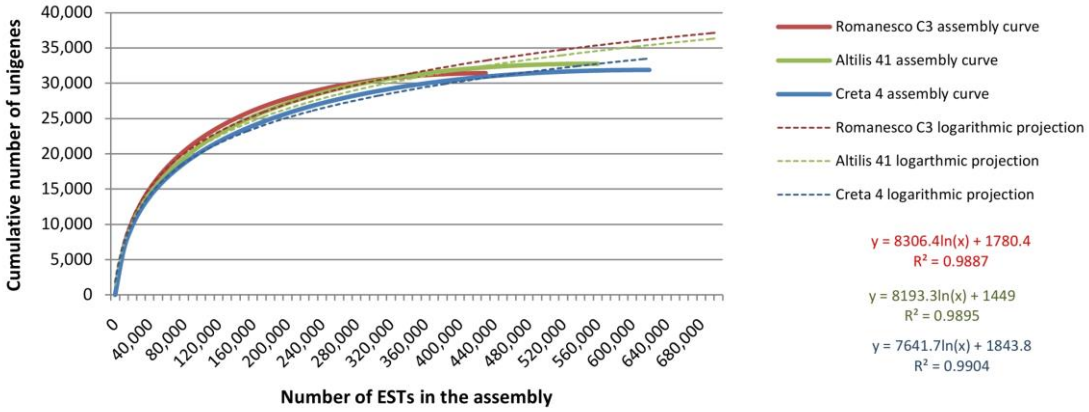

Cumulative number of unigenes
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
0
40,000
80,000
120,000
160,000
200,000
240,000
280,000
320,000
360,000
400,000
440,000
480,000
520,000
560,000
600,000
640,000
680,000
Number of ESTs in the assembly
Romanesco C3 assembly curve
Altilis 41 assembly curve
Creta 4 assembly curve
Romanesco C3 logarithmic projection
Altilis 41 logarthmic projection
Creta 4 logarithmic projection
y = 8306.4ln(x) + 1780.4
R² = 0.9887
y = 8193.3ln(x) + 1449
R² = 0.9895
y = 7641.7ln(x) + 1843.8
R² = 0.9904

Figure 4

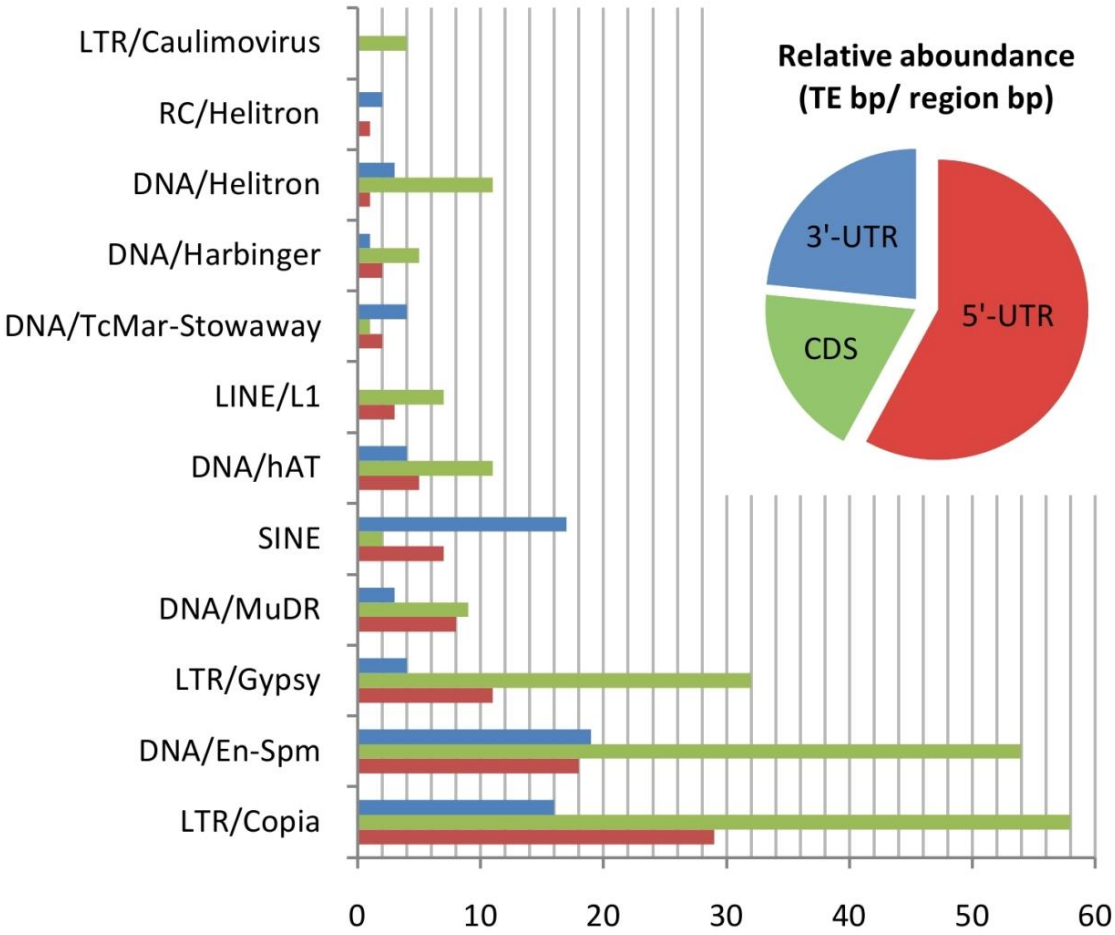

LTR/Caulimovirus
RC/Helitron
DNA/Helitron
DNA/Harbinger
DNA/TcMar-Stowaway
LINE/L1
DNA/hAT
SINE
DNA/MuDR
LTR/Gypsy
DNA/En-Spm
LTR/Copia
0
10
20
30
40
50
60
Relative aboundance
(TE bp/ region bp)
3'-UTR
5'-UTR
CDS

Figure 5

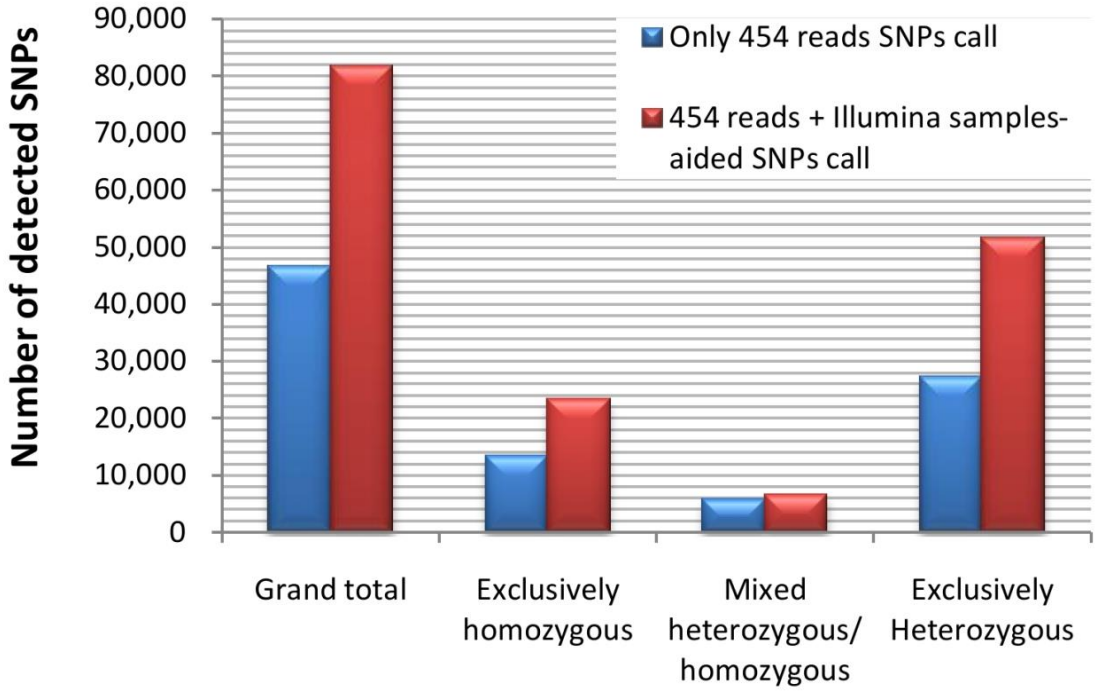

Number of detected SNPs
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Only 454 reads SNPs call
454 reads + Illumina samples-aided SNPs call
Grand total
Exclusively homozygous
Mixed heterozygous/ homozygous
Exclusively Heterozygous

Figure 6

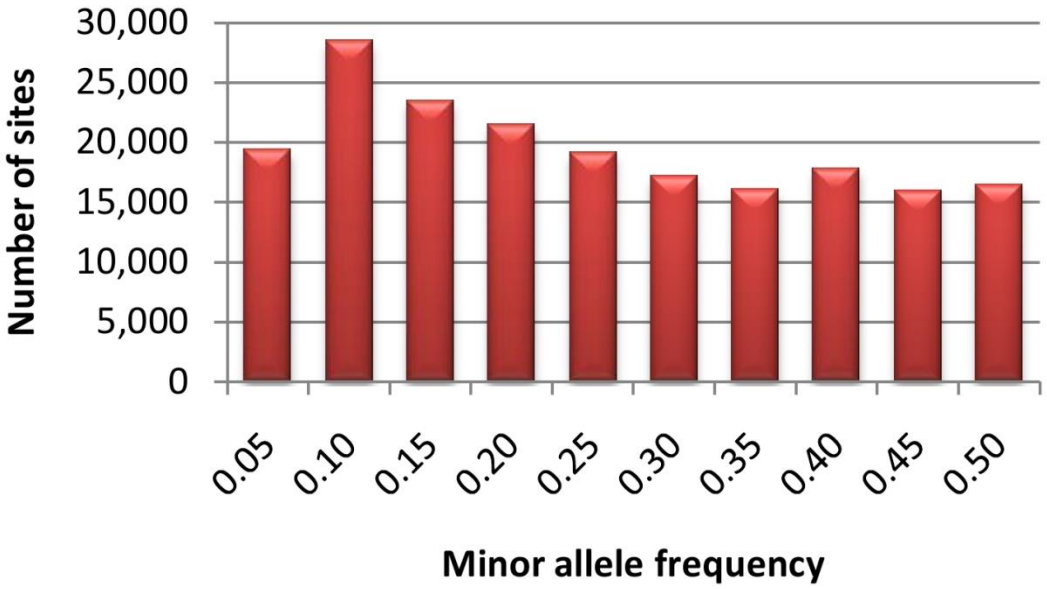
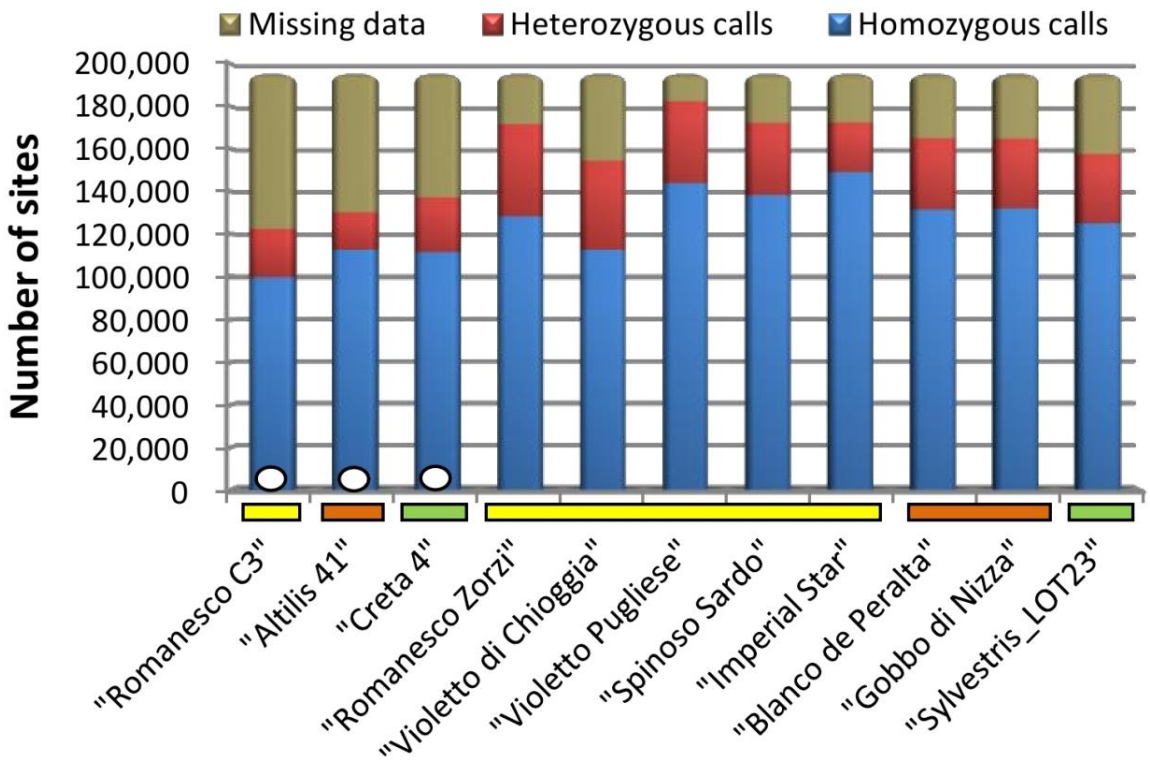

30,000
25,000
20,000
15,000
10,000
5,000
0
Number of sites
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
Minor allele frequency

Figure 7


Missing data
Heterozygous calls
Homozygous calls
Number of sites
200,000
180,000
160,000
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
"Romanesco C3"
"Altilis 41"
"Creta 4"
"Romanesco Zorzi"
"Violetto di Chioggia"
"Violetto Pugliese"
"Spinoso Sardo"
"Imperial Star"
"Blanco de Peralta"
"Gobbo di Nizza"
"Sylvestris_LOT23"