

## PERMANENT GENETIC RESOURCES ARTICLE

# Establishing genomic tools and resources for *Guizotia abyssinica* (L.f.) Cass.—the development of a library of expressed sequence tags, microsatellite loci, and the sequencing of its chloroplast genome

HANNES DEMPEWOLF,\* NOLAN C. KANE,\* KATHERINE L. OSTEVIK,\* MULATU GELETA,†  
MICHAEL S. BARKER,\* ZHAO LAI,§ MEGAN L. STEWART,\* ENDASHAW BEKELE,¶  
JOHANNES M. M. ENGELS,\*\* QUENTIN C. B. CRONK\*†† and LOREN H. RIESEBERG\*

\*The Biodiversity Research Centre and Department of Botany, 3529-6270 University Blvd, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4, †Department of Plant Breeding and Biotechnology, Swedish University of Agricultural Sciences, Box 101, SE-230 53 Alnarp, Sweden, §Department of Biology, Indiana University, Bloomington, IN 47405, USA, ¶Addis Ababa University, PO Box 1176, Addis Ababa, Ethiopia, \*\*Bioversity International, 00057 Maccaresse, Rome, Italy, ††Centre for Plant Research, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4

## Abstract

We present an EST library, chloroplast genome sequence, and nuclear microsatellite markers that were developed for the semi-domesticated oilseed crop noug (*Guizotia abyssinica*) from Ethiopia. The EST library consists of 25 711 Sanger reads, assembled into 17 538 contigs and singletons, of which 4781 were functionally annotated using the *Arabidopsis* Information Resource (TAIR). The age distribution of duplicated genes in the EST library shows evidence of two paleopolyploidizations—a pattern that noug shares with several other species in the Heliantheae tribe (Compositae family). From the EST library, we selected 43 microsatellites and then designed and tested primers for their amplification. The number of microsatellite alleles varied between 2 and 10 (average 4.67), and the average observed and expected heterozygosities were 0.49 and 0.54, respectively. The chloroplast genome was sequenced *de novo* using Illumina's sequencing technology and completed with traditional Sanger sequencing. No large re-arrangements were found between the noug and sunflower chloroplast genomes, but 1.4% of sites have indels and 1.8% show sequence divergence between the two species. We identified 34 tRNAs, 4 rRNA sequences, and 80 coding sequences, including one region (trnH-psbA) with 15% sequence divergence between noug and sunflower that may be particularly useful for phylogeographic studies in noug and its wild relatives.

**Keywords:** chloroplast genome, EST library, genetic resources, *Guizotia abyssinica*, microsatellites, paleopolyploidy

Received 2 December 2009; revision received 24 March 2010; accepted 2 April 2010

## Introduction

Noug (*Guizotia abyssinica*) is a species in the Compositae that is used traditionally as an oilseed crop in Ethiopia, Eritrea and less extensively in India and several other countries in Africa and South Asia (Getinet & Sharma

1996). It has been recognized as a semi-domesticated crop (Dempewolf *et al.* 2008) and shares many characteristics with its wild relatives. Noug has been categorized as a 'neglected and underutilized species' (Getinet & Sharma 1996) and has received little attention from the scientific community. Consequently, it suffers from a lack of improvement through modern breeding efforts and an absence of basic genomic resources. In North America, it is popular as bird-feed and is being sold under various names such as thistle or niger seed. Phylogenetic

Correspondence: Hannes Dempewolf, Fax: (1) 604-822-6089;  
E-mail: handem@biodiversity.ubc.ca

analyses aiming at fully resolving the origin of the domesticated lineage have remained so far unsuccessful, although it appears that *Guizotia scabra* (Vis.) Chiov. ssp. *scabra*, *Guizotia scabra* (Vis.) Chiov. ssp. *schimperii* (Sch. Bip. in Walp.) J. Baagøe and *Guizotia villosa* Sch. Bip. in Walp. all have contributed to the origin of *Guizotia abyssinica* (Bekele *et al.* 2007; Geleta *et al.* unpublished data).

Our aim was to develop genomic tools and resources for population genetic studies, phylogeographic and evolutionary analyses, research on mating systems, studies of gene flow between the crop and its wild relatives, as well as to aid modern breeding efforts.

The generation of a library of expressed sequence tags (ESTs) is often the first step in developing genomic resources for non-model organisms. EST databases can be used for many different purposes, including genome-wide studies of gene expression and selection, the study of gene family evolution or simply for providing sequence data for molecular marker development (Bouck & Vision 2007). We generated an EST database for noug, assembled unigenes, assessed functional categories for those noug unigenes that we were able to annotate, screened the database for the evidence of past genome duplications, and developed simple sequence repeat (SSR) markers, also known as microsatellites from several unigenes. The development of SSRs from ESTs has become the method of choice for many researchers, as it is a more time- and cost-efficient alternative to more traditional approaches, such as library construction, enrichment, and screening. Previous efforts to characterize the genetic diversity of noug populations, using anonymous genetic markers, revealed the presence of high levels of intra- and inter-population diversity (Geleta *et al.* 2007, 2008), but could not fully resolve the origin of domesticated lineages. We aim to utilize the EST-derived microsatellite markers to more closely study the level and partitioning of genetic diversity in noug and arrive at a better understanding of phylogeographic patterns.

We also sequenced the chloroplast genome of noug using Illumina's sequencing technology. The importance of the plastid genome for phylogenetics, DNA barcoding, studies of photosynthesis and, more recently, transplastomics (Bungard 2004; Grevich & Daniell 2005; Jansen *et al.* 2005, 2007) has led to the sequencing of an increasingly large number of whole chloroplast genomes, using both traditional and next-generation sequencing methods (Jansen *et al.* 2005; Cronn *et al.* 2008). The small size and low repeat content of chloroplast genomes make them particularly amenable to sequencing with short-read next-generation platforms such as the Illumina Genome Analyzer (IGA).

Noug's closest crop relative, sunflower (*Helianthus annuus* L.), has a wealth of genomic resources available owing to its global importance as an oilseed crop and its

status as model species for research on speciation. This is also true for lettuce (*Lactuca sativa* L.) which is a more distantly related crop in the Compositae. The sunflower and lettuce chloroplast genomes have both been sequenced and compared (Timme *et al.* 2007). By sequencing the chloroplast genome of noug and comparing it with the plastid genomes of sunflower and lettuce, we were able to assess the level of Compositae chloroplast genome divergence at a finer scale than previous analyses and to increase our understanding of Compositae plastid genome evolution. Furthermore, the chloroplast is an important source of markers for phylogenetic and phylogeographic analyses. Making the full chloroplast genome sequence of noug available empowers researchers to assess the usefulness of a wide range of chloroplast DNA markers for such studies in noug, the Heliantheae (the tribe that includes noug and sunflower), and the Compositae as a whole.

## Methods

### *EST library development and analyses*

Seeds for noug were obtained from USDA's Western Regional Plant Introduction Station, Pullman, WA, and were germinated and grown in the University of British Columbia's greenhouses. USDA-ARS accession PI 508077 from Ethiopia was chosen for sequencing, as it is an Ethiopian accession that is readily available through the USDA's National Genetic Resources Program. Noug RNA was extracted from leaf and root material of 4-week-old seedlings using the Spectrum Plant Total RNA kit from SIGMA. RNase-free DNase I on-column digestion (Qiagen) was performed to further purify the RNA. Total RNA was quantified using a Nanodrop, and its quality was verified using a Bioanalyzer. To obtain full-length, low-redundancy cDNA libraries, it was necessary to employ a normalization strategy to remove high-abundance cDNA transcripts. The following protocol was developed from the manuals of the Clontech Creator SMART cDNA Library Construction Kit (catalogue number 634903) and the Evrogen TRIMMER-DIRECT cDNA normalization kit (catalogue number NK002). Approximately 1–1.5 µg of total RNA was reverse transcribed to first-strand cDNA. Following the first-strand cDNA synthesis, cDNA amplification was performed with a hot start of 95 °C for 1 min followed by 15 cycles of 95 °C for 7 s, 66 °C for 20 s, and 72 °C for 5 min. The QIAquick PCR Purification Kit (Qiagen) was used to purify ds-cDNA, which was eluted with 20 µL to a final concentration of around 100–200 ng/µL. Approximately 800–1200 ng purified ds-cDNA was used as starting material for normalization. The best normalization result was achieved using 0.5 µL double-strand nuclease (DSN)

enzyme; therefore, a 0.5  $\mu$ L DSN normalization tube was used for normalized cDNA for the first and second amplification. After the second normalized cDNA amplification, 50  $\mu$ L of amplified normalized cDNA was used for proteinase K digestion following the Clontech Creator SMART cDNA Library Construction Kit manual. After transformation, twenty clones were randomly selected for checking insert size.

ESTs were sequenced by the Compositae Genome Project (<http://compgenomics.ucdavis.edu/>) using ABI 3730 machines at the Joint Genome Institute, Walnut Creek, CA. Phred basecalling, masking and trimming was conducted using the CGPdb bioinformatic pipelines (<http://cgpdb.ucdavis.edu/cgpdb2/>). The 25 711 noug transcriptome reads were then submitted to NCBI GenBank (GE551264.1 GI:211701855 to GE576974.1 GI:211733899). Prior to assembly of noug transcriptome reads, vector and low-quality sequences were removed using Seqclean (<http://compbio.dfci.harvard.edu/tgi/software/>) with the UniVec database (<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>). To reduce the impact of repetitive elements on assembly quality, repeats were masked with the library-less repeat masker RBR (Malde *et al.* 2006) using the default settings. Contigs were assembled for the EST collection using the program TGICL with default settings (<http://compbio.dfci.harvard.edu/tgi/software/>) (Quackenbush *et al.* 2000), and a unigene file containing 17 538 assembled contigs and singletons was created. Gene ontology (GO) annotations for the unigenes were obtained through discontinuous MegaBlast searches against *Arabidopsis thaliana* transcripts from the *Arabidopsis* Information Resource TAIR (TAIR 7 released 25 April 2007) (Rhee *et al.* 2003) for the best hit with at least 100 bp and an *e* value of  $1 \times 10^{-10}$ . Using a Pearson Chi-squared test with simulated *P*-value, computed from 100 000 Monte Carlo simulations in R (R Development Core Team 2009), we tested for significant differences between the number of genes in each set of GOSlim categories between noug and a set of previously analysed, pooled EST databases from 18 other Compositae species (Barker *et al.* 2008) (Table S1, Supporting information).

To search for evidence for paleopolyploidy events in the EST library, the general methods of Barker *et al.* (2008) were followed. Briefly, duplicate gene pairs were identified, reading frames assigned, the rate of substitutions per synonymous site ( $K_s$ ) calculated, and phylogenies for each gene family constructed. To identify significant features in the age distribution of duplicated genes, a boot-strapped K-S goodness of fit test (Cui *et al.* 2006) was applied to assess if the distribution deviated from a simulated null. A maximum likelihood mixture model (McLachlan *et al.* 1999) was used to fit a range of 1–5 normal distributions to the data because peaks

produced by paleopolyploidy events are expected to approximately follow a Gaussian distribution (Barker *et al.* 2008). The mixture model with the lowest Bayesian information criterion (BIC) was selected as the best fit.

#### Microsatellite marker development

One thousand four hundred and thirty-three simple sequence repeats (SSRs) were initially identified from unigenes in the EST library using the online software tool SSRIT (Temnykh *et al.* 2001), as well as a custom Perl script (N. Kane, unpublished). Tetra- and Tri-nucleotides were preferentially selected, as their alleles are commonly easier to distinguish and therefore scoring is facilitated. Primers were designed using PRIMER3 v 0.4.0 (Rozen & Skaletsky 2000).

PCR was performed in 15  $\mu$ L reaction volumes containing  $1 \times$  HF PCR buffer (Finnzymes), 1.5 mM  $MgCl_2$ , 200  $\mu$ M of each dNTP, 0.05  $\mu$ M forward primer, 0.5  $\mu$ M reverse primer, 0.5  $\mu$ M of dye-labelled universal primer, 0.15 U Phusion DNA polymerase, and 5–10 ng genomic DNA. Fragments were amplified using one of two PCR touch-down programs:

- 1 TD-50: An initial denaturation step at 95 °C for 2 min was followed by 9 cycles of 94 °C for 30 s, 60 °C for 30 s (temp decreased by 1 °C for every cycle), and 72 °C for 45 s, followed by 29 cycles of 94 °C for 30 s, 50 °C for 30 s, 72 °C for 45 s, and finally followed by a final extension at 72 °C for 20 min.
- 2 TD-55: An initial denaturation step at 95 °C for 2 min was followed by 9 cycles of 94 °C for 30 s, 65 °C for 30 s (temp decreased by 1 °C for every cycle) and 72 °C for 45 s, followed by 29 cycles of 94 °C for 30 s, 55 °C for 30 s, 72 °C for 45 s, and finally followed by a final extension at 72 °C for 20 min.

The repeat motif and primer sequences of 43 selected microsatellites are shown in Table 2. A test set of 20 individuals from a population that was collected in the Ethiopian highlands in 2007 was genotyped using these SSR markers. Basic characteristics such as the number of alleles, size range of alleles as well as observed and expected heterozygosity were calculated using Arlequin 3.11 (Excoffier *et al.* 2005) (Table 2). All loci were tested for Hardy–Weinberg equilibrium (HWE) using the Guo & Thompson (1992) approach, as implemented by Arlequin 3.11 (1 000 000 steps in Markov Chain; 100 000 dememorization steps; 0.05 significance level) (Excoffier *et al.* 2005). Linkage disequilibrium was assessed between all pairs of loci using the Slatkin & Excoffier (1996) approach, as implemented by Arlequin 3.11 (5000 permutations; 5 initial conditions; 0.05 significance level) (Excoffier *et al.* 2005) (Table 2). Information on the

unigene origin is included in Table S2 (Supporting information).

#### *Chloroplast DNA sequencing and analyses*

In preparation for sequencing the chloroplast genome, ten noug plants (USDA-ARS accession PI 508077 from Ethiopia) were grown under controlled environmental conditions in a growth chamber (Conviron E15, Winnipeg, MB, Canada) with a daily regime of 20 °C and 16-h light. Chloroplast isolations followed the general protocol of SIGMA's Chloroplast Isolation Kit (CP-ISO). Plants were kept in the dark for 24 h before harvesting 20 g of leaf tissue. Mid-rib veins were removed, and the leaf tissue was cut into small pieces and placed in the chloroplast isolation buffer (CIB; 0.33 M Sorbitol, 50 mM Tricine-OH pH 7.9, 2 mM EDTA, 1 mM MgCl<sub>2</sub>, 1 mM DTT, and 0.1% BSA) at a ratio of 4 mL/g of tissue. Tissue was macerated using a small blender (Philips, Amsterdam, The Netherlands). The macerate was filtered through two layers of Miracloth (Calbiochem, San Diego, CA, USA). The filtrate was collected and evenly divided among eight centrifuge tubes. All steps were carried out at 4 °C and low light conditions unless otherwise noted. To remove unwanted cell debris, the tubes were centrifuged (2100R; Thermo electron corporation, Waltham, MA, USA) at 200 g for 3 min. The supernatant was transferred to a clean tube and centrifuged again at 1000 g for 7 min to pellet the chloroplasts. The supernatant was discarded and the pellet was re-suspended in 1 mL of CIB. Chloroplasts were purified by centrifugation at 3200 g for 15 min through a 40/80% Percoll (Sigma-Aldrich, St. Louis, MO, USA) and CIB gradient. Purified intact chloroplasts were harvested from the interface between 40% and 80%. The band was re-suspended in three volumes of CIB (without BSA) and centrifuged at 1700 g for 1 min. The pellet was re-suspended in 0.5 mL of CIB (without BSA) and stored in the dark on ice. Chloroplast DNA was extracted by adding 400 µL of Lysis Buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 0.1% SDS, 0.1 M NaCl) to 200 µL of isolated purified chloroplasts and vortexed briefly. The mixture was incubated for 10 min at 65 °C. To the lysate, 130 µL of protein precipitation buffer (3 M potassium/5 M acetate) was added and incubated on ice for 5 min. The lysate mixture was then centrifuged (Eppendorf 5414D, Westbury, NY, USA) at 20 000 g for 5 min. The supernatant was transferred to a fresh tube, and 1.5 volumes of binding buffer (1 M Guanidine hydrochloride) were added and mixed. This mixture was loaded onto a silica column (Epochbiolabs, Sugarland, TX, USA) and centrifuged for 1 min at 6000 g. The membrane was washed twice with 500 µL of 70% ethanol (20 000 g, 2 min). The membrane was then incu-

bated for 5 min with 100 µL of elution buffer (TE Buffer) and centrifuged at 6000 g for 1 min.

Library construction and Illumina sequencing were performed at the Genome Sciences Centre (GSC) at the British Columbia Cancer Agency in Vancouver, Canada, using an Illumina Genome Analyser GA-II. A single lane of one flow cell was used for sequencing. Reads of 36 bp were generated and 8.2 million reads passed the default quality filtering. The initial Illumina assembly was used together with the sunflower cp genome to design 26 primer pairs, 19- to 30-bp long for gap filling and sequence confirmation by standard Sanger sequencing.

We assembled the 36-bp sequences using both VCAKE with the commands: -k 36 -n 19 -m 15 -v 10 -t 5 -e 22 (Jeck *et al.* 2007) and VELVET (Zerbino & Birney 2008). Initial assemblies using VELVET used a hash length of 19, minimum contig length of 100 bp, and minimum average coverage of 6×. These contigs were combined and extended using CAP3 (Huang 1996), and were subsequently aligned against the sunflower chloroplast genome (NC\_007977) using BLAST. Contigs with at least 10-bp identical overlap were joined. Remaining gaps and regions of low complexity were sequenced using Sanger sequencing. The full-length chloroplast sequence was annotated using DOGMA (Dual Organellar GenoMe Annotator) (Wyman *et al.* 2004), with additional information about splice sites and open reading frames provided from comparisons with the well-annotated *Helianthus* chloroplast genome (Timme *et al.* 2007). The resulting annotation was illustrated using OGDRAW (Lohse *et al.* 2007). Genome rearrangements, insertions and deletions were illustrated using zPicture (Ovcharenko *et al.* 2004).

## Results and Discussion

### *Analyses of EST library*

The 25 711 noug transcriptome reads were assembled in to 17 538 contigs and singletons (unigene file available from <http://msbarker.com/>). Of this set of unigenes, we were able to functionally annotate 4781 unigenes using the Gene Ontology database at TAIR (Rhee *et al.* 2003). The percentage of genes that were included in each GO-Slim category for the noug EST database is displayed in Table 1. To facilitate comparisons, Table 1 also shows the percentage of genes for each GOSlim category for a pooled set of EST databases from 18 other members of the Compositae (Barker *et al.* 2008; Laitinen *et al.* 2005). Those EST libraries had been prepared using the same normalization procedures as described previously. A Chi-squared test revealed that the order of GOSlim categories is significantly different ( $\chi^2 = 2317.652$ ;  $P = 0.0001$ ), when ordered according to the number of annotated genes in each category. Overall, the

**Table 1** GOSlim annotations for the noug EST database and a pooled sample of other Compositae EST databases (Barker *et al.* 2008; Laitinen *et al.* 2005)

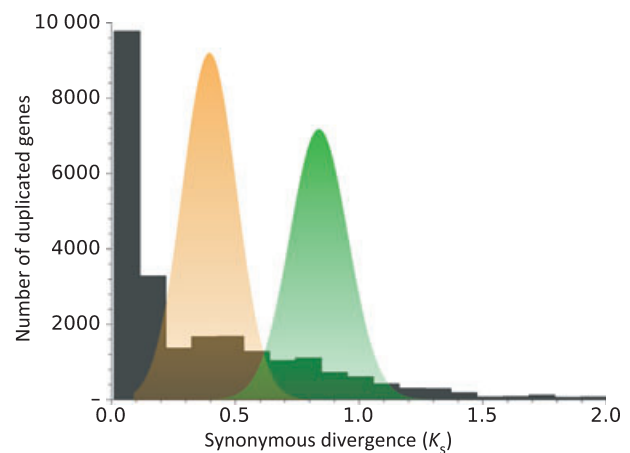
GOSlim categories	Noug	Pooled Compositae
	Gene count (%)	Gene count (%)
GO cellular component		
Other intracellular components	16.70	17.61
Other cytoplasmic components	14.21	10.66
Chloroplast	12.04	10.05
Other membranes	10.01	14.75
Plastid	6.22	3.61
Plasma membrane	8.22	2.07
Nucleus	6.49	9.38
Unknown cellular components	7.39	17.99
Cytosol	3.89	1.94
Ribosome	2.55	2.39
Mitochondria	3.68	4.68
Other cellular components	2.71	0.66
Cell wall	2.04	0.87
Extracellular	1.48	0.36
ER	1.19	1.48
Golgi apparatus	1.19	1.49
GO molecular function		
Other enzyme activity	14.46	12.58
Transferase activity	9.93	11.15
Other binding	10.59	7.93
Hydrolase activity	8.70	13.00
Kinase activity	4.98	6.29
Nucleotide binding	9.15	5.87
Protein binding	8.87	7.06
Unknown molecular functions	9.33	9.25
Transporter activity	5.02	6.32
DNA or RNA binding	5.63	6.79
Structural molecule activity	3.93	2.81
Other molecular functions	3.43	2.94
Nucleic acid binding	3.07	2.18
Transcription factor activity	2.74	5.44
Receptor binding or activity	0.17	0.39
GO biological process		
Other cellular processes	25.04	22.99
Other metabolic processes	22.72	24.24
Protein metabolism	9.26	9.79
Response to abiotic or biotic stimulus	5.29	3.39
Unknown biological processes	7.65	10.77
Transport	5.23	5.24
Response to stress	4.85	3.00
Developmental processes	4.32	3.63
Other biological processes	4.94	2.79
Cell organization and biogenesis	4.04	5.09
Signal transduction	2.08	2.20
Transcription	2.12	3.80
Electron transport or energy pathways	1.85	1.80
DNA or RNA metabolism	0.61	1.27

proportional differences are relatively small and probably reflect variation among species in the tissues and developmental stages employed for sequencing rather than differences in gene content (Tables 1 and S1).

Our analyses of the age distribution of duplicated genes in the noug EST library revealed two major peaks that likely result from paleopolyploidization (Fig. 1). The youngest peak is centred at  $K_s \sim 0.4$ – $0.5$ , whereas the older peak's median is located at  $K_s \sim 0.8$ . The young peak was previously found in all six species of the sunflower genus, *Helianthus*, and Barker *et al.* (2008) suggested that it resulted from a paleopolyploidization at the base of the Heliantheae tribe, to which *Guizotia abyssinica* also belongs. The presence of the peak in *Guizotia* supports this hypothesis. The older peak is shared by all members of the Compositae that have so far been analysed (Barker *et al.* 2008), indicating its position near the base of the phylogeny or before the divergence of extant Compositae.

#### Microsatellite analyses

From the total set of 1433 microsatellites identified in the noug ESTs library, 206 were screened and 43 were selected on the basis of consistent amplification and evidence of polymorphism in our test population. The one exception was GA082 which only showed evidence of polymorphism when tested on a different population (data unpublished). All 43 loci should be suitable for a range of applications in the molecular ecology and breeding, especially phylogeographic and population genetic studies of noug and its wild relatives. Excluding GA082, the number of microsatellite alleles varied between 2 and 10 (average 4.67) and the average observed and expected heterozygosities were 0.49 and 0.54, respectively



**Fig. 1** Histogram of noug gene duplication ages and fitted mixture model analyses. The yellow peak represents the putative Heliantheae paleopolyploidization and the green peak represents the putative basal Compositae paleopolyploidization.

**Table 2** Characteristics of microsatellites

Locus	Repeat motif	Forward Primer sequence	Reverse primer sequence	Num. of alleles	Allele size ranges	Observed Heterozyg.	Expected Heterozyg.	HWE	Linked loci	PCR cond.
GA003	gat	CGCCCTAAAGCTACTTTCCTCC	CACACTGCACACTAGGAACCTTC	2	399–402	0.05	0.05	Yes	GA127, GA238	TD-55
GA012	gat	CAGTAAAGCTCGGTATCTCCAAGTT	AGAAGATCTCGTCAGCAGAAACAG	2	263–275	0.2	0.18	Yes	GA107, GA138	TD-55
GA013	ctt	GGTAATGGTAATGGAGGTTCTGG	CCTCATCAGAGTCTTCGGGTTAT	9	424–455	0.9	0.86	Yes	None	TD-55
GA018	agc	GTTCAGCCCATGATCATAAT	CTATCTCTATCTCGTGGGTTTTG	2	353–358	0.3	0.43	Yes	GA183, GA186, GA205	TD-50
GA029	atc & tc	CCATCATCAATGGCGTTACTC	GTCTCGTTC TAGAAGCTTCATCCT	3	270–276	0.35	0.47	Yes	GA108, GA143, GA186, GA214, GA217	TD-55
GA035	tga	GAITTCAGGTGAAGGA GAAGAG	GCCCTCCCTACAACATACTTGATA	3	301–307	0.35	0.30	Yes	GA107, GA144, GA217	TD-55
GA037	ta & gaa	GGTGTITTTGTAGTGGTCTGTC	GACTAGCCAGAAACCGAAGAATC	2	347–350	0.55	0.51	Yes	GA081	TD-55
GA054	ta	AACGGTTTAGGAGACCTTGG	TCACCTGGCTCAGACTTGTT	3	247–265	0.25	0.31	Yes	GA182, GA210, GA204, GA205	TD-55
GA055	ct	CCTGAAACAAGCCCAACAA	CAGTACATCGCGGAGAGAGG	3	194–200	0.25	0.41	No	GA191, GA205	TD-55
GA077	tc	TCAGCCAAAACATTCCAAAGC	AAACAACGGCTAAAACCGA	2	487–490	0.05	0.14	Yes	GA108, GA139, GA150, GA229, GA205, GA238, GA246	TD-55
GA081	tc	AATCTCGATTGGCTGAGTGG	AGGAAGTTGGGGCTTCGTAA	3	437–441	0.5	0.48	Yes	GA037, GA242	TD-55
GA082	tc	TGTCCGTATGAAACCCATTGA	CAATGATCATGGGACTGCT	1	197–197	0	N/A	N/A	N/A	TD-55
GA107	cct	ATCACCCCTCTCTCCAAAACCCAT	GAAGATCTAAATCCAGCTCCCTG	4	220–232	0.2	0.35	No	GA012, GA035, GA108, GA183	TD-55
GA108	cca	ATGGCCTCCACCTTCCTCT	GAGTGATAATCCGGTGCTAAGACT	4	194–209	0.15	0.58	No	GA029, GA107, GA077, GA188, GA214, GA220	TD-55
GA117	cac	CCCTTCATCCAAITTTAACGAC	AGGTCTAATCCCAGCCCTCTCTAAT	2	336–339	0.25	0.22	Yes	GA210, GA127, GA214, GA242	TD-55
GA127	cct	CAATCTGCAACTACTGCCAATACC	CCAGTCAGAAACCCCTTGATCACTA	2	213–216	0.05	0.05	Yes	GA003, GA117	TD-55
GA138	aag	ATCAACTTCCCATAATACCTCTGG	CTTCTCTGTCACCTCTTTTGGAC	5	363–378	0.65	0.57	Yes	GA012, GA035, GA108, GA183	TD-55
GA139	gaa	GTACATCCCAACTTTACCATCCAC	CTCTACAACCAACACCACTTCC	7	223–241	0.75	0.69	Yes	GA077, GA238	TD-55
GA143	tga	GGATGGTGTACTTCTTCTGACCT	TAGCCGCGTAACATACGAGTCT	7	296–312	0.75	0.79	Yes	GA029, GA182, GA165, GA172, GA190	TD-55
GA144	agtt	GGTCCCAAAAACCAATATGATG	CTAGGGCTTGTACCAACACCTTAAA	5	331–347	0.35	0.39	Yes	GA035	TD-50
GA150	acc	GTAATGACTTGTGAGGAACACGAC	GGGTTGGAGGTACAGTGAAGAT	8	279–298	0.6	0.81	No	GA162, GA182, GA077, GA204	TD-55
GA156	aag	CCAGTTGTGAGAAITTCACCGTGT	GAGCTCCAGGTCTCTAGGGTTATC	3	158–173	0.7	0.54	Yes	GA220	TD-55
GA162	cca	AGCCACTCTCTTGTGTTACC	CAAGTTCTGGTGGTGGTATG	3	134–140	0.3	0.27	Yes	GA150, GA210	TD-55

Table 2 Continued

Locus	Repeat motif	Forward Primer sequence	Reverse primer sequence	Num. of alleles	Allele size ranges	Observed Heterozy.	Expected Heterozyg.	HWE	Linked loci	PCR cond.
GA165	gat	GGGTACCTACGTACTGGA AACAAAG	TCCTTTGGAAAAATCCCTTCC	4	280–289	0.7	0.67	Yes	GA143, GA205	TD-50
GA172	tca	AAGAACAAGGGAGAGTGGAT	AGGAGTTGTGAGGACAAAATG	10	233–354	0.75	0.80	Yes	GA143, GA214	TD-50
GA182	aaagt & ctaag	GAAAGACAACGACTGGAATG	TGTTTCCTTAAAGGTACC	7	363–393	0.65	0.79	Yes	GA054, GA150, GA143, GA204, GA205, GA220	TD-50
GA183	atg	ATAGGGTTAGGGTTCCATGT	CCTCTTCTTCATCATCAATCG	9	280–317	0.6	0.82	No	GA107, GA018, GA186, GA192	TD-50
GA186	atg	CTCCCAAGAGAATCAAACAG	GTCATTCTGCGCAATAACTC	7	414–438	0.7	0.64	Yes	GA029, GA018, GA183	TD-50
GA188	cac	GTGCTTCCCCTACTCAITTA	GGCAGTTTCATCCATGTACT	5	339–351	0.85	0.72	Yes	GA108	TD-50
GA190	cca	CACCTCAGTTGTCACCTTCT	GAGAGTGGCTGAATGGATTA	8	364–396	0.85	0.75	Yes	GA143, GA204	TD-55
GA191	tca	CCCACCAACCTATATCTTC	GTGGAAACAGAACTCCAT	5	265–277	0.65	0.76	Yes	GA055	TD-55
GA192	cac	AACACCAAGATCAGTGGCT	CACCTATCTCCATTTGCC	5	247–262	0.55	0.67	Yes	GA183, GA204	TD-50
GA204	tga	GGAAGAAGAAGAGGATGGT	CAACATTTACCAGCGTTCTC	5	212–224	0.6	0.65	Yes	GA054, GA150, GA182, GA190, GA192	TD-55
GA205	cat	CCTGGCCTTCTCTAATCTT	GGTCATGATGGTGATGATG	6	378–408	0.5	0.75	No	GA054, GA182, GA055, GA018, GA077, GA165, GA220	TD-55
GA210	atc	ACAACACCACAACACTACTCC	GGTGGACTGATTTGAAGAGA	6	340–366	0.6	0.58	Yes	GA054, GA117, GA162	TD-55
GA214	cca	ATATCGTTAGAGTTCGTGGG	CGGTTCTTGTGCTTGTACTTC	6	394–412	0.5	0.50	Yes	GA029, GA108, GA117, GA172	TD-55
GA217	acc & cca	CACCACCACCTACCTACCTA	GATTGTGAGGGAGAACAAAGA	6	290–305	0.7	0.79	Yes	GA029, GA035	TD-55
GA220	acc	CATAGCATCCTCTCCACCT	CCTTTACATCCTTTCTTCCC	4	409–421	0.65	0.63	Yes	GA108, GA156, GA182, GA205, GA228	TD-55
GA228	cct & ata & tg	GTTTCCCTCACCTCTTTGAT	CATGGATCTGAAGACAAAACC	7	165–474	0.6	0.77	No	GA220	TD-55
GA229	gaa	GTAACATGAGCATCCCACAT	GTGAAAAGATCAGCAGTCCAT	5	284–299	0.7	0.68	Yes	GA077, GA238	TD-55
GA238	aga	ATCACAGTAGCACCAAAATCC	CATAAATCTCCCCACATGAC	3	237–243	0.75	0.56	Yes	GA003, GA139, GA229, GA077, GA246	TD-50
GA242	cac	CAGATTCTCTCCACAAAAG	GAGTGTCTATGAGCTTTGCC	3	479–485	0.35	0.30	Yes	GA081, GA117	TD-55
GA246	tct	CAATACTCGTCTCCTCTGCG	GGTAAACAATATCGGTGAGC	5	298–310	0.5	0.63	Yes	GA077, GA238	TD-55
Average				4.67		0.49	0.54			



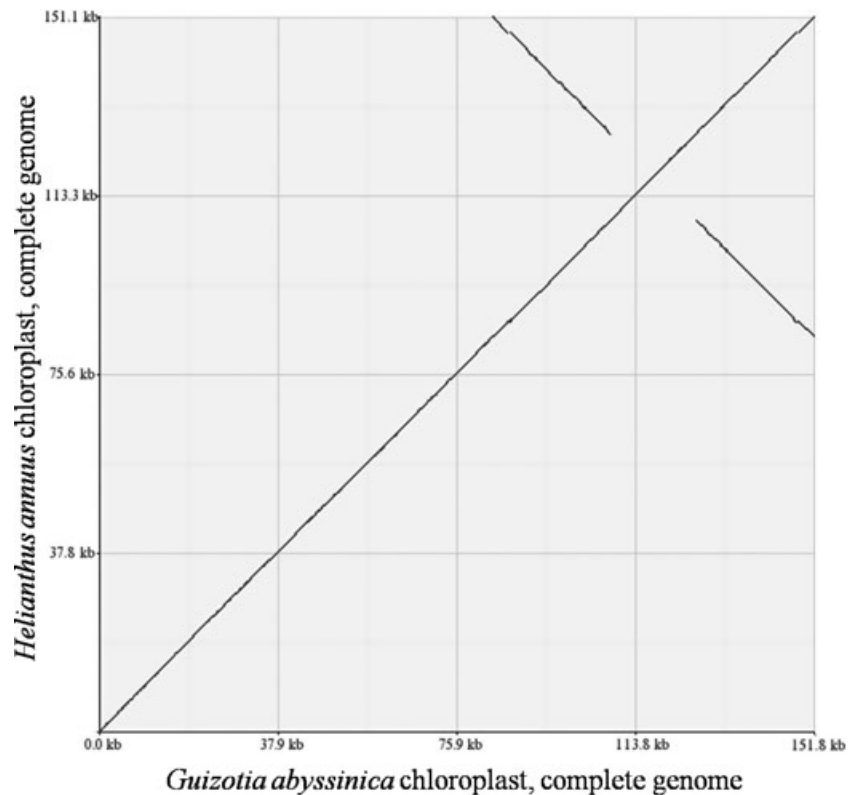
All 80 annotated coding sequences in the *H. annuus* chloroplast were identified in the *G. abyssinica* chloroplast genome, as well as all 34 tRNAs and 4 rRNA sequences (Fig. 2). There were no substantial rearrangements found between *H. annuus* and *G. abyssinica*, which were entirely collinear other than numerous small insertions and deletions (Fig. 3). Aligned sequences of the two chloroplasts showed over 96% identity, with 1.8% of the remaining sites showing different nucleotide sequences and 1.4% of sites having gaps because of small insertions and deletions. The sequence divergence and length differences between the two species are listed in Table 3 for a representative set of coding genes and intergenic regions. Similar information for shorter coding sequences as well as non-coding regions such as tRNA, rRNA, introns, and intergenic regions is provided in Table S3 (Supporting information). Of particular note are *rbcL*, *atpB*, *rpoC1*, *rpoB*, *matK*, and *trnH-psbA*, which have been used for phylogenetic analyses (APG II 2003; Shaw *et al.* 2005) and DNA barcoding (Lahaye *et al.* 2008) in the past. Of these sequences, *trnH-psbA* shows the highest divergence (15%), an order of magnitude faster than most other loci, indicating its potential utility in these and other taxa in the Helianthae, which with 2500 species comprises approximately 1% of all flowering plants. Additionally, the rapidly evolving intergenic regions identified by Timme *et al.* 2007 (Timme *et al.* 2007) show promise. Although they evolve slightly slower than *trnH-*

*psbA*, they are substantially longer, thus providing more characters. They are flanked by conserved regions, enabling a single set of primers to be useful in even widely divergent taxa (Timme *et al.* 2007).

Whole chloroplast sequencing using short reads is simplified when a high-quality scaffold for a closely related species may be leveraged, as in the present case. The low level of rearrangements in Compositae chloroplast genomes further simplified the scaffolded assembly. Unfortunately, some taxa show relatively high levels of chloroplast genome rearrangements (e.g. Campanulaceae; Cosner *et al.* 1997) that make the use of existing scaffolds problematic. However, the *de novo* assembly of fairly large contigs (100–4000 bp) in our study indicates that useful sequence information can be recovered with this approach even without a scaffold. The use of paired-end reads with variable insert sizes would probably enable the assembly of larger *de novo* contigs and provide improved assemblies.

Comparisons between *Guizotia* and *Helianthus* and *Lactuca* confirm previous research showing a lack of major rearrangements in the Compositae. One interesting difference is that the 456-bp deletion in *ycf2* identified in *Helianthus* by Timme *et al.* (2007) is not found in *Guizotia*, which is more similar to *Lactuca* in this respect, indicating that this deletion is probably limited to a smaller portion of the Helianthae. In most other major respects, the noug chloroplast is much more similar to *Helianthus*, as

**Fig. 3** Dotplot comparison showing conserved regions found in both *Guizotia* (x axis) and *Helianthus* (y axis) chloroplast genomes.



**Table 3** Protein coding genes and important non-coding sequences. Comparison of all long (> 1,000 bp) protein sequences in *Guizotia* and *Helianthus* chloroplast genomes. Dn and Ds were calculated using PAML, and the Arabidopsis best hit was identified using BLAST

Name	Dn	Ds	Dn/Ds	% identity	Indels	Length (noug)	Length (sun-flower)	Type of sequence
<i>rpoA</i>	0.007	0.014	0.471	0.991	0	1008	1008	Coding
<i>psbA</i>	0.001	0.059	0.020	0.987	0	1062	1062	Coding
<i>psbD</i>	0.003	0.009	0.287	0.996	0	1062	1062	Coding
<i>ndhA</i>	0.005	0.043	0.113	0.987	0	1092	1092	Coding
<i>ndhH</i>	0.001	0.052	0.021	0.989	0	1182	1182	Coding
<i>psbC</i>	0.002	0.045	0.044	0.987	0	1422	1422	Coding
<i>accD</i>	0.008	0.057	0.148	0.985	1	1452	1443	Coding
<i>rbcL</i>	0.005	0.076	0.071	0.982	0	1458	1458	Coding
<i>atpB</i>	0.004	0.023	0.164	0.992	0	1497	1497	Coding
<i>ndhD</i>	0.013	0.049	0.255	0.980	0	1503	1503	Coding
<i>matK</i>	0.047	0.069	0.674	0.980	0	1515	1503	Coding
<i>atpA</i>	0.004	0.020	0.218	0.992	0	1527	1527	Coding
<i>psbB</i>	0.005	0.039	0.117	0.988	0	1527	1527	Coding
<i>ndhB</i>	0.002	0.006	0.270	0.996	0	1533	1533	Coding
<i>rpoC1</i>	0.004	0.038	0.113	0.989	0	2091	2091	Coding
<i>psaB</i>	0.005	0.028	0.184	0.991	0	2205	2205	Coding
<i>ndhF</i>	0.013	0.073	0.178	0.974	1	2226	2232	Coding
<i>psaA</i>	0.002	0.028	0.086	0.992	0	2253	2253	Coding
<i>rpoB</i>	0.003	0.031	0.094	0.991	0	3183	3183	Coding
<i>rpoC2</i>	0.007	0.046	0.154	0.985	0	4089	4089	Coding
<i>ycf1</i>	0.015	0.007	2.028	0.960	7	5040	5118	Coding
<i>ycf2</i>	0.004	0.007	0.589	0.996	10	6831	6396	Coding
<i>psbA-trnH</i>	–	–	–	0.842	7	390	387	Non-coding
<i>ndhC-trnV</i>	–	–	–	0.922	7	1024	883	Non-coding
<i>trnL-rpl32</i>	–	–	–	0.858	9	733	781	Non-coding
<i>trnY-rpoB</i>	–	–	–	0.935	6	1150	1167	Non-coding

expected. For instance, although there are several small insertions and deletions when *accD* is compared between the two taxa, neither contains the 25-aa insertion found in *Lactuca* (Timme *et al.* 2007). In most coding regions, however, all three taxa are nearly identical. Chloroplast sequence variation has been shown to work well for phylogeographic studies in many plant species. The prospect of using whole chloroplast genomes instead of single sequence fragments will allow for much more comprehensive assessments of phylogeographic patterns and hence enable researchers to take these studies to the next level.

### Acknowledgements

The EST library for noug was developed by the Compositae Genome Project, which is supported by a U.S. National Science Foundation Plant Genome grant to LHR. We thank the JGI team for sequencing and A. Kozik and R. Michelmore for processing the sequences and submitting them to GenBank. Microsatellite development was supported by a Canadian International Development Agency project grant to HD, JE and LHR, coordinated by Bioversity International. The chloroplast genome sequencing was supported by a Natural Sciences and Engineering Research Council (NSERC, Canada) Discovery grant to QCB Cronk.

### References

- APG II (2003) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society*, **141**, 399–436.
- Barker MS, Kane NC, Matvienko M *et al.* (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution*, **25**, 2445–2455.
- Bekele E, Geleta M, Dagne K *et al.* (2007) Molecular phylogeny of the genus *Guizotia* (Asteraceae) using DNA sequences derived from ITS. *Genetic Resources and Crop Evolution*, **54**, 1419–1427.
- Bouck A, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology*, **16**, 907–924.
- Bungard RA (2004) Photosynthetic evolution in parasitic plants: insight from the chloroplast genome. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, **26**, 235–247.
- Cosner ME, Jansen RK, Palmer JD, Downie SR (1997) The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Current Genetics*, **31**, 419–429.
- Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research*, **36**, e122.

- Cui L, Wall PK, Leebens-Mack JH *et al.* (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Research*, **16**, 738–749.
- Dempewolf H, Rieseberg LH, Cronk QC (2008) Crop domestication in the Compositae: a family-wide trait assessment. *Genetic Resources and Crop Evolution*, **55**, 1141.
- Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.
- Geleta M, Bryngelsson T, Bekele E, Dagne K (2007) Genetic diversity of *Guizotia abyssinica* (L.f.) Cass. (Asteraceae) from Ethiopia as revealed by random amplified polymorphic DNA (RAPD). *Genetic Resources and Crop Evolution*, **54**, 601–614.
- Geleta M, Bryngelsson T, Bekele E, Dagne K (2008) Assessment of genetic diversity of *Guizotia abyssinica* (L.f.) Cass. (Asteraceae) from Ethiopia using amplified fragment length polymorphism. *Plant Genetic Resources*, **6**, 41–51.
- Getinet A, Sharma SM (1996) Niger, *Guizotia abyssinica* (L.f.) Cass. In: *Promoting the Conservation and Use of Underutilized and Neglected Crops* (IPGRI) Vol. no. 5, 59 p, IPGRI, Rome, Italy.
- Grevich JJ, Daniell H (2005) Chloroplast genetic engineering: recent advances and future perspectives. *Critical Reviews in Plant Sciences*, **24**, 83–107.
- Guo S, Thompson E (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361–372.
- Huang X (1996) An improved sequence assembly program. *Genomics*, **33**, 21–31.
- Jansen RK, Raubeson LA, Boore JL *et al.* (2005) Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods in Enzymology*, **395**, 348–384.
- Jansen RK, Cai Z, Raubeson LA *et al.* (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19369–19374.
- Jeck WR, Reinhardt JA, Baltrus DA *et al.* (2007) Extending assembly of short DNA sequences to handle error. *Bioinformatics*, **23**, 2942–2944.
- Lahaye R, van der Bank M, Bogarin D *et al.* (2008) DNA barcoding the floras of biodiversity hotspots. *PNAS*, **105**, 2923–2928.
- Laitinen RA, Immanen J, Auvinen P *et al.* (2005) Analysis of the floral transcriptome uncovers new regulators of organ determination and gene families related to flower organ differentiation in *Gerbera hybrida* (Asteraceae). *Genome Research*, **15**, 475–486.
- Lohse M, Drechsel O, Bock R (2007) OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics*, **52**, 267–274.
- Malde K, Schneeberger K, Coward E, Jonassen I (2006) RBR: library-less repeat detection for ESTs. *Bioinformatics (Oxford, England)*, **22**, 2232–2236.
- McLachlan GJ, Peel D, Basford KE, Adams P (1999) The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*, **4**, 1–14.
- Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L (2004) zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Research*, **14**, 472–477.
- Quackenbush J, Liang F, Holt I, Pertea G, Upton J (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research*, **28**, 141–145.
- R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. Vienna, Austria.
- Rhee SY, Beavis W, Berardini TZ *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, **31**, 224–228.
- Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology (Clifton, NJ)*, **132**, 365–386.
- Shaw J, Lickey EB, Beck JT *et al.* (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, **92**, 142–166.
- Slatkin M, Excoffier L (1996) Testing for linkage disequilibrium in genotypic data using the EM algorithm. *Heredity*, **76**, 377–383.
- Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Research*, **11**, 1441–1452.
- Timme RE, Kuehl JV, Boore JL, Jansen RK (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.
- Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics (Oxford, England)*, **20**, 3252–3255.
- Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

## Supporting Information

Additional supporting information may be found in the online version of this article.

**Table S1** GOSlim category comparison between noug and a sample of Compositae

**Table S2** Unigene origins of microsatellites

**Table S3** Non-coding sequences and short protein-coding sequences

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.