Gene discovery in Atlantic Forest Plant Species Using GR-RSC Simplified Genomes

Marcella A A Detoni[1], Raony G C C L Cardenas[1], Marcela Uliano-Silva[1], and Mauro F Rebelo[1]

[1]**Instituto de Biofísica Carlos Chagas Filho. Universidade Federal do Rio de Janeiro. Rio de Janeiro. RJ. Brazil.**

ABSTRACT

The Atlantic Forest is one of the most import biodiversity hotspots in the world, nevertheless, its 20,000 plant species are poorly characterized genetically, what could undermine conservation efforts and bioprospection of natural products. We used a genome reduction using restriction site conservation (GR-RSC) technique to minimize sequencing effort and build in a short period a databank of gene sequences from 35 plant species from the Atlantic Forest in a private natural protected area in Southwest Brazil. After Illumina sequencing and standard bioinformatics, we produced more than 66 million super reads, of which 11 million (17%) were annotated using Diamond and UNIREF90 database and 55 million were 'No hit'. We picked 17 enzymes from 2 secondary metabolite synthesis pathways that are both important representatives of biological processes for plants and also of industrial interest, to test the usefulness of the databank we created for gene discovery. All 17 genes were detected in at least one of the 35 species and all species exhibited at least one of the genes. Eight of the 35 species exhibited all 17 genes. These results shows that genome simplification by restriction enzyme can be applied to preliminary screen thousands of species in tropical forests, generating useful databanks for scientific and entrepreneurial activities both in conservation biology and bioprospection.

Keywords: Terpene, Bioprospection, Conservation

INTRODUCTION

History tells that 1 % of species have provided the necessary resources for the development of all civilizations Beattie et al. (2005). Considering that estimates of non-microbial biodiversity vary from 2 to over 50 million species Scheffers et al. (2012) that contain each thousands of genes, one could wonder how much we could achieve if our planet's biodiversity was fully undisclosed. The challenge to unravel our planet's diversity, however, is enormous considering that less than two million species have been morphologically described over the last 250 years. Additionally, with identification rates at around 15,000 species per year, it may take centuries before we can name all species on Earth Costello et al. (2013). If we accept Pimm and Raven (2000) a projection of 40 % extinction rate by deforestation of all species in topical forests by 2100, then species will go extinct before we have had a chance to acknowledge that they have ever existed. In fact, an estimate made by Dirzo et al. (2014) suggests that we are already loosing between 11 and 58 thousand species a year that we never knew about. Defaunation leads to loss of genetic diversity, which is the ultimate scale of biodiversity. Records of DNA sequences may prevent genetic biodiversity extinction, besides playing an important role in conservation biology and bioprospection of natural products. High-throughput sequencing methods and bioinformatics can automatically, precisely, and rapidly identify genes in DNA sequences that can lead to gene discovery, without the need of having any previous knowledge on the organism morphology. Over the last 10 years, the cost of DNA sequencing has dropped significantly from US$ 100 million to US$1 thousand for an entire human genome Wetterstrand (2014). However, massive genome sequencing and assembly still represent a challenge from both technical and economical points of views Sboner et al. (2011). The large amount of repetitive sequences in most genomes requires high sequencing coverage, use of different protocols and bioinformatics algorithms to fully assemble a genome. Alternatively, the use of genome reduction using restriction site conservation - GR-RSC methods can provide a fast, cheap and simplified version of the genome of innumerous complex organisms Etter et al. (2011). If we assume that the chances of a given thermostable and methylation sensitive restriction enzymes (RE) to reach repetitive and non-translated areas of the genome are higher than reaching coding genes, then RE may be used to create a representative databank of genes from a large number of species. Such databank, if suitable for gene discovery, will be a powerful aid for massive conservation and bioprospection efforts. In this study, we evaluated the use of RE followed by genome sequencing as a strategy to unveil genetic biodiversity information on 35 plant genomes from a private natural reserve in the Atlantic Forest biome in southeast Brazil.

Table 1. Access to their genetic information was granted by the Brazilian National Research Council (CNPq) (permit: 010567/2015-1) and according to the Brazilian law (13.123/2012).

Sample ID	Family	Species	Sample DNA concentration ng.µL-1	Sample DNA quality 260/280
1	Solanaceae	*Solanum pseudoquina*	302	1.84
2	Fabaceae	*Inga edulis*	198	1.86
3	Euphorbiaceae	*Alchornea sidifolia*	160	1.84
4	Asteraceae	*Achyrocline alata*	117	1.74
5	Asteraceae	*Mikania micranta*	113	1.78
6	Verbenaceae	*Lantana trifolia*	109	1.79
7	Piperaceae	*Piper gaudichaudianum*	99.7	1.58
8	Asteraceae	*Baccharis semiserrata*	96.7	1.83
9	Zingiberaceae	*Hedychium coronarium*	94	1.81
10	Meliaceae	*Cedrela fissilis*	91.6	1.81
11	Fabaceae	*Piptadenia gonoacantha*	91.2	1.81
12	Lauraceae	*Ocotea odorífera*	73.2	1.81
13	Asteraceae	*Sphagneticola trilobata*	71	1.58
14	Solanaceae	*Solanum swartzianum*	70	1.74
15	Annonaceae	*Xylopia brasiliensis*	64.6	1.78
16	Meliaceae	*Guarea macrophylla*	57.6	1.76
17	Meliaceae	*Cabralea canjerana*	53.7	1.64
19	Lauraceae	*Nectandra leucantha*	51.2	1.8
20	Salicaceae	*Casearia obliqua*	41	1.78
22	Onagraceae	*Ludwigia octovalvis*	42.8	1.84
24	Celastraceae	*Maytenus aquifolia*	41.5	1.76
26	Myrtaceae	*Myrceugenia myrcioides*	36.7	1.61
28	Apiaceae	*Centella asiática*	33	1.84
29	Salicaceae	*Casearia sylvestris*	29	1.65
30	Rubiaceae	*Psychotria vellosiana*	27	1.86
31	Euphorbiaceae	*Maprounea guianensis*	26	1.69
35	Solanaceae	*Solanum americanum*	23.5	1.79
36	Verbenaceae	*Stachytarpheta cayennennsis*	23.3	1.83
38	Rutaceae	*Zanthoxylum rhoifolium*	19.9	1.84
39	Fabaceae	*Senna multijuga*	18.5	1.87
42	Asteraceae	*Jaegeria hirta*	14	1.91
43	Fabaceae	*Mimosa pudica*	13.4	1.78
45	Solanaceae	*Solanum castaneum*	11	1.76
47	Malvaceae	*Sida rhombifolia*	7.8	1.61
48	Primulaceae	*Myrsine umbellata*	7.38	1.69

MATERIAL AND METHODS

A previous study of 752 Atlantic Forest plant species found in the private natural reserve Legado das Águas Flores et al. (2015) were considered in this study for choosing the species to be sequenced. Exsiccates of all plants are stored in the herbarium of São Paulo University, in Brazil. Species were selected based on availability of previous information in the literature regarding their conservation status, biotechnology interest, genetic and genomic data. For this purpose, we surveyed text databases over the internet (Web of Science, Google Scholar, Scientific Electronic Library Online (SciELO), thesis records deposited at the Coordination for the Improvement of Higher Education in Brazil (CAPES) website, Open Public Resource for Innovation Cartography (LENS) and the National Center for Biotechnology Information (NCBI). A total of 35 species (Table 1) were selected for further investigation. Access to their genetic information was granted by the Brazilian National Research Council (CNPq) (permit: 010567/2015-1) in accordance with the Brazilian law (13.123/2012). Plant soft parts (leaves, 10 cm2) from 35 species (Table 1) were collected by local experts (bushman) in July 2015 in Legado das Águas (Reservas Votorantim), the largest private Natural Reserve in the remaining area of the Atlantic forest in Brazil, within a region of 31,000 ha in Juquiá, Miracatu and Tapiraí municipalities in São Paulo, Southeast Brazil (Figure 1). DNA extraction of 48 leaf samples was performed with DNeasy Plant Mini Kit (Qiagen), following the manufacturer's recommendations. DNA quantification of each sample was done by Qubit (Invitrogen) and quality control was evaluated by NanoDrop 2000 Spectrophotometer (Thermo Scientific).

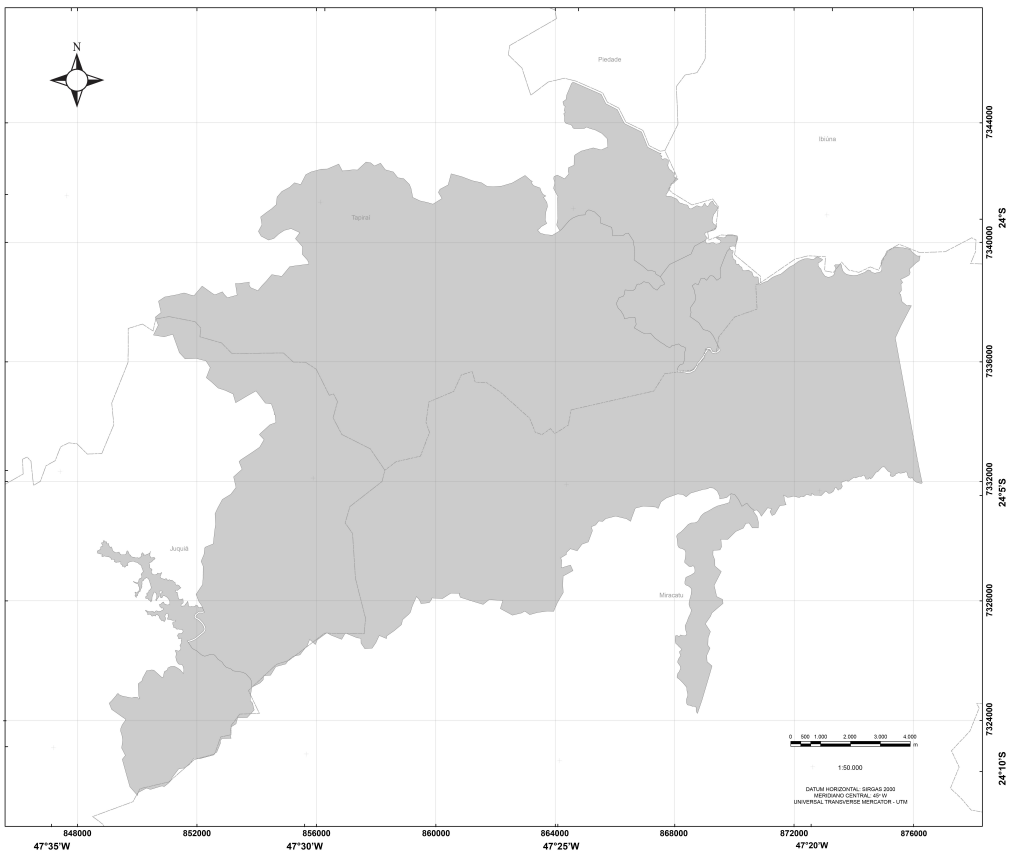

Figure 1. Localization of the Preserved Area Legado das Águas in Southeast Brazil, 47°33'W 24 °05'S. Total Area is 30,764,55 hectares.

A Genotype By Sequencing – GBS protocol Elshire et al. (2011) was chosen as the strategy for genome simplification. GBS Libraries were built with 150 ng of high quality DNA (260/280 rate higher than 1.6) from samples of each species using PST I as restriction enzyme. DNA integrity was assessed by electrophoresis on 1 % agarose gel. Barcode adapters were added and dimer free fragments of expected size (200-400 pb) were obtained with 30 µL final reaction volume in which 5 µL consisted of the set of adapters (0.05 pmol); 1.6 µL of T4 DNA ligase (NEB. MK0202L); Buffer (10X); 1 µL enzyme T4 DNA and 18 µL of water. The ligase reaction was carried out for 2 h at 18 °Cand inactivation for 30 sec at 65 °C. DNA libraries were amplified followed by sequencing. A pooled sample was created with the addition of a 10 µL of each of the 35 species' DNA. The pooled sample was then purified using PCR Purification Kit (Qiagen) and eluted to a final volume of 30 µL. Amplification was carried in a 15 µL aliquote of the pooled DNA sample using 1X Master Mix Taq (New England Biolabs), 25 pmol of each of random primers in a final reaction volume of 50 mL, with the following conditions: 10 min at 98 °Cinitial denaturation; 5 min at 72 °Cof initial extension; 18 cycles of 30 sec at 98 °Cdenaturation, 30 sec at 65 °Cannealing and 30 sec at 72 °Cextension; with 5 min at 75 °Cof final extension. No DNA negative controls were added. PCR products were purified usin magnetic beads (Agencourt AMPure XP - BECKMAN COULTER). The GBS library quality was evaluated using the Agilent 2100 Bioanalyzer with High Sensitivity DNA kit and considered suitable for sequencing when primer dimers and adapters (between 100 and 150 bp) were virtually absent and most fragments were between 200-450 bp. Library quantification by real-time PCR was carried-out using a KAPA Biosystems Quantification Kit (KAPA). Samples were then diluted to 2 nM (30 µL) and sequenced using the HiSeq platform 2500 (Illumina). Sequences were demultiplexed (allowing one mismatch in barcode) using Sabre program (https://github.com/najoshi/sabre) and then primers were removed (10 % mismatch) with Cutadapt (https://pypi.python.org/pypi/cutadapt). We further used MaSuRCA (Zimin et al., 2013) to remove redundancy and build Super reads (extended by unique paths). Quorum parameters were altered to allow k-mers of 32, which is, bellow the standard k-mer 50 used for mounting genomes with high coverage. All Super Reads were used in Diamond BLASTX program Buchfink et al. (2014) mode against Uniref90 database Suzek et al. (2015) in which the search targets are grouped by domain, reducing redundancy and increase sing the likelihood of annotation, as well as speed. The Diamond output is an alignment of binary files that can be converted to tabulate or SAM. The tabular format is identical to the Blast + with '-outfmt' 6 Std option (Bethesda (2008) available at the online) which is the default tabular result. For annotation transfer we used the relationship of Uniref90 Fasta identifiers

with Genbank ID that can be found in ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/. Gene Ontology (GO) terms were summarized for each library using the GO categorizer http://www.animalgenome.org/tools/catego/ with each being counted only once. GO terms were related to Enzyme Commission Number (EC). Annotated genes were imported to a proprietary mySQL server based in Amazon Web Servers to create a database of Atlantic Forest species genes. The utility of the database for gene discovery was tested conducting a search in the resulting databank for terpene producing enzymes. Terpenes are the largest class of secondary metabolites, with over 22 thousand individual compounds Dionísio et al. (2009) that are extensively applied as flavor and fragrance compounds Cabaleiro et al. (2012). They exhibit a variety of structures and their activities may largely vary from roles as diverse as communication to defense. Terpenes synthesis is compartmentalized so that monoterpenes and diterpenes are produced through 1-deoxy-D-xylulose-5-phosphate (DXP) pathway whereas sesquiterpenes are produced through mevalonic acid (MVA) pathway. Many enzymes are involved in terpene production, but some are exclusive of these 2 pathways. We searched for their Enzyme Commission Number (EC), Gene Ontology correspondence and gene sequence available at Genbank. Three enzymes related to the MVA pathway mevalonate kinase (EC: 2.7.1.36), phosphomevalonate kinase (EC: 2.7.4.2) and diphosphomevalonate decarboxylase (EC:4.1.1.33) are located in the cytoplasm and mitochondria Pang et al. (2006) whereas six enzymes of the DXP pathway pathway 1-deoxy-D-xylulose-5-phosphate reductoisomerase 1 (EC: 1.1.1.267); 2C-methyl-D-erythritol 4-phosphate cytidylyltransferase (EC: 2.7.7.60); 2C-methyl-D-erythritol 2.4-cyclodiphosphate synthase (EC: 4.6.1.12); 1-hydroxy-2C-methyl-2-(E)-butenyl 4-diphosphate synthase (EC: 1.17.7.1); 1-hydroxy-2C-methyl-2-(E)-butenyl 4-diphosphate reductase 1 (EC: 1.17.1.2); isopentenyl diphosphate delta isomerase 1 (EC: 5.3.3.2); farnesyl diphosphate synthase (EC: 2.5.1.10); dimetilaliltranstransferase (dihydrofolate synthetase activity) (EC: 2.5.1.15); dimetilaliltranstransferase (spermidine synthase activity) (EC: 2.5.1.16); dimetilaliltranstransferase (adenosyltransferase activity c-diamide) (EC: 2.5.1.17); dimetilaliltranstransferase (glutathione transferase activity) (EC: 2.5.1.18); dimetilaliltranstransferase (3-phosphoshikimate 1-carboxyvinyltransferase) (EC: 2.5.1.19) are located in the plastids Rohmer (2003); Guevara-García et al. (2005); Ganjewala and Kumar (2008); Kirby and Keasling (2009). These were the enzyme used to test the database suitability for bioprospection and conservation studies.

RESULTS

Based on our literature review, we selected 35 species with both economical potential to be explored and conservation interest. We were able to collect samples from each of them and have their DNA extracted and sequenced. Two libraries were produced for all samples collected. One library had 24 species and produced 210,771,131 reads with an average sequence length of 101 bp and 51 % GC content. The other library had 11 species and produced 231,624,922 reads with an average sequence length of 116 bp and 51 % GC content. The median number of super reads per species was 1,022,532 (min 5 and max 7,725,847). Table 2 shows the average length of the super reads in each species. The median number of annotated super reads per species was 168,660 (min 3 and max 953,598) and the median of No hits was 832,960 (min 2 and max 6,772,249). Super reads were annotated with Gene Ontology (GO) terms. Figure 2 shows the distribution of reads in the main GO grouping categories under the 3 most general categories: Biological process, with 24 grouping GO terms and most reads under catalytic activity (GO:0003824), transporter activity (GO:0005215) and structural molecule activity (GO:0005198); Cellular component with 20 grouping GO terms with most reads under Binding (GO:0005488), Membrane (GO:0016020), Cell part (GO:0044464), Membrane part (GO:0044425), Organelle (GO:0043226) and Virion part (GO:0044423); and Molecular functions with 15 grouping GO terms with most reads under regulation of biological process (GO:0050789), multi-organism process (GO:0051704), cellular process (GO:0009987), metabolic process (GO:0008152) and response to stimulus (GO:0050896).

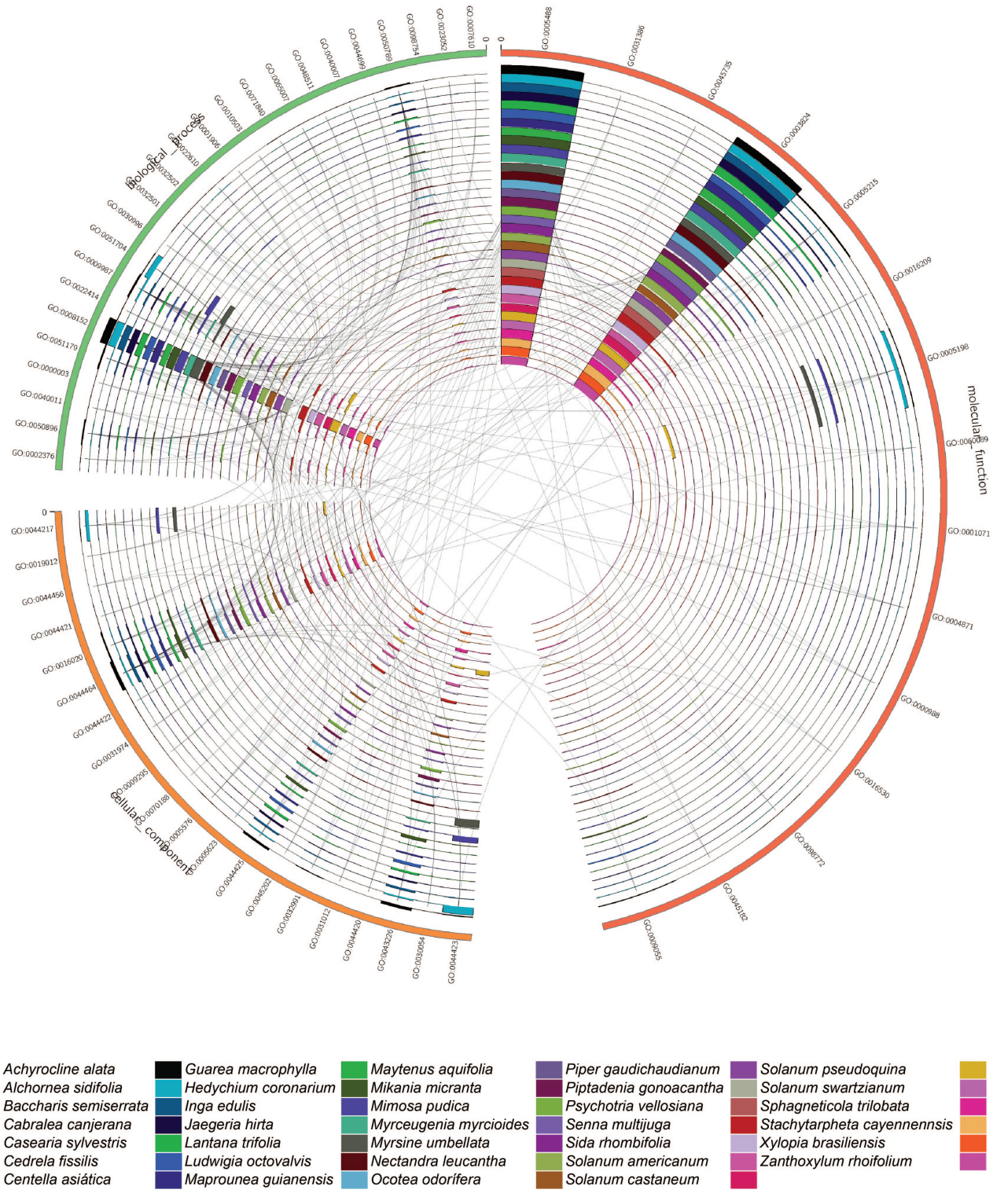

Figure 2. Major and minor Gene Ontology categories found in the simplified genomes. Each line represents one of the 35 species sequenced in this study. Line thickness represents the abundance of reads belonging to each major GO category. The 3 most related GO terms are connected to each other with hairlines.

Table 2. Number of super reads (with average, minimum and maximum size) created with Masurca and annotated by Diamond

Species (Sample ID)	Super Reads Count	Average Size (bp)	Maximum Length (bp)	minimum length (bp)	Annotated by Diamond*	GO Terms to GO Slim	No Hits
Achyrocline alata (4)	198,455	117.3	686	67	35,002	2,796	163,453
Alchornea sidifolia (3)	25,964	114.6	447	67	11,580	529	14,384
Baccharis semiserrata (8)	307,382	116.8	547	67	57,732	2,344	249,650
Cabralea canjerana (17)	1,020,353	116.9	1370	47	223,205	3,393	797,148
Casearia sylvestris (29)	183,487	88.1	209	47	33,107	3,013	150,380
Cedrela fissilis (10)	1,596,354	115.6	686	67	382,396	3,026	1,213,958
Centella asiatica (28)	698,296	82.8	475	31	125,019	4,419	573,277
Guarea macrophylla (16)	3,110,227	114.4	618	49	865,910	4,376	2,244,317
Hedychium coronarium (9)	525,040	116.4	513	49	105,596	2,318	419,444
Inga edulis (2)	21,550	122.5	490	67	8,260	538	13,290
Jaegeria hirta (42)	519,784	81.1	453	31	90,483	3,881	429,301
Lantana trifolia (6)	6,499	139.6	492	67	5,848	4,376	651
Ludwigia octovalvis (22)	1,325,140	115.4	754	47	275,443	3,498	1,049,697
Maprounea guianensis (31)	1,989,256	84.8	467	31	362,206	5,516	1,627,050
Maytenus aquifolia (24)	702,477	117.8	692	47	169,833	3,471	532,644
Mikania micrantha (5)	1,068,167	116.0	495	67	156,466	3,130	911,701
Mimosa pudica (43)	392,566	90.8	242	47	90,414	3,532	302,152
Myrceugenia myrcioides (26)	2,359,885	81.0	486	31	416,146	5,067	1,943,739
Myrsine umbellata (48)	399,367	78.8	452	31	63,124	4,000	336,243
Nectandra leucantha (18)	788,846	115.5	530	47	162,452	2,880	626,394
Ocotea odorifera (12)	1,967,917	117.3	551	49	224,255	3,544	1,743,662
Piper gaudichaudianum (7)	1,659,849	116.5	539	49	344,309	3,526	1,315,540
Piptadenia gonoacantha (11)	354,135	115.8	490	67	71,238	2,527	282,897
Psychotria vellosiana (30)	5	83.6	103	69	3		2
Senna multijuga (39)	470,083	93.0	431	45	122,166	3,152	347,917
Sida rhombifolia (47)	1,511,937	89.2	478	33	389,652	5,136	1,122,285
Solanum americanum (35)	810,209	90.0	471	31	233,245	4,605	576,964
Solanum castaneum (45)	1,680,001	76.4	497	31	338,058	5,504	1,341,943
Solanum pseudoquina (1)	30,008	116.9	552	67	13,314	4,064	16,694
Solanum swartzianum (14)	1,108,612	116.1	535	49	292,622	4,064	815,990
Sphagneticola trilobata (13)	399,394	115.7	444	67	95,303	2,454	304,091
Stachytarpheta cayennennsis (36)	2,854,573	83.3	485	31	601,948	5,375	2,252,625
Xylopia brasiliensis (15)	1,346,027	117.2	532	49	167,487	3,558	1,178,540
Zanthoxylum rhoifolium (38)	7,725,847	76.1	472	31	953,598		6,772,249

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2316v1 | CC BY 4.0 Open Access | rec: 26 Jul 2016, publ: 26 Jul 2016

Table 3. List of enzymes involved in terpene synthesis found for each species (N=32) according to bioinformatics analysis. Numbers in the first row refer to samples ID in table 1. Samples 20, 30, and 38 in Table 1 were not analyzed for lack of sequences and thus are not included here. Black dots indicate that the enzymes have been identified in the respective species. The correlation of the enzyme code and its name is presented as follows: 2.7.1.36: mevalonate kinase; 2.7.4.2: phosphomevalonate kinase; 4.1.1.33: diphosphomevalonate decarboxylase; 1.1.1.267: 1-deoxy-D-xylulose-5-phosphate reductoisomerase 1; 2.7.7.60: 2C-methyl-D-erythritol 4-phosphate cytidylyltransferase; 4.6.1.12: 2C-methyl-D-erythritol 2.4-cyclodiphosphate synthase; 1.17.7.1: 1-hydroxy-2C-methyl-2-(E)-butenyl 4-diphosphate synthase; 1.17.1.2: 1-hydroxy-2C-methyl-2-(E)-butenyl 4-diphosphate reductase 1; 5.3.3.2: isopentenyl diphosphate delta isomerase 1; 2.5.1.10: farnesyl diphosphate synthase; 2.5.1.15: dimetilaliltranstransferase (dihydrofolate synthetase activity); 2.5.1.16: dimetilaliltranstransferase (spermidine synthase activity); 2.5.1.17: dimetilaliltranstransferase (adenosyltransferase activity c-diamide); 2.5.1.18: dimetilaliltranstransferase (glutathione transferase activity); 2.5.1.19: dimetilaliltranstransferase (3-phosphoshikimate 1-carboxyvinyltransferase.

Enzyme code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	22	24	26	28	29	31	35	36	39	42	43	45	47	48
2.7.1.36	•			•	•	•				•	•	•		•	•	•			•		•	•		•	•	•	•	•	•	•	•
2.7.4.2	•		•			•					•		•	•	•	•				•	•			•		•	•	•	•	•	
4.1.1.33	•				•	•	•				•	•		•	•	•	•				•			•	•	•	•		•	•	•
2.2.1.7	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
1.1.1.267						•	•	•	•	•	•	•			•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	
2.7.1.148	•		•		•	•	•		•	•	•			•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•
2.7.7.60	•					•	•	•	•	•	•			•	•	•	•	•	•	•	•	•		•		•	•	•	•	•	•
4.6.1.12	•				•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•
1.17.7.1	•				•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
1.17.1.2	•			•		•	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
5.3.3.2				•	•	•		•	•	•		•			•	•		•	•	•	•	•		•	•	•		•	•	•	•
2.5.1.10	•			•	•	•	•					•		•	•	•	•	•	•	•	•			•	•	•	•	•	•	•	•
2.5.1.15	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•		•		•	•
2.5.1.16	•					•	•			•	•			•		•	•	•	•	•	•			•	•	•		•		•	•
2.5.1.17	•				•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•		•		•	•	•	•
2.5.1.18	•		•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
2.5.1.19	•			•	•	•	•		•	•		•	•	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•

Table 3 shows the enzymes involved with terpene synthesis found for each species, according to bioinformatics analysis. A visual analysis of the table shows that most genes were found in most species. We were able to identify all the enzymes in the mevalonic acid (MVA) pathway (EC: 2.7.1.36; EC: 2.7.4.2; EC:4.1.1.33) in *Alchornea sidifolia* (3), *Solanum pseudoquina* (1), *Piptadenia gonoacantha* (11), *Solanum swartzianum* (14) and *Xylopia brasiliensis* (15) and but not all the enzymes in the 1-deoxy-D-xylulose-5-phosphate (DXP) pathway. On the other hand, we were able to identify all the enzymes in the DXP pathway (EC: 1.1.1.267; EC: 2.7.7.60; EC: 4.6.1.12; EC: 1.17.7.1; EC: 1.17.1.2; EC: 5.3.3.2; EC: 2.5.1.10; EC: 2.5.1.15; EC: 2.5.1.16; EC: 2.5.1.17; EC: 2.5.1.18; EC: 2.5.1.19) in *Baccharis semiserrata*(8), *Cabralea canjerana* (17), *Cedrela fissilis* (10), *Centella asiatica* (28), *Hedychium coronarium*(9), *Ludwigia octovalvis* (22), *Maytenus aquifolia* (24), *Jaegeria hirta* (42), *Mimosa pudica*(43), *Nectandra leucantha* (18), *Ocotea odoriferous* (12) and *Piper gaudichaudianum* (7) but none of the enzymes in the MVA pathway. All enzymes in both MVA and DXP pathways were identified in *Guarea macrophylla* (16), *Lantana trifolia* (6), *Maprounea guianensis* (31), *Myrceugenia myrcioides* (26), *Senna multijuga* (39), *Sida rhombifolia* (47), *Solanum castaneum* (45) and *Stachytarpheta cayennennsis* (36).

DISCUSSION

To investigate the potential of restriction enzyme simplified genomes to produce useful DNA information for either conservation or bioprospection purposes, we extracted and sequenced DNA from 35 species selected according to information obtained from the literature, creating the largest and most diverse genetic database of Atlantic forest species available so far. Gene discovery was our main goal in this study, thus we did not aim to assemble *de novo* these simplified genomes at this time. One of the problems in testing the applicability of a databank like the one we have produced relies on the fact that there is no genome size estimation for any of the 35 species whose sequences were available in the Genbank. Indeed, plant genomes vary widely in size, even in the same family, but not necessarily in the same genus. A closer look at genome projects involving species in the same families as the ones worked in this study unveils this variability. The Solanum genus in the Solanacea family has many important edible species such as Eggplant (*Solanum melongena*, 833 Mb - PRJDB1505); Tomato (*Solanum pimpinellifolium*, 688 Mb - PRJNA72351) and Potato (*Solanum tuberosum*, 705 Mb -PRJNA225997), all with similar genome sizes. But Tabaco, another genus in the Solanacea family, has plant species with similarly larger genomes than the ones in the Solanum genus (*Nicotiana otophora*, 2689 Mb - PRJNA208212; *Nicotiana tabacum*, 3732 Mb - PRJNA319578). The same seems to occur in the Fabaceae family, in which Peanuts species in the Arachis genus (*Arachis duranensis*, 1068 Mb - PRJNA258023 and *Arachis ipaensis*, 1349 Mb - PRJNA258025) have larger genomes than Beans in the Vigna genus (*Vigna angularis*, 379 Mb - PRJNA261643 and *Vigna radiata*, 459 Mb - PRJNA243847) or Cacao (*Theobroma cacao*, 346 Mb - PRJNA51633). The pattern is similar for all other families that have genomes available in Genbank. The only genus that we sequenced for which there is available information on genome size is Solanum. Based on the seven species listed in Supplementary Material 1 (last seven species listed), we were able to roughly estimate the genome size of Solanum species to be around 766 Mb. Than, by multiplying the number of reads obtained for each of the four species of the Solanus genus we sequenced for their corresponding average length (Table 2), we obtained an estimate coverage of 16% for *S. americanum*, 30% for *S. castaneum*, 0.7% for *S. pseudoquina* and 29% for *S. swartzianum*. The basic gene ontology description we provide here shows the abundance and diversity of enzymes found in these 35 species, indicating the richness of Brazilian biodiversity and the suitability of the databank for a wide variety of studies. Even though genome simplification and genome reduction using restriction site conservation - GR-RSC- have been previously used mainly to identify homologous loci across species, these techniques have also been proposed for phylogeny studies and breeding efforts Dockter et al. (2013). We believe this databank is useful, for example, for reforestation studies in the Atlantic Forest. To demonstrate the usefulness of this database for specific genetic studies that have either conservation or biotechnological perspectives, we accessed the abundance and diversity of enzymes in the MVA and DXP pathways. Overall, 17 enzymes (table 3) were analyzed. At least one of the 17 genes was found in any of the 35 species and all species exhibited at least one of the 17 genes. Some species, like *Guarea macrophylla* (6), *Lantana trifolia* (16), *Myrceugenia myrcioides* (26), *Maprounea guianensis* (31), *Sida rhombifolia* (47), *Solanum castaneum* (45) and *Stachytarpheta cayennennsis* (36) exhibited all of them, suggesting that these species can produce sesquiterpenes and triterpenes through the MVA pathway; and monoterpenes and diterpenes through the DXP pathway. The fact that we did not find some enzymes of a given (MVA or DXP) pathway when other enzymes from the same pathway have been found, could have indicated that GBS genome simplification was selecting some genes over others. GBS restriction enzymes are sensitive to methylation and reduce genome complexity by avoiding cutting the methylated repetitive regions Elshire et al. (2011). Since all 17 investigated enzymes have been identified at least once in at least one species, it is unlikely that the limitation was caused by genome simplification and we could expect that an increase in sequencing coverage would allow us to find all enzymes in all species. Even though, in our case, the difference in the coverage that we estimated to vary from 0.7% to 30% (of the total genome) in the 4 Solanum species that we sequenced, did not influence the discovery of most genes of both MVA and DXP pathways in all of them. We could expect high conservation of gene sequences in these pathways, but minor changes in the sequences that could be observed among species or even individuals, could be enough to prevent RE cutting and eliminate the genes from the constructed library. One of the most accepted estimates of the average cost of successfully developing

a new molecular entity, including R&D spending on failed drug, is that of DiMasi et al. (2003), who established the price in US$802 million in 2000, with a time for development varying between 7 and 12 years. However, OECD states in its "The Bioeconomy to 2030: Designing a Policy Agenda" report OECD (2009) that approximately 75% of the future economic contribution of biotechnology and large environmental benefits are likely to come from agricultural and industrial applications, even thought today, over 80% of research investments in biotechnology by the private and public sectors go to health-related applications. For these applications, the pure regulatory costs could be around millions of dollars, being higher for genetically modified plant varieties (up to US$ 13.5 million per variety) than for the open release of genetically modified microorganisms (approximately US$ 3 million per release), as well as for bioremediation to clean up polluted soils OECD (2009). However, there is little dispute on the value of investing in biotechnology. A recent report by Tripp and Grueber (2011) shows that sequencing of the human genome cost US$ 10 billion at the time but the economic impact 10 years later of this effort was estimated to be around US$ 1 trillion. The Brazilian Atlantic Rain-forest has been consistently acknowledged as the most diverse biome in Brazil and one of the most diverse in the world, with estimates of over 20,000 plant and fungi species Mittermeier et al. (1998); Myers et al. (2000); Forzza et al. (2012); Zappi et al. (2015). Since the sequencing of the Human Genome in 2001, DNA sequencing technology has evolved 10-times faster than the silicon chip technology as predicted by Moore's law Morey et al. (2013) and the price of sequencing one entire human genome has dropped from US$ 95 million in 2001 to US$ 1 thousand in 2015 Wetterstrand (2014). It is possible that soon we will be able to sequence all the existing biodiversity in the Atlantic Forest and by doing so we will save gene diversity from extinction. These are the bases for establishing a biodiversity-based bioeconomy with biotechnology.

CONCLUSIONS

Genome simplification by restriction enzyme was useful to minimize sequencing effort and allow us to create the most comprehensive gene library of the Atlantic Forest to date in a short time and with reasonable cost/benefit. Genome simplification did not impair the ability to observe genes of interest in all sequenced species showing that this methodology can be applied to preliminary screen thousands of species in tropical forests, generating useful databanks for scientific and entrepreneurial activities both in conservation biology and bioprospection.

ACKNOWLEDGMENTS

This study was financed by Reservas Votorantim that is also the proprietary of the sequences databank. Authors are thankful to Miguel M. F. de Jesus, the bushman that helped with species sampling.

REFERENCES

Beattie, A. J., Barthlott, W., Elisabetsky, E., Farrel, R., Kheng, C. T., Prance, I., Rosenthal, J., Simpson, D., Leakey, R., Wolfson, M., Kate, K., Editor, R., and Laird, S. (2005). CH. 10 New Products and Industries from Biodiversity. *Ecosystems and Human Well-Being: Current State and Trends*, pages 273–295.

Bethesda, M. D. (2008). BLAST® Command Line Applications User Manual. Options for the command-line applications.

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60.

Cabaleiro, N., de la Calle, I., Bendicho, C., and Lavilla, I. (2012). Fast screening of terpenes in fragrance-free cosmetics by fluorescence quenching on a fluorescein-bovine serum albumin probe confined in a drop. *Analytica Chimica Acta*, 719:61–67.

Costello, M. J., May, R. M., and Stork, N. E. (2013). Can We Name Earth ' s Species Before They Go Extinct ? *Science*, 339(January):413–416.

DiMasi, J. A., Hansen, R. W., and Grabowski, H. G. (2003). The price of innovation: New estimates of drug development costs. *Journal of Health Economics*, 22(2):151–185.

Dionísio, A., Molina, G., Bicas, J. L., Dias, S., and Pastore, G. (2009). Fungal biotransformation of terpenes.

Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N. J. B., and Collen, B. (2014). Defaunation in the Anthropocene. *Science*, 345(6195):401–406.

Dockter, R. B., Elzinga, D. B., Geary, B., Maughan, P. J., Johnson, L. a., Tumbleson, D., Franke, J., Dockter, K., and Stevens, M. R. (2013). Developing molecular tools and insights into the Penstemon genome using genomic reduction and next-generation sequencing. *BMC genetics*, 14(1):66.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. a., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 6(5):1–10.

Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. a., and Johnson, E. a. (2011). Local de novo assembly of rad paired-end contigs using short sequencing reads. *PLoS ONE*, 6(4).

Flores, T. B., Colletta, G. D., Souza, V. C., Ivanauskas, N. M., Tamashiro, J. Y., and Rodrigues, R. R. (2015). *Guia ilustrado para identificação das plantas da Mata Atlântica*. Oficina de Textos.

Forzza, R. C., Baumgratz, J. F. a., Bicudo, C. E. M., Canhos, D. a. L., Carvalho, A. a., Coelho, M. a. N., Costa, A. F., Costa, D. P., Hopkins, M. G., Leitman, P. M., Lohmann, L. G., Lughadha, E. N., Maia, L. C., Martinelli, G., Menezes, M., Morim, M. P., Peixoto, A. L., Pirani, J. R., Prado, J., Queiroz, L. P., Souza, S., Souza, V. C., Stehmann, J. R., Sylvestre, L. S., Walter, B. M. T., and Zappi, D. C. (2012). New Brazilian Floristic List Highlights Conservation Challenges. *BioScience*, 62(1):39–45.

Ganjewala, D. and Kumar, S. (2008). An Account of Cloned Genes of Methyl-erythritol-4- phosphate Pathway of Isoprenoid Biosynthesis in Plants CDP-ME Phytoene Phytal-PP. *Curr. Issues.Mol. Biol.*, 11:35–46.

Guevara-García, A., San Román, C., Arroyo, A., Cortés, M. E., de la Luz Gutiérrez-Nava, M., and León, P. (2005). Characterization of the Arabidopsis clb6 mutant illustrates the importance of posttranscriptional regulation of the methyl-D-erythritol 4-phosphate pathway. *The Plant cell*, 17(2):628–643.

Kirby, J. and Keasling, J. D. (2009). Biosynthesis of plant isoprenoids: perspectives for microbial engineering. *Annu Rev Plant Biol*, 60:335–55.

Mittermeier, R. A., Myers, N., Thomsen, J. B., da Fonseca, G. A. B., and Olivieri, S. (1998). Biodiversity Hotspots and Major Tropical Wilderness Areas: Approaches to Setting Conservation Priorities. *Conservation Biology*, 12(3):516–520.

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L., and Cocho, J. A. (2013). A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism*, 110(1-2):3–24.

Myers, N., Mittermier, R. A., Mittermier, C. G., Fonseca, G. A. B., Kents, J., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. a., and Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403:853–8.

OECD (2009). The Bioeconomy to 2030: Designing a Policy Agenda. Technical report, OECD.

Pang, Y., Shen, G. A., Berg??s, T., Cardier, H., Wu, W., Sun, X., and Tang, K. (2006). Molecular cloning, characterization and heterologous expression in Saccharomyces cerevisiae of a mevalonate diphosphate decarboxylase cDNA from Ginkgo biloba. *Physiologia Plantarum*, 127(1):19–27.

Pimm, S. L. and Raven, P. (2000). Biodiversity: Extinction by numbers. *Nature*, 403(6772):843–845.

Rohmer, M. (2003). Mevalonate-independent methylerythritol phosphate pathway for isoprenoid biosynthesis. Elucidation and distribution. *Pure and Applied Chemistry*, 75(2-3):375–388.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., and Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome biology*, 12(8):125.

Scheffers, B. R., Joppa, L. N., Pimm, S. L., and Laurance, W. F. (2012). What we know and don' t know about Earth' s missing biodiversity. *Trends in Ecology & Evolution*, 27(9):501–510.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., and Wu, C. H. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932.

Tripp, S. and Grueber, M. (2011). Economic Impact of the Human Genome Project. Technical report, Battelle Memorial Institute.

Wetterstrand, K. (2014). DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP). Technical report, NIH.

Zappi, D. C., Ranzato Filardi, F. L., Leitman, P., Souza, V. C., Walter, B. M. T., Pirani, J. R., Morim, M. P., Queiroz, L. P., Cavalcanti, T. B., Mansano, V. F., and Forzza, R. C. (2015). Growing knowledge: An overview of Seed Plant diversity in Brazil. *Rodriguesia*, 66(4):1085–1113.