Torales et al. BMC Genomics 2013, 14:705
http://www.biomedcentral.com/1471-2164/14/705

BMC Genomics

RESEARCH ARTICLE Open Access

De novo assembly and characterization of leaf transcriptome for the development of functional molecular markers of the extremophile multipurpose tree species *Prosopis alba*

Susana L Torales[1*†], Máximo Rivarola[2,5†], María F Pomponio[1], Sergio Gonzalez[2], Cintia V Acuña[2], Paula Fernández[2,5], Diego L Lauenstein[3], Aníbal R Verga[3], H Esteban Hopp[2,4], Norma B Paniego[2,5] and Susana N Marcucci Poltri[2]

Abstract

Background: *Prosopis alba* (Fabaceae) is an important native tree adapted to arid and semiarid regions of north-western Argentina which is of great value as multipurpose species. Despite its importance, the genomic resources currently available for the entire *Prosopis* genus are still limited. Here we describe the development of a leaf transcriptome and the identification of new molecular markers that could support functional genetic studies in natural and domesticated populations of this genus.

Results: Next generation DNA pyrosequencing technology applied to *P. alba* transcripts produced a total of 1,103,231 raw reads with an average length of 421 bp. *De novo* assembling generated a set of 15,814 isotigs and 71,101 non-assembled sequences (singletons) with an average of 991 bp and 288 bp respectively. A total of 39,000 unique singletons were identified after clustering natural and artificial duplicates from pyrosequencing reads. Regarding the non-redundant sequences or unigenes, 22,095 out of 54,814 were successfully annotated with Gene Ontology terms. Moreover, simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs) were searched, resulting in 5,992 and 6,236 markers, respectively, throughout the genome. For the validation of the the predicted SSR markers, a subset of 87 SSRs selected through functional annotation evidence was successfully amplified from six DNA samples of seedlings. From this analysis, 11 of these 87 SSRs were identified as polymorphic. Additionally, another set of 123 nuclear polymorphic SSRs were determined in silico, of which 50% have the probability of being effectively polymorphic.

Conclusions: This study generated a successful global analysis of the *P. alba* leaf transcriptome after bioinformatic and wet laboratory validations of RNA-Seq data.
The limited set of molecular markers currently available will be significantly increased with the thousands of new markers that were identified in this study. This information will strongly contribute to genomics resources for *P. alba* functional analysis and genetics. Finally, it will also potentially contribute to the development of population-based genome studies in the genera.

Keywords: *Prosopis alba*, Fabaceae, Pyrosequencing, Transcriptome assembly, SSRs, SNPs, Functional annotation

* Correspondence: storales@cnia.inta.gov.ar
†Equal contributors
[1]Instituto de Recursos Biológicos, IRB, Instituto Nacional de Tecnología Agropecuaria (INTA Castelar), CC 25, Castelar B1712WAA, Argentina
Full list of author information is available at the end of the article

Background

The genus *Prosopis* Linnaeus emend Burkart, a member of the subfamily Mimosoideae within the family Fabaceae, comprises 44 species divided into 5 sections: *Prosopis, Anonychium, Strombocarpa, Monilicarpa and Algarobia* [1]. This genus is spread around the world in arid and semiarid regions, including North and South America, North and Central Africa, Near East and the Caribbean region. The main center of diversity for *Prosopis* genus is located in Argentina [1] with 27 species. Of these species, 21 belonging to *Algarobia* section [2], which are distributed in the phytogeographic provinces of Chaco, Monte, and Espinal [3]. They cover over one million square kilometers, which represents approximately one third of the total country area [4].

One of the most important features of this genus is its natural capacity to produce fertile interspecific hybrids [5-7]. This generates a syngameon complex integrated by species and subspecies which form a continuum [8]. This complex includes six taxonomic species that play a significant role in Argentina: *P. alba, P. hassleri, P. nigra, P. ruscifolia, P. chilensis and P. flexuosa.*

The members of this complex develop deep roots that give these plants several advantaged. For instance, these deep roots reduce competition for water with herbaceous species, improve water balance of the system, provide nutrients to the subsurface layers and in some cases make the plants fairly independent of rainfall [9].

Their fruits are pods and may contain large amounts of sugar and protein which offer optimal energy for its use as fodder and for human consumption. They can also be used for firewood and charcoal, as well as for other products (honey, pollen, gums, etc.) [10]. Also, “algarrobos” can be an alternative of livestock-forestry systems [11].

Within this group of “algarrobos”, *P. alba* known as "white algarrobo" displays the widest geographical distribution. This species grows in areas under average annual precipitations of 500 to 1200 mm, which are summer dominant, with extreme temperatures between 48°C maximal absolute, up to −10°C absolute minimum [12]. *P. alba* comprises groups with different morphological characteristics, such as variations in leaves and fruits, and inhabits different ecological zones [13]. Also, these morphological groups have distinct adaptation mechanisms to drought stress [14].

In Argentina, this native species is mainly used for saw timber (wood flooring and furniture) and the whole wood consumed comes from the native forests in “Parque Chaqueño” (Argentina) [15].

Besides, all “algarrobos”, including *P. alba*, may play a role on the recovery of degraded ecosystem [16]; hence re-population with these species generates favourable conditions for natural recovery of the whole ecosystem.

Up to date, few genomic data exist on *Prosopis* genus. A total of 1,467 expressed sequence tag (EST) from *Prosopis juliflora* has been deposited in the NCBI EST database [17]. There are also a limited number of molecular markers published: six microsatellites isolated from *Prosopis chilensis* [18] and 12 from a bulk sample of American algarrobos introduced to Australia [19].

To the best of our knowledge, here we report the largest contribution to sequence information of *Prosopis* spp. generated through new generation sequencing technologies. The results from the assembly and functional annotation of *P. alba* leaf transcriptome are presented, along with SSR and SNP motif mining. Nuclear and chloroplast SSR and SNP were discriminated in the analysis. Finally, this work generated a collection of 11 nuclear-SSR primer pairs validated for its application to diversity studies in *P. alba* and another set of 123 nuclear polymorphic SSRs determined in silico, of which 50% have the probability of being effectively polymorphic. The overall workflow of the project is represented in the Additional file 1.

Results and discussion

Transcriptome sequencing and assembly

An Rna-seq from a leaf bulk sample of three different individuals was performed using 454 GS FLX Titanium technology (Roche). The use of Rna-seq generated 464 Mb of sequence data from 1,103,231 reads with an average length of 421 bp, ranging from 21 to 692 bp. The sequences were subjected to filtering for adaptors, primer sequences and low-quality sequences. After this filtering, 39,711 reads were removed resulting in 1,063,520 high quality reads (96% of the first raw sequences). De novo transcriptome assembly was performed using Newbler Software v. 2.5 (Roche, IN, USA). With this assembly, 788,737 full length reads (74%) and 203,682 partial number of reads (19.3%) were assembled into 15,814 isotigs (equivalent to unique RNA transcripts) with an average length of 991 bp, ranging from 19 bp to 12,995 bp and N50 length of 1,124 bp (Figure 1A). In addition, the isotigs originated from the same contig-graph were grouped into 12,610 isogroups. These isogroups are equivalent to genomic loci (they contains the group of isotigs mapped to corresponding isogroups) with an average of 1.2 transcripts per locus, which potentially reflect multiple splicing variants. Most of the isogroups (86%) had only one isotig each. A total of 71,101 reads (6.6%) did not assemble into isotigs; therefore they were singletons. These singletons had an average length of 288 bp (Figure 1B) and, 32,991 of these singletons (85%) were shorter than 500 bp. All singletons were clustered using CD-HIT-454 algorithm to eliminate artificial duplicates. After clustering, we obtained 39,000 unique singletons longer than 200 bp. Then, all unigenes (whose

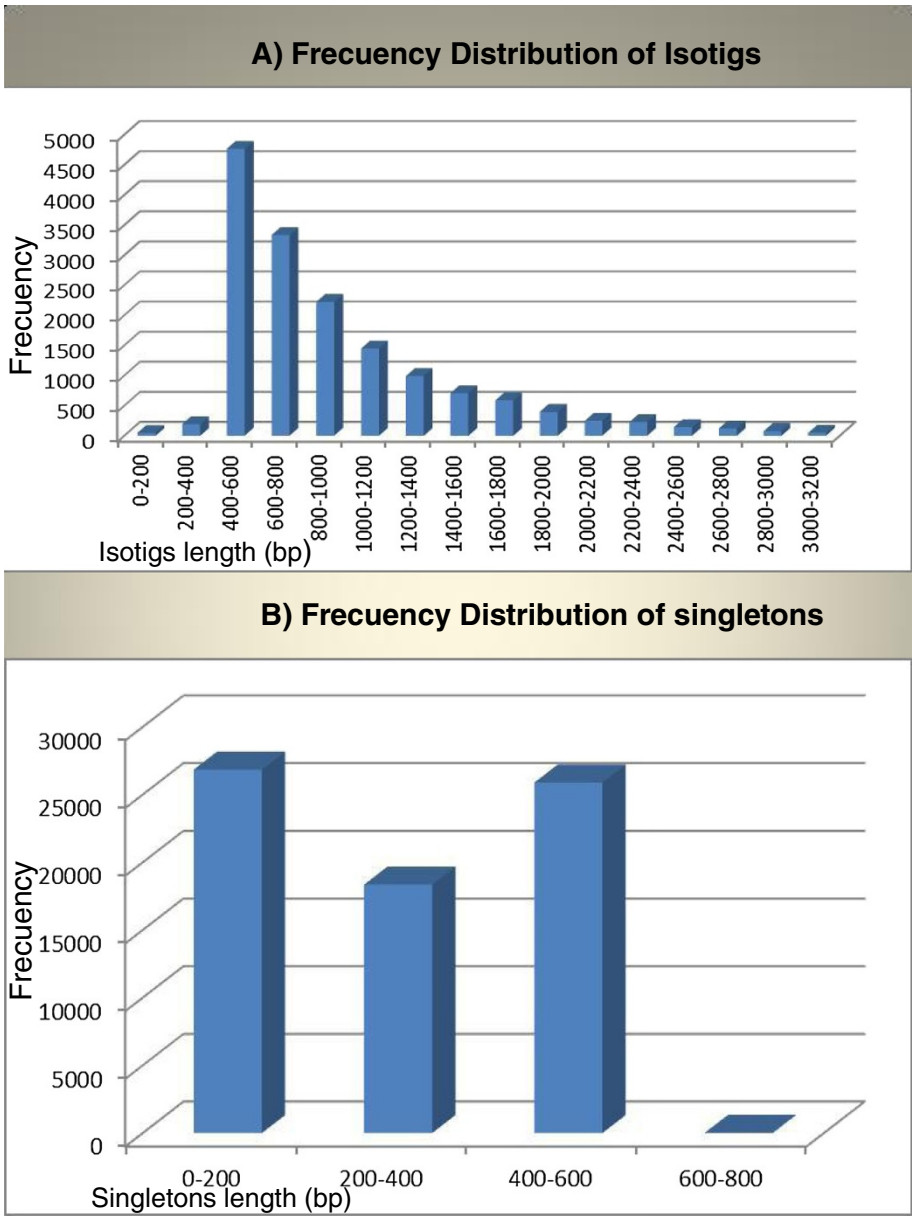

Figure 1 Frequency distribution of isotigs (A) and singletons (B) lengths. The histograms represent the number of isotig and singleton sequences in relation to their length.

length exceeded 200 bp (54,814 in total) were kept for further analysis (see Additional file 2, Table 1). The average length of *P. alba* isotigs was larger than those assembled in other native tree species *Nothofagus nervosa*, which had an average length of 765 bp [20]. The average length found in the present research was also larger than those of other non-model organisms ranging from 197 to 707 bp, [21-25]. This can be explained because of the use of the new 454 GS FLX Titanium (Roche) run that probably allowed us to obtain better and longer reads. Isotigs were integrated by different number of subcontigs generated in the assembly process. These isotigs were integrated by 2 to 25 contigs with an average of 11 contigs assembled into each isotig. These results are similar to the average numbers obtained in the 454 *N. nervosa* transcriptome analyses (mean=9) [20] and larger than other 454 transcriptome analyses (mean=2.1), such as those of *Pinus contorta* and microalgae *Dunaliella tertiolecta* [23,24].

Functional annotation

For assigning putative functions to the *P. alba´*s transcriptome, BLASTX searches [26] were performed aligning the assembled sequences to the 1,958,459

Table 1 Transcriptome functional annotation summary of *P. alba*

	Number of sequences		
	Isotigs (15,814)	**Singletons (39,000)**	**Combined (54,814)**
Viridiplantae DB			
Sequences with positive BLAST matches	14,664 (93%)	22,899 (59%)	37,563
Sequences annotated with Gene Ontology (GO) terms	10,107 (64%)	11,988 (31%)	22,095
Sequences without detectable BLAST matches	1,150 (7%)	16,101 (41%)	17,251
Sequences assigned to already known Enzyme Commission category	2,191 (14%)	2,347 (6%)	4,538

Number and percentages of 454 sequences in the assembled isotigs, singletons and total unigenes (combined) with significant matches against a Viridiplantae protein database.

protein sequences from a custom-made Viridiplantae database. A total of 14,664 isotigs and 22,899 singletons showed significant BLASTX matches (with an expectation value<1e-10) (Table 1). A higher percentage of isotigs (93%) than singletons (59%) had BLASTX hits, probably due to the good quality of isotigs (68% longer than 600 bp), short lengths of singletons and the high e-value cut-off applied. Previous reports on de novo transcriptome assemblies of eukaryotes described lower percentage of isotigs, ranging from 20 to 40%, such as those described for lanville fritillary butterfly, a coral larval, lodgepole pine and microalgae [21-24]. In total, 37,563 unique sequences had at least one BLAST hit in the search, while the 17,251 remaining sequences (i.e. 32%) (Table 1) were orphans. However, these orphan sequences may still be informative for identifying putative biological functions which may be considered as *P .alba* specific.

After the analyses of seven completely sequenced genomes, the average number of genes encoded in a plant nuclear genome was estimated in approximately 30 thousands [27]. Our annotated dataset with 12,610 isogroups, which can be used to estimate the number of gene locus, and 39,000 unique singletons most likely represent a good proportion of the *P. alba* gene catalogue.

BLASTX hits and top hits in terms of the total number of hits to all unigenes were mostly found with *Glycine max* (hits 80,668), *Vitis vinifera* and *Medicago truncatula* (Figure 2).

Gene Ontology (GO) term annotation and metabolic pathway mapping

Using a full local installation of Blast2GO [28] and the InterProScan suite [29], we retrieved gene ontology (GO) terms and enzyme commission numbers (EC) for the *P. alba* unigenes (Additional file 2).

From the Blast2GO and InterProScan programs, a total of 43,389 GO terms were assigned to 22,095 unigenes (including 10,107 annotated isotigs and 11,988 annotated singletons). Among all the GO terms extracted, 14,422 (33%) belong to the Biological Process class, 19,077 (44%) fit the Molecular Function class and 9,890 (23%) belong to the Cellular Component class.

The "Biological Process" (BP) GO category comprises different types of metabolic processes, which in turn are the most represented subcategories. Indeed, there are 10,140 sequences associated with metabolic processes (GO level 2) and 8,376 sequences related to cellular processes. These results may indicate that the analysed tissues were undergoing extensive metabolic activities [30]. These findings were expected, since the metabolic network in plants is very extensive compared to other organisms [31]. Within the sequences associated with biological processes, we found GO terms associated with primary and secondary metabolism. In this respect, the primary metabolites are known to be essential for plant survival and the secondary metabolites are described as playing roles in plant protection. Several genes involved in other important biological processes were also identified (Figure 3A). Among these genes, we can mention the ones associated with cellular processes, establishment of localization, biological regulation, biogenesis, developmental processes and signalling, to name a few. Another category worth mentioning is "response to stimulus" (BP category, level 2). We found 1,319 sequences in association with this category, which includes candidate genes involved in the resistance to biotic and abiotic stimulus.

In terms of molecular function, the top three GO terms found were related to the following categories: binding 13,007 (46%), catalytic activities, 11,708 (41%) and transporters 1,312 (5%) (Figure 3B).

A detailed analysis (level 2) at the cellular component category sorted all transcripts from *P. alba* into 5 groups. Of these groups, the most representative categories were: cell (9,577), organelle (4,100) and macromolecular complex component (1,699) (Figure 3C).

Of the 22,095 sequences annotated with GO terms, 4,538 were assigned with EC numbers (2,191 isotigs and 2,347 singletons) (Table 1).

Figure 4 displays the most represented enzymes in all sequences: transferase activity (37%), hydrolase activity (35%) and oxidoreductase activity (13%). The large

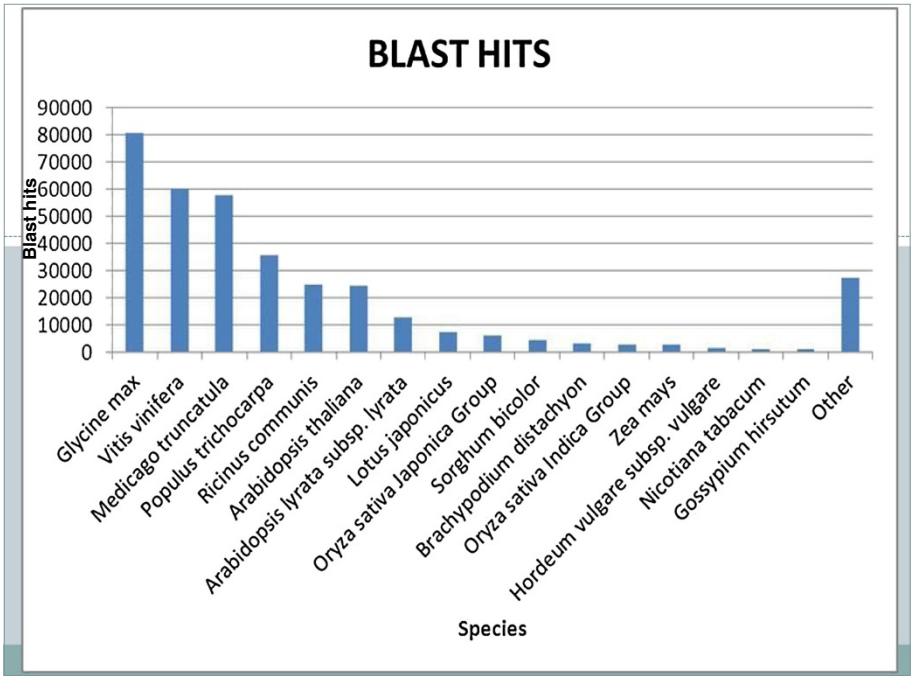

Figure 2 Hit species distribution of BLASTX matches of *P. alba* unigenes. Proportion of *P. alba* unigenes (isotigs and singletons) with similarity to sequences from Viridiplantae protein database.

number of annotated enzymes within these three groups suggests the presence of genes associated to pathways of secondary metabolite synthesis [30,32,33].

To further enhance the annotation of the transcriptome dataset, all genes with GO terms were mapped to metabolic pathways using KEGG automatic annotation server (KAAS) [34]. From this analysis, 125 unique enzymes commission (EC) numbers were assigned to 22,095 genes, of which 31 unique enzymes were assigned to 50 metabolic pathways (Table 2) (Additional file 2).

Table 2 Top metabolic pathways in *P. alba*

*Kegg metabolics patways	Number of unigenes
Purine metabolism	485
Thiamine metabolism	476
Nitrogen metabolism	69
Oxidative phosphorylation	63
Phenylpropanoid biosynthesis	55
Lysine degradation	51
Starch and sucrose metabolism	42
Phenylalanine metabolism	36
Cyanoamino acid metabolism	31
Methane metabolism	28
Fatty acid biosynthesis	25
Other pathways	180

*KEGG: Kyoto Encyclopedia of Genes and Genomes.
This table shows the KEGG metabolic pathways of plants that were well represented by unique sequences of *Prosopis alba*. The numbers of unigenes involved are described.

Regarding the analysis using KAAS, we found 485 transcripts involved in purine metabolism. This metabolic pathway is of fundamental importance in the growth and development of plants [35]. For instance, purine is involved in building blocks for nucleic acid synthesis and is also an energy source, as well as a precursor for the synthesis of primary products and secondary products [36,37]. Additionally, 476 genes associated with thiamine metabolism were detected. These genes are of particular interest to the *Prosopis* genus since the thiamine metabolism is involved in abiotic stress response through the Ca^{2+} salicylic acid and related signaling pathways [38].

When the metabolic pathways from *P. alba* was compared with other tree species (*N. nervosa*), we were able to observe 45 shared pathways. From these pathways, 8 were differential from *Prosopis* and 10 from *Nothofagus* (data not shown).

Assessment of leaf transcriptome assembly

For the assessment of the quality and completeness of the *P alba* transcriptome, a pair-wise reciprocal BLASTP was performed using the gene catalogues from *M. truncatula*, *G. max* and our *P alba* unigene dataset. All protein-coding loci were translated into their correct amino acid

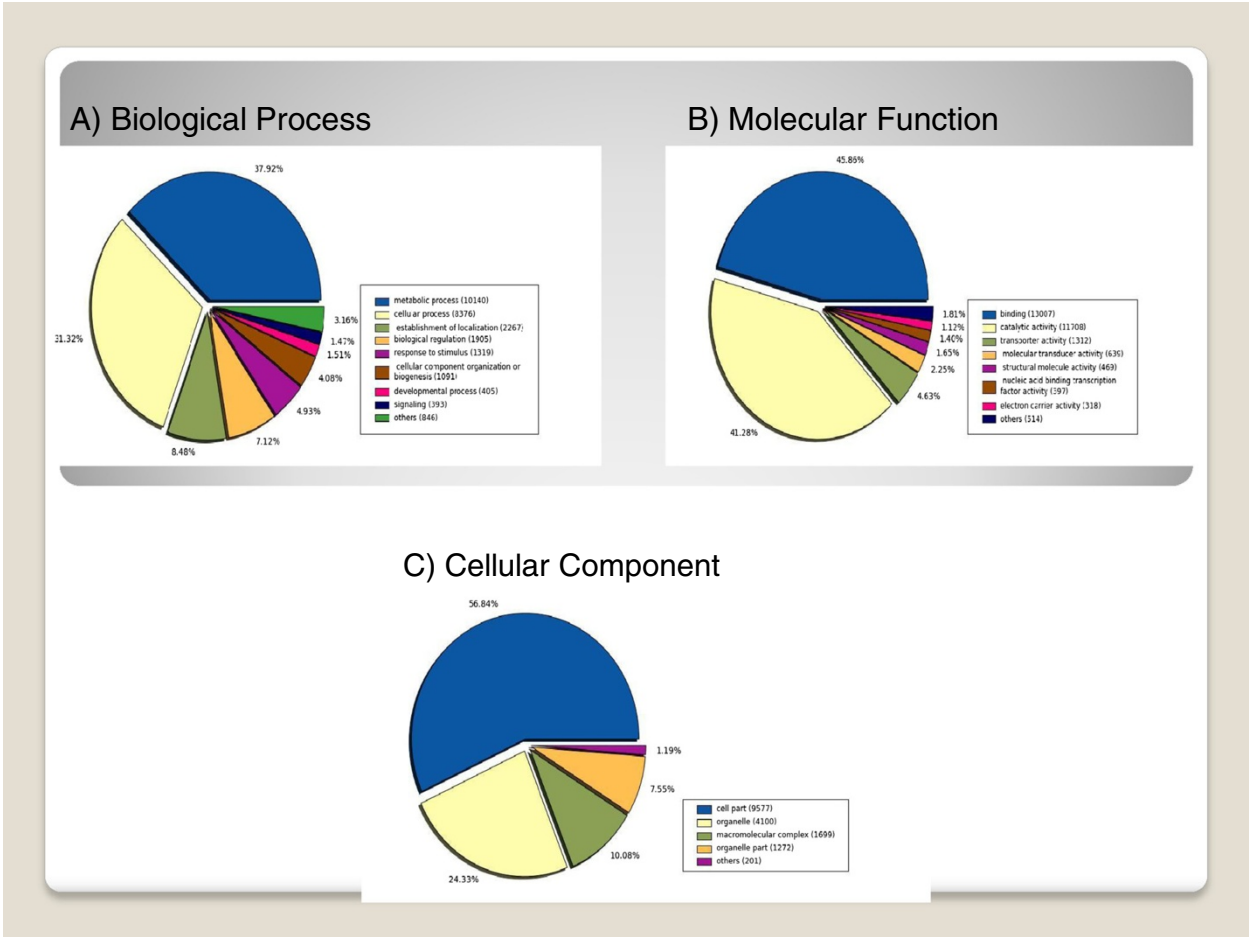

Figure 3 Gene Ontology (GO) assignment of 37,563 *P. alba* unigenes in level 2. The total number of unigenes annotated (22,095) for each main category are: 14,422 for "Biological Process" **(A)**, 19,077 for "Molecular Function" **(B)** and 9,890 for "Cellular Component" **(C)**.

sequences and a BLASTP analysis was performed using the default parameters and an E-value cut-off of 10^{10}. Sequences that were the reciprocal best hits among all 3 species were considered as the orthologue set and, taking this into account, 2500 *Prosopis* unigenes fell in this category. For the following analysis, the same strategy was followed but the comparison was carried out just between two species. The comparisons were performed with two legumes that have a complete genome assembly, *M. truncatula* and *G. max*. The results of these comparisons showed 4,872 and 4,703 unigenes from *P. alba* when compared to *M. truncatula* and *G. max* respectively. As test of the stringency of our strategy, we compared *M. truncatula* and *G. max* with each other and obtained a set of 15,219 orthologues as approximately expected with this strategy. In addition, to evaluate the distribution of all the *P. alba* isotigs along the eight chromosomes of *M. truncatula* and the 20 chromosomes of *G. max*, a protein-protein analysis was performed with "Promer". This program translates all sequences into its six-open reading frames and makes an alignment [39]. The results of this analysis were plotted using a window size of 100kb through their genomic sequences (Figure 5). Transposon and gene densities in each species were distributed along all chromosomes of *M. truncatula* and *G. max*. As expected, in *G. max*, sequences of genes were most distributed in chromosome regions with low density of transposons. Genes belonging to *P. alba* distributed along all chromosomes in both species (Figure 5). However, sequences of genes were less represented in chromosome 6 of *M. truncatula*. This could also be seen in the central green lines (ring 4) that show the distribution of the 2,500 unigene homologous to *G. max*, *P. alba* and *M. truncatula* altogether.

Again, chromosome 6 of *M. truncatula* had fewer mRNA sequences homologous to *G. max* as well as with *P. alba*. When comparing *M. truncatula* genome with *Lotus japonica* [40], similar results were obtained. These

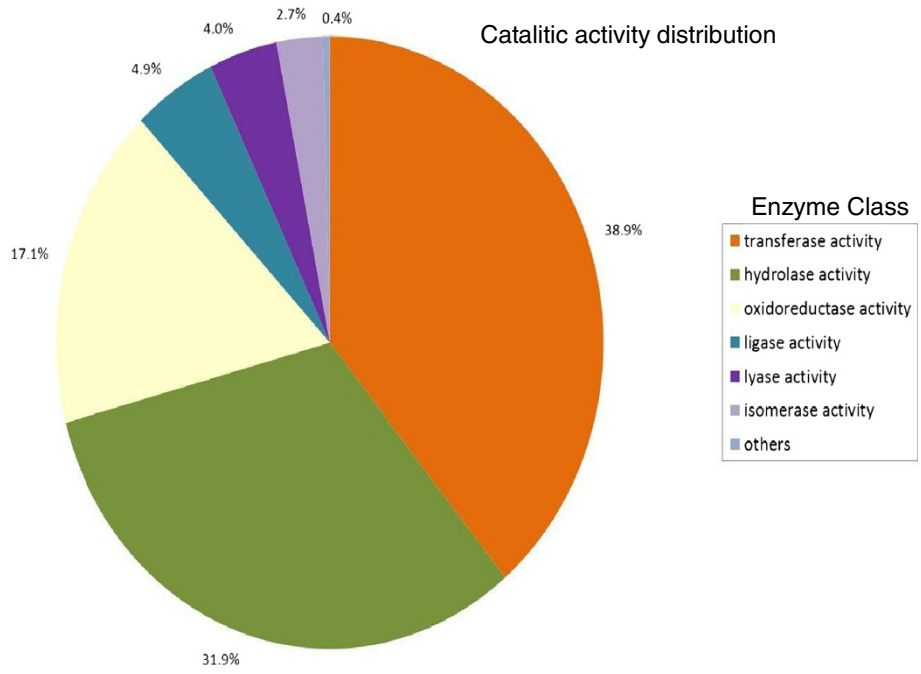

Figure 4 Catalytic activity distribution in annotated *P. alba* unigenes.

findings demonstrate the lack of marker-based synteny with pea [41] and the abundance of nucleotide-binding site-Leu-rich repeat genes [42]. The unusual high proportion of heterochromatin in this chromosome as it was previously reported [43] may explain why we found less homologous mRNA sequences in chromosome 6 of *M. truncatula*.

SNP detection

Single Nucleotide Polymorphisms (SNPs) were identified through the analysis of the multiple alignments produced during the assembly process. The criterion for this analysis was reducing the probability of false positive identification (see Materials and Methods).

The analysis of 15,814 isotigs resulted in the identification of 7,134 putative SNPs. After applying filters, we obtained a total of 6,236 high confidence SNPs from 1,834 isotigs (average = 3.4 SNPs per isotig). Of these SNPs, 70 belonged to 14 chloroplast isotig and 6 contig sequences (see Additional file 3). The 15,814 isotigs that were mined for SNPs identification comprised 15,665 KB of *P. alba* transcriptome, with 1 SNP per 2,512 bases on average. The SNPs density in "algarrobo" was similar to that found in *Capsicum annuum* trancriptome (1 SNP per 2,253 bases) [44]. In both cases the criteria to identified SNPs and the number of individual analysed were analogous. However, in oak [45] and in *Eucalyptus grandis* [46] the SNP frequency was much higher than in *P. alba,* (1 SNP every 471 bp and 1 SNP every 192 bp, respectively). These differences can mainly be attributed to the number of individuals that were sequenced in the other forest species (21 individuals of oaks and more than 200 individuals of *Eucalyptus grandis*). Within the identified SNPs, transitions (65%) were far more frequent than transversions (35%) (Table 3). A similar number of A/G and C/T transitions together with equivalent values of the four transversion categories (A/T, A/C, G /T, C/G) were found. These results are in accordance with the findings described in *Cucurbita pepo* SNPs [30].

Single sequence repeats (SSRs) detection

Using the SSR webserver from the Genome Database for Rosaceae (GDR), we identified and characterized several SSRs (microsatellites) motifs as potential molecular markers in the *Prosopis* putative unigenes collection generated in this work.

The criterion used for the SSR selection was based on the minimum number of repeats (see Materials and Methods). These settings resulted in the identification of 5,956 nuclear SSRs within 54,768 unigenes, i.e. SSR frequency of 11% taking into account multiple repeat occurrences in a same unique locus. This frequency was comparable to that reported in *Nothofagus* (15%) [20] and lower than in oak (19 and 24%) [45,47]. A total of 4,990 (9%) nuclear unigenes contained at least one SSR, suggesting that they are distributed throughout the whole leaf transcriptome. Additionally, 4,593 SSRs

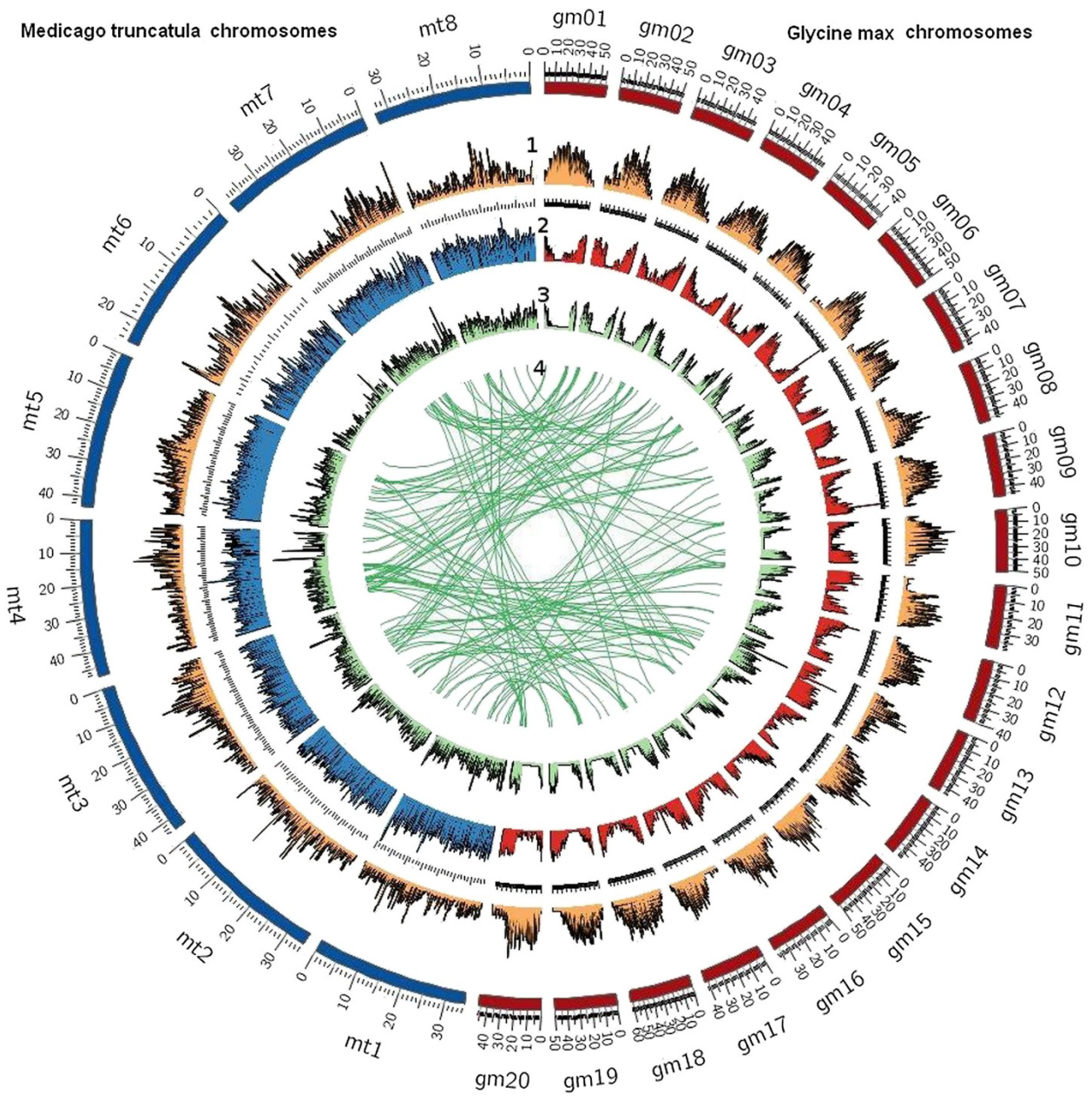

Figure 5 *P. alba* contigs/isotigs mapped to the *M. truncatula* and *G. max* genomes. The 8 and 20 individual chromosomes of *M. truncatula* and *G. Max*, respectively, are shown in the outer circle (blue and red, respectively). Relative transposon and gene densities (by 100kb region) on each chromosome are successively displayed inward as follows: (i) Ring 1: transposon density in *M. truncatula* and *G. max* (orange), (ii) Ring 2: gene density in *M. truncatula* (blue) and *G. max* (red), (iii) Ring 3: *P. alba* gene density relative to gene densities of *M. truncatula* and *G. max* (green) (iv) Ring 4: sequences homologous to *G. max*, *P. alba* and *M. truncatula* altogether (central green lines).

(77 %) had sufficient flanking sequences to allow the design of appropriate unique primers to generate PCR products within the range of 100 to 300 bp. Detailed information of the SSRs that were discovered in this research is described in Additional file 3.

As expected, the most frequent type of microsatellite corresponded to trimeric repeats (41%), while much lower frequencies were found for dimeric motifs (29%), tetra- (20%), penta- (5%) and hexanucleotide repeats (5%) (Figure 6). Similar results were found in *Nothofagus* and oak [20,45].

Seventy two percent of the sequences had only one SSR (72%) and 20 % had two. Of the unique SSR, 44% were of trimeric motif followed by 28% of di- and 19%

Table 3 Single nucleotide polymorphism (SNPs) statistics

SNPs	Number	SNPs	Number
Transitions		Transversions	
A<–>G	2,076 (51%)	A<–>C	548 (25%)
C<–>T	2,004 (49%)	C<–>G	610 (28%)
		G<–>T	570 (26%)
		A<–>T	428 (20%)
Total	4,080 (65%)	Total	2,156 (35%)

Class and number of transitions and transversions are shown for putative high quality single nucleotide polymorphism (SNPs) identified in *P. alba* transcriptome.

of tetranucleotide motifs. The SSRs were highly distributed over the sequences; which provides a useful tool for different genetic studies (Figure 6).

The topography of SSR distribution was analyzed for the presence of SSR within predicted UTRs and coding sequence regions (See Materials and Methods). About 44% of the SSR sequences were inside ORF sequences, being most of them trinucleotide and hexanucleotides repeats (58%). In the UTRs, the dinucleotides motifs were more frequent (35%) comparable to those reported in other trees such as oak (27%) [45], *Nothofagus* (40%) [20] and pines (65-75%) [48].

Eighty one percent of the repeated sequences found in ORF had a combination of length motif and repeat number multiple of three, (i.e., $(TC)_9$, $(GA)_9$, $(CTCC)_3$, $(CGCCC)_3$) which did not modify the reading frame, (i.e. nine motifs of two nucleotides =18 bp = 6 codons). This could be explained on the basis of the selective disadvantage of non-trimeric SSR variants in coding regions, possibly causing frame-shift mutations [49].

Finally, we also identified SSRs belonging to gene families associated with the production of cellulose, the lignin biosynthetic pathway [50] and with transcription factors. As for the production of cellulose, we detected genes such as cellulose synthases (CesA), glycosyl transferases and sucrose synthase. The identified genes related to the lignin biosynthetic pathway [50] were cinnamoyl alcohol dehydrogenase as well as cinnamoyl reductase-like protein. From the genes associated with transcription factors, zinc finger proteins and some antioxidants (for example, gdp-mannose pyrophosphorylase) were identified. Stress related sequences such as heat shock proteins and zeaxanthin epoxidase were also identified in the reference *P. alba* transcriptome, Interestingly, zeaxanthin epoxidase is a precursor of abscisic acid (ABA) that is involved in response to abiotic stress, including tolerance to heavy metals.

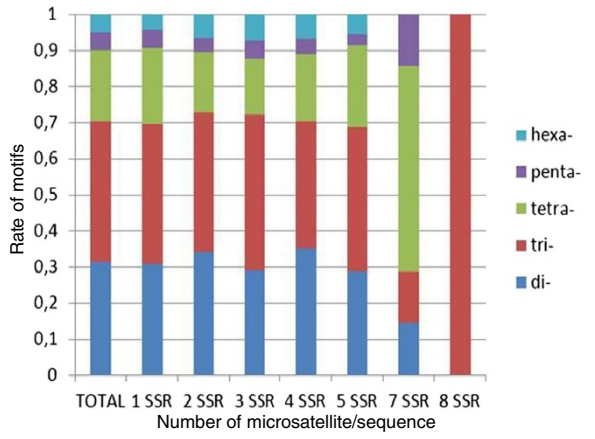

Figure 6 Frequencies of SSRs in *P. alba* unigenes. Frequencies of di- tri- tetra- , penta- and hexa nucleotide SSRs in unigenes containing one to eight SSRs per locus. Of the sequences, 4294 had one SSR, 602 contained two, 116 had three, 18 had four, six had five SSRs and one sequence showed 8 SSRs.

Validation of the predicted microsatellite markers

Eighty seven microsatellites were selected according to their sequence length, GC content and functional annotation. As for functional annotation, we selected those related to the following categories: stress, calcium metabolism, peroxidases, myb and zinc finger proteins, among other putative functions (51% were located in predicted ORFs). The 87 loci were tested for PCR amplification in six individuals. All of them (100%) were effectively amplified validating the quality of the assembly and the utility of the SSRs herein identified. Similar results were obtained in *Nothofagus* by applying the same strategy for the assembly and the in silico searches for SSRs. Similar research carried out using Illumina sequencing technology in sesame showed that about 90% primer pairs successfully amplified DNA fragments [51]. However, the rate of SSR validation was lower (64.9%) when the marker mining was done using EST produced by Sanger technology [52], possibly because of the low-quality of the EST sequences. The lower rate could also be explained by the use of primer sequences derived from chimerical cDNA clones.

About 15% (11 SSR) of the tested *Prosopis* SSRs were polymorphic and showed at least one individual differing in allelic composition, belonging to 9 different loci positions ie: three SSRs belonged to the same isotig 00930. This relative low percentage of polymorphic loci could be due to the small sample size tested (six seedlings), the selected target loci (focused principally in abiotic stress) and their presence in predicted ORF. The percentage was slightly lower than that of a similar study in *Nothofagus* (20%). Furthermore, even higher percentages of polymorphism were identified in other reports: 30% in *Phoenix dactylifera L* date palm (assessed in 12 cultivars) [53], 46% in *E. globulus* (evaluated in 8 samples) [52] and 80% in sesame (assayed in 24 samples) [51]. Five of the polymorphic SSRs found in this work were located within predicted ORF and showed repeat motifs

Table 4 Polymorphic SSR primer pairs derived from *P. alba* unigenes

Primer name	Marker ID name gene bank accession no.	Motif	ORF	Primer sequence 5´-3´	Amplicon Lenght	Seq description
I-P00930b	isotig00930b	acc (5)	Y	F:GCAACAGCACTGCTTCAAA	260-268	Zinc finger protein magpie-like
				R:AAAATAGCGCCATAGTTTGCTC		
I-P00930c	isotig00930c	gtc (4)	Y	F:TATGGCGCTATTTTTGGAGG	236-240	Zinc finger protein magpie-like
				R: TCATGCTCCTCACAATCTGC		
I-P00930d	isotig00930d	aac (6)	Y	F:TCGAGATTTTCTTGGGGTTG	176-178	Zinc finger protein magpie-like
				R: AAATTCCCTCCTCCTCCAAA		
I-P03211	isotig03211a	aat (4)	Y	F: TTGCTTCAGAAAGCTGCTCA	190-198	Uncharacterized protein loc100815794
				R: AACCCTCGAAGATGATGGTG		
I-P03325a	isotig03325a	ca (5)	N	F: CGTGCATGAATGTCACAGAC	226-230	Peroxidase
				R:AGGGTGAGATCAGAAGGCAA		
I-P06286b	isotig06286b	tc (5)	N	F:TGACAACCCATCTTCTTCTTCA	206-216	Myb transcription factor myb138
				R:ATTTGCACAAGGGTAAAGATGG		
I-P06639	isotig06639a	at (5)	N	F: CATCCCGTTCAAGTCCAAGT	226-230	Aquaporin pip11
				R: AGCCCCCTTCCAACTTCTAA		
I-P07653	isotig07653a	gtt (4)	Y	F: AGTGATGATTCGGATCCTGG	216-220	zf-hd homeobox protein
				R:GAGAGACGAGGACTTGGTGC		
I-P10500	Isotig10500a	ttc (6)	N	F: CTCCGACAGATTCAGCATCA	260-275	Pentatricopeptide repeat-containing protein
				R: TTCTTTCAAACTCGCCATCA		
S-P1DKSFA	GR7D2IN01DKSFA	ttta (3)	N	F: GTTTACCCATTGCAGGTCGT	162-168	Calcium-binding mitochondrial carrier protein s −1
				R: CCCCATATGCAGAATCACCT		
S-P1EPIV2	GR7D2IN01EPIV2	taa (4)	N	F: TAAGCATTCATAGCCAGCCC	290-298	Peroxidase 73
				R: GACCAGGTCCTGTTTACCGA		

The included data are: primer names, marker ID names, motif and number of repeats, position in ORF, forward and reverse primer sequences (5´3´), amplicon length (bp), BLASTX similarity matches (Putative Function).

multiple of three (Table 4), therefore maintaining the coding region in frame [49].

Polymorphic SSR predicted

From the 2,352 nuclear SSRs detected in contigs/isotigs (which means they have several reads that allow to determine putative polymorphism), a total of 1,995 (85%) had defined primers for PCR reactions, belonging to 1,622 different isogroups (unigenes).

In order to predict nuclear polymorphic SSRs, we carried out in silico PCR for each of the sequences from the different contigs/isotigs. For this purpose, all 1,063,520 high quality reads were used. To achieve a higher success rate, another set of primers was also were designed closer to the SSR motifs in order to capture short length reads included in contigs.

At least 123 nuclear polymorphic SSRs were detected by PCR in silico; which only includes isotigs integrated by three or more reads and whose product size generated by in silico PCR differed in at least two base length (Additional file 3). An apparent underestimation of nuclear polymorphisms in silico was observed when considering that from the 9 clear polymorphic SSRs coming from isotigs (Table 4), of which, only two of them (22%) resulted polymorphic under the criteria used in silico. However, from the 87 SSRs that were amplified in vitro, 69 belonged to isotigs and only six had enough reads to be considered in this in silico analysis; which resulted in only three effectively polymorphic SSRs in silico (50%). Therefore, it can be predicted that 52 SSRs out of the 123 new SSRs that were detected in silico will be effectively polymorphic in vitro. This result could be an interesting survey of potential useful SSRs and could contribute significantly to the SSRs available in other reports [18,19].

A total of 135 GO terms were allocated to the 116 nuclear isogroups containing the polymorphic SSRs that were identified in silico in this research. They were assigned under the categories “Biological Process” (39 terms), “Molecular Function” (55 terms), and “Cellular Component” (41 terms). The most represented subcategories assigned under “Biological Process” at third-level terms were: “primary metabolic process” (16.4%), “cellular metabolic process” (15.6%) and “macromolecule metabolic

process" (14%). In addition, many of the terms that classified as "Molecular Function" were represented by genes in the following subcategories: "hydrolase activity" (20%), "nucleic binding" (18%), "ion binding" (16%) and "protein binding" (13%). In addition, seven metabolic pathways were represented by at least one sequence, with its corresponding EC number. This makes these functional markers especially useful for population and evolution analyses of *P. alb*a.

Chloroplasts mining

We detected 44,079 chloroplast reads through alignment analyses to related chloroplast (cp) sequences. After an alignment analysis with the legume *Vigna radiata* chloroplast genome, 56 contigs composed of 59,040 bp were generated, spanning a total of 129,208 bp that belong to the *Prosopis* cp genome. The chloroplast reads of *P. alba* with 59,040 bp represented ~40 % of the total cp genome of *V. radiata* (151,271 bp) [54]. There were 55 intra scaffold gaps in *P. alba* cp genome with a mean sequence gap size of 1252 bp.

A total of 14 isogroups harboring 36 SSRs were also found: 18 with designed primers, 17 with different BLASTX hits related to chloroplast metabolism (oxidoreductase, ribosomal, etc.) and four polymorphic in silico (Additional file 3). Chloroplast SSRs were previously described in several plants such as *Pinus radiata*, *Oryza sativa, Nicotiana* [55-57]. More recently several other organisms were also characterized for these SSRs: *Eucalyptus globulus, G. max, V. radiata, M. truncatula, V. vinifera* among others. All of these Chloroplast SSRs have been deposited in a data base http://www.mcr.org.in/chloromitossrdb/ [58]. Also, in rice around 4.5% of the chloroplast genome has been covered by SSRs [59].

Methods

RNA preparation and cDNA library synthesis

Total RNA was extracted from leaves of seedling collected from natural populations of *P.alba* from different provinces of Argentina: population 1 from Campo Durán (province of Salta), population 2 from Isla Cuba (province of Formosa) and population 3 from Chañar Bajada (province of Santiago del Estero)

The RNA extraction method used in this research is the one described by Chang el al, (1993) [60]. Briefly, one gram of fresh tissue was ground to a fine powder under liquid nitrogen. Then, after two extractions with chloroform, the RNA was precipitated with $LiCl_2$, extracted again with chloroform and finally precipitated with ethanol. The resultant RNA was resuspended in 50 μl of DEPC treated water. The RNA was quantified using a Nanodrop 1,000 spectrophotometer and its quality was measured with a 2,100 Bioanalyzer (Agilent Technologies Inc.). Then, it was subjected to purification using the Poly (A) Purist kit (Ambion) and its quality was once more assessed with the 2,100 Bioanalyzer. cDNA was synthesized using cDNA Kit (Roche) and used to construct a shotgun library for pyrosequencing technology (Roche). The *Prosopis* cDNA library was subjected to a 1/3 of plate production run on the 454-GS-FLX sequencing instrument. This run was conducted at INDEAR (Rosario Biotechnology Institute, Rosario, Argentina).

Transcript assembly and analysis

The sequences were subjected to filtering for adaptors, primer sequences and low-quality sequences After removing the low quality sequences, the curated raw 454 read sequences were assembled into contigs, isotigs and isogroups using Newbler Assembler software 2.5p1 (Roche, IN, USA). The reads identified like singletons (i.e., reads not assembled into isotigs) after assembly were subjected to CD-HIT-454 clustering algorithm at 95% identity cutoff, which eliminates redundant sequences [61].

BLASTX (e-value cut off ≤ 10e-10) searches were performed against a Viridiplantae protein database first. Then, the sequences with no hits were used to perform a successive BLASTX against the NCBI nr protein database in order to make an assessment of the putative identities of the sequences. Unigenes (>200 bp) were deposited at the National Centre for Biotechnology Information (NCBI) Transcriptome Shotgun Assembly (TSA) Database under BioProject: 218545 TSA- SUB336788.

Annotation and mapping routines were run with BLAST2GO [28], which assigns Gene Ontology ([62], http://www.geneontology.org) annotation, KEGG maps (Kyoto Encyclopedia of Genes and Genomes, KASS) and an enzyme classification number (EC number) using a combination of similarity searches and statistical analysis [34]. In addition to BLAST2GO, the full suite of InterProScan was ran with default parameters. InterProScan combines different protein signature recognition methods native to the InterPro member databases into one resource that searches for the corresponding InterPro and GO annotations.

Chloroplast assembly analysis was carried out using AMOScmp [63]. To search for chloroplast sequences, BLASTN and TBLASTX (BLASTN e-50 TBLASTX e-10) were performed. The analysis was based on similarity with and without translation to 109 chloroplast genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/Chloroplasts/plastids/).

Comparative genomics

Circos software tool [64] was used to visualize *P. alba* sequences with *M. truncatula* and *G. max´*s genomes, through circular concentric ideograms layout to facilitate the display data as scatter, line and histogram plots for each different sample. Promer analysis was performed

and filtered by using a window size of 100kb through their genome sequences. Homologous sequences for the three species were determined when reciprocal TBLASTX best hits were found for the three "genes" tested.

SNP identification

In order to perform matching, alignment of DNA sequences and searching for putative SNPs, the SSAHAsnp Program (Sequence Search and Alignment by Hashing Algorithm) was used [65]. The criterion designed to reduce the probability of false positive identification was that the minority allele (the second most common nucleic allele) should be found in at least 4 sequences and that at least the 10% of reads had an SNP from total coverage, which should be at least 8x.

SSR identification

In order to identify SSRs for all possible combinations of dinucleotide, trinucleotide, tetranucleotide and pentanucleotide repeats, we performed a run using the SSR webserver (GDR) (http://www.rosaceae.org/bio/content?title=&url=/cgi-bin/gdr/gdr_ssr). This webserver uses the GETORF algorithm (EMBOSS Package) and selects the longest ORF as the putative coding region. This webserver also uses Primer 3 (v.0.4.0) [66] to design primer pairs. The criteria used for the SSR selection based on the minimum number of repeats was as follows: five for dinucleotide, four for trinucleotide, three for tetranucleotide and three for penta and hexanucleotide motives.

The presence of expressed repetitive DNA was revealed using the BLASTN (e-value cut off ≤10e-10) searches against all Viridiplantae Repbase.

SSR validation

For validation of SSR primers, total DNA was extracted from young leaves of six *Prosopis alba* seedlings from three native populations described previously (two for each one). For DNA extraction, the Dneasy Plant mini kit (Quiagen) was used following the manufacturer´s instructions. Regular primers were synthesized (AlphaDNA, Montreal, CA, USA) and used for PCR (polymerase chain reaction) amplification. PCR reactions consisted of 20 ng of total DNA, 0.25 μM of each primer, 3 mM of $MgCl_2$, 0.2 mM of each dNTP, 1X of PCR buffer and 1 U of Platinum Taq polymerase (Invitrogen). All PCRs were performed with the following conditions: a denaturation step of 2 min at 94°C, a regular touchdown PCR ranging from 60°C to 50°C with 28 cycles at the touchdown temperature of 50°C (45 s at 92°C, 45 s at 50°C and 45 s at 72°C). The final extension step was of 10 min at 72°C and then the temperature conditions were adjusted for each particular microsatellite. Samples were mixed with denaturing loading buffer, incubated for 5 min at 95°C, and separated on a 6% polyacrylamide gel. Amplification products were stained using the DNA silver staining procedure of Promega (USA) following the manufacturer's instructions. Details of primers sequences, SSR location and amplicon sizes are described in Table 4.

Conclusions

The *P.alba* transcriptome database obtained and characterized here represents a major contribution for *Prosopis sp.* genomics and genetics. It will be useful for discovering genes of interest and genetic markers, which could allow to investigate functional diversity in natural populations. These tools will also lead to conduct comparative genomics studies with other *Prosopis* species taking advantage of their remarkable ecophysiological differences. This work highlights the utility of transcriptome high performance sequencing as a fast and cost effective way for obtaining rapid information on the coding of genetic variation in *Prosopis* genus. This study allowed us to: (i) obtain 1,103,231 transcript raw reads and 54,814 unigene sequences from *P.alba*, (ii) identify putative function in 37,563 unigenes for the genus, (iii) identify 700 putative stress-response genes, (iv) discover 4,593 genomic SSRs with designed primers, validate 87 and detect 11 polymorphic SSRs, several of them related to response to stress, (v) identify probably 52 effectively polymorphic after in silico analysis, and (vi) identify 6,158 higher confidence nuclear SNPs, some of them related to the production of cellulose, together with the lignin biosynthetic pathway and with stress, among others.

Additional files

Additional file 1: Schematic representation of the overall sequencing and annotation workflow of *Prosopis alba* transcriptome. The steps and sets of sequences involved in transcriptome sequencing, assembly of reads, annotation using protein databases, the statistical thresholds, filters, genetic marker discovery and characterization.

Additional file 2: Annotation. This table provides information on the annotation of isotigs and singletons, GO information and the enzymes putatively encoded by the RNA sequences, based on homology prediction and their associated pathways. This includes KEGG maps, enzyme names, and sequences ID.

Additional file 3: In silico SSRs and SNPs derived from *Prosopis alba* leaf transcriptome. The data describe the 5,996 SSR and 6236 SNPs. Sheet SSR: Included are unigenes names, Isogroup, marker ID, Sequence Lenght (bp), SSR Polym: (In silico "IS" or "IS(2)"=two sequences in pcr in silico, PCR Amplification: "POLYM", "COMPLEX PATTERN", " MONOM", >mw=molecular weight out of range) SSR description: # SSRs per seq, repeat lenght, motif, # Repeats, SSR position (start, stop) ORF definition (start, stop, SSR in ORF) primers description: sequence of forward and reverse primers, expected product size (bp), similarity matches, E value, similarity mean, #GO, GO terms, Enzimes codes and their chloroplast belonging. Sheet SNP: Included are unigenes names, Isogroup, marker ID, SNP position, SNP, # of mapping reads with SNP, # total reads coverage on the SNP position, # of mapping SNP reads vs consensus, % of reads with SNP, similarity matches, E value, similarity mean, #GO, GO terms, Enzimes codes , chloroplast belonging.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SLT organized the research, provided funds, contributed to RNA extraction, data analysis and wrote this manuscript. MR coordinated and carried out bioinformatics analyses and contributed to the manuscript, MFP contributed to RNA extraction and carried out SSR validation. SG carried out the bioinformatics analyses. CVA contributed to the analyses involving BLAST, SSR characterization and with the manuscript revision. PF contributed to RNA extraction. DLL contributed to write the project and manuscript, ARV provided the biological material for transcriptome sequencing and contributed to manuscript revision. HEH conceived this study and contributed to the revision of the manuscript. NBP provided funding, participated in the design of the bioinformatics study and reviewed the manuscript. SNMP provided funding, contributed to research design, data analysis, contributed to the manuscript and its review. All authors approved the final manuscript.

Acknowledgments
CVA and MFP thank PROMEF for their fellowships. This research was supported by the INTA-PE 242421, PPR 242001 and PPR 245001. Special thanks to Julia Sabio for the critical English edition of the manuscript.

Author details
[1]Instituto de Recursos Biológicos, IRB, Instituto Nacional de Tecnología Agropecuaria (INTA Castelar), CC 25, Castelar B1712WAA, Argentina. [2]Instituto de Biotecnología, CICVyA, Instituto Nacional de Tecnología Agropecuaria (INTA Castelar), CC 25, Castelar B1712WAA, Argentina. [3]Instituto de Fisiología y Recursos Genéticos Vegetales (IFRGV), Centro de Investigaciones Agropecuarias (CIAP), Instituto Nacional de Tecnología Agropecuaria (INTA), Camino 60 Cuadras, km 5.5, X5020ICA, Córdoba, Argentina. [4]Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina. [5]CONICET, Buenos Aires, Argentina.

Received: 1 July 2013 Accepted: 7 October 2013
Published: 14 October 2013

References

1. Burkart A: **A monograph of the genus Prosopis (Leguminosae subfam. Mimosoideae). [Part 1.].** *Journal of the Arnold Arboretum* 1976, **57**(Suppl 3):219–249.
2. Palacios R, Carmaran C, Iglesias L, Picca P, Torregrosa S, Gonzales S: *Taxonomía numérica (descriptores).* Prosopis en Argentina, UNC-UBA-FAO Press; Buenos Aires- Córdoba: 1988:91–96.
3. Cabrera A: *Regiones Fitogeográficas Argentinas. In Enciclopedia Argentina de Agricultura y Jardinería. Volume II, Fascicle 1.* Buenos Aires: ACME Press; 1976.
4. Mottura M: *Development of microsatellites in Prosopis spp. and their application to study the reproduction system. PhD Thesis. Institute of Forest Genetics and Forest Tree Breeding, Faculty of Forest Sciences and Forest Ecology.* Germany: Georg-August University of Göttingen; 2006.
5. Saidman BO: **Isoenzymatic studies of alcohol dehydrogenase and glutamate oxalacetate transaminase in four South American species of Prosopis and their natural hybrids.** *Silvae Genetica* 1986, **35:**3–10.
6. Saidman BO, Vilardi JC: **Analysis of the genetic similarities among seven species of Prosopis (Leguminosae: Mimosoideae).** *Theoretical and Applied Genetics (TAG)* 1987, **75:**109–116.
7. Saidman BO, Bessega CF, Ferreira LI, Julio N, Vilardi JC: **The use of genetic markers to assess population structure and relationships among species of the genus Prosopis (Leguminosae).** *Boletín de la Sociedad Argentina de Botánica* 2000, **34**(Suppl 3):315–324.
8. Verga A: *Genetic study of Prosopis chilensis y Prosopis flexuosa (Mimosaceae) in the dry Chaco of Argentina. PhD Thesis.* Universität Göttingen, Alemania; 1995.
9. Demaio P, Karlin UO, Medina M: *Árboles Nativos del Centro de Argentina.* Buenos Aires, Argentina: L.O.L.A. Press; 2002.
10. Karlin U: *Recursos forrajeros naturales del Chaco Seco: Manejo de Leñosas.* Córdoba, Argentina: II Reunión de Intercambio Tecnológico en Zonas Áridas y Semiáridas; 1983:78–96.
11. Bregaglio M, Karlin U, Oirini R: **Efecto del desmonte selectivo sobre la regeneración de la masa forestal y la producción de pasturas, en el chaco árido de la provincia de Córdoba, Argentina.** *Multequina* 2001, **10:**17–24.
12. Karlin UO, Coirini R, Catalán L, Zapata R: **Especies arbóreas y arbustivas para zonas áridas y semiáridas de América Latina.** *Serie Zonas Áridas y Semiáridas N°12 OEA* 1997:41–51.
13. Verga A, Navall M, Joseau J, Royo O, Degano W: **Caracterización morfológica de los algarrobos (Prosopis sp.) en las regiones fitogeográficas Chaqueña y Espinal norte de Argentina.** *Quebracho* 2009, **17:**31–40.
14. López Lauenstein D, Luna C, Verga A: *Respuesta al estrés hídrico en dos grupos morfológicos de Prosopis alba. En resúmenes de la XXVIII Reunión Argentina de Fisiología Vegetal.* La Plata, Argentina; 2010.
15. SAyDS: *Informe Regional Parque Chaqueño.* Argentina: Primer Inventario Nacional De Bosques Nativos, Proyecto Bosques Nativos y Áreas Protegidas BIRF 4085-AR; 2007.
16. Verga A: **Algarrobos como especies para forestación. Una estrategia de mejoramiento.** *SAGPyA Forestal* 2000, **17:**2–9.
17. George S, Venkataraman G, Parida A: **Identification of stress-induced genes from the drought-tolerant plant Prosopis juliflora (Swartz) DC. through analysis of expressed sequence tags.** *Genome* 2007, **50**(Suppl 5):470–478.
18. Mottura MC, Finkeldey R, Verga AR, Gailing O: **Development and characterization of microsatellite markers for Prosopis chilensis and Prosopis flexuosa and cross-species amplification.** *Molecular Ecology Notes* 2005, **5**(Suppl 3):487–489.
19. Bessega CF, Pometti CL, Miller JT, Watts R, Saidman BO, Vilardi JC: **New Microsatellite Loci for Prosopis alba and P. chilensis (Fabaceae).** *Applications in Plant Sciences* 2013, **1**(Suppl 5):1200324.
20. Torales SL, Rivarola M, Pomponio MF, Fernández P, Acuña CV, Marchelli P, Gonzalez S, Azpilicueta MM, Hopp HE, Gallo L, Paniego NB, Poltri SNM: **Transcriptome survey of Patagonian southern beech Nothofagus nervosa (= N. Alpina): assembly, annotation and molecular marker discovery.** *BMC genomics* 2012, **13**(Suppl 1):291.
21. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Molecular ecology* 2008, **17**(Suppl 7):1636–1647.
22. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx.** *BMC Genomics* 2009, **10**(Suppl 1):219.
23. Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA: **Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery.** *BMC genomics* 2010, **11:**180.
24. Rismani-Yazdi H, Haznedaroglu BZ, Bibby K, Peccia J: **Transcriptome sequencing and annotation of the microalgae Dunaliella tertiolecta: pathway description and gene discovery for production of next-generation biofuels.** *BMC genomics* 2011, **12**(Suppl 1):148.
25. Pazos-Navarro M, Dabauza M, Correal E, Hanson K, Teakle N, Real D, Nelson MN: **Next generation DNA sequencing technology delivers valuable genetic markers for the genomic orphan legume species, Bituminaria bituminosa.** *BMC genetics* 2011, **12**(Suppl 1):104.
26. Gish W, States DJ: **Identification of protein coding regions by database similarity search.** *Nature genetics* 1993, **3**(Suppl 3):266–272.
27. Logacheva MD, Kasianov AS, Vinogradov DV, Samigullin TH, Gelfand MS, Makeev VJ, Penin A: **De novo sequencing and characterization of floral transcriptome in two species of buckwheat (Fagopyrum).** *BMC genomics* 2011, **12**(Suppl 1):30.
28. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**(Suppl 18):3674–3676.
29. Zdobnov EM, Apweiler R: **InterProScan - an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**(Suppl 9):847–848.
30. Blanca J, Cañizares J, Roig C, Ziarsolo P, Nuez F, Picó B: **Transcriptome characterization and high throughput SSRs and SNPs discovery in Cucurbita pepo (Cucurbitaceae).** *BMC genomics* 2011, **12**(Suppl 1):104.
31. Aharoni A, Galili G: **Metabolic engineering of the plant primary-secondary metabolism interface.** *Current opinion in biotechnology* 2011, **22**(Suppl 2):239–244.
32. Garzón-Martínez GA, Zhu ZI, Landsman D, Barrero LS, Mariño-Ramírez L: **The Physalis peruviana leaf transcriptome: assembly, annotation and gene model prediction.** *BMC genomics* 2012, **13:**151.
33. Li Y, Luo H-M, Sun C, Song J-Y, Sun Y-Z, Wu Q, Wang N, Yao H, Steinmetz A, Chen S-L: **EST analysis reveals putative genes involved in glycyrrhizin biosynthesis.** *BMC Genomics* 2010, **11**(Suppl 1):268.

34. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Res* 2007, **35**(Web Server issue):182–185.
35. Zrenner R, Stitt M, Sonnewald U, Boldt R: **Pyrimidine and purine biosynthesis and degradation in plants.** *Annual review of plant biology* 2006, **57**:805–836.
36. Stasolla C, Katahira R, Thorpe TA, Ashihara H: **Purine and pyrimidine nucleotide metabolism in higher plants.** *Journal of plant physiology* 2003, **160**(Suppl 11):1271–1295.
37. Boldt R, Zrenner R: **Purine and pyrimidine biosynthesis in higher plants.** *Physiol Plant* 2003, **117**(Suppl 3):297–304.
38. Goyer A: **Thiamine in plants: aspects of its metabolism and functions.** *Phytochemistry* 2010, **71**(Suppl 14–15):1615–1624.
39. Delcher AL, Phillippy A, Carlton J, Salzberg SL: **Fast algorithms for large-scale genome alignment and comparison.** *Nucleic acids research* 2002, **30**(Suppl 11):2478–2483.
40. Young N, Cannon S, Sato S, Kim D, Cook D, Town C, Roe B, Tabata S: **Sequencing the Genespaces of Medicago truncatula and Lotus japonicus.** *Plant Physiol* 2005, **137**:1174–1181.
41. Choi H-K, Mun J-H, Kim D-J, Zhu H, Baek J-M, Mudge J, Roe B, Ellis N, Doyle J, Kiss GB, Young ND, Cook DR: **Estimating genome conservation between crop and model legume species.** *Proc Natl Acad Sci U S A* 2004, **101**(Suppl 43):15289–15294.
42. Zhu H, Cannon S, Young N, Cook D: **Phylogeny and genomic organization of the TIR and non-TIR NBS-LRR resistance gene family in Medicago truncatula.** *Mol Plant Microbe Interact* 2002, **15**(Suppl 6):529–539.
43. Kulikova O, Geurts R, Lamine M, Kim D-J, Cook DR, Leunissen J, De Jong H, Roe BA, Bisseling T: **Satellite repeats in the functional centromere and pericentromeric heterochromatin of Medicago truncatula.** *Chromosoma* 2004, **113**(Suppl 6):276–283.
44. Ashrafi H, Hill T, Stoffel K, Kozik A, Yao J, Chin-Wo SR, Van Deynze A: **De novo assembly of the pepper transcriptome (Capsicum annuum): a benchmark for in silico discovery of SNPs, SSRs and candidate genes.** *BMC genomics* 2012, **13**(Suppl 1):571.
45. Ueno S, Le Provost G, Léger V, Klopp C, Noirot C, Frigerio J-M, Salin F, Salse J, Abrouk M, Murat F, Brendel O, Derory J, Abadie P, Léger P, Cabane C, Barré A, De Daruvar A, Couloux A, Wincker P, Reviron M-P, Kremer A, Plomion C: **Bioinformatic analysis of ESTs collected by Sanger and pyrosequencing methods for a keystone forest tree species: oak.** *BMC genomics* 2010, **11**(Suppl 1):650.
46. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome.** *BMC genomics* 2008, **9**(Suppl 1):312.
47. Durand J, Bodénès C, Chancerel E, Frigerio J-M, Vendramin G, Sebastiani F, Buonamici A, Gailing O, Koelewijn H-P, Villani F, Mattioni C, Cherubini M, Goicoechea PG, Herrán A, Ikaran Z, Cabané C, Ueno S, Alberto F, Dumoulin P-Y, Guichoux E, De Daruvar A, Kremer A, Plomion C: **A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study.** *BMC Genomics* 2010, **11**(Suppl 1):570.
48. Chagné D, Chaumeil P, Ramboer A, Collada C, Guevara A, Cervera MT, Vendramin GG, Garcia V, Frigerio J-M, Echt C, Richardson T, Plomion C: **Cross-species transferability and mapping of genomic and cDNA SSRs in pines.** *TAG* 2004, **109**(Suppl 6):1204–1214.
49. Metzgar D, Bytof J, Wills C: **Selection against frameshift mutations limits microsatellite expansion in coding DNA.** *Genome research* 2000, **10**(Suppl 1):72–80.
50. Humphreys JM, Chapple C: **Rewriting the lignin roadmap.** *Current opinion in plant biology* 2002, **5**:224–229.
51. Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, Lv H, Zhang X: **Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers.** *BMC genomics* 2011, **12**(Suppl 1):451.
52. Acuña CV, Fernandez P, Villalba PV, García MN, Hopp HE, Marcucci Poltri SN: **Discovery, validation, and in silico functional characterization of EST-SSR markers in Eucalyptus globulus.** *Tree Genetics & Genomes* 2012, **8**(Suppl 2):289–301.
53. Zhao Y, Williams R, Prakash CS, He G: **Identification and characterization of gene-based SSR markers in date palm (Phoenix dactylifera L.).** *BMC Plant Biol* 2012, **12**(Suppl 1):237.
54. Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T, Jomchai N, Tragoonrung S: **The chloroplast genome sequence of mungbean (Vigna radiata) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships.** *DNA research* 2010, **17**(Suppl 1):11–22.
55. Powell W, Morgante M, Mcdevitt R, Vendramin G, Rafaslki J: **Polymorphic Simple Sequence Repeat Regions in Chloroplast Genomes: Applications to the Population Genetics of Pines.** *Proc Natl Acad Sci* 1995, **92**(Suppl 17):7759–7763.
56. Cato SA, Richardson TE: **Inter- and intraspecific polymorphism at chloroplast SSR loci and the inheritance of plastids in Pinus radiata D.** *Don. Theoretical and Applied Genetics* 1996, **93**(Suppl 4):587–592.
57. Provan J, Corbett G, Waugh R, McNicol JW, Morgante M, Powell W: **DNA fingerprints of rice (Oryza sativa) obtained from hypervariable chloroplast simple sequence repeats.** *Proceedings Biological sciences* 1996, **263**(Suppl 1375):1275–1281.
58. Sablok G, Mudunuri SB, Patnana S, Popova M, Fares MA, La Porta N: **ChloroMitoSSRDB: Open Source Repository of Perfect and Imperfect Repeats in Organelle Genomes for Evolutionary Genomics.** *DNA research* 2013, **20**(Suppl 2):127–133.
59. Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM: **Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions.** *Bioinformatics* 2007, **23**(Suppl 1):1–4.
60. Chang S, Puryear J, Cairney J: **A simple and efficient method for isolating RNA from pine trees.** *Plant Molecular Biology Reporter* 1993, **11**(Suppl 2):113–116.
61. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**(Suppl 13):1658–1659.
62. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature genetics* 2000, **25**(Suppl 1):25–29.
63. Pop M, Phillippy A, Delcher AL, Salzberg SL: **Comparative genome assembly.** *Briefings in bioinformatics* 2004, **5**(Suppl 3):237–248.
64. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome research* 2009, **19**(Suppl 9):1639–1645.
65. Ning Z, Caccamo M, Mullikin JC: **SSAHAsnp: A polymorphism D detection tool on a whole genome scale.** *IEEE Computational Systems Bioinformatics Conference* 2005:251–254.
66. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods in molecular biology* 2000, **132**(Suppl 3):365–386.

doi:10.1186/1471-2164-14-705
Cite this article as: Torales *et al.*: De novo assembly and characterization of leaf transcriptome for the development of functional molecular markers of the extremophile multipurpose tree species *Prosopis alba*. *BMC Genomics* 2013 **14**:705.