MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) doi: 10.1111/1755-0998.12605

Development of genomic tools in a widespread tropical tree, *Symphonia globulifera* L.f.: a new low-coverage draft genome, SNP and SSR markers

SANNA OLSSON,* PEDRO SEOANE-ZONJIC,† ROCÍO BAUTISTA,† M. GONZALO CLAROS,† SANTIAGO C. GONZÁLEZ-MARTÍNEZ,*‡ IVAN SCOTTI,§ CAROLINE SCOTTI-SAINTAGNE,§ OLIVIER J. HARDY¶ and MYRIAM HEUERTZ*‡¶

Department of Forest Ecology and Genetics, INIA Forest Research Centre (INIA-CIFOR), Carretera de A Coruña km 7.5, E-28040 Madrid, Spain, †Departamento de Biología Molecular y Bioquímica, and Plataforma Andaluza de Bioinformática, Universidad de Málaga, calle Severo Ochoa 34, E-29590 Campanillas, Málaga, Spain, ‡UMR1202 BioGeCo, INRA, Univ. Bordeaux, 69 route d'Arcachon, F-33610 Cestas, France, §INRA, UR629 URFM, Ecologie des Forêts Méditerranéennes, Site Agroparc, Domaine Saint Paul, F-84914 Avignon Cedex 9, France, ¶Faculté des Sciences, Evolutionary Biology and Ecology, Université Libre de Bruxelles, Av. F.D. Roosevelt 50, CP 160/12, B-1050 Brussels, Belgium

Abstract

Population genetic studies in tropical plants are often challenging because of limited information on taxonomy, phylogenetic relationships and distribution ranges, scarce genomic information and logistic challenges in sampling. We describe a strategy to develop robust and widely applicable genetic markers based on a modest development of genomic resources in the ancient tropical tree species *Symphonia globulifera* L.f. (Clusiaceae), a keystone species in African and Neotropical rainforests. We provide the first low-coverage (11X) fragmented draft genome sequenced on an individual from Cameroon, covering 1.027 Gbp or 67.5% of the estimated genome size. Annotation of 565 scaffolds (7.57 Mbp) resulted in the prediction of 1046 putative genes (231 of them containing a complete open reading frame) and 1523 exact simple sequence repeats (SSRs, microsatellites). Aligning a published transcriptome of a French Guiana population against this draft genome produced 923 high-quality single nucleotide polymorphisms. We also preselected genic SSRs in silico that were conserved and polymorphic across a wide geographical range, thus reducing marker development tests on rare DNA samples. Of 23 SSRs tested, 19 amplified and 18 were successfully genotyped in four *S. globulifera* populations from South America (Brazil and French Guiana) and Africa (Cameroon and São Tomé island, F_{ST} = 0.34). Most loci showed only population-specific deviations from Hardy–Weinberg proportions, pointing to local population effects (e.g. null alleles). The described genomic resources are valuable for evolutionary studies in *Symphonia* and for comparative studies in plants. The methods are especially interesting for widespread tropical or endangered taxa with limited DNA availability.

Keywords: Clusiaceae, draft genome, microsatellites, single nucleotide polymorphisms, transcriptomic

Received 19 July 2016; revision received 30 September 2016; accepted 4 October 2016

Introduction

The study of evolutionary processes in tropical plants is often hampered by poor knowledge on distribution ranges of taxa, scarcity of known discriminant phenotypic characters, and lack of genetic information, as well as by logistical hurdles in the transfer of biological material from field to bench (Raven & Wilson 1992; Goodwin *et al.* 2015). Incomplete knowledge on taxonomy and distribution ranges represents challenges for sampling design and marker choice. For example, the presence of cryptic species can increase the expected phylogenetic depth of the study group (e.g. Turchetto-Zolet *et al.* 2013; Heuertz *et al.* 2014). If access to fresh plant material is difficult, herbarium vouchers can represent complementary sources of DNA, although low DNA concentration and sometimes heavy degradation restrain this option (Ribeiro & Lovato 2007; Särkinen *et al.* 2012). Such complications can be mitigated by a careful choice of materials for adapted marker design and validation. We here illustrate how the *ad hoc* choice of a few samples for transcriptomic and genomic sequencing in a tropical tree

Correspondence: Sanna Olsson, Fax: +34 913476767; E-mail: sanna.olsson@helsinki.fi and Myriam Heuertz, Fax: +33 (0)55 7122881, E-mail: myriam.heuertz@inra.fr

species resulted in the discovery of high-quality single nucleotide polymorphisms (SNPs) and the development of robust microsatellite markers, which are applicable across a wide range of sampling regions and heterogeneous sample qualities.

Genetic marker development can be greatly facilitated by the availability of a collection of genomic sequences or a draft genome (Primmer 2009; Ekblom & Wolf 2014). The sequencing of a low-coverage draft genome is now within reach even in relatively small projects, opening promising avenues in evolutionary research in nonmodel organisms (Ekblom & Wolf 2014). De novo assembly of large genomes from low-coverage data remains challenging; for example due to high heterozygosity or repetitive elements (Claros *et al.* 2012; Schatz *et al.* 2012), and functional characterization can be difficult (Tagu *et al.* 2014). Nevertheless, low-coverage genome sequencing has repeatedly allowed the characterization of genome repeat content, the discovery and annotation of single- or low-copy genes and the development of organellar or nuclear genetic markers such as microsatellites or amplicons for population genetic or phylogenetic inference (Straub *et al.* 2011; Leese *et al.* 2012; Blischak *et al.* 2014; Gardner *et al.* 2016). Draft genomes are especially useful when aligned with transcriptome sequences, making it possible to characterize intron–exon boundaries and thus accurately annotate genes of interest, for example candidates of gene expression studies, and to design hybridization baits for targeted enrichment sequencing studies (Weitemier *et al.* 2014).

Microsatellite markers (simple sequence repeats, SSRs) continue being popular and robust tools for population genetic analyses in nonmodel organisms due to their abundance, high polymorphism, codominant inheritance, ease of use, moderate cost and modest requirements on DNA quality (Nybom *et al.* 2014). Homoplasy and null alleles hamper their analysis but can be addressed using specific tools (Van Oosterhout *et al.* 2006; Chapuis & Estoup 2007) and can even represent sources of phylogenetic information (e.g. Barthe *et al.* 2012). Transcriptome-based SSRs are a popular alternative to anonymous genomic SSRs. They often retain phylogenetic signals better, making them especially suitable for the study of species complexes (Tabbasam *et al.* 2014). Also, because they are based on expressed sequences, they may contain signatures of natural selection, informing on adaptive processes (e.g. Bradbury *et al.* 2013; Xia *et al.* 2014).

In this study, we describe the development of a low-coverage draft genome in the tropical tree *Symphonia globulifera* L.f. (Clusiaceae), a keystone tree in African and Neotropical rainforests and illustrate how it can be used in conjunction with a sequenced transcriptome to discover single nucleotide polymorphism markers (SNPs) and to develop robust polymorphic SSR markers. Alignment of a transcriptome with a draft genome makes it possible to identify intron–exon boundaries and to detect ambiguously aligned transcripts, which is helpful for marker development. Choosing remote source populations for genome and transcriptome (Cameroon for the genome and French Guiana for the transcriptome in our case) makes it possible to screen for robust and polymorphic markers in silico, reducing the number of marker tests on rare DNA extracts. The targeting of short regions, multiplex genotyping and the use of whole-genome amplification of test samples allowed us to reduce the amount of template DNA and to enhance genotyping success on heterogeneous DNA qualities. The approach we describe is especially interesting for researchers working with valuable population samples, such as those interested in population and conservation genetics of widespread tropical or endangered species. The *S. globulifera* genome draft presented here is an important resource for further genetic marker development in *Symphonia*, to elucidate the particular biogeographic and evolutionary history of the genus, which besides harbouring the widespread *S. globulifera*, has undergone a radiation in Madagascar (Perrier de la Bâthie 1951; Dick *et al.* 2003; Dick & Heuertz 2008; Budde *et al.* 2013). *Symphonia globulifera* is used for its timber (trade name 'manil', 'manni', 'ossol' (Gabon) or 'boarwood', Oyen 2005); thus, the genome draft can be useful to explore the genetic basis of wood properties (Clair *et al.* 2003) and other adaptive traits (e.g. González-Martínez *et al.* 2007). Furthermore, *S. globulifera*, other *Symphonia* species and related Clusiacae are commonly used in traditional medicine and represent thus a reservoir of bioactive molecules (e.g. Boiteau 1986; Boiteau *et al.* 1999; Fromentin *et al.* 2015) the investigation of which can greatly benefit from the availability of a draft genome.

Material and methods

Study species

Symphonia globulifera L.f. (Clusiaceae) is a 25- to 40-m-tall late successional tree in evergreen mixed humid forests with a geographical range from Guinea Bissau to Tanzania in continental Africa (but not Madagascar) and from Mexico to Brazil in America. It has a wide ecological amplitude and grows in forests from sea level to 2600 m (Oyen 2005). *Symphonia* is a very old genus with a probable origin in continental Africa or Madagascar (Dick *et al.* 2003). The oldest fossil pollen records were found in Nigeria and date to the mid-Eocene (ca. 45 Ma, Jan du Chêne & Salami 1978). The genus consists of a further 16–23 species, all endemic to Madagascar (Perrier de la Bâthie 1951; Abdul-Salim 2002).

Tissue collection and DNA extraction

For genome sequencing, cambium was sampled from a single individual from Nkong Mekak in Cameroon (sample ID MH2382, Lat 2.77°, Lon 10.54°, 433 m a.s.l.). The published transcriptome we used was obtained from leaves and stems of two seedlings from a French Guiana population (Brousseau *et al.* 2014). Plant material (cambium or leaves) for SSR genotyping was collected from 31 to 32 randomly sampled adult individuals in each of four populations (total $N = 125$ individuals, Table 1) in South America [Paracou (Lat 5.30°, Lon −52.88°), French Guiana, and Ituberá (Lat −13.80°, Lon −39.18°), Brazil] and Africa [Nkong Mekak (Lat 2.80°, Lon 10.54°), Cameroon, and island of São Tomé (Lat 0.27°, Lon 6.55°), São Tomé and Principe] and dried in silica gel. The samples from Paracou represented two separate morphotypes (Baraloto *et al.* 2007). SSR testing prior to genotyping was performed on five individuals, one from each mentioned population and one additional individual from Benin (Lat 6.39°, Lon 2.62°). For genome sequencing, DNA was extracted using the DNeasy Plant mini Kit (Qiagen, the Netherlands) separately in ten reactions and then combined. The extracted high molecular weight DNA was directly used for Illumina sequencing. As DNA was of insufficient quantity for subsequent 454 pyrosequencing, it was whole-genome amplified by multiple strand displacement (MDA, using REPLI-g mini kit, Qiagen) in ten separate reactions increasing concentration 10- to 100-fold prior to pooling for library construction. MDA has been chosen because it is a whole-genome amplification technology commonly used in human genetics with a demonstrated low bias when using minute DNA quantities, for example a chimera rate as low as 2% on single cells (Murphy *et al.* 2012; Huang *et al.* 2015). For SSR genotyping, samples were extracted using the DNeasy Plant mini Kit or the Invisorb DNA Plant HTS 96 kit (Stratec Molecular, Germany). Because DNA samples were of heterogeneous quantity and quality due to variable sources of plant material and because they were also used for other studies, they were whole-genome amplified (REPLI-g) prior to SSR-typing (except samples for initial SSR tests).

Genome sequencing, de novo assembling and annotation

Illumina HiSeq 2000 sequencing (2 × 100 bp) of one paired-end library (½ lane) with 300- to 500-bp insert size was performed at GATC Biotech, Konstanz, Germany. Paired-end Roche 454 Titanium FLX+ sequencing of two mate pair libraries of 3 kb (¼ lane, and subsequently, ½ lane because the initial ¼ lane data were of insufficient quality) and 7-kb inserts (¼ lane) was performed at the Ultrasequencing Unit of the Supercomputing and Bioinnovation Center of the University of Málaga, Spain.

Raw reads were preprocessed and filtered using SeqTrimNext [a next-generation sequencing-evolved version of SeqTrim (Falgueras *et al.* 2010)]. This included trimming of adapters, removal of PCR duplicates and filtering sequences with short insert size, low-quality base calling, empty inserts or possible contaminants (including microorganisms, organelles and plasmids; Figure 1). Three de novo assembling strategies were tried to select the best one. The first strategy (Fig. 1A) was based on a hybrid approach combining Illumina and 454 reads. Reads were independently assembled using Ray (Boisvert *et al.* 2012) with different *k*-mers, SOAP2 (Luo *et al.* 2012), MaSurCA (Zimin *et al.* 2013) or CABOG (Celera Assembler with Best Overlap Graph, version 7.1 by Miller *et al.* 2008). SOAP2 assembling was considered the best as it produced a genome size closer to the expected (1522 Mbp, Ewédjè 2012), and most (>96%) original reads mapped to the seeding scaffolds (results not shown). Therefore, seeding scaffolds obtained from SOAP2 were used downstream, in the gap-filling step (bottom of Fig. 1). In the second strategy (Fig. 1B), the 454 reads were assembled into seeding scaffolds with CABOG and Newbler (Margulies *et al.* 2005), which were then rescaffolded using SOAP2 and the Illumina reads. This strategy was abandoned because the coverage of the 454 reads was too low to produce satisfactory results. In the third assembling strategy (Fig. 1C), Illumina reads served to build seeding scaffolds using SOAP2 which were then split into pseudo-long reads with EMBOSS' tool Splitter (Rice *et al.* 2000). The pseudo-reads were then assembled and scaffolded with the 454 reads using CABOG and Newbler. CABOG produced more extended scaffolds and was eventually chosen over Newbler. In the final step, the resulting scaffolds were subjected to gap-filling with GAPCLOSER 1.12 (http://soap.genomics.org.cn/soapdenovo.html) using the Illumina reads to provide the final scaffolds conforming the draft genome. Gene prediction in final scaffolds was performed using MAKER v2.31.6 (Campbell *et al.* 2014) trained with the *S. globulifera* scaffolds, full-length plant proteins from UniProtKB and the transcriptome assembly described by Brousseau *et al.* (2014). MAKER annotations were saved in GFF3 format and imported into the genome browser Gbrowse 2.54 (Donlin 2009) to allow visualization, browsing and querying. Genomic SSRs were predicted using MREPS (Kolpakov *et al.* 2003). Finally, raw reads and validated SSRs were mapped on scaffolds with Bowtie (Langmead *et al.* 2009).

For a functional overview of the *S. globulifera* draft genome, predicted protein sequences were annotated with Full-LengtherNEXT (as described in Carmona *et al.* 2015) and Sma3 (Muñoz-Mérida *et al.* 2014) to provide

Table 1 Population genetics statistics for polymorphic microsatellite loci described in this study, tested for 125 *Symphonia globulifera* samples from four populations

Locus	*N*	*A*	*P*	A_R	H_E	F_{IS}	Null	*N*	*A*	*P*	A_R	H_E	F_{IS}	Null
	Ituberá, Brazil							Nkong Mekak, Cameroon						
1582	31	2	0	2	0.495	−0.31	<0.001	20	4	0	3.15	0.583	−0.656*	<0.001
3131	31	2	0	2	0.500	−0.497	<0.001	10	4	0	3.99	0.647	−0.086	0.037
3984	31	2	0	2	0.503	−0.818**	<0.001	31	2	0	2	0.508	−0.935**	<0.001
4464	31	3	1	2.29	0.413	−0.095	<0.001	31	3	2	2.48	0.236	−0.093	<0.001
5489	31	2	0	1.99	0.337	−0.054	<0.001	31	1	0	1	0.000	NA	0.001
6387	19	2	0	2	0.209	−0.067	<0.001	21	4	0	3.79	0.645	0.161	0.092
6636	30	3	0	2.3	0.508	−0.659*	<0.001	9	5	2	5	0.791	−0.133	<0.001
6783	31	1	0	1	0.000	NA	0.001	25	3	1	2.34	0.189	0.154	<0.001
7189	25	4	0	3.11	0.577	−0.542*	<0.001	22	6	0	4.38	0.654	−0.475*	<0.001
7694	29	4	2	3.47	0.635	−0.144	<0.001	30	7	0	4.75	0.720	0.169	0.048
9610	31	3	0	1.94	0.124	0.485	0.099	31	3	1	2.5	0.539	−0.201	<0.001
9990	25	3	0	2.36	0.530	−0.923**	<0.001	28	3	0	2.32	0.527	−0.931**	<0.001
10829	31	5	1	3.07	0.399	−0.218	<0.001	31	3	1	1.94	0.124	−0.039	<0.001
10904	29	4	1	2.77	0.254	0.186	<0.001	31	1	0	1	0.000	NA	0.001
14623	28	1	0	1	0.000	NA	0.001	31	6	2	4.19	0.562	0.256	0.106
15834	30	3	0	2.3	0.371	0.104	0.028	31	4	1	2.65	0.212	0.242	0.081
15979	30	3	0	2.2	0.213	−0.097	<0.001	25	5	0	4.1	0.578	−0.11	<0.001
16615	31	6	0	3.91	0.599	0.194	0.088	31	5	0	4.25	0.713	0.234	0.101
Multilocus average	29.11	2.94	0.28	2.317[a]	0.370[c]	−0.291	0.012	26.06	3.83	0.556	3.102[ab]	0.457[c]	−0.167	0.026
	Paracou, French Guiana							São Tomé, São Tomé and Principe						
1582	32	1	2	1	0.000	NA	0.001	31	2	0	1.29	0.032	0	<0.001
3131	32	4	1	3.27	0.603	0.225	0.063	22	5	2	4.77	0.779	0.187	0.061
3984	32	1	0	1	0.000	NA	0.001	29	3	1	2.78	0.550	−0.131	<0.001
4464	32	5	2	3.59	0.619	−0.267	<0.001	31	2	0	1.92	0.204	0.844*	0.186
5489	32	2	0	1.99	0.329	0.147	0.041	31	1	0	1	0.000	NA	0.001
6636	32	3	1	2.27	0.335	0.162	0.047	30	4	0	3.49	0.648	0.023	0.035
6387	32	8	4	5.44	0.698	0.242	0.103	19	4	2	3.47	0.677	0.227	0.069
6783	32	2	0	1.88	0.173	−0.088	<0.001	30	2	0	1.3	0.033	0	<0.001
7189	32	9	0	7.33	0.872	0.250	0.117	22	11	3	8.64	0.880	0.176	0.083
7694	31	6	1	4.83	0.756	0.062	0.062	30	6	1	4.63	0.538	0.447*	0.172
9610	30	4	1	3.11	0.349	−0.053	<0.001	31	NA	1	NA	NA	NA	NA
9990	30	7	4	5.47	0.770	0.048	<0.001	30	3	0	2.66	0.493	−0.43	<0.001
10829	32	5	1	3.58	0.439	0.076	<0.001	30	1	0	1	0.000	NA	0.001
10904	32	7	4	5.81	0.825	0.093	0.024	31	1	0	1	0.000	NA	0.001
14623	30	4	0	2.6	0.460	−0.015	<0.001	30	2	2	1.99	0.305	0.237	0.067
15834	32	5	1	2.83	0.233	0.198	0.068	30	4	2	2.48	0.189	−0.058	<0.001

Table 1 (Continued)

Locus	*N*	*A*	*P*	A_R	H_E	F_{IS}	Null	*N*	*A*	*P*	A_R	H_E	F_{IS}	Null
15979	32	8	1	4.28	0.613	0.238	0.088	30	3	0	2.3	0.437	−0.07	<0.001
16615	31	7	1	5.51	0.767	0.118	0.046	31	1	0	1	0.000	NA	0.001
Multilocus average	31.56	4.89	1.333	3.655[b]	0.491[c]	0.102	0.037	28.78	3.24	0.778	2.689[ab]	0.339[c]	0.112	0.040

N, number of successfully amplified samples; *A*, total number of alleles in the population; *P*, number of alleles private to the population; A_R, allelic richness (for a standard sample of 18 individuals); H_E, expected heterozygosity; F_{IS}, fixation index (* indicates a F_{IS} significantly different from zero after Bonferroni adjustment with P val < 0.05; **P val < 0.01); null, expected null allele frequency; NA, not available; avg = average.

[abc]Comparison of A_R and H_E between populations: values with different letters are significantly different for the tested statistic as by a Wilcoxon signed-rank test with Holm multiple test correction.

protein orthologs, putative gene names, descriptions, Gene Ontology terms (GO terms), enzyme codes (EC numbers) for putative enzymes, and the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway in which they can be involved.

In silico SNP discovery using a published transcriptome and the new draft genome

We conducted in silico discovery of high-quality SNPs using the 454 pyrosequencing reads from a published transcriptome (Brousseau *et al.* 2014) from a French Guiana population mapped against new genome assembly (from strategy 3) from a Cameroonian individual. Transcriptomic sequence reads were mapped against intron-masked genomic scaffolds longer than 4000 bp using the software BWA-MEM (parameters −t 24 −B 2 −O 1, Li & Durbin 2009). Variant calling was performed using samtools and bcftools (Li *et al.* 2009). GATK (McKenna *et al.* 2010) was used to filter for dense SNP clusters (–clusterSize 3 –clusterWindowSize 20). Variants were subsequently filtered to exclude indels and multiallelic SNPs. Minimum quality threshold was set to 30. Only SNPs that had a minimum coverage of eight reads were considered. We validated heterozygous SNPs with at least three reads for the alternative allele, and homozygous SNPs for the alternative allele if they had fewer than three reads of the reference allele, to avoid paralogs.

In silico preselection of polymorphic SSRs

The software MsatCommander v. 1.0.8 (Faircloth 2008) was used to screen for di-, tri-, tetra-, penta- and hexanucleotide repeats in the transcriptome from Brousseau *et al.* (2014) with the minimum repeat number set to four. Primer3 (Rozen & Skaletsky 2000) implemented in MSATCOMMANDER was used for automatic design of untagged primers with default settings, that is for putative amplified regions from 80 to 470 bp, retaining only SSRs with successful primer design. Transcripts that contained SSRs were annotated via BLASTx searches (with a cut-off *E*-value of $\leq 10^{-3}$), combining information from GenBank's nonredundant protein database (nr) and the UniProtKB protein database using the methods and scripts by De Wit *et al.* (2012). We discarded transcripts of organellar origin and those containing nucleotide mismatches, based on information reported in Brousseau *et al.* (2014). Likely coding sequences were extracted using the Transdecoder script from the Trinity package (Grabherr *et al.* 2011), which predicts partial and complete open reading frames (ORFs). The location of SSRs was recorded as within or outside the predicted ORF. If no ORF was defined for a contig, it was interpreted as a probable untranscribed region (UTR). The software

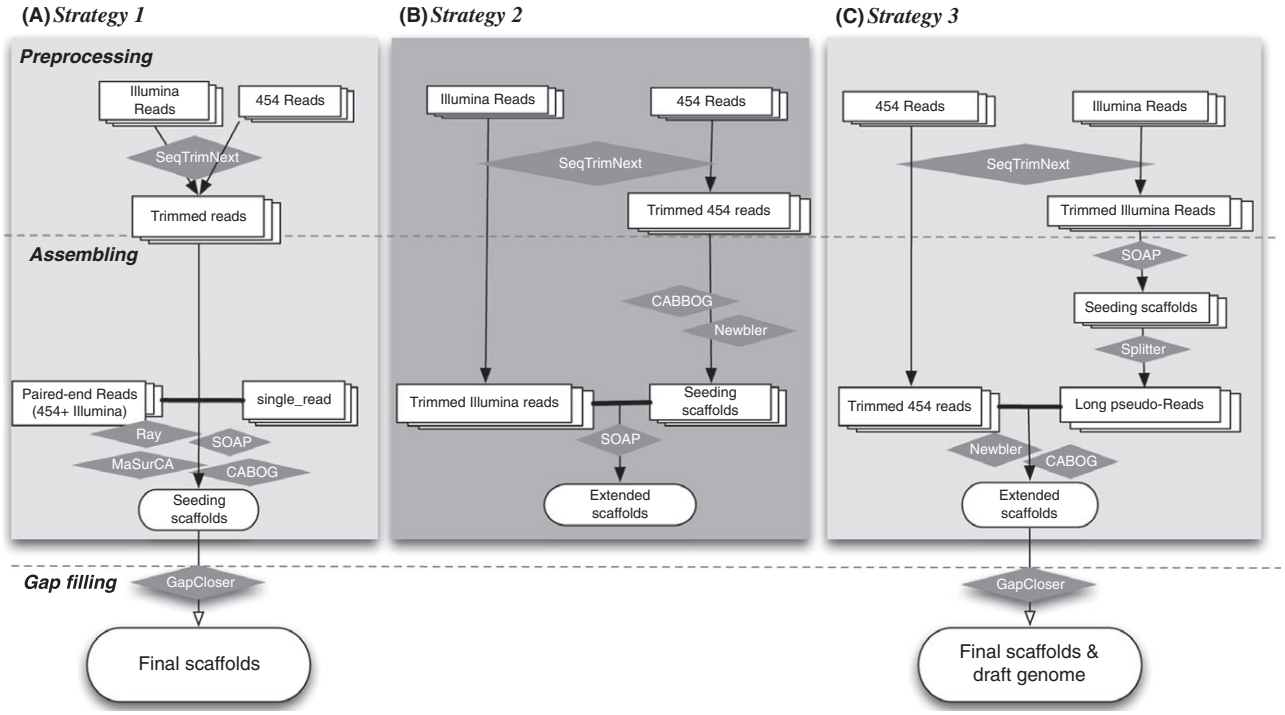

Fig. 1 Assembling strategies to obtain the *Symphonia globulifera* draft genome. After the preprocessing step, strategy 1 (A) follows a hybrid approach where all reads were assembled together using several assemblers, strategy 2 (B) produced seeding scaffolds from 454 reads that were extended using Illumina reads, and strategy 3 (C) produced the seeding scaffolds from Illumina paired-end reads that were then extended using 454 paired-end reads (see details in the main text). The final step consisted in filling gaps in scaffolds based on all Illumina reads only for strategies 1 and 3 (which proved best).

Exonerate (Slater & Birney 2005) was then used to align the transcriptome against the draft genome assemblies from strategies 1 and 3 using the est2genome option. Alignments with candidate SSRs that fulfilled the listed selection criteria were visually checked for alignment quality using PHYDE® v1.0 (Müller *et al.* 2006), and SSRs were kept only if they were polymorphic between transcriptome and genome. SSRs were discarded if the two genomic assemblies were incongruent. SSRs for which primers were located in intron–exon junctions had their primers redesigned based on the genome sequence using Primer3. Transcripts were sorted according to motif complexity, presence of UTR and predicted amplicon length. As UTRs are more variable than coding regions, SSRs located in the UTRs were selected preferentially. Interrupted and compound repeat motifs were avoided and SSRs with trinucleotide or longer repeat motives were preferred for ease of scoring. A total of 23 putatively polymorphic SSRs were selected to be tested in the laboratory.

SSR testing and genotyping

For genotyping, we used the three-primer labelling technique described by Micheneau *et al.* (2011), where the 5′ end of each forward primer is extended with an 18- to 19-bp unique sequence called Q-tag (Q1, Q2, Q3 or Q4) and a third primer corresponding to the specific Q-tag with a 5′ fluorescent label (PET, FAM, VIC, or NED) is added in PCR. Costs are reduced using the same Q-tag for several SSRs and through PCR multiplexing. OligoAnalyzer 3.1 (Integrated DNA Techonologies) was used to select a suitable Q-tag for each locus, ensuring absence of self- or heterodimers with a strong ΔG (≤ -9 kcal/mol) and secondary structures with melting temperatures (Tm) higher than the annealing temperature. As Q1 did not pass these criteria for any SSR, only Q2–Q4 were used.

The 23 selected SSRs were first amplified individually on five test samples using the Q-tagged forward and corresponding reverse primers, with the following PCR conditions: 1 μL buffer (10 $\times$), 0.4 μL $MgCl_2$ (25 mM), 0.3 μL dNTPs (10 mM each), 0.2 μL of each primer (0.01 mM), 0.05 μL Taq polymerase (BioTaq DNA Polymerase, 5 U/μL, Bioline), 1 μL of diluted template DNA (of ca. 10 ng/μL) and ddH_2O to make a final volume of 10 μL. Amplifications were performed as follows: 94 °C (4 min), 40 cycles of 94 °C (30 s), 56 °C (45 s), 72 °C (1 min) and a final extension at 72 °C for 10 min. PCR

products were run on a 1% agarose gel and stained with ethidium bromide. To verify the amplification of the target region, PCR products from several samples were Sanger-sequenced prior to genotyping.

Successfully amplifying loci were then amplified individually, now including 0.15 μL (0.002 mM) labelled Q-tag primer with the following program: 94 °C (4 min), 30 cycles of 94 °C (30 s), 56 °C (45 s), 72 °C (1 min), then 10 cycles of 94 °C (30 s), 53 °C (45 s), 72 °C (45 s) and a final extension at 72 °C for 10 min. The 19 SSRs that amplified in most of the test samples were sorted into three multiplexes (Table 2) and amplified in the four

Table 2 Details on the nineteen polymorphic SSR markers developed for *Symphonia globulifera* including primer sequences and GenBank accession numbers

Locus	Repeat	Q-tail	Mix	Primer sequence (5′–3′)	Putative function	UTR	GenBank Acc. no.
1582	$(ATC)_4$	Q3	1	F: **Q3**-GTGGTGGGATTGCTGCTATT R: TGGCAAGGAACAAGTGAAGA	Aquaporin TIP1,6	Yes	KR363116
2978	$(GGT)_4$	Q4	1	F: **Q4**-GGTGGAGGAGAAGGAGCAG R: CACCGTAACCACCACCTTG	No match	Probably	KR363117
3131	$(ACC)_5$	Q3	1	F: **Q3**-TCGAAGAAGAAAGCATTTACGTG R: ATGAGTACGTTCCAGGGCG	No match	Probably	KR363118
3984	$(ACC)_4$	Q2	1	F: **Q2**-TTACGTGCAAGAAGATTCACG R: ACCACAACCCGCTCATACAC	No match	Probably	KR363119
4464	$(CTT)_9$	Q3	3	F: **Q3**-CCGCTTGAATCTTCAATTTCTC R: AACGAACTTGGTGGTCTTGG	No match	Probably	KR363120
5489	$(GGATT)_4$	Q2	3	F: **Q2**-AGAAGGACTTGACGGTGCC R: GGAGCGGAAAGTGGACTCG	SKP1	Yes	KR363109
6387	$(AAT)_5$	Q2	1	F: **Q2**-ACGGGGATCAGATCGAGTTT R: TCACACATAACAGAATTTGCAATC	Predicted protein	Probably	KR363121
6636	$(GGTTT)_5$	Q2	1	F: **Q2**-CAGTGGGATGAAACCGAAAT R: CCCGTAACTTTGACCCAACA	NAC domain-containing protein	Yes	KR363110
6783	(GCT)^4	Q2	1	F: **Q2**-AATACGCAGAGATGGGCAC R: GAATGCTCGGGTTCAAATGC	No match	Probably	KR363111
7189	$(AAG)_4$	Q3	1	F: **Q3**-CCGACTTCACATCCCTAAACC R: GACCGAGATGCTTGATTCCC	No match	Yes	KR363112
7694	$(GTT)_7$	Q3	2	F: **Q3**-GGCACTAATCCGGAAACCAG R: TCTCCACGAAAGCTCAGGTC	Cyclin p3	Yes	KR363122
9610	$(ATC)_6$	Q2	2	F: **Q2**-GGGAGCAAGAAGCACTGTC R: TGATGAGGCTTGATTGGCG	No match	Probably	KR363123
9990	$(GCT)_7$	Q3	2	F: **Q3**-TCGTTGCTTTACCGAACTCC R: CCATCCATATCGAAGATGACG	Pseudouridine-5′-monophosphatase	Probably	KR363124
10829	$(AGC)_7$	Q2	3	F: **Q2**-ACTATGGTTTGGGTCCCGTC R: ACTCCCTGGCAAAGAACCC	Transcription factor MYC2	Yes	KR363113
10904	$(AGC)_6$	Q2	1	F: **Q2**-ATCTCTCCTCCCAGTGCAG R: GGCTCAAGGCAACTTGGTC	Predicted protein	Yes	KR363114
14623	$(CTT)_5$	Q2	2	F: **Q2**-TAGGTGGGGAGAAGGATGC R: TAAGGGAAGGAGGTGAACGA	Predicted protein	Probably	KR363125
15834	$(AGCG)_7$	Q2	3	F: **Q2**-GGGTTGGTGGATCGAGTACC R: AAGAGCATAGCGCTTGACG	No match	Probably	KR363126
15979	$(GGT)_7$	Q4	1	F: **Q4**-GCTTTTGTCTCGGCACTTGT R: CTCCAAACCGACTAGGACCA	Cation exchanger	Probably	KR363115
16615	$(AAC)_7$	Q4	2	F: **Q4**-GCCGAAAACCACCAAACC R: CGGAAGCTATAGGAAGGGATT	ATP-dependent RNA helicase	Probably	KR363127

Locus 2978 was omitted from population genetics analyses due to low level of polymorphism. Repeat: Number of repeats found in the sequence that corresponds to the accession number. UTR: inclusion of the SSR region in UTRs. Flurochrome in 5′ was 6-FAM for Q2 (TAGGAGTGCAGCAAGCAT), VIC for Q3 (CACTGCTTAGAGCGATGC) and NED for Q4 (CTAGTTATTGCTCAGCGGT).

populations using the Qiagen Multiplex PCR Kit (Qiagen) following the protocol described by Micheneau *et al.* (2011). The multiplex PCR conditions were: 1.5 μL Master Mix, 1.5 μL of primer mix (consisting of 0.7 μM of each F-primer and 2 μM of each R-primer), 0.15 μL of each Q-tag primer (10 μM), 1.5 μL of template DNA and ddH_2O to make a final volume of 15 μL. Amplified fragments were separated using an ABI 3730 DNA Analyzer (Applied Biosystems, Carlsbad, CA, USA), and allele sizes were determined with reference to the Gene-Scan—500 LIZ® Size Standard (Applied Biosystems) using Peak Scanner 2 software (Applied Biosystems).

SSR population genetic analysis

We estimated allelic richness (A_R), Nei's expected heterozygosity (H_E) and fixation index (F_I) for each locus and population using SPAGeDi version 1.4 (Hardy & Vekemans 2002). Differences between populations for A_R and H_E were assessed with paired Wilcoxon signed-rank tests with Holm multiple test correction as implemented in the *stats* package in *R* (R Development Core team 2014). Genetic differentiation between populations and for each population pair was estimated as F_{ST} and R_{ST} and tested using permutation tests in SPAGeDi. The effect of allele size mutations on differentiation was tested by permuting allele sizes among alleles (Hardy *et al.* 2003). The frequency of null alleles was estimated for each locus per population using FreeNA (Chapuis & Estoup 2007). As genotyping errors can bias SSR analysis (e.g. Bonin *et al.* 2004; Hoffman & Amos 2005), we estimated the genotyping error rate in our data by replicating the genotyping on five randomly chosen samples from each population (16.7%).

Results

De novo assembling and characterization of genomic sequences

The genome assembly was based on a total of ca. 190 million reads, which reduced to ca. 175 million useful reads after preprocessing to remove reads with indeterminations, contamination, without insert, etc. (Table 3). Table 4 summarizes some statistics on the assemblies obtained from strategies 1 and 3 using the useful single- and paired-end reads. The assembly from strategy 1 provided the best set of seeding scaffolds (with fewer gaps and Ns than strategy 3) but was improved less at the gap-filling stage, resulting in more scaffolds, more Ns, smaller N50 (length N for which 50% of all bases in the sequences are in a scaffold of length L < N) and smaller scaffold length on average. Therefore, even if more reads were mapped on final scaffolds of strategy 1 (final rows of Table 4), the final scaffolds from strategy 3 (Fig. 1C) were chosen as the best *S. globulifera* genome draft and were uploaded to the Dryad Digital Repository with link http://dx.doi.org/10.5061/dryad.78ng1. The length of the genome draft was 1.027 Gbp, which covers an expected 67.5% of the genome size estimated from flow cytometry data (Ewédjè 2012), and N50 was 500. It must be highlighted that the genomic coverage of the reads in Table 3 was low, 11× for Illumina reads and 0.1× for 454 reads, which explains the high number of scaffolds and the low N50 value (Table 4).

As genome annotators normally perform better with long sequences, we only annotated the 543 scaffolds longer than 10 kb and 22 additional scaffolds onto which one of the 19 validated SSRs mapped (mapping of loci

Table 3 Number of sequence reads considered for the de novo assembly of the *Symphonia globulifera* draft genome

Number of reads/ sequencing method	Illumina	454 3 kb (¼)	454 3kb (½)	454 7 kb (¼)
Raw reads	180 957 764	358 890	765 233	394 629
Rejected reads	6 832 186	40 063	69 780	165 922
Repeated read	—	13 850	43 466	27 039
Short insert	1 885 321	24 138	24 275	42 097
Indeterminations	383 164	232	211	78
Unexpected vector	6	286	680	7061
Empty insert	328 078	284	137	19 588
Contamination	4 235 617*	1273†	1011†	70 059*
Preprocessed paired reads	170 831 276	99 429	277 440	63 297
Preprocessed single reads	3 294 302	119 969	146 663	102 113
Preprocessed read mean size (bp)	91	140	170	158

*The main source of contamination was the genomic DNA of *E. coli* DH10B, causing the rejection of 648 034 Illumina reads and 69 553 Roche/454 reads.

†The main source of contamination was plastid DNA, most likely from the *S. globulifera* plastid. The sequences bear strong similarity to plastid DNA from other Malphighiales, for example the *Manihot esculenta* plastome.

Table 4 Assembly statistics on the de novo assembly of the *Symphonia globulifera* draft genome following the nomenclature of Figure 1

Assembling characteristic	Strategy 1	Strategy 3
Seeding scaffolds	9 107 943	8 632 000
Internal gaps in seeding scaffolds	608 183	1 052 592
#Ns in internal gaps	11 172 420	27 774 946
Final scaffolds	9 107 943	2 653 526
Internal gaps in final scaffolds	300 901	738 935
#Ns in internal gaps	1 501 653	104 746
N50 (nt)	221	500
N90 (nt)	117	192
Average length (nt)	218	387
Longest scaffold (nt)	53554	67855
Total assembly size (nt)	1 979 465.250	1 027 372 851
Mapping rate (Illumina)	96%	90%
Mapping rate (454)	29%	2%

5489, 1582 and 14 623 was partial and matched with two different scaffolds, Table 5). These 565 scaffolds accounted for 7 568 702 bp (0.74% of the genome draft size). A total of 1523 exact SSRs were predicted on them, including 367 di-, 483 tri-, 296 tetra- and 377 larger-than-tetranucleotide repeats. The most abundant repeat motif was AT. The 565 scaffolds were also searched for gene models with higher than 45% homology. This revealed 1046 putative genes (File S1, Supporting information), or gene fragments, most of them (782, 74.8%) producing transcripts longer than 500 and up to 6050 bp (sequences of the 1046 putative transcripts are provided in File S2, Supporting information). The sequences of annotated scaffolds, gene and SSR predictions and positions of validated SSRs have been integrated in Gbrowse, a genomic browser tool, and can be accessed and downloaded at SymphoniaDB (http://www.scbi.uma.es/symphoniaDB/). Functional annotation (description, best ortholog, GO terms, EC codes, pathways and gene names) was obtained using FullLengtherNext for 676 of the 1046 putative genes [complete annotations in File S3 (Supporting information), Full-LengtherNEXT Summary in File S4 (Supporting information)]. The main species that served to annotate the *S. globulifera* genes belonged to the order Malpighiales which includes the Clusiaceae: *Ricinus communis* (Euphorbiaceae, 81 annotations), *Populus trichocarpa* (Salicaceae, 55

Table 5 Position of the nineteen microsatellite loci described in Table 2 on the *Symphonia globulifera* genome draft

Locus name	Scaffold name	Scaffold Length (nt)	Type*	Start	End	Putative ortholog AC#	Putative ortholog description†
1582	deg7180003511049	500	Partial	500	317		
1582	deg7180004106732	100	Partial	70	100		
2978	deg7180003296107	851	Full	647	743		
3131	scf7180005380020	2807	Full	1543	1731		
3984	scf7180005408043	1267	Full	720	953		
4464	scf7180005327514	2278	Full	533	765		
5489	deg7180004279264	157	Partial	76	157		
5489	deg7180003785113	236	Partial	236	135		
6387	deg7180003307943	900	Full	227	597	F6HZ58	Putative uncharacterized protein of *Vitis vinifera*
6636	scf7180005337707	1800	Full	324	525	Q9SK55	Transcription factor JUNGBRUNNEN 1 of *Arabidopsis thaliana*
6783	scf7180005319115	2860	Full	1016	1188		
7189	scf7180005428261	1696	Full	168	327		
7694	scf7180005340183	1700	Full	1210	1566	Q9SHD3	Cyclin-U2-1 of *Arabidopsis thaliana*
9610	scf7180005308769	6300	Partial	6195	6300		
9990	scf7180005314014	3497	Full	174	287	F4JTE7	(DL)-glycerol-3-phosphatase 1 of *Arabidopsis thaliana*
10829	scf7180005312076	4168	Full	2773	3150	Q39204	Transcription factor MYC2 of *Arabidopsis thaliana*
10904	scf7180005307787	8400	Full	2916	3186	A0A067LD58	Uncharacterized protein of *Jatropha curcas*
14623	deg7180005250159	300	Partial	1	243		
14623	deg7180003594156	371	Partial	1	43		
15834	deg7180003616557	386	Full	223	381		
15979	scf7180005338180	1600	Full	289	637	O04034	Cation/calcium exchanger 5 of *Arabidopsis thaliana*
16615	deg7180003119103	1048	Full	129	218	Q38953	Probable pre-mRNA-splicing factor ATP-dependent RNA helicase of *Arabidopsis thaliana*

*Full: the complete sequence of the microsatellite is found identical on the scaffold. Partial: 95%–99% of the microsatellite sequence is found on the scaffold. Shorter matches are not shown.
†Orthologs are specified only if the predicted coding sequence is >45% identical to the orthologous protein.

annotations) and *Jatropha curcas* (Euphorbiaceae, 38 annotations). In the annotated genome fraction, (i) the most abundant molecular functions are ATP Binding, metal ion binding and DNA binding; (ii) the main biological processes are transcription, regulation of transcription, protein transport and protein ubiquitination; and (iii) the gene products are mainly located in nucleus, membrane and plastid (Fig. 2).

The 676 predicted genes with an identified orthologue represented 641 different orthologue IDs. Moreover, 231 genes (File S4, Supporting information) were predicted to produce a complete protein, coding for a total of 221 different proteins. The size of the coding region of complete proteins ranged from 177 bp (a member of the ribosomal protein L33 family) to 13 721 bp (calpain-type cysteine protease) as determined by Sma3, with a mean

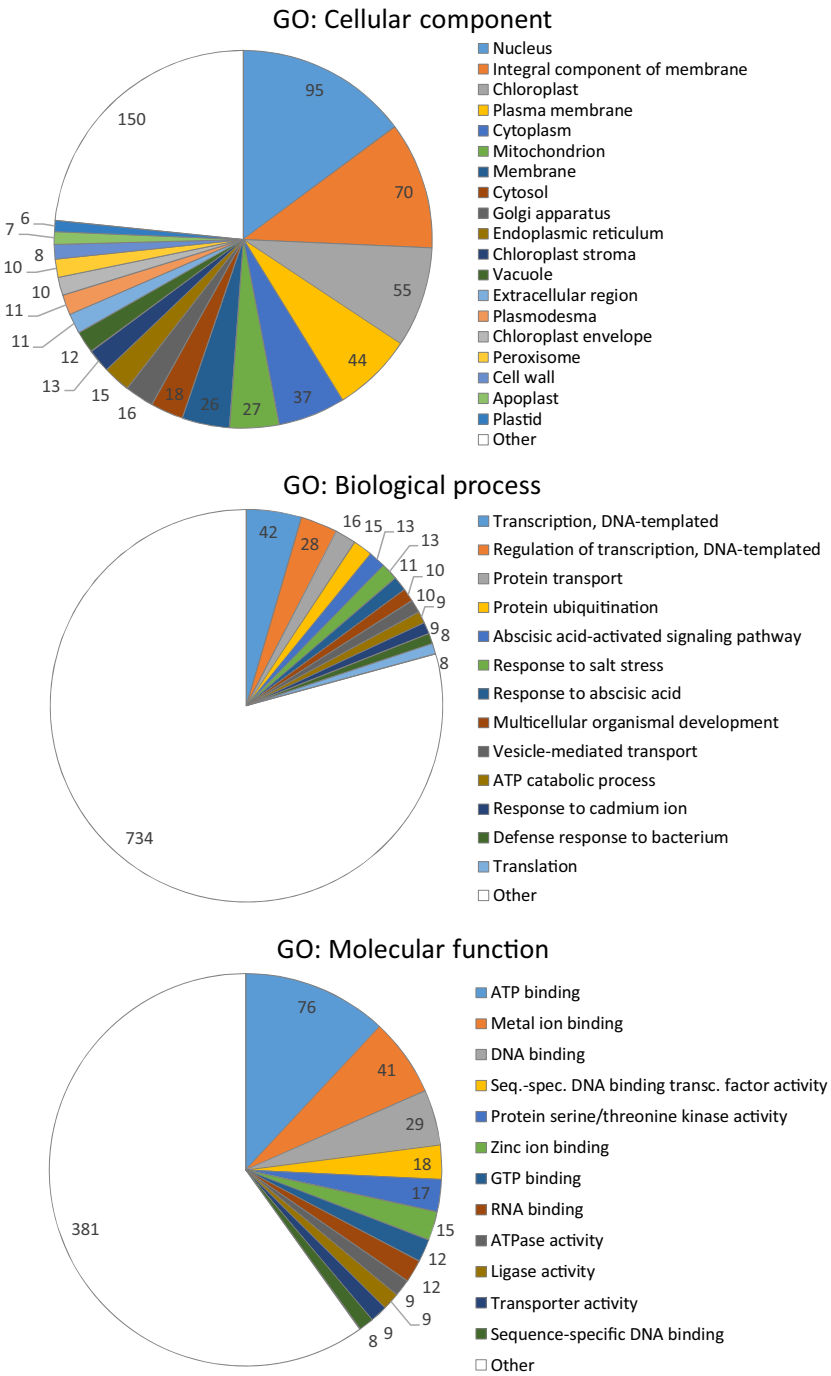

Fig. 2 Overview of GO-terms annotation of the 565 selected *Symphonia globulifera* scaffolds. GO-terms correspond to annotation of 676 predicted genes. [Colour figure can be viewed at wileyonlinelibrary.com]

size of 1423 bp and a median size of 2345 bp. The nine smallest gene predictions did not contain any intron (File S5, Supporting information).

In silico discovery of high-quality genic SNPs

We mapped the transcriptomic reads against the 7041 genome scaffolds longer than 4000 bp, called SNPs and applied strict quality filters. This led to the description of 923 nonclustered high-quality SNPs, 411 of them heterozygotes, thus variable within French Guiana, and 512 of them homozygotes for the alternative allele, thus susceptible to differentiate *S. globulifera* populations from the two continents. The resulting variant call file is available as File S6 (Supporting information).

Development and validation of SSR markers

SSR screening on the transcriptome identified 84 di-, 449 tri-, 38 tetra-, 12 penta- and 11 hexa-nucleotide repeats. Of the 594 SSR-harbouring transcripts, a third (195) did not get any significant BLAST hits. A total of 257 transcripts included ORFs, 264 did not contain any probable ORFs, and in 73 transcripts, the SSR was located outside the predicted ORFs. At least one nucleotide mismatch (Brousseau *et al*. 2014) was found in 83 transcripts, which were discarded along with four putative organellar transcripts. The remaining 507 transcripts were aligned to the draft genome (strategies 1 and 3), which revealed that 141 SSRs were not polymorphic, 53 had several variable motifs within the same amplification region, three were located in the end of the alignment, and nine could not be used because they did not align with the genomic sequences. Twenty-three loci were selected for validation in the laboratory, nineteen of which were retained for population-level genotyping. The loci were arranged into three multiplexes which successfully amplified ten, five and four loci (Table 2). All loci were polymorphic, but one locus (number 2978) displayed low polymorphism and was therefore discarded in the population genetic analysis.

The location of the nineteen validated SSRs (GenBank Accession nos. KR363109–KR363127) on the scaffolds of the retained genome draft (strategy 3) is shown in Table 5. Some SSRs mapped on very small scaffolds (≤300 nt, loci 5489, 14623, 15834 for example) and sometimes they matched with two such scaffolds, others mapped on longer scaffolds that contained a near gene but none mapped on the >10-kb scaffolds retained *a priori* for genome annotation. Unsurprisingly, the annotations for SSRs were consistent between transcriptomic (Table 2) and genomic (Table 5) sequences for both known (7694, 9990, 10829, 15979 and 16615) and predicted (6387, 10904) genes. Some functional predictions from the transcriptome were not confirmed on the genome (1582, 5489, 14623) because the SSRs lay on small genome contigs for which no annotations were obtained.

Population genetic analysis

Genotypes of 125 *S. globulifera* individuals from four populations were obtained at 18 SSRs (Tables 1 and 6). The total number of alleles observed at the eighteen loci was 129. The lowest polymorphism was observed at locus 5489 with two alleles and a total heterozygosity of $H_E = 0.383$, the highest polymorphism occurred at 7189 with 13 alleles and $H_E = 0.833$. Tests for Hardy–Weinberg genotypic proportions at the within-population level (Table 1) revealed 10 significant tests after Bonferroni correction ($P < 0.05$) out of 61 relationships tested (18 loci in four populations excluding 11 locally monomorphic loci). In Paracou, all loci were in Hardy–Weinberg equilibrium (HWE). Deviation from HWE was population-specific for most loci. For example, loci 4464 and 7694 had a significant heterozygote deficit only in São Tomé, indicating local occurrence of null alleles (with estimated frequencies of 0.172 and 0.186). On the other hand, some loci had a locally negative F_{IS} (e.g. 3984 and 9990) indicating an excess of heterozygosity. This could be due to local co-amplification of paralogous gene copies (see Discussion); however, excess of heterozygosity was only found in Ituberá and Nkong Mekak. Allelic richness was lowest in Ituberá ($A_R = 2.32$) and highest in Paracou ($A_R = 3.67$, values significantly different, $P < 0.05$, Table 1). Expected heterozygosity did not show significant differences between populations. Among-population differentiation varied from $F_{ST} = 0.086$ between morphotypes in Paracou to $F_{ST} = 0.522$ between Paracou and São Tomé, for an overall $F_{ST} = 0.317$ among the four populations (all values $P < 0.001$, File S7, Supporting information).

Replication of the genotyping in 20 individuals produced mismatches (allelic dropout or size differences) in 14 of 200 genotypes in Mix 1 (7%), in two of 100 genotypes in Mix2 (2%) and in 0 of 80 genotypes in Mix3, resulting in an overall genotyping error rate (i.e. the number of mismatches divided by the total number of replicated genotypes) of 4.2%. The highest error rate was detected in locus 15979 (0.15, three conflicting genotypes) followed by 3984, 6636, 6783 and 7189 with an error rate of 0.10 (Table 6).

Discussion

A first genome draft for Symphonia globulifera

De novo assembling of plant genomes remains a challenging step in genetic marker development, especially

Table 6 Characterization of the eighteen microsatellite loci genotyped in four *Symphonia globulifera* populations

Locus	Alleles	H_E	Allele range (bp)	F_{IS}	Variance of allele size (bp)	Error rate
1582	4	0.531	364–379	0.319**	3.2	0.05
3131	7	0.770	187–208	0.187*	36.7	0.05
3984	4	0.492	244–253	−0.267*	9.1	0.1
4464	8	0.579	221–251	0.334***	20.8	0
5489	2	0.383	217–222	0.583***	4.8	0
6387	10	0.693	373–405	0.292***	21.8	0.05
6636	7	0.636	200–225	0.015	51.3	0.1
6783	3	0.110	182–194	0.048	3.1	0.1
7189	13	0.833	170–204	0.049	42.7	0.1
7694	11	0.749	324–384	0.221***	51	0.05
9610	8	0.681	157–186	0.354**	6.1	0
9990	7	0.632	122–152	−0.327***	112.5	0.05
10829	7	0.677	374–398	0.614***	12.9	0
10904	8	0.401	279–300	0.392***	19.7	0
14623	8	0.684	329–362	0.584***	62.7	0
15834	9	0.260	156–186	0.157	10.5	0
15979	8	0.515	365–386	0.106	11.3	0.15
16615	8	0.805	210–231	0.468***	43.6	0

H_E, expected heterozygosity or gene diversity; F_{IS}, fixation index (* indicates a F_{IS} significantly different from zero after Bonferroni adjustment with $P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

for large repetitive plant genomes that can be highly heterozygous (Nielsen *et al.* 2011; Claros *et al.* 2012). The assembly of highly similar gene family members and genes with multiple domains is particularly difficult, especially when sequence coverage is low as in our case (Schatz *et al.* 2012). As data quality as well as genome size, GC content, library size, level of polymorphism and repeat content all affect the outcome of genome assemblers and no tool consistently gives best results (Ekblom & Wolf 2014), it is important to test different assembling strategies given the available data (Card *et al.* 2014). Methods using *de Brujin graphs* generally perform well on large genomes (Zhang *et al.* 2011). SOAP2, which has been specifically designed for assemblies based on short reads, was found to be the best strategy for our data. In our case, assembling strategy 2 failed because a sufficient coverage of the genome was not reached with 454 reads, given the low sequencing effort and relatively large estimated genome size of *S. globulifera* (1522 Mbp, Ewédjè 2012). Our study, with a sequencing effort of 11× genome coverage and final genome draft covering 67.5% of the estimated genome size in *S. globulifera*, represents a sequencing effort about an order of magnitude lower and a much simpler library construction strategy than genome projects in commercially important plants (coverage ca. 60–300×), for example Varshney *et al.* (2013) for chickpea, Xu *et al.* (2013) for sweet orange or Bombarely *et al.* (2012) for tobacco.

A conservative annotation of only the longest (>10 kb) scaffolds in our study provided 1046 gene predictions in as few as 7.57 Mb, suggesting that long scaffolds are clearly enriched in genes (one gene per 7.2 kb). Moreover, even if the assembling of the 67.5% of the estimated genome is quite fragmented (Table 4) and we only analysed a 0.74% of the genome draft, 231 gene predictions seem to contain the complete intron–exon pattern that code for a complete protein, that is, one complete protein per 32.8 kb. These results are in agreement with other studies, demonstrating that a fragmented low-coverage draft genome without complete annotation is very helpful for marker development (Straub *et al.* 2011; Leese *et al.* 2012; Blischak *et al.* 2014). Therefore, this resource would allow for the development of sequence capture or resequencing markers giving access to full haplotypic information including intron information, a considerable improvement over SNP markers in random genomic regions (e.g. Remington 2015). It is noteworthy that gene annotations were mostly obtained from phylogenetically distant organisms of the Euphorbiaceae or Salicaceae families, highlighting the low prior availability of genomic resources for the Malpighiales order of flowering plants (e.g. Vanneste *et al.* 2014).

The *Symphonia globulifera* draft genome developed here will be useful for marker development to elucidate the particular evolutionary history of the *Symphonia* genus. In addition, the new genomic resources will greatly benefit the research on bioactive molecules. *Symphonia globulifera* and related taxa have been widely used in traditional medicine (see Introduction and Fromentin *et al.* 2015 for a review). Several antimicrobial compounds have been identified from *S. globulifera*, most of them unique to this species (Fromentin *et al.* 2015). The

potential use of these chemical compounds range from HIV-inhibitory (Gustafson *et al.* 1992) to antiplasmodial activities, useful against malaria (Ngouela *et al.* 2006; Marti *et al.* 2010). The discovery of the genes involved in their biosynthesis and any genetic variation within them could guide the discovery of new related compounds (e.g. Burgarella *et al.* 2012).

Discovery and development of genic markers

Marker development in tropical nonmodel species is often hampered by lacking genomic resources and limited tissue material. In this study, we showed that choosing individuals from different populations for high-throughput sequencing and aligning transcriptomic sequences to a draft genome makes it possible to detect intron–exon junctions and to discover and preselect loci in silico that are conserved and polymorphic over a wide range. This approach led to the discovery of 923 high-quality genic SNPs for the tropical tree *S. globulifera* which are useful for the future design of a SNP genotyping chip (considering however the narrow SNP ascertainment panel). We also developed and tested 19 novel microsatellite markers. In silico selection of polymorphic markers saves DNA in laboratory testing, which is often crucial for research in tropical or endangered species. In the case of SSR development, filtering steps and visual screening of alignment quality enabled us to eliminate misaligned regions resulting in a high success rate of well-amplifying polymorphic loci. In our case, SSR testing and genotyping on valuable samples with low-quantity DNA extracts relied on prior whole-genome amplification with a method based on multiple chain displacement, a low-bias technology (Murphy *et al.* 2012; Huang *et al.* 2015) that has previously proven efficient in SSR development with minute DNA quantities (Dracatos *et al.* 2006). Genic markers such as the SNPs and SSRs characterized in this paper are transcribed and thus subjected to natural selection. They are therefore potentially interesting for investigating genetic signals of selection and for genotype–phenotype or genotype–environment association studies (e.g. González-Martínez *et al.* 2007, Jaramillo-Correa *et al.* 2015).

Evaluation of microsatellite markers

Our SSR genotyping error rate was higher than the rate obtained in other studies, 4.2% per replicated genotype, vs. 0.13–0.74% (Hoffman & Amos 2005) or 0.26% (Frantz *et al.* 2006), perhaps reflecting that we report here newly developed markers among which the most reliable ones can be selected. Testing and reporting genotyping error rates is not established practice, and only few studies were available for comparison. Hoffman & Amos (2005) found common misinterpretation of allele banding patterns with confusion between homozygote and adjacent allele heterozygote genotypes, and they detected a positive correlation of error rate with locus polymorphism and product size. Such trends were not observed in our data. Conversely, we observed cases of allele dropout (e.g. Frantz *et al.* 2006), where, due to competition in the PCR, one of the allelic copies in a heterozygous sample amplifies weakly and remains undetected in the scoring. Another reason for inconsistent genotypes in our study was unambiguous size variation between replicated genotypes, in the absence of stutter bands, secondary peaks or other misleading noise. This incongruence could be due to early PCR errors and/or amplification of paralogous gene copies (Sharma *et al.* 2009). If the microsatellite locus is located in a duplicated genetic region or a member of a gene family, sometimes one of the regions and sometimes another could amplify. If paralogous copies amplified at the same time, this would lead to an apparent excess of heterozygosity and thus a negative F_{IS}, as observed for some loci in Ituberá and Nkong Mekak. Interestingly though, the loci with the highest error rate did not deviate from Hardy–Weinberg proportions more frequently than others. The risk of unspecific PCR can be high in plants where large gene families and abundant pseudogenes with nearly identical sequences occur due to recent genome duplication events and transposon activity (Schnable *et al.* 2009). Also, the genome content varies across individuals, with a set of stable *core* genes and accessory *shell* or *cloud* genomic elements, together constituting the pan-genome (Marroni *et al.* 2014). In an ancient species like *S. globulifera*, there could thus be substantial genomic variation across populations and individuals, perhaps contributing to local variation in the specificity of PCR. Our results highlight that special attention should be paid to target single-copy regions in marker design and that error rates should be reported to facilitate selecting the most reliable loci for follow-up studies.

The new transcriptomic SSRs with 2–13 alleles/locus across four African and American populations seemed to be much less polymorphic than the genomic markers used in previous studies on *S. globulifera*: using five genomic SSRs Budde *et al.* (2013) identified a total of 111 alleles (9–43 alleles/locus) and an average heterozygosity of H_E = 0.860 in African populations and Dick & Heuertz (2008) reported a total of 132 alleles (19–41 alleles/locus) and H_E = 0.860 in American populations. As the flanking regions of genic markers are less variable than those of anonymous genomic markers, they are better transferable between closely related species (e.g. Dufresnes *et al.* 2014; Huang *et al.* 2014). Therefore, the SSRs developed here will likely be useful in studies involving other *Symphonia* species.

Population genetic analysis

We demonstrated a good amplification and genotyping success across the novel SSR markers in four *S. globulifera* populations from Africa and America. The global genetic differentiation in our data set, F_{ST} = 0.317, was larger than F_{ST} = 0.135 in Africa (Budde *et al.* 2013) or F_{ST} = 0.138 in America (Dick & Heuertz 2008), which can largely be explained by the lower SSR polymorphism in our study (Jost 2008). Genetic differentiation was of the same order of magnitude between African or American populations than between continents, which reflects disjunct sampling ranges within continents: the volcanic São Tomé island *vs.* the African mainland (see also Budde *et al.* 2013) and the Guiana shield *vs*. the Brazilian Atlantic Forest. All populations had private alleles, but a substantial proportion of alleles were shared among populations, which could be due either to common ancestry or to homoplasious mutations (Ellegren 2004). Homoplasious mutations are expected to erase the phylogeographic signal at moderately polymorphic markers in an ancient species like *S. globulifera*, but a phylogeographic signal was nevertheless detected for some population pairs (R_{ST} values, File S7, Supporting information). The two different morphotypes in Paracou (French Guiana) displayed low but significant genetic differentiation, although fixation indices did not suggest deviation from random mating. Morphotypes might thus represent incipient diverging lineages (e.g. Feder *et al.* 2012). The high genetic diversity in Paracou could additionally be explained by a complex biogeographic history of this Guiana shield population (Scotti-Saintagne *et al.* 2013), which shows cpDNA and SSR similarity with Amazonian populations (Dick & Heuertz 2008). The low diversity of Ituberá in the Brazilian Atlantic Forest was unexpected given that this area has been identified as a climatically stable genetic diversity hotspot (Carnaval *et al.* 2009). Phylogeographic studies did however highlight that evolutionary relationships between the disjunct Amazon and Atlantic forest ranges can be highly taxon specific (e.g. Costa 2003).

Acknowledgements

This study was funded by the Spanish Ministry of Science and Innovation (MICINN) under the project AFFLORA (CGL2012-40129-C02-02), cofunded by the ERDF (European Regional Development Fund) and Plan Andaluz de Investigación, Desarrollo e Innovación under the grant P10-CVI-6075, the French National Research Agency (ANR) under the project FLAG (ANR-12-ADAP-0007) and the Research Council of Norway (203822/E40). This work has benefited from support of a grant from Investissement d'Avenir grants of the ANR (CEBA:ANR-10-LABX-25-01). Research Permits for sample collection were issued by the Ministry of Science and Innovation, Cameroon. Genotyping services were carried out at the Spanish 'Parque Científico' in Madrid (CSIC—PCM). We acknowledge computing resources and expertise used at the CSC—Finnish IT Center for Science, the Finnish grid infrastructure (FGI), and computing resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) centre of the University of Malaga. MH acknowledges a 'Ramón y Cajal' fellowship (RYC2009-04537) from the Spanish Ministry of Science and Innovation (MICINN), and a Marie-Curie Intra-European fellowship (PIEF-GA-2012-329088). We thank Carmen García Barriga, Zaida Lorenzo, Henri Caron and Valérie Troispoux for assistance with laboratory work and Katharina B. Budde, Katrin Heer, Peter Mambo, Bonaventure Sonké, Gilles Dauby and Saintomer Cazal for help with fieldwork.

References

Abdul-Salim K (2002) *Systematics and Biology of Symphonia L.f. (Clusiaceae)*. PhD thesis, Harvard University, Boston, Massachusetts.

Baraloto C, Morneau F, Bonal D, Blanc L, Ferry B (2007) Seasonal water stress tolerance and habitat associations within four Neotropical tree genera. *Ecology*, **88**, 478–489.

Barthe S, Gugerli F, Barkley NA, Maggia L, Cardi C, Scotti I (2012) Always look on both sides: phylogenetic information conveyed by simple sequence repeat allele sequences. *PLoS ONE*, **7**, e40699. doi: 10.1371/journal.pone.0040699.

Blischak PD, Wenzel AJ, Wolfe AD (2014) Gene prediction and annotation in Penstemon (Plantaginaceae): a workflow for marker development from extremely low-coverage genome sequencing. *Applications in Plant Sciences*, **2**, 1400044.

Boisvert S, Raumond F, Godzaridis E, Laviolette F, Corbeil J (2012) Ray Meta: scalable *de novo* metagenome assembly and profiling. *Genome Biology*, **13**, R122.

Boiteau P (1986) *Précis de matière médicale malgache*. Agence de Coopération Culturelle et Technique, Paris, France.

Boiteau P, Boiteau M, Allorge-Boiteau L (1999) *Dictionnaire des noms malgaches de végétaux*. C. Alzieu, Grenoble.

Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller LA, Martin GB (2012) A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. *Molecular Plant Microbe Interactions*, **25**, 1523–1530.

Bonin A, Bellemain E, Eidesen PB, Pompanon F, Brochmann C, Taberlet P (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Bradbury D, Smithson A, Krauss SL (2013) Signatures of diversifying selection at EST-SSR loci and association with climate in natural Eucalyptus populations. *Molecular Ecology*, **22**, 5112–5129.

Brousseau L, Tinaut A, Duret C, Lang T, Garnier-Gere P, Scotti I (2014) High-throughput transcriptome sequencing and preliminary functional analysis in four Neotropical tree species. *BMC Genomics*, **15**, 238.

Budde KB, González-Martínez SC, Hardy OJ, Heuertz M (2013) The ancient tropical rainforest tree *Symphonia globulifera* L.f. (Clusiaceae) was not restricted to postulated Pleistocene refugia in Atlantic Equatorial Africa. *Heredity*, **111**, 66–76.

Burgarella C, Navascués M, Zabal-Aguirre M *et al.* (2012) Recent population decline and selection shape diversity of taxol-related genes. *Molecular Ecology*, **21**, 3006–3021.

Campbell MS, Holt C, Moore B, Yandell M (2014) Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, **48**, 4.11.1–4.11.39.

Card DC, Schield DR, Reyes-Velasco J *et al.* (2014) Two low coverage bird genomes and a comparison of reference-guided versus de novo genome assemblies. *PLoS ONE*, **9**, e106649.

Carmona R, Zafra A, Seoane P *et al.* (2015) ReprOlive: a database with linked data for the olive tree (*Olea europaea* L.) reproductive transcriptome. *Frontiers in Plant Science*, **6**, 625.

Carnaval AC, Hickerson MJ, Haddad CFB, Rodrigues MT, Moritz C (2009) Stability predicts genetic diversity in the Brazilian Atlantic forest hotspot. *Science*, **323**, 785–789.

Chapuis M-P, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, **24**, 621–631.

Clair B, Jaouen G, Beauchêne J, Fournier M (2003) Mapping radial, tangential and longitudinal shrinkages and relation to tension wood in discs of the tropical tree *Symphonia globulifera*. *Holzforschung*, **57**, 665–671.

Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N (2012) Why assembling plant genome sequences is so challenging. *Biology*, **1**, 439–459.

Costa LP (2003) The historical bridge between the Amazon and the Atlantic Forest of Brazil: a study of molecular phylogeography with small mammals. *Journal of Biogeography*, **30**, 71–86.

De Wit P, Pespeni MH, Ladner JT *et al.* (2012) The simple fool's guide to population genomics via RNA-seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, **12**, 1058–1067.

Dick CW, Heuertz M (2008) The complex biogeographic history of a widespread tropical tree species. *Evolution*, **62**, 2760–2774.

Dick CW, Abdul-Salim K, Bermingham E (2003) Molecular systematic analysis reveals cryptic Tertiary diversification of a widespread tropical rain forest tree. *The American Naturalist*, **162**, 691–703.

Donlin MJ (2009) Using the generic genome browser (GBrowse). *Current Protocols in Bioinformatics*, **28**, 9.9: 9.9.1–9.9.25.

Dracatos PM, Dumsday JL, Olle RS *et al.* (2006) Development and characterization of EST-SSR markers for the crown rust pathogen of ryegrass (*Puccinia coronata* f.sp. lolii). *Genome*, **49**, 572–575.

Dufresnes C, Brelsford A, Béziers P, Perrin N (2014) Stronger transferability but lower variability in transcriptomic- than in anonymous microsatellites: evidence from Hylid frogs. *Molecular Ecology Resources*, **14**, 716–725.

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, **7**, 1026–1042.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435–445.

Ewédjè EBK (2012) *Biologie de la reproduction, phylogéographie et diversité de l'arbre à beurre Pentadesma butyracea Sabine (Clusiaceae)—implications pour sa conservation au Bénin*. PhD thesis, Université Libre de Bruxelles, Brussels, Belgium.

Faircloth BC (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources*, **8**, 92–94.

Falgueras J, Lara AJ, Férnandez-Pozo N *et al.* (2010) SEQTRIM: a high-throughput pipeline for preprocessing any type of sequence reads. *BMC Bioinformatics*, **11**, 38.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Frantz AC, Pourtois JT, Heuertz M *et al.* (2006) Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. *Molecular Ecology*, **15**, 3191–3203.

Fromentin Y, Cottet K, Kritsanida M, Michel S, Gaboriaud-Kolar N, Lallemand M-C (2015) *Symphonia globulifera*, a widespread source of complex metabolites with potent biological activities. *Planta Medica*, **81**, 95–107.

Gardner EM, Johnson MG, Ragone D, Wickett NJ, Zerega NJC (2016) Low-coverage, whole-genome sequencing of *Artocarpus camansi* (Moraceae) for phylogenetic marker development and gene discovery. *Applications in Plant Sciences*, **4**, 1600017.

González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007) Association genetics in *Pinus taeda* L. I. Wood property traits. *Genetics*, **175**, 399–409.

Goodwin ZA, Harris D, Filer D *et al.* (2015) Widespread mistaken identity in tropical plant collections. *Current Biology*, **25**, R1066–R1067.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Gustafson KR, Blunt JW, Munro MHG *et al.* (1992) The guttiferones, HIV-inhibitory benzophenones from *Symphonia globulifera*, *Garcinia livingstonei*, *Garcinia ovalifolia* and *Clusia rosea*. *Tetrahedron*, **48**, 10093–10102.

Hardy OJ, Vekemans X (2002) SPAGEDI: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Hardy OJ, Charbonnel N, Fréville H, Heuertz M (2003) Microsatellite allele sizes: a simple test to assess their significance on genetic differentiation. *Genetics*, **163**, 1467–1482.

Heuertz M, Duminil J, Dauby G, Savolainen V, Hardy OJ (2014) Comparative phylogeography in rainforest trees from Lower Guinea, Africa. *PLoS ONE*, **9**, e84307.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, **14**, 599–612.

Huang D, Yiqing Z, Mingda J *et al.* (2014) Characterization and high cross-species transferability of microsatellite markers from the floral transcriptome of *Aspidistra saxicola* (Asparagaceae). *Molecular Ecology Resources*, **14**, 569–577.

Huang L, Ma F, Chapman A, Lu S, Xie XS (2015) Single-cell whole-genome amplification and sequencing: methodology and applications. *Annual Review of Genomics and Human Genetics*, **16**, 79–102.

Jan du Chêne RE, Salami MB (1978) Palynology and micropaleontology of the Upper Eocene of the well nsukwa-1 (Niger Delta, Nigeria). *Archives des Sciences*, **13**, 5–9.

Jaramillo-Correa J-P, Rodríguez-Quilón I, Grivet D *et al.* (2015) Molecular proxies for climate maladaptation in a long-lived tree (*Pinus pinaster* Aiton, Pinaceae). *Genetics*, **199**, 793–807.

Jost L (2008) G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015–4026.

Kolpakov R, Bana G, Kucherov G (2003) MREPS: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acid Research*, **31**, 3672–3678.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Leese F, Brand P, Rozenberg A *et al.* (2012) Exploring Pandora's box: potential and pitfalls of low coverage genome surveys for evolutionary biology. *PLoS ONE*, **7**, e49202.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Luo R, Liu B, Xie Y *et al.* (2012) SOAPDENOVO2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **1**, 18.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in open microfabricated high density picoliter reactors. *Nature*, **437**, 376–380.

Marroni F, Pinosio S, Morgante M (2014) Structural variation and genome complexity: Is dispensable really dispensable? *Current Opinion in Plant Biology*, **18**, 31–36.

Marti G, Eparvier V, Moretti C *et al.* (2010) Antiplasmodial benzophenone derivatives from the root barks of *Symphonia globulifera* (Clusiaceae). *Phytochemistry*, **71**, 964–974.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MAPREDUCE framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Micheneau C, Dauby G, Bourland N *et al.* (2011) Development and characterization of microsatellite loci in *Pericopsis elata* (Fabaceae) using a cost-efficient approach. *American Journal of Botany*, **98**, e268–e270.

Miller JR, Delcher AL, Koren S *et al.* (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**, 2818–2824.
Müller K, Quandt D, Müller J, Neinhuis C (2006) PhyDE®: Phylogenetic Data Editor, version 0.995. http://www.phyde.de.
Muñoz-Mérida A, Viguera E, Claros MG, Trelles O, Pérez-Pulido AJ (2014) SMA3s: a three-step modular annotator for large sequence datasets. *DNA Research*, **21**, 341–353.
Murphy SJ, Cheville JC, Zarei S *et al.* (2012) Mate pair sequencing of whole-genome-amplified DNA following laser capture microdissection of prostate cancer. *DNA Research*, **19**, 395–406.
Ngouela S, Lenta BN, Noungoue DT *et al.* (2006) Anti-plasmodial and antioxidant activities of constituents of the seed shells *of Symphonia globulifera* Linn f. *Phytochemistry*, **67**, 302–306.
Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.
Nybom H, Weising K, Rotter B (2014) DNA fingerprinting in botany: past, present, future. *Investigative Genetics*, **5**, 1–35.
Oyen LPA (2005) *Symphonia globulifera* L.f. [Internet] Record from PROTA4U (eds Louppe D, Oteng-Amoako AA, Brink M). PROTA (Plant Resources of Tropical Africa / Ressources végétales de l'Afrique tropicale), Wageningen, Netherlands. Available at: http://www.prota4u.org/search.asp (accessed 17 October 2016).
Perrier de la Bâthie H (1951) 136e Famille Guttifères (Guttiferae). In: *Flore de Madagascar et des Comores (Plantes vasculaires)*(Humbert H), vol. Suppl. 49. Imprimerie Officielle, Muséum National d'histoire Naturelle, Antananarivo, Paris.
Primmer C (2009) From conservation genetics to conservation genomics. *Annals of the New York Academy of Science*, **1162**, 357–368.
R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Raven PH, Wilson EO (1992) A 50-year plan for biodiversity surveys. *Science*, **258**, 1099–1110.
Remington DL (2015) Alleles vs. mutations: understanding the evolution of genetic architecture requires a molecular perspective on allelic origins. *Evolution*, **69**, 3025–3038.
Ribeiro RA, Lovato MB (2007) Comparative analysis of different DNA extraction protocols in fresh and herbarium specimens of the genus *Dalbergia*. *Genetics and Molecular Research*, **6**, 173–187.
Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276–277.
Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **6**, 365–386.
Särkinen T, Staats M, Richardson JE, Cowan RS, Bakker F (2012) How to open the treasure chest? Optimizing DNA Extraction from herbarium specimens. *PLoS ONE*, **7**, e43808.
Schatz MC, Witkowski J, McCombie RW (2012) Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biology*, **13**, 243.
Schnable PS, Ware D, Fulton RS *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
Scotti-Saintagne C, Dick CW, Caron H *et al.* (2013) Phylogeography of a species complex of lowland Neotropical rain forest trees (*Carapa, Meliaceae*). *Journal of Biogeography*, **40**, 676–692.
Sharma RK, Bhardwaj P, Negi R, Mohapatra T, Ahuja PS (2009) Identification, characterization and utilization of unigene derived microsatellite markers in tea (*Camellia sinensis* L.). *BMC Plant Biology*, **9**, 53.
Slater GS, Birney E (2005) Automated generation of heuristics for biological comparison. *BMC Bioinformatics*, **6**, 31.
Straub SCK, Fishbein M, Livshultz T *et al.* (2011) Building a model: developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics*, **12**, 211.
Tabbasam N, Zafar Y, Mehboob-ur-Rahman (2014) Pros and cons of using genomic SSRs and EST-SSRs for resolving phylogeny of the genus *Gossypium*. *Plant Systematics and Evolution*, **300**, 559–575.
Tagu D, Colbourne JK, Nègre N (2014) Genomic data integration for ecological and evolutionary traits in non-model organisms. *BMC Genomics*, **15**, 490.
Turchetto-Zolet AC, Pinheiro F, Salgueiro F, Palma-Silva C (2013) Phylogeographical patterns shed light on evolutionary process in South America. *Molecular Ecology*, **22**, 1193–1213.
Van Oosterhout C, Weetman C, Hutchington WF (2006) Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Notes*, **6**, 255–256.
Vanneste K, Baele G, Maere S, Van de Peer Y (2014) Analysis of 241 plant genomes supports a wave of successful genome duplication in association with the Cretaceous-Palaeogene boundary. *Genome Research*, **24**, 1334–1347.
Varshney RK, Song C, Saxena R *et al.* (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology*, **31**, 240–246.
Weitemier K, Straub SCK, Cronn RC *et al.* (2014) HYB-SEQ: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, **2**, 1400042.
Xia H, Zheng X, Chen L *et al.* (2014) Genetic differentiation revealed by selective loci of drought-responding EST-SSRs between upland and lowland rice in China. *PLoS ONE*, **9**, e106352.
Xu Q, Chen L-L, Ruan X *et al.* (2013) The draft genome of sweet orange (*Citrus sinensis*). *Nature Genetics*, **45**, 59–66.
Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B (2011) A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PLoS ONE*, **6**, e17915.
Zimin A, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. *Bioinformatics*, **29**, 2669–2677.

M.H., S.C.G.-M. and S.O. designed the study; M.H., O.H., I.S. and C.S.-S. provided materials; S.O., P.S.-Z., R.B., M.G.C. and M.H. analysed the data; S.O. and M.H. wrote the first draft of the manuscript; and all authors contributed to improving the manuscript.

Data accessibility

Raw sequence reads were deposited in the Short Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) under BioProject PRJNA288627 and SRA Accession no. SAMN03835263 (SRX1080598 and SRX1080599 for 454 reads and SRX1081174 for Illumina reads). The genome draft (fasta format) and microsatellite genotypes are archived in the Dryad Digital Repository with link http://dx.doi.org/10.5061/dryad.78ng1. Annotated genomic scaffolds can be accessed at http://www.scbi.uma.es/symphoniaDB/. Sanger sequences of validated SSRs were deposited at GenBank under Accession nos. KR363109–KR363127.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

File S1 Maker gene prediction and annotation results of the 565 *Symphonia globulifera* scaffolds selected for annotation.

File S2 Nucleotide sequences in fasta format of transcripts of the 1046 predicted genes in *Symphonia globulifera*.

File S3 Gene annotation by Full-LengtherNEXT of predicted *Symphonia globulifera* genes.

File S4 Summary of Full-LengtherNEXT annotation results of predicted *Symphonia globulifera* genes.

File S5 Detailed description obtained with Sma3 of the genes predicted to code for a complete protein in *Symphonia globulifera*.

File S6 Variant call file containing describing high-quality genic SNPs in *Symphonia globulifera* with the new genome draft used as reference.

File S7 Genetic differentiation among *Symphonia globulifera* populations at 18 SSR loci.