Microsatellites mined in Globe Artichoke EST database: linkage analysis and relation to gene function

Scaglione D[1], Acquadro A[1], Portis E[1], Lanteri S[1], Taylor CA[2], Knapp SJ[2]

Motivation

Italy is the world leading producer of the globe artichoke (Cynara cardunculus var. scolymus). Despite its economical relevance, the knowledge of its genomic organization is limited, thus hampering the application of marker-assisted breeding programs.. In order to fill in this gap, the development of new molecular markers is required to make possible the construction of dense genetic maps suitable for the identification of QTLs (Quantitative Traits Loci). Microsatellite markers (SSRs, simple sequence repeats) can be easily mined by analysing single-pass sequence data, and their map positioning might serve as primary scaffold for future genome sequencing projects. A Cynara cardunculus EST database (36321 ESTs released in NCBI by the Compositae Genome Project - CGP) was assembled and mined for the identification of SSR which can be employed as putative functional marker loci to easily tag corresponding functional genes. Furthermore, by a Gene Ontology analysis, we highlighted which gene categories preferentially contain microsatellites.

Methods

A customised bioinformatic pipeline was built up. The 36321 ESTs were cleaned by the SeqClean script querying the UniVec database and assembled using the TGICL script The resulting sequences were annotated by a batch BlastX process against the Arabidopsis thaliana proteins database (TAIR8), considering as a threshold E-value 10e-29. Unigenes were analysed for the presence of microsatellite using SSR Identification Tool (SSRIT) perl script, adopting the following parameters: 5 repeats for dinucleotide, 4 for tri-, tetra-, and pentanucleotide and 3 for esanucleotide. Flanking primers were designed by means of BatchPrimer3 web tool. Three hundred primer pairs were selected for a first screening using a 28-genotypes panel and subsequently mapped on a C. cardunculus genetic map obtained by the cross of a genotype of globe artichoke ("Romanesco C3") with one of cultivated cardoon (C. cardunculus var. altilis, "A41"). Parsing the BlastX output, together with the data obtained by an ORF (Open Reading Frame) predictor, we estimated the position of SSRs along the transcripts. The Arabidopsis-based annotation allowed the categorization of each

[1] DIVAPRA Plant Genetics and Breeding, University of Turin, via L. da Vinci 44, 10095 Grugliasco (Turin), Italy [2] Center for Applied Genetic Technologies, University of Georgia, 111 Riverbend Rd., 30605 Athens, Georgia (U.S.A.)

unigene at different hierarchical levels of the Gene Ontology (GO). Combining the afore-mentioned data with a Fisher's exact test we were able to identify specific gene categories in which specific SSR were highly represented, with regard to motifs (di- to exa-nucleotidic) and positions (CDS, 5'- and 3'- UTRs).

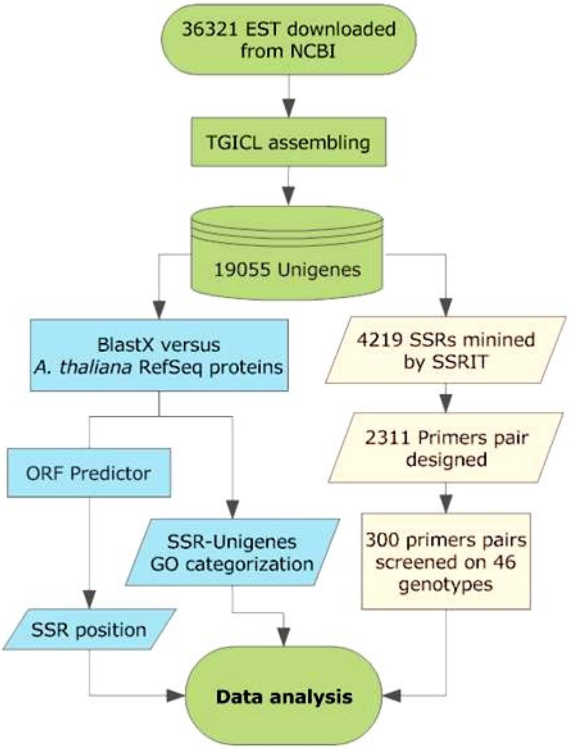
Results

A total of 19055 unigenes was generated by the assembly process, while the SSRIT script harvested 4219 microsatellite in 3308 unigenes. Sufficient flanking sequences for primers design were present for 2311 SSRs in 1974 unigenes. A total of 238 primer pairs (out of the 300 tested) produced clear PCR amplicons, of which 236 were polymorphic; the estimated PIC (polymorphic information content) values ranged from 0.035 to 0.891, with an average of 0.660. Polymorphic markers segregating in the "Romanesco C3" (globe artichoke) x "A41" (cultivated cardoon) F1 progeny were mapped, leading to the construction of a SSR-based consensus map. Each parental map appreciably increased its coverage, merging to a number of linkage groups (17) equal to the haploid chromosomal number of C. cardunculus (n=17). By analyzing the GO terms linked to specific microsatellite motifs, and their relative position, significant enrichments in the dataset were discovered. The majority of them belong to the 'nucleic acids binding' as well as 'metabolism and gene expression' GO terms, with regard to ATC/GAT in the coding regions (CDS) and AG/CT repeats in the 5'-UTR. On the other hand, AG/CT motif in CDS seemed to be involved in the 'response to stress' and 'DNA damage'. These data highlighted a close relation between specific microsatellites and gene function, thereby confirming their role as cis-regulatory elements.

Contact e-mail

alberto.acquadro@unito.it

Image


36321 EST downloaded from NCBI
TGICL assembling
19055 Unigenes
BlastX versus *A. thaliana* RefSeq proteins
4219 SSRs minined by SSRIT
ORF Predictor
2311 Primers pair designed
SSR-Unigenes GO categorization
300 primers pairs screened on 46 genotypes
SSR position
Data analysis