MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) **14**, 1281–1295 doi: 10.1111/1755-0998.12276

Development of genomic resources for *Nothofagus* species using next-generation sequencing data

V. A. EL MUJTAR,*†‡ L. A. GALLO,* T. LANG§ and P. GARNIER-GÉRɆ‡

**Unidad de Genética Ecológica y Mejoramiento Forestal, Instituto Nacional de Tecnología Agropecuaria (INTA) EEA Bariloche, Modesta Victoria, 4450 (8400), Bariloche, Río Negro, Argentina, †INRA, UMR 1202 Biodiversity Genes & Communities, F- 33610 Cestas, France, ‡UMR1202 Biodiversity Genes & Communities, University of Bordeaux, Bordeaux, F-33400 Talence, France, §Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan 666303, China*

Abstract

Using next-generation sequencing, we developed the first whole-genome resources for two hybridizing *Nothofagus* species of the Patagonian forests that crucially lack genomic data, despite their ecological and industrial value. A de novo assembly strategy combining base quality control and optimization of the putative chloroplast gene map yielded ~32 000 contigs from 43% of the reads produced. With 12.5% of assembled reads, we covered ~96% of the chloroplast genome and ~70% of the mitochondrial gene content, providing functional and structural annotations for 112 and 52 genes, respectively. Functional annotation was possible on 15% of the contigs, with ~1750 potentially novel nuclear genes identified for *Nothofagus* species. We estimated that the new resources (13.41 Mb in total) included ~4000 gene regions representing ~6.5% of the expected genic partition of the genome, the remaining contigs potentially being nongenic DNA. A high-quality single nucleotide polymorphisms resource was developed by comparing various filtering methods, and preliminary results indicate a strong conservation of cpDNA genomes in contrast to numerous exclusive nuclear polymorphisms in both species. Finally, we characterized 2274 potential simple sequence repeat (SSR) loci, designed primers for 769 of them and validated nine of 29 loci in 42 individuals per species. *Nothofagus obliqua* had more alleles (4.89) on average than *N. nervosa* (2.89), 8 SSRs were efficient to discriminate species, and three were successfully transferred in three other *Nothofagus* species. These resources will greatly help for future inferences of demographic, adaptive and hybridizing events in *Nothofagus* species, and for conserving and managing natural populations.

Keywords: 454 genome sequencing, hybridization, *Nothofagus nervosa*, *Nothofagus obliqua*, single nucleotide polymorphisms identification, species identification, SSR loci

Received 18 February 2014; revision received 23 April 2014; accepted 28 April 2014

Introduction

Nothofagus nervosa (Phil.) Dim. et Mil. and *Nothofagus obliqua* (Mirb.) Oerst. ssp. *obliqua* are two closely related species among the six endemic species of the genus *Nothofagus* occurring in Argentina. Both species have economical value on the international market due to their relatively fast growth and high-quality wood. They also constitute ecologically important native species of temperate forest ecosystems. Clear differences along rainfall and altitudinal gradients suggest that *N. obliqua* is better adapted to drought stress but less well adapted to low temperatures than *N. nervosa* (Veblen *et al.* 1996). During the last century, overexploitation of forest resources, combined with recurrent fires and agricultural settlement, has greatly reduced their original distribution range (Lara *et al.* 1999). Therefore, conservation and domestication programmes were initiated in Argentina in the early 1990s (Gallo *et al.* 2009), with the aims of acquiring basic knowledge about genetic diversity and biological processes shaping its distribution and proposing conservation measures.

Nothofagus nervosa and *N. obliqua* extend in Argentina across a surface of around 79 600 ha and 33 900 ha, respectively (Sabatier *et al.* 2011), and both species distributions follow west–east-oriented lake basins, probably as a consequence of the last glaciations (Glasser *et al.* 2008). Natural interspecific hybridization events have been inferred in sympatric areas (Gallo *et al.* 1997) and have likely occurred at different evolutionary stages in

Correspondence: Pauline Garnier-Géré, Fax: +33-5-57-12-2881; E-mail: Pauline@pierroton.inra.fr

lake watersheds (Marchelli & Gallo 2001). Previously, postglacial history has been studied with chloroplast DNA (cpDNA) (Marchelli *et al.* 1998; Azpilicueta *et al.* 2009), evolutionary forces shaping genetic variation patterns have been discussed with isozymes (Marchelli & Gallo 2001; Azpilicueta & Gallo 2009), and intersimple single repeats and random amplified polymorphic DNA have helped assessing genetic diversity (Mattioni *et al.* 2002). Geneflow and fine-scale genetic diversity studies, however, require more informative markers such as simple sequence repeats (SSRs), which have largely been applied in the last decade in plant population genetics (Varshney *et al.* 2005; Wang *et al.* 2009) despite significant challenges in their development. Commonly used strategies, either de novo genomic library construction (Zane *et al.* 2002) or transfer of known available SSRs from related species (Rossetto 2001); have been applied to South American *Nothofagus* species, but only a few SSRs were obtained (Azpilicueta *et al.* 2004; Marchelli *et al.* 2008; Soliani *et al.* 2010). Moreover, studying demographic and selection histories in *Nothofagus* species' natural populations from molecular diversity patterns calls for much larger amounts of genomic resources.

Next-generation sequencing (NGS) technologies and bioinformatics tools now allow producing genomic resources at reasonable prices and schedules (Mardis 2008), with the increasing development of single nucleotide polymorphisms (SNPs) and SSRs in the last few years in nonmodel species (e.g. Abdelkrim *et al.* 2009; Kumar *et al.* 2012; Zalapa *et al.* 2012; Montes *et al.* 2013). For Fagaceae species, a draft reference genome has been released in 2014 for *Castanea mollisima* (www.hardwoodgenomics.org/), although gene databases and other genome sequencing projects have been under development in the last few years (Neale & Kremer 2011; Neale *et al.* 2013). For Nothofagaceae species, two NGS projects have been reported using either the transcriptome for *N. nervosa* (Torales *et al.* 2012) or the genome for *Nothofagus solandri* (Smissen *et al.* 2012). However, both studies rely on one individual, limiting the detection and development of polymorphic markers.

The general aim of our work was to generate genomic resources for *N. nervosa* and *N. obliqua* in order to aid the study of demographic, adaptive and hybridization processes in their natural range. We used total DNA (i) to complement the available transcriptome resources by obtaining both organelle and nuclear genomic data, (ii) to get both noncoding or nongenic sequences that would potentially be less affected by selection and coding sequences that could be a preferred source of candidate genes for adaptive traits and (iii) to develop SSR and SNP markers for future studies. We report in particular the full gene map of the *Nothofagus* chloroplast genome and identification of 52 mitochondrial genes including protein, rRNA and tRNA genes. We also identified 2274 potential SSRs, 769 of which allowed primer design and are available to validation and transferability to other *Nothofagus* species. We validated nine of them, showed that eight efficiently discriminate both species and transferred three of them to other *Nothofagus* species (*N. antarctica*, *N. dombeyi* and *N. pumilio*). Finally, we identified quality SNPs that allowed preliminary estimates of diversity among organelle and nuclear genomes and divergence among species.

Material and methods

Sample collection and DNA extraction

Fresh leaf material was sampled for *Nothofagus obliqua* and *N. nervosa* from two sites that were located at the opposite ends (west–east) of one of the most important watersheds for these species in Argentina (Sabatier *et al.* 2011) (Table 1). These sites were chosen because they exhibit contrasted longitudinal patterns of genetic diversity, *N. obliqua* showing the greatest genetic diversity in eastern populations under more xeric conditions, and *N. nervosa* being more variable in western humid locations (Azpilicueta *et al.* 2013). DNA was extracted according to Marchelli *et al.* (1998), and their quality was assessed by band intensity and integrity from electrophoresis on a 0.8% agarose gel stained with 1× SYBR Safe (Invitrogen, USA). DNA was quantified with Qubit fluorometer and dsDNA BR Assay Kit (Invitrogen, USA) for 454 sequencing samples and with BioPhotometer (Eppendorf, USA) for SSR validation samples.

Table 1 Geographic location of sampled populations

Species	Site codes	Latitude	Longitude	Altitude	N_{NGS}	N_{SSR}
Nothofagus nervosa	1-Nn	40° 09′ 00′	71° 21′ 00″	980	9	20
N. obliqua	1-No	40° 09′ 00′	71° 21′ 00″	850	5	22
N. nervosa	2-Nn	40° 10′ 12″	71° 40′ 12″	940	8	22
N. obliqua	2-No	40° 10′ 12″	71° 40′ 12″	670	5	20

N_{NGS}, number of individuals sampled for next-generation sequencing, included in N_{SSR}; N_{SSR}, number of individuals sampled for simple sequence repeat validation.

454 sequencing and de novo assembly

DNA was pooled for each species using around half of the individuals sampled (17 in *N. nervosa* and 10 in *N. obliqua* from both sampling sites, Table 1) in very similar quantities to achieve 4 μg of DNA per species. The samples were sent to the DNA Services of the University of Illinois (www.biotech.illinois.edu/htdna) for library preparation and tagging (one for each species), and sequencing on a 454 GS-FLX titanium pyrosequencer (Roche) with a 3/8th run. Different steps were followed with the objective of partitioning the original pool of reads into chloroplast, mitochondrial or nuclear genomes (Fig. 1). First, a de novo assembly was performed on reads from both species simultaneously using the MIRA pipeline version 3.4.0 (Chevreux *et al.* 2004), with options appropriate to consider nonuniform distribution of genomic data (Option A switches: --job=denovo,genome,accurate,454 --fasta --notraceinfo -AS:urd=off 454_SETTINGS -AS:mrpc=1 -OUT:sssip=yes). The –notraceinfo option deals with reads previously extracted and clipped by the program sff_extract v0.3.0 (switches -A -c) provided in MIRA. Five de novo assemblies were produced with the same options to test for assembly stability, as we had noted differences in the structure of the largest contigs from preliminary trials. Various assembly statistics (number of reads assembled, mean–maximum contig length, mean base quality, etc.) obtained from MIRA and the script my_process_contigs.pl (https://github.com/ranjit58/NGS/) were used to choose the best among different assemblies.

To fill gaps in organelle sequences, another MIRA assembly was performed with alternative parameters (Option B switches: --job=denovo,genome,accurate,454 -asta -CL:pec=off:ascdc=off -AS:urd=off 454_SETTINGS -AS:mrpc=1 -OUT:sssip=yes). Clipping parameters (pec and ascdc) were deactivated as recommended in MIRA for obtaining longer contigs when working with low coverage data. Reads extraction was still carried out with the program sff_extract v0.3.0, but clipping was performed using the traceinfo file that integrates quality information. TABLET 1.11.05.03 (Milne *et al.* 2010) was

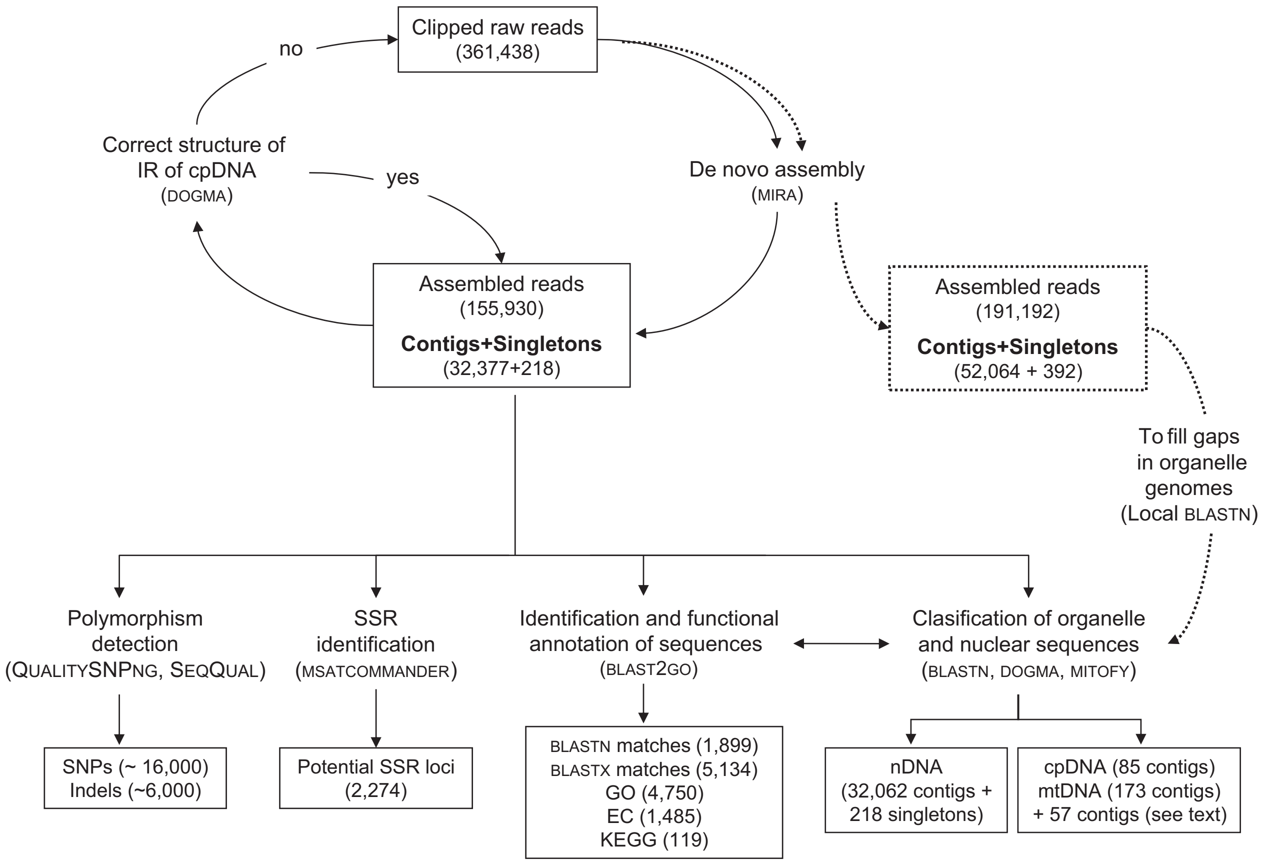

Fig. 1 Summary of the strategy followed for assembly, annotation and classification of genomic resources. Numbers in brackets are from the best MIRA assembly (Notho_clip4). Dotted arrows and box are from the alternative MIRA assembly (Notho_new12). Text in brackets indicates used softwares. See text for assembly parameters and Gene Ontology, Enzyme Commission and Kyoto Encyclopedia of Genes and Genomes descriptions.

used for assembly visualization. The same procedure (Options A and B) was also used to obtain separate assemblies for each species and compare them to their joint assembly.

Functional annotation

The program BLAST2GO (Conesa & Götz, 2008) was used to annotate contigs and singletons from the best assembly: BLASTN and BLASTX searches were performed with *E*-values below 10^{-10} against the NCBI 'nr' nucleotide and protein databases, using the Blast Description Annotator option to extract informative BLAST results for each sequence, and followed by attribution of Gene Ontology (GO) terms, Enzyme Commission (EC) categories and InterproScan and mapping of annotations to the Kyoto Encyclopedia of Genes and Genomes (KEGG). Eleven contigs larger than 8000 bases were analysed at NCBI (http://blast.ncbi.nlm.nih.gov/) because they exceeded the length limit imposed by BLAST2GO. The BLAST tools were also used *a posteriori* to estimate how much of transcriptome-like sequences from Torales *et al.* (2012) were included in our genomic data and to parse results from BLASTX on the most recent 'nr' protein database by excluding 'noninformative' annotations (such as 'unknown', 'U/uncharacterized', 'hypothetical', 'proteins').

Organelle contigs were first identified from BLASTN results and then annotated using the Web-based annotation packages DOGMA (Wyman *et al.* 2004) and MITOFY (Alverson *et al.* 2010). Both packages build upon sequence similarity to protein, rRNA and tRNA genes in other organelle genomes in the database from 15 to 23 plant species in DOGMA and MITOFY, respectively. MITOFY additionally implements tRNAscan-SE (Lowe & Eddy 1997) to corroborate tRNA boundaries identified by BLASTN. Because BLAST does not explicitly look for start and stop codons in protein coding genes, they must be defined by the user, as well as intron/exon boundaries; so we manually adjusted the structure and annotation of each gene in BIOEDIT (Hall 1999), using as references the organelle genomes of *Castanea mollisima*, a species with available data belonging to Fagaceae family closely related to Nothofagaceae. Frameshift or false start/stop codon derived from sequencing or alignment errors were checked with TABLET 1.11.05.03 (Milne *et al.* 2010) and corrected. We then used CODONCODE ALIGNER v.4.0.4 (CodonCode Corporation) to optimize alignments among cpDNA contigs and GENOMEVX (Conant & Wolfe 2008) to generate the gene map of the *Nothofagus* chloroplast genome. The strategy above was used both on the joint species assembly and separate assemblies. Additionally, the chloroplast gene sequence of each species was obtained by mapping their raw reads separately against the obtained *Nothofagus* cpDNA genome consensus sequence (from both species) using MIRA (Option C switches: --job=mapping,genome,accurate,454 --fasta -AS:nop=1: urd=off -SB:lb=yes:bbq=30).

BLASTN and gene map analyses were performed using CGView Server (Grant & Stothard 2008) to compare gene content and structure among cpDNA genomes of *Nothofagus* species and among *Nothofagus* and two Fagaceae species (*C. mollisima* and *Quercus rubra*).

SNP discovery

Assembly files were further post-treated using both QualitySNPng (Nijveen *et al.* 2013) and our own pipeline (SeqQual, bioperl scripts described in Table S1 and code available in Appendix S13, Supporting information) with the general aim of masking/excluding sequence data of poorer quality and identify SNPs with a high quality probability.

QualitySNPng includes information on base quality scores and combines them with additional criteria to propose an integrated strategy that filters assembled NGS data and detects polymorphisms. Three main sequential filtering options are followed, using depth, minor allele frequency (maf), base quality at a potential SNP and in neighbouring positions, and haplotype definitions as additional support for SNP alleles' reliability (see details at http://www.bioinformatics.nl/QualitySNPng). Default parameters were used except for minimal number of reads per allele and haplotypes (set to 2) to adjust to the relatively low coverage of the current project. Briefly, SeqQual is based on an initial extraction of contig alignments (fasta files) from the MIRA assembly (ace file) where poor-quality alignments read ends have been masked. Various post-treatments steps are chosen and applied in batch across contigs that can be visualized in a sequence editor at each step of the process. Here, the pipeline extracts each nucleotide original quality score from raw data and uses it to mask poor-quality bases in alignment files. The occurrence of homopolymer-linked false insertion–deletions (indels) is a serious issue for 454 data (Balzer *et al.* 2011). However, ignoring all indels or homopolymer regions will surely exclude true indels (e.g. insertions surrounded by different nucleotides) and underestimate diversity for genomic data, so we also postprocessed all contigs by only masking deletions potentially due to repeated bases (here with a stringent filter with param=2 based on Gilles *et al's.* (2011) homopolymer definition, see Table S1, Supporting information). A final script produces a range of SNP statistics across post-treated contig alignments (different types of polymorphisms counts, depth and maf, shared and exclusive alleles between species, preliminary divergence statistics (e.g. G_{ST}', Hedrick (2005)). Various statistics' thresholds are then combined for filtering data. Both pipelines differ in the treatment of homopolymers because QualitySNPng can exclude repeated regions based on the repeat number, while SeqQual can extract indels of higher quality not due to homopolymers.

Identification, characterization, validation and transfer of SSR loci

MSATCOMMANDER version 0.8.2 (Faircloth 2008) was used for searching SSRs with recommended criteria for primer design and detection of their annealing sites across contigs and singletons. The screening was made simultaneously for dinucleotides (repeat length ≥ 6) and tri-, tetra-, penta- and hexanucleotides (repeat length ≥ 4). Mononucleotides (homopolymers) were excluded due to their higher probability of sequencing and genotyping errors (Gilles *et al.* 2011). Primer testing was performed first by choosing 29 SSR loci among those with sufficiently long flanking regions for primer design, covering simple and compound SSRs with 5 to higher than 50 repeats depending on SSR types. Three individuals per species were then amplified for each locus and polymerase chain reaction (PCR) products were visualized by electrophoresis on 2% agarose gels. In a second step, 18 loci that showed single-banded and strong amplifications for all samples per species were retained and genotyped with an ABI 3730 XL DNA Analyzer (Applied Biosystems, USA) at the Genotyping Services of CNIA (INTA, Argentina). Finally, nine loci showing allele profiles without ambiguities were chosen and genotyped for 42 individuals per species belonging to the populations of the 454 sequencing project (Table 1). Touchdown (TD) PCR was used for amplifying fragments containing each selected SSR locus according to alternative conditions (Table S2, Supporting information). The SSR profiles were examined and scored using GENEMARKER version 1.95 (SoftGenetics, USA).

Major alleles of the nine genotyped SSRs were sequenced in both species to confirm their initial description (see details in Table S2, Supporting information). Transferability of the nine developed SSRs to other *Nothofagus* species was tested with eight individuals from *Nothofagus antarctica*, *N. dombeyi* and *N. pumilio.* Slightly different TD-PCR programmes and concentrations of $MgCl_2$ were used (Table S2, Supporting information), and Sanger sequences were obtained as already described to confirm amplifications of targeted loci.

SSR data analyses and species or hybrids identification

Number of alleles (N_a), number of effective alleles (N_e), observed (H_O) and expected heterozygosity (H_E) were estimated with GENALEX V6.5 (Peakall & Smouse 2012). Exact tests of Hardy–Weinberg equilibrium were performed with the software ARLEQUIN version 3.5 (Excoffier *et al.* 2005). Null alleles and genotyping errors were checked with MICROCHECKER version 2.2.3 (Van Oosterhout *et al.* 2004). Inbreeding coefficients and null allele frequencies were estimated simultaneously with INEST version 1.0 (Chybicki & Burczyk, 2009) under the individual inbreeding model (IIM) considering its highest accuracy and precision. SSR genotypes' assignment to different clusters was tested with STRUCTURE v2.3.3 (Falush *et al.* 2007), using 10 replicates of an admixture model allowing for correlated allele frequencies with K ranging from 1 to 10, a burn-in period of 100 000 iterations and a post-burn-in period of 200 000 iterations, following recommendations by Gilbert *et al.* (2012). Twelve putative hybrids based on phenotypic data (i.e. showing intermediate traits of leaves and bark) were also sampled in the central region of the same watershed, and belonged to neighbouring populations where individuals from Table 1 originated, in order to test the usefulness of the developed SSR markers for identifying species and detecting hybrids.

Results

Assembly and functional annotation

A total of 361 438 reads were obtained from the 454 3/8th run, with an average read length of 313 base pairs (bp). Mean contig length, mean base quality and other statistics used to compare assemblies were very similar (Table S3, Supporting information). One of the five assembly replicates showed a clear and incorrect chimeric assembly after one of the inverted repeats (IRa), with truncated ycf1, ndhF and rpl32 located after trnNGUU (see Material and methods above and Results below) so it was discarded. The assembly giving the correct structure and the best statistics (Notho_clip4) among all others was finally retained (Table S3, Supporting information and Fig. 1), with an average consensus quality score of 42. Forty-three per cent of the reads were assembled in 32 377 contigs and 218 singletons totalizing 13.42 Mb, with an average of 4.78 reads per contig (up to 5033) and a mean length of 411.7 bp (range: 40–45 364) (Table 2). Combining reads of both species improved the assembly quality based on the maximum contig length, compared with species separate assemblies (Table S3, Supporting information).

With an E-value below 10^{-10}, ~16% (5134) of the contigs had significant BLASTX matches (Table 2), 3.7% only for singletons. The most represented species among BLASTX best hits (focusing on informative annotations, see Material and methods) were *Theobroma cacao*, *Vitis vinifera*, and *Medicago truncatula*, *Arabidopsis thaliana* and *Populus trichocarpa* being among the 12 most hit species (Table S4, Supporting information). GO terms were

Table 2 General assembly and functional annotation statistics of *Nothofagus* contigs

	Total contigs + singletons	Nuclear contigs	All cpDNA contigs	Main cpDNA contigs	mtDNA contigs
Total number	32 595	32 061	86	5	173
Average consensus quality	42	42	41.5	76.2	57.82
GC content (%)	38.17	38.16	35.21	33.83	45.23
N50 contig length (bp)	527	319	421	—	1266
Nb of contigs in N50	7094	16 052	43	—	87
Mean length (bp)	411.7	399.3	1894.7	26 000.2	2203.39
Minimal length (bp)	40	40	74	1475	56
Maximal length (bp)	45 364	11 847	45 364	45 364	19 950
Sum of length (bp)	13 419 398	12 845 815	162 944	130 012	381 186
Average number of reads	4.78	4.25	122.30	2046.00	49.29
Maximal number of reads	5033	3489	5033	5033	508
Total number of assembled reads	155 930	136 230	10 518	10 230	8527
Average coverage (mean–max across contigs)	2.1–143.5	2.1–143.5	3.3–33.2	19.5–33.2	5.0–12.3
Maximum coverage (mean–max across contigs)	3.3–275	3.3–275	7.6–236	77.4–236	10.5–58
Nb of seq. with BLASTX hits	5134	4991	32	—	93
Mean length of seq. with BLASTX hits (bp)	653.3	620.6	616.4	—	2444.2
Nb of seq. with GO terms	4750	4621	31	—	82
Nb of seq. with EC codes	1485	1434	15	—	34
Nb of seq. matching Torales data	6116	—	—	—	—
With BLASTX hits matching Torales data	1430	—	—	—	—
With no BLASTX hits matching Torales data	4686	—	—	—	—
Total Nb of identified seq. with transcribed regions	9820	—	—	—	—
Nb of unigenes	2350	—	—	—	—
Nb of unigenes not matching Torales data	1750	—	—	—	—
Mean length of unigenes (bp)	1400	—	—	—	—

EC, Enzyme Commission; GO, Gene Ontology; Nb, number.
N50 contig length: the length for which the cumulative sum of contig lengths equal or higher than this value corresponds to 50% of total contig length sum. BLASTX matches considering an *e*-value threshold of $<10^{-10}$ and nr databases b2g_nov12/b2g_sep13 and excluding contigs above 8000 bp. 'Torales' refers to Torales *et al's.* (2012) transcriptome data. See text and Table S4 (Supporting information) for GO and EC descriptions. 'Total contigs + singletons' is either the sum or average of statistics across all contigs and singletons, including also organelle contigs that could not be clearly assigned to either cpDNA or mtDNA.

assigned to 92.5% of sequences with BLASTX matches (4743 contigs and seven singletons in Table 2, consensus sequences in Appendix S5 (Supporting information), and annotation description, GO terms and EC codes in Table S4, Supporting information). Over 3000 sequences were assigned to 'Molecular Functions' (MF) and 'Biological Processes' (BP), and <1500 to 'Cellular Components' (CC). GO assignment at level 2 for BP showed that apart from 'metabolic process' and 'cellular process' (~2500 sequences), 'response to stimulus' also included a few hundred contigs (Fig. S1, Supporting information), many of them being assigned to the 'response to stress' term at level 3 (Fig. S1-A and E, Supporting information). Enzyme Category was assigned to 1485 (~29%) contigs (Table 2), which were mapped to 119 KEGG pathways.

For 6116 sequences (including 26 singletons), a very high homology was observed with 557 isotigs and 3008 singletons from Torales *et al's.* (2012) transcriptome data (Table 2). Along these, 4686 were not included in the 5134 sequences with BLASTX hits, giving a total of 9820 identified sequences that would include gene transcribed regions (Table 2). Moreover, among sequences with BLASTX results, 3704 did not have any match in the transcriptome data, suggesting that they could represent novel identified gene regions for *Nothofagus nervosa* and *N. obliqua* (Table 2 and Table S4, Supporting information). Based on GO terms and sequence description (Table S4, Supporting information), and excluding redundant annotations, we obtained ~2350 unigenes, ~1750 of which were not characterized in the published *N. nervosa* transcriptome (Table 2 and Table S4, page 2, Supporting information). Applying BLASTN among all contig-annotated sequences led to much less redundancy than when using annotations, only 0.03% showing a positive match with high similarity. Thus, different contigs probably correspond to different sequenced parts of the same gene regions.

In total, 316 contigs totalizing 573 583 bp (Table 2) were attributed to organelles (including 10 contigs larger than 8000 bp, seven for mitochondrial DNA and three for cpDNA) and classified according to BLASTN, DOGMA and MITOFY mifoty analyses into: (i) cpDNA contigs (86), (ii) mtDNA contigs (173) and (iii) organelle contigs that could not be clearly assigned to either cpDNA or mtDNA genome (57).

Five of 86 cpDNA consensus sequences represented more than 99% of the genome according to DOGMA annotation, and the remaining gaps could not be filled despite attempts to manually improve their assembly with CODONCODE. Including also the 57 ambiguous organelle sequences did not improve the length of any contig. Therefore, we considered the five original contigs the best that we could obtain and used them to reconstruct the chloroplast genome map (Fig. 2 and Table 2). An alternative MIRA assembly (Notho_new12, see Material and methods and Fig. 1) was, however, useful for completing two gaps (98 and 24 bp). It allowed recovering a few larger contigs than those from the Notho_clip4 assembly using local BLASTN, although being less accurate globally (see average consensus quality in Table S3, Supporting information). The length of two other gaps remained unknown, although they are likely to be small based on the gap length above. Finally, 96% of the predicted chloroplast genome size was recovered, in reference to *Castanea mollisima* (Table 3). The 112 chloroplast-annotated genes included ribosomal RNA (4), transfer RNA (30) and protein (78) genes (Table 3). The detailed gene map organization of cpDNA for *Nothofagus* species shows one large single copy (LSC), one small single copy (SSC) and two inverted regions (IRs), with introns detected for 18 genes and duplication observed for 17 genes (Fig. 2 and Table 3). The only pseudogene detected was rpl22 whereas ndhD and rps19 showed alternative start codons ACG and GTG, respectively (detailed description and consensus sequences of all genes given in Table S6 and Appendix S7, Supporting information). The same cpDNA genome consensus sequences and gene content or structure were obtained from either the joint or separate *Nothofagus* species assemblies (Appendix S8 and Fig. S2, Supporting information). Their gene map did not show any structural changes when aligned and compared against those of two other Fagaceae species (Fig. S3, Supporting information).

Table 3 Comparison of major structural features between *Nothofagus* spp. and *Castanea mollisima* chloroplast genomes

	Nothofagus spp.	*C. mollissima*
Total observed size (bp)	155 513*	160 799
LSC size (bp)	85 484*	90 432
SSC size (bp)	17 865*	18 995
IR size (bp)	26 082	25 686
Number of genes	112 (infA)	111
Number of duplicated genes	17	17
Number of genes with introns	18	18
Pseudogene	rpl22	rpl22
Alternative start codon	ndhD → ACG rps19 → GTG	ndhD → ACG

LSC, large single copy; SSC, small single copy; IR, inverted repeats.

*See gaps locations in Fig. 2.

Functional and structural annotations were also obtained for 46 full-length mitochondrial genes (three ribosomal RNA, 13 transfer RNA and 30 protein genes) and for six other protein genes with partial sequences (Table S6, Supporting information). Ten genes showed introns, and the alternative start codon ACG was detected in three genes (Table S6, Supporting information). Overall, these 52 genes represent around 70% of the gene content reported across 24 plant species (Lei *et al.* 2013) and are within the average observed across eukaryotes (40–50 genes, Burger *et al.* 2003).

SNP detection and preliminary diversity analyses

We obtained 14% more polymorphisms from the 32 377 contigs with SeqQual+SNP-statistics (23 922) compared with QualitySNPng (20 436) using their less stringent quality filters, but the convergence between both pipelines increased with more stringent filters (Table 4, filter b vs. f and c,d,e vs. g,h,i). This illustrates the consistency of both approaches and the critical need for applying quality filters on raw data as only around 7% of originally detected polymorphisms were retained when integrating quality. Fewer polymorphisms per bp were detected on average in organelle than in nuclear contigs, the five cpDNA contigs used to reconstruct the chloroplast genome map showing values below 1.5% (Fig. 3A and see solid ovals in 3B). Moreover, the proportion of SNPs among all polymorphisms was only 4.3% or 8.5% in organelles compared with 74.7% or 81.7% in nuclear contigs (Table 4, filters b and f). As expected, the number of polymorphisms decreased with more stringent filters (Table 4, filters c,d,e and g,h,i), but the lower diversity trend in organelles vs. nuclear contigs was conserved. Masking globally low-quality regions in QualitySNPng decreases sharply the number of polymorphisms (see filters f/h vs. g/i), whereas the targeted masking of indels due to homopolymers by SeqQual+SNP-statistics for all bases allows to identify not only true potential indels but

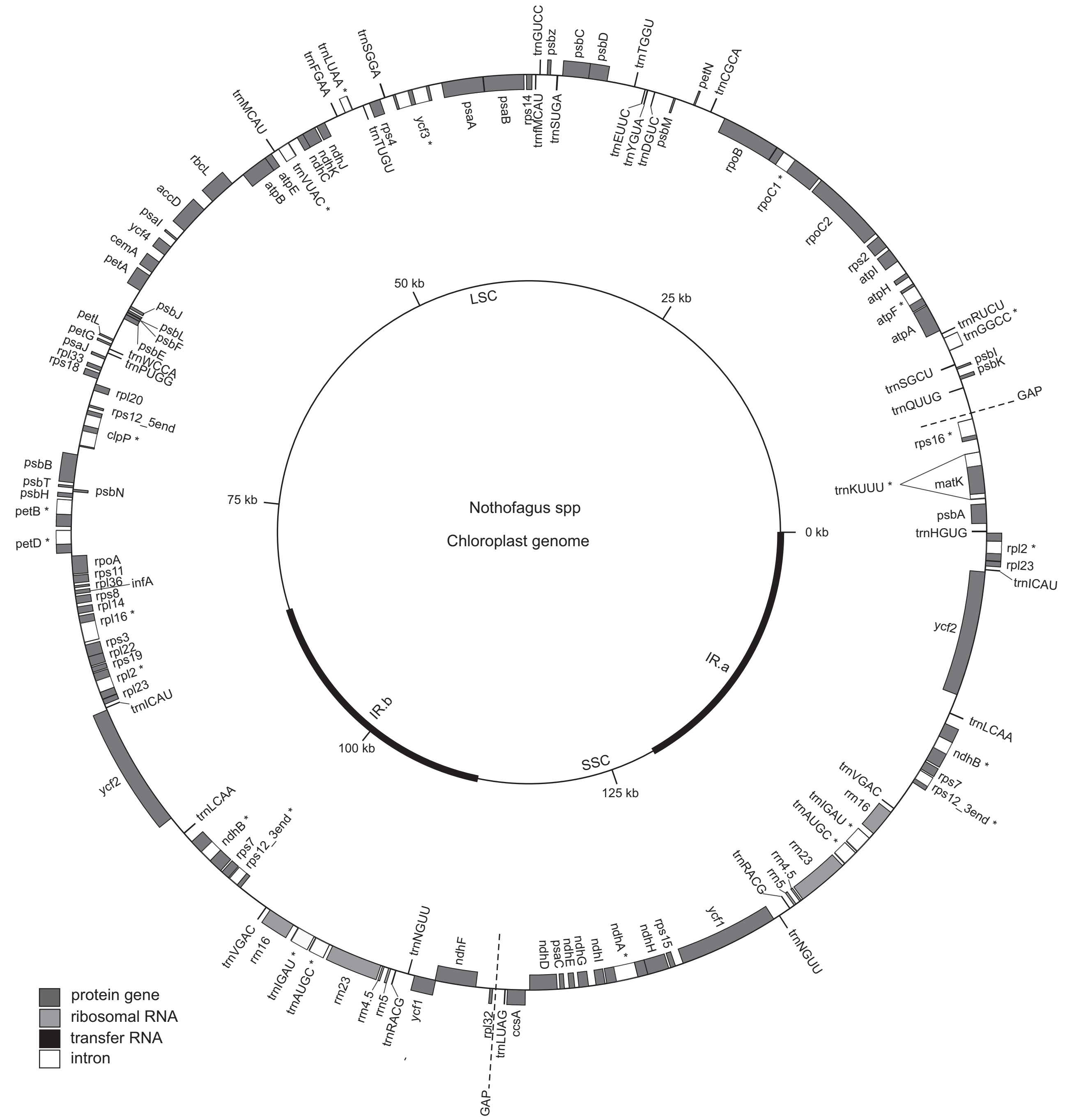

Fig. 2 Circular gene map of the joint chloroplast genome of both Nothofagus species.

also a small number of additional SNPs (see filter b vs. d and c vs. e). There was a clear trend for nuclear contigs to be less polymorphic as their length increase (Fig. 3A), which could be consistent with a higher assembly efficiency (longer contigs) for more conserved regions of the genome. The short contigs with relatively high polymorphism rates have been manually checked and likely result from paralogs assembly; however, the trend is maintained if we exclude values above 4%. For organelles, the few short contigs with the highest values (>1.5%) correspond to mtDNA or ambiguous organelle contigs (Fig. 3B, dashed oval), and they are also probably due to incorrect assembly of paralogs or repeat regions.

A lower divergence among species was observed for organelle compared with nuclear genomes (Table 5). In nuclear contigs, different alleles were observed between

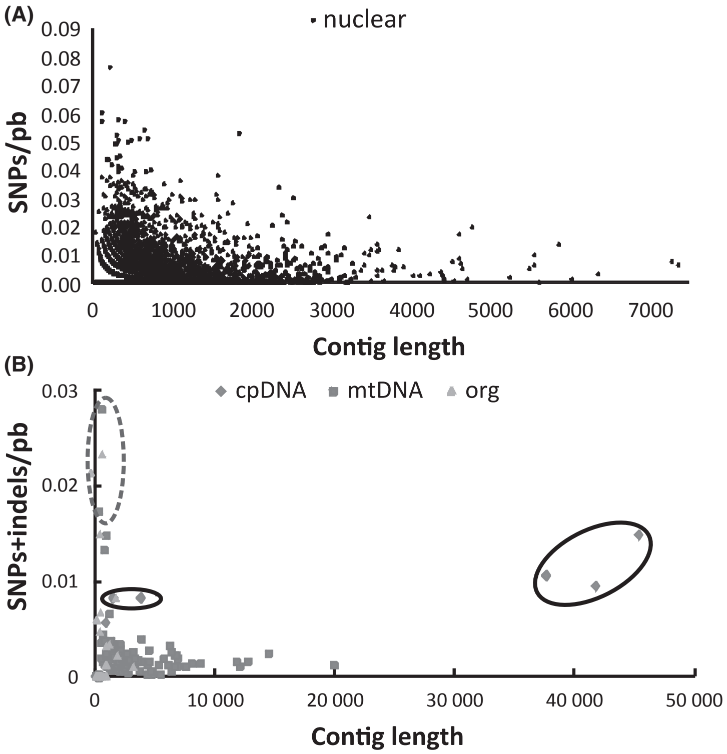

Fig. 3 Number of high-quality filtered polymorphisms per base pair (bp) across contigs ranked by their length in bp (data from qual20, man = 2, nb of reads = 4). (A) Nuclear contigs (showing only single nucleotide polymorphisms for clarity). (B) Organelle contigs.

Table 4 Comparison of polymorphisms' detection between organelle and nuclear contigs using two softwares and different filters

		Total	Organelle				Nuclear			
Filter	Method and filter description	Nb of SNPs + indels	Nb of SNPs + indels	Nb of SNPs	Nb of indels	% of SNPs	Nb of SNPs + indels	Nb of SNPs	Nb of indels	% of SNPs
a	Total nb without any filter	300 894	16 794	1720	15 074	10.24	284 100	166 502	117 598	58.61
	SeqQual+SNP-statistics (qual20, man ≥ 2, nb of reads ≥ 4)									
b	Total nb	23 922	2152	182	1970	8.46	21 770	17 781	3989	81.68
c	b+ excluding contigs with nb of reads <6	12 663	1847	131	1716	7.09	10 816	8507	2309	78.65
d	b+ homopolymers filter	21 806	1309	184	1125	14.06	20 497	17 857	2640	87.12
e	c+ homopolymers filter	11 137	1066	133	932	12.48	10 071	8536	1535	84.76
	QualitySNPng (qual20, conf5, nHQ2, man ≥ 2, nb of reads ≥ 4)									
f	Total nb	20 436	2489	108	2381	4.34	17 947	13 403	4544	74.68
g	f+ excluding contigs with nb of reads <6	13 382	2392	92	2300	3.85	10 990	8043	2947	73.18
h	f+ repeats regions set as LQ (*w* = 6, rep = 5)	16 622	1113	95	1018	8.54	15 509	12 544	2965	80.88
i	h+ excluding contigs with nb of reads <6	10 508	1059	80	979	7.55	9449	7493	1956	79.30

qual20, 454 data quality score 20; man, minimal number of reads per allele; LQ, low-quality regions; conf5, score confidence according to QualitySNPng; nHQ2, minimal number of high-quality reads for allele; SNPs, single nucleotide polymorphisms; nb, number.
Numbers for filter f to i are reliable SNPs according to QualitySNPng.
Nb of indels for filters d and e are potentially true indels.

Table 5 Comparison of divergence between *Nothofagus nervosa* and *N. obliqua* based on organelle and nuclear contigs' analyses using SeqQual+SNP-statistics filters (all numbers estimated with base quality score ≥20 and homopolymer indels excluded)

	Organelle				Nuclear			
Filter description and divergence statistics	Range of reads (nb)	Average nb of reads	Nb of SNPs	Nb indels	Range of reads (nb)	Average nb of reads	Nb of SNPs	Nb indels
all-Nn ≥ 2 all-No ≥ 2, no singleton overall (man ≥ 2)	4–200	15.9	104	929	4–243	10.0	8583	1195
Nb of shared polymorphisms	4–120	15.4	25	492	4–243	12.2	2403	536
Nb of exclusive polymorphisms	4–200	16.4	79	437	4–243	9.0	6180	659
Nb of exclusive polymorphism with G_{ST} > 0.5	4–20	6.0	13	76	4–37	5.4	4053	320
Nb of G_{ST} = 1	4–12	5.8	5	29	4–37	5.0	2480	168
all-Nn ≥ 3 all-No ≥ 3, no singleton overall (man ≥ 3)	6–200	19.6	23	292	6–229	20.1	1040	297
Nb of shared polymorphisms	6–120	17.5	6	202	6–209	22.9	491	221
Nb of exclusive polymorphisms	7–200	23.6	17	90	6–229	17.0	549	76
Nb of exclusive polymorphism with G_{ST} > 0.5	7–20	9.1	2	6	6–37	7.8	354	20
Nb of G_{ST} = 1	10–12	9.0	1	1	6–37	7.1	192	4

all-Nn/all-No, minimal number of reads per species; man, minimal number of total reads per allele; SNPs, single nucleotide polymorphisms; nb, number.

species for ~18% and ~29% of the SNPs, depending on the depth, whereas they were different for <5% in organelle contigs (Table 5). For contigs represented by few reads, the allele frequency estimation has got a very large variance, so many alleles in this case might not be 'true' fixed alleles, but the range of read number for this comparison is similar so likely to produce a similar bias. Among the few SNPs in organelles with higher divergence (13 and 2 for both filter sets, respectively) all belonged to mtDNA, suggesting high conservation for cpDNA contigs among species. The proportion of shared polymorphisms is comparatively much higher in organelle (50%) than in nuclear regions (30%), and these are likely to remain shared even with a greater sampling effort. Consistently, exclusive polymorphism proportions (SNPs+indels), those with $G_{ST} > 0.5$ in particular, are only 2.5%/8.6% in organelles depending on the filters compared with higher values of 28%/45% in nuclear contigs (Table 5).

Characterization of SSR loci and polymorphism detection

We detected 2274 SSRs located in 2190 contigs and 84 singletons (Table S9, Supporting information), 1923 being simple repeats (85%) and 351 complex repeats (15%, including compound and interrupted compound repeats). Dinucleotide repeats (1060 representing 55%) were the most frequently detected motifs among simple SSRs, followed by 664 trinucleotide repeats (35%). Primers could be designed for 764 contigs and five singletons (Table S9, Supporting information). Among the 29 loci chosen (see Material and methods), 18 gave good amplifications across both species. Among these, nine could easily be genotyped in 42 individuals per species (Table 1), the rest being discarded because of complex peak patterns. These loci were polymorphic for at least one species, with the exception of Notho228 that was fixed in both but for different alleles (Table S10, Supporting information). Among the eight polymorphic loci, *N. nervosa* showed 19 exclusive alleles (i.e. not found in *N. obliqua*), 11 with frequencies higher than 10% and *N. obliqua* had 37 exclusive alleles, 16 with frequencies higher than 10% (Tables 6 and S10, Supporting information). Sanger sequences of major alleles for the nine loci confirmed the presence of SSRs and their divergence between species (Table S11, Appendix S12, Supporting information). No evidence of large allele dropout was detected among these SSRs, while highly significant shortage of heterozygote genotypes with alleles of one repeat unit difference was detected for the locus Notho218 in *N. obliqua*, suggesting genotyping errors due to stuttering. However, homozygote genotypes for the low-frequency allele (159) of this locus were confirmed by re-genotyping and sequencing (Table S11, Appendix S12, Supporting information).

A higher polymorphism was observed in *N. obliqua* (e.g. 4.89 alleles on average) than in *N. nervosa* (2.89 alleles) (Table 6). Significant heterozygote deficiencies were observed in both species, especially in *N. obliqua* (Table 6), but no evidence of inbreeding was detected when accounting for the presence of possible null alleles

Table 6 Summary of SSR genetic diversity statistics for *Nothofagus nervosa* and *N. obliqua*

Locus	N	N_a	N_e	N_{Ex}	H_O	H_E	HW test	Null
N. nervosa								
Notho_224	42	1	1.000	0	0.000	0.000	—	0.095 (0.068)
Notho_218	37	2	1.583	2	0.486	0.373	0.08215	0.037 (0.032)
Notho_228	25	1 [2]	1.000 [1.277]	1	0.000	0.000 [0.222]	—	0.124 (0.087)*
Notho_219	42	5	2.562	2	0.500	0.617	0.09939	0.093 (0.049)
Notho_214	42	5 [6]	1.536 [2.032]	5	0.190	0.353 [0.514]	0.00041	0.157 (0.064)*
Notho_227	30	4 [5]	2.192 [2.818]	4	0.367	0.553 [0.656]	0.06014	0.128 (0.069)*
Notho_226	42	1	1.000	0	0.000	0.000	—	0.094 (0.067)
Notho_216	41	3	2.024	1	0.561	0.512	0.86132	0.046 (0.036)
Notho_204	42	4	1.898	4	0.500	0.479	0.08565	0.036 (0.031)
Mean over loci	38.111	2.889 [3.222]	1.644 [1.799]	2.11	0.289	0.321 [0.375]		
SE over loci	2.118	0.564	0.190	1.73	0.081	0.085		
F_{IS} overloci							0.013 (0.015)	
N. obliqua								
Notho_224	42	3	1.606	2	0.405	0.382	0.68886	0.047 (0.038)
Notho_218	42	2 [3]	1.100 [1.463]	2	0.000	0.092 [0.320]	0.00048	0.163 (0.074)*
Notho_228	35	1 [2]	1.000 [1.224]	1	0.000	0.000 [0.186]	—	0.102 (0.072)*
Notho_219	42	16 [17]	8.205 [9.239]	13	0.667	0.889 [0.902]	0.00207	0.118 (0.044)*
Notho_214	42	3	1.487	3	0.310	0.331	0.69847	0.069 (0.049)
Notho_227	42	6 [7]	1.918 [2.681]	6	0.143	0.485 [0.634]	0.00000	0.253 (0.066)*
Notho_226	42	2	1.049	1	0.048	0.047	100.000	0.084 (0.062)
Notho_216	42	3	1.423	1	0.238	0.301	0.04750	0.087 (0.057)
Notho_204	36	8 [9]	4.679 [4.604]	8	0.167	0.797 [0.794]	0.00000	0.350 (0.064)*
Mean overloci	40.556	4.889 [5.750]	2.496 [2.753]	4.11	0.220	0.369 [0.433]		
SE overloci	0.959	1.567	0.807	3.90	0.072	0.105		
F_{IS} overloci							0.023 (0.028)	

N, sample size; N_a, allele number; N_{Ex}, exclusive allele number; N_e, number of effective alleles = $1/\Sigma pi^2$; H_O, observed heterozygosity = No. of heterozygotes/N; H_E, expected heterozygosity = $1-\Sigma pi^2$, pi is the frequency of the ith allele for population i; HW, P-value of the Hardy–Weinberg equilibrium exact test; Null, null allele frequency estimate, standard error in parentheses; F_{IS}, inbreeding coefficient; SSR, simple sequence repeat.
Data in brackets are for statistics after corrections for null alleles.
*Loci with significant null allele frequency.

when estimating F_{IS} values (Table 6). Among the six loci with homozygote excess, three showed missing data ranging from 4.5% to 14.3% (Notho228/227/204 in Table 6), where null alleles could be due to mutations within primer annealing regions, as genotyping and DNA quality had been previously validated. For the other three loci with no missing data (Notho218/214/219), departures from neutrality and the presence of undetected hybrids among individuals need to be considered. Finally, tests of transferability in other species yielded good amplifications of three SSRs in eight individuals of *Nothofagus antarctica*, *N. dombeyi* and *N. pumilio*, with polymorphism and new alleles within and among species for two of them (Table S10, Supporting information).

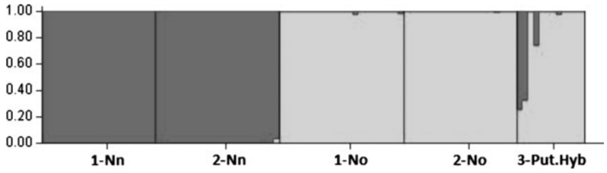

Fig. 4 Assignment probabilities of *Nothofagus nervosa* (Nn), *N. obliqua* (No) and putative hybrids (Put.Hyb) individuals across two STRUCTURE genetic clusters. For each individual, vertical lines are partitioned into two segments that represent its probability to belong to each cluster.

Species and hybrids identification

Data on the nine validated SSRs for 84 individuals from *N. nervosa* and *N. obliqua* (Table 1), and 12 putative hybrids (see Material and methods), were used in the STRUCTURE analysis. Individuals *a priori* belonging to different species were clearly assigned to different clusters, the model with $K = 2$ showing the highest probability (Figs 4, S4 and S5, Supporting information). Average inferred ancestry for both parental species (0.997) was very close to extreme opposite values, indicating clear species divergence. Among the 12 putative hybrids, three showed a probability of belonging to either cluster (species) between 0.125 and 0.875 (Fig. 4), consistently with the expected range of values for first and higher generation hybrids (Guichoux *et al.* 2013). The other nine were assigned to the *N. obliqua* cluster, suggesting that they could be offspring of successive backcrosses with that species, and/or that intraspecies phenotypic variation within both species could be higher and might not be reliable enough and should be integrated when defining traits for identifying hybrids.

Discussion

Genomic resources have greatly increased in forest tree species in the last decade, but have only been influenced recently by next-generation sequencing technologies (Neale & Kremer 2011). Using a 454 pyrosequencing run on genomic DNA from 54 gametes of two hybridizing species of *Nothofagus*, we developed the first whole-genome resources for these important species of the South American forests ecosystems complementing the recently sequenced transcriptome of *Nothofagus nervosa* (Torales *et al.* 2012): SNPs and SSRs markers, novel-annotated candidate genes and noncoding genomic sequences for nuclear genomes of *Nothofagus* species, and organelle genomes content and structure. A high-quality goal for data production and assembly was pursued and illustrated by successful amplification of SSR loci, validation of their sequences and a very small proportion (0.03%) of contigs with annotations blasting onto themselves. These 454-derived genomic resources will be of great help to better understand demographic, adaptive and hybridizing processes among *Nothofagus* species as discussed below.

Preliminary analyses in both *Nothofagus* species sampled across the same watershed region showed a very low diversity in cpDNA genomes overall (compared with the nuclear genome), with a large proportion of detected polymorphisms being shared. This is consistent with previous reports of shared cpDNA haplotypes between Nothofagaceae species indicative of chloroplast capture (Azpilicueta *et al.* 2009; Premoli *et al.* 2012). Pseudogenization of rpl22 was detected for *N. nervosa* and *N. obliqua* according to the loss of this gene reported for Fagaceae species (Jansen *et al.* 2011). The gene order description of the chloroplast genome for two *Nothofagus* species will also serve as a reference for phylogenetics, phylogeography and speciation studies among Fagaceae and Nothofagaceae (e.g. Yang *et al.* 2013).

The new resources include nongenic parts of the genome likely less affected by selective effects and thus useful for future demographic inference studies. They include also novel candidate gene regions (for abiotic and biotic stresses in particular, see Table S4, Supporting information) that will help unravelling adaptive molecular processes in relation to variation in environmental conditions. Assuming ~30 000 genes for an average plant eudicot genome (Sterck *et al.* 2007), and applying the level of redundancy observed in annotated contigs to all genic sequences identified (9820, see Results), we may have targeted overall up to ~4000 gene regions in the *Nothofagus* genome, hence around 13% of the expected gene number. Besides, with a sequenced length estimate of ~1400 bp per gene (so ~5.7 Mb in length, see Table S4, end of page 2, Supporting information), which would cover at least around 50% of their expected length given sizes of 2–3 kb in plants (e.g. Bevan & Walsh 2005; rice.plantbiology.msu.edu), we have access to ~6.5% of *Nothofagus* total genic partition. The rest of the sequences available (7.7 Mb, 60% of the total length assembled) would represent between 1% and 2% of the nongenic partition of the genome, assuming a putative size of ~500 to ~850 Mb (see Fagaceae values at data.kew.org/cvalues/) and 10% to 18% of gene regions (i.e. 3 kb*30000 genes giving 90 Mb). This is consistent with a less efficient assembly expected in nongenic regions containing more retrotransposons or highly repeated regions.

Preliminary polymorphism number estimates showed a large variation across nuclear contigs and a mixture of potentially highly divergent regions with regions including shared polymorphisms between species. The large proportion of exclusive polymorphisms and their patterns across the genome in particular could be confirmed and further studied by undertaking larger scale resequencing projects either at the levels of genes or genotypes. Resequencing projects focused on candidate genes for particular ecologically important traits could also help identify SNPs of adaptive significance and monitor their genetic diversity in restoration or conservation programmes. Combining candidate genes and organelle genome data will therefore be useful for better characterizing the directionality of hybridization and introgression between *N. obliqua* and *N. nervosa* in the context of their speciation history (Burgess *et al.* 2005; Sun & Lo 2011).

We derived amplification primers for 769 SSRs that can be a useful resource for future marker development. We could assign 13% of them to gene regions in contrast to noncoding regions, providing SSRs that would allow focusing on different evolutionary factors in population genetics inference. We also showed that nine of 29 SSRs could be easily validated in a large sample of individuals in both species. Applying the same success rate to all putative loci would yield more than 150 useful SSRs. Considering the nine developed markers, both the average number of alleles and expected heterozygosity were higher in *N. obliqua* compared with *N. nervosa.* Overall, however, diversity was lower here than in the study using seven SSRs by Azpilicueta *et al.* (2013), probably because of their higher number of populations covering a larger geographic sampling. For studying hybridization patterns among species in thoroughly sampled mixed stands, our new markers would be useful as eight of nine combine a high number of species-specific alleles (i.e. different alleles fixed in both species) in relatively high frequency (>10%) and the large number of allelic configurations that would possibly allow telling apart first- and second-generation hybrids. This clearly differs from the seven SSRs of Azpilicueta *et al.* (2013), which exhibit a majority of shared alleles between species, with exclusive alleles being rare and localized to particular populations. The differences in the strategies applied for developing these seven SSRs (see Azpilicueta *et al.* 2004; Marchelli *et al.* 2008; Soliani *et al.* 2010) could explain their patterns due to more conserved regions targeted among species, while we derived SSRs from both species' combined assembly and thus did not exclude more divergent regions among species.

The resources developed here will therefore also be a valuable tool to develop polymorphic markers in several species, including those growing in other continents, hence assisting their conservation, restoration and management in natural environments.

Acknowledgements

The authors thank Paula Marchelli and María Marta Azpilicueta for providing DNA for 454 sequencing and for genotyping of developed markers; and Susana Torales who kindly provided transcriptome resources of *N. nervosa* for comparison with genome data. The 454 sequencing work was funded by the EVOLTREE network of Excellence (EU contract no. 016322) and SSR genotyping by PNFOR 44321 from INTA (Argentina) and Programa de Mejoramiento de Especies Forestales (PROMEF) BIRF 7520-AR.

References

Abdelkrim J, Robertson BC, Stanton J-AL, Gemmell NJ (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques*, **46**, 185–192.

Alverson AJ, Wei X, Rice DW *et al.* (2010) Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution*, **27**, 1436–1448.

Azpilicueta MM, Gallo LA (2009) Shaping forces modelling genetic variation patterns in the naturally fragmented forests of a South-American Beech. *Biochemical Systematics and Ecology*, **37**, 290–297.

Azpilicueta MM, Caron H, Bodénès C, Gallo LA (2004) SSR markers for analysing South American *Nothofagus* species. *Silvae Genetica*, **53**, 240–243.

Azpilicueta M, Marchelli P, Gallo L (2009) The effects of Quaternary glaciations in Patagonia as evidenced by chloroplast DNA phylogeography of Southern beech *Nothofagus obliqua*. *Tree Genetics & Genomes*, **5**, 561–571.

Azpilicueta MM, Gallo LA, van Zonneveld M *et al.* (2013) Management of Nothofagus genetic resources: definition of genetic zones based on a combination of nuclear and chloroplast marker data. *Forest Ecology and Management*, **302**, 414–424.

Balzer S, Malde K, Jonassen I (2011) Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics*, **27**, 304–309.

Bevan M, Walsh S (2005) The Arabidopsis genome: a foundation for plant research. *Genome Research*, **15**, 1632–1642.

Burger G, Gray MW, Franz Lang B (2003) Mitochondrial genomes: anything goes. *Trends in Genetics*, **19**, 709–716.

Burgess KS, Morgan M, Deverno L, Husband BC (2005) Asymmetrical introgression between two *Morus* species (*M. alba*, *M. rubra*) that differ in abundance. *Molecular Ecology*, **14**, 3471–3483.

Chevreux B, Pfisterer T, Drescher B *et al.* (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, **14**, 1147–1159.

Chybicki IJ, Burczyk J (2009) Simultaneous estimation of null alleles and inbreeding coefficients. *The Journal of Heredity*, **100**, 106–113.

Conant GC, Wolfe KH (2008) GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics*, **24**, 861–862.

Conesa A, Götz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, **2008**, 1–12.

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.

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

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Gallo L, Marchelli P, Breitembucher A (1997) Morphological and allozymic evidence of natural hybridization between two southern beeches (*Nothofagus* spp.) and its relation to heterozygosity and height growth. *Forest Genetics*, **4**, 15–23.

Gallo LA, Marchelli P, Chauchard L, Peñalba MG (2009) Knowing and doing: research leading to action in the conservation of forest genetic diversity of Patagonian temperate forests. *Conservation Biology*, **23**, 895–898.

Gilbert KJ, Andrew RL, Bock DG *et al.* (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. *Molecular Ecology*, **21**, 4925–4930.

Gilles A, Meglécz E, Pech N *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **12**, 245.

Glasser NF, Jansson KN, Harrison S, Kleman J (2008) The glacial geomorphology and Pleistocene history of South America between 38°S and 56°S. *Quaternary Science Reviews*, **27**, 365–390.

Grant JR, Stothard P (2008) The CGView Server: a comparative genomics tool for circular genomes. *Nucleic Acids Research*, **36**, W181–W184.

Guichoux E, Garnier-Géré P, Lagache L *et al.* (2013) Outlier loci highlight the direction of introgression in oaks. *Molecular Ecology*, **22**, 450–462.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

Jansen RK, Saski C, Lee S-B, Hansen AK, Daniell H (2011) Complete plastid genome sequences of three Rosids (Castanea, Prunus, Theobroma): evidence for at least two independent transfers of rpl22 to the nucleus. *Molecular Biology and Evolution*, **28**, 835–847.

Kumar S, Banks TW, Cloutier S (2012) SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics*, **2012**, Article ID: 831460. doi: 10.1155/2012/831460

Lara A, Rutherford P, Montory C *et al.* (1999) Mapeo de la Eco-región de los bosques Valdivianos. *Boletín Técnico Fundación Vida Silvestre Argentina*, **51**, 1–27.

Lei B, Li S, Liu G *et al.* (2013) Evolution of mitochondrial gene content: loss of genes, tRNAs and introns between *Gossypium harknessii* and other plants. *Plant Systematics and Evolution*, **299**, 1889–1897.

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**, 955–964.

Marchelli P, Gallo L (2001) Genetic diversity and differentiation in a southern beech subjected to introgressive hybridization. *Heredity*, **87**, 284–293.

Marchelli P, Gallo L, Scholz S, Ziegenhagen B (1998) Chloroplast DNA markers reveal a geographical divide across Argentinean southern beech *Nothofagus nervosa* (Phil.) Dim. et Mil. distribution area. *Theoretical and Applied Genetics*, **97**, 642–646.

Marchelli P, Caron H, Azpilicueta MM, Gallo LA (2008) Primer note: a new set of highly polymorphic nuclear microsatellite markers for *Nothofagus nervosa* and related South American species. *Silvae Genetica*, **57**, 82–85.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387–402.

Mattioni C, Casasoli M, Gonzalez M, Ipinza R, Villani F (2002) Comparison of ISSR and RAPD markers to characterize three Chilean *Nothofagus* species. *Theoretical and Applied Genetics*, **104**, 1064–1070.

Milne I, Bayer M, Cardle L *et al.* (2010) Tablet-next generation sequence assembly visualization. *Bioinformatics*, **26**, 401–402.

Montes I, Conklin D, Albaina A *et al.* (2013) SNP discovery in European anchovy (*Engraulis encrasicolus*, L) by high-throughput transcriptome and genome sequencing. *PLoS One*, **8**, e70051.

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Reviews Genetics*, **12**, 111–122.

Neale DB, Langley CH, Salzberg SL, Wegrzyn JL (2013) Open access to tree genomes: the path to a better forest. *Genome Biology*, **14**, 120.

Nijveen H, van Kaauwen M, Esselink DG, Hoegen B, Vosman B (2013) QualitySNPng: a user-friendly SNP detection and visualization tool. *Nucleic Acids Research*, **41**, 587–590.

Peakall R, Smouse P (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*, **28**, 2537–2539.

Premoli AC, Mathiasen P, Acosta MC, Ramos VA (2012) Phylogeographically concordant chloroplast DNA divergence in sympatric Nothofagus s.s. How deep can it be? *The New Phytologist*, **193**, 261–275.

Rossetto M (2001) Sourcing of SSR markers from related plant species. In: *Plant Genotyping: The DNA Fingerprinting of Plants* (ed. Henry R), pp. 211–224. CAB International, Oxford, UK.

Sabatier Y, Azpilicueta M, Marchelli P *et al.* (2011) Distribución natural de *Nothofagus alpina* y *Nothofagus obliqua* (Nothofagaceae) en Argentina, dos especies de primera importancia forestal de los bosques templados norpatagónicos. *Boletín de la Sociedad Argentina de Botánica*, **46**, 131–138.

Smissen RD, Morse CW, Prada D, Ramón-Laca A, Richardson S (2012) Characterisation of seven polymorphic microsatellites for Nothofagus subgenus Fuscospora from New Zealand. *New Zealand Journal of Botany*, **50**, 227–231.

Soliani C, Sebastiani F, Marchelli P, Gallo L, Vendramin GG (2010) Development of novel genomic microsatellite markers in the southern beech *Nothofagus pumilio* (Poepp. et Endl.) Krasser. *Molecular Ecology Resources*, **10**, 404–408.

Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van de Peer Y (2007) How many genes are there in plants (.. and why are they there)? *Current Opinion in Plant Biology*, **10**, 199–203.

Sun M, Lo EYY (2011) Genomic markers reveal introgressive hybridization in the Indo-West Pacific mangroves: a case study. *PLoS One*, **6**, e19671.

Torales SL, Rivarola M, Pomponio MF *et al.* (2012) Transcriptome survey of Patagonian southern beech *Nothofagus nervosa* (= *N. Alpina*): assembly, annotation and molecular marker discovery. *BMC Genomics*, **13**, 291.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535–538.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology*, **23**, 48–55.

Veblen T, Donoso C, Kitzberger T, Rebertus A (1996) Ecology of southern Chilean and southern Argentinean Nothofagus forests. In: *Ecology and Biogeography of Nothofagus Forests* (eds Veblen T, Hill R, Read J), pp. 293–353. Yale University Press, New Haven, Connecticut.

Wang M, Barkley N, Jenkins T (2009) Microsatellite markers in plants and insects. Part I. Applications of biotechnology. *Genes, Genomes and Genomics*, **3**, 54–67.

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, **20**, 3252–3255.

Yang J-B, Tang M, Li H-T, Zhang Z-R, Li D-Z (2013) Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evolutionary Biology*, **13**, 84.

Zalapa JE, Cuevas H, Zhu H *et al.* (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany*, **99**, 193–208.

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology*, **11**, 1–16.

L.A.G. proposed the hybridizing species for this study. V.A.E.M. and P.G.G. conceived the experiment and organized the funding with the help of L.A.G. V.A.E.M. carried out bioinformatic and population genetic analyses with the help of P.G.G. and L.T. L.T. and P.G.G. developed the SeqQual pipeline. V.A.E.M. performed the SSR development, sequencing and validation. L.A.G. selected putative hybrids from the field. V.A.E.M. and P.G.G. wrote the manuscript. All authors approved the final manuscript.

Data Accessibility

Original data at NCBI SRA (http://www.ncbi.nlm.nih.gov/sra): SRX382841 (*Nothofagus nervosa*) and SRX382843 (*N. obliqua*). For joint and separate species assemblies and corresponding files (.ace, call parameters, contigs statistics and fasta), see the DRYAD database (http://datadryad.org/) entry doi:10.5061/dryad.35v02.

Database of nuclear and organelle-annotated genes is available in Table S4, Appendix S5, Table S6 and Appendix S7 (Supporting information).

Aligned pseudomolecules (single sequences with gaps represented as Ns) for *Nothofagus* spp., *N. nervosa* and *N. obliqua* cpDNA genomes are available in Appendix S8 (Supporting information).

Database of 769 identified putative SSRs loci is available in Table S9 (Supporting information).

Sanger sequences of SSRs loci are available in Appendix S12 (Supporting information).

Bioperl scripts from the SeqQual pipeline are available in Appendix S13 (Supporting information).

Supporting Information

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

Figure S1 Contigs consensus sequence numbers across Gene Ontology (GO) categories for level 2 (A, B, C) and level 3 (D, E, F): A- and D- for Biological Process (BP), B- and E- for Cellular Components (CC), C- and F- For Molecular Functions (MF).

Figure S2 BLASTN comparison for circular cpDNA genomes of *Nothofagus* spp (from assembly of both species) against *Nothofagus nervosa* and *N. obliqua*, plotting with CGView Server and including gene map information.

Figure S3 BLASTN comparison for circular cpDNA genomes of *Nothofagus* spp (from assembly of both species) against two Fagaceae species (*Castanea mollisima* and *Q. rubra*), plotting with CGView Server and including gene map information.

Figure S4 Mean of the estimated ln probability (of the data given *K*) from the STRUCTURE analysis with *K* ranging from 1 to 10.

Figure S5 Values for the Δ*K* criterium (based on Evanno *et al.* 2005) for *K* = 2–10, showing support for *K* = 2, and using the STRUCTURE HARVESTER tool.

Table S1 SeqQual pipeline program list used in this study

Table S2 PCR primers and conditions

Table S3 Detailed statistics of the different MIRA assemblies compared

Table S4 Annotation results of contigs and singletons from BLAST2GO analysis

Table S6 Annotated organelle genes of *Nothofagus*

Table S9 Putative SSR loci derived from the *Nothofagus* genome sequencing project

Table S10 Allele frequencies and sample size by locus and species

Table S11 SSR motifs of genotyped markers in both species as revealed from Sanger sequencing

Appendix S5 Nucleotide sequences of BLASTX-annotated contigs and singletons.

Appendix S7 Annotated chloroplast genes in gff3 format for the simultaneous assembly of both *Nothofagus* species.

Appendix S8 Aligned pseudomolecules (single sequences with gaps represented as Ns) of *Nothofagus* spp., *N. nervosa* and *N. obliqua* cpDNA genomes.

Appendix S12 Sanger sequences of amplified SSR loci.

Appendix S13 SeqQual-BioperlScripts-Notho454.