Ants of French Guiana: 16S rRNA sequence dataset

Gaëtan Rongier‡, Audrey Sagne‡, Sandrine Etienne‡, Frederic Petitclerc‡, Gaelle Jaouen‡, Jerome Murienne§, Jerome Orivel‡

‡ UMR Écologie des Forêts de Guyane (AgroParisTech, CIRAD, CNRS, INRAE, Université de Guyane, Université des Antilles), Kourou, French Guiana

§ Laboratoire Evolution et Diversité Biologique (EDB UMR5174) CNRS, Université Paul Sabatier Toulouse 3, IRD, Toulouse, France

Corresponding author: Gaëtan Rongier (rongiergaetan15@gmail.com), Jerome Orivel (jerome.orivel@cnrs.fr)

Academic editor: Brian Lee Fisher

Abstract

This dataset represents a reference library of DNA sequences for ants from French Guiana. A total of 3931 new sequences from the 16S rRNA gene has been generated. The reference library covers 344 species distributed in 57 genera. Overall, 3920 sequences have been assigned at the species level and 11 at the genus level. All these sequences were submitted to DDBJ/EMBL/GenBank databases in the Bioproject: PRJNA779056: 16S French Guiana Ants (Hymenoptera: Formicidae), sequence identifier KFFS00000000.

Keywords

DNA sequencing, 16S rRNA, molecular identification, Formicidae, NGS, Neotropics

Introduction

The current biodiversity crisis calls for efforts to reach more rapid biodiversity characterisation. Indeed, our global knowledge of biodiversity is still largely unknown, with ca. 80% of species to be described and more than 20 years, on average, for the description of a new species following its discovery (Fontaine et al. 2012). Moreover, the classical taxonomical identification of specimens relies most often on subtle morphological criteria and expert knowledge, which is confronted by the shortage of taxonomists (Engel et al. 2021). Such issues are not only valid for the description of extant species diversity, but also to answer ecological questions, such as how communities are assembled and how they respond to global change. In this context, DNA barcoding has proved its effectiveness and has been successfully applied in taxonomical and ecological studies (Kress et al. 2015). DNA barcoding allows identifying specimens

at species level using a short sequence of DNA as species tag (Hubert and Hanner 2015). If DNA barcoding is a robust and rapid technology with applications in many scientific areas from taxonomy to ecology, its accuracy and reliability relies on the completeness of a reference library.

With more than 16,000 described species to date (California Academy of Science 2022 (AntWeb)), ants constitute a moderately diversified group amongst insects. They are, however, a major component of terrestrial ecosystems, being ecologically dominant in all strata and involved in key ecological functions (Lori et al. 2010). Within tropical forests, ants can make up to 25% of the total animal biomass (Hölldobler and Wilson 1990). Their study provided so far important insights into community ecology, global biodiversity patterns or impacts of global change, which make them of the keystone taxa for studying ecological patterns and processes (Lori et al. 2010).

French Guiana, the largest French overseas territory, is located in the Guiana shield on the north-eastern coast of South America. Covered with primary forest on more than 90% of its surface, it is part of the largest block of tropical forest worldwide, hosting a large diversity of species. As an example, the recent checklist of ants from French Guiana highlighted the presence of 659 valid species and subspecies from 84 genera and 12 subfamilies, representing ca. 10% of the ant diversity known in the Neotropical realm (Franco et al. 2019). Here, we provide a large dataset of ribosomal DNA sequences for ants of the region using a short DNA marker (16S rRNA gene, 250-300 bp in length) that can be used to describe and monitor ant biodiversity in the Neotropical area using Next Generation Sequencing methods.

Methods

Sampling

Geographic range

Ants were sampled from 2013 onwards in a diversity of sites covering most of the major forest habitats represented in French Guina (Guitet et al. 2015b) and topography: terra-firme (29 plots distributed in 9 sites), swamp (16 plots / 6 sites), white-sand forests (11 plots / 5 sites), transitional forests (1 plot) on slope of inselberg, costal savannah (10 plots / 5 sites), cloud forest (9 plots / 2 sites) and pastures (4 plots / 2 sites) (Fig. 1).

Collecting method

Sampling was performed following the Ants of Leaf Litter Protocol (Agosti and Alonso 2000). At each site, 0.12-ha plots (30 m × 40 m) were established in the different habitats locally represented. Within each plot, 20 sampling points were established on a grid, with a 10 m distance between each point. At each point, two sampling methods were used: pitfall traps and mini-winkler (Bestelmeyer et al. 2000). Pitfall traps were 6 cm diameter

containers placed in the ground with an opening at surface level, partially filled with a soap and salt water solution and left open for 72 h. At the same sampling point, 1 m^2 of leaf litter was also sifted and then placed in mini-winkler extractors for a period of 48 h (Bestelmeyer et al. 2000).

Sample processing

Specimens were preserved in 95% ethanol and then sorted to morphospecies in the lab. One individual of each morphospecies was then mounted for morphological identification to species using taxonomic resources available in the literature and the expertise of taxonomy specialists. Voucher specimens were deposited in the Laboratorio de Mirmecologia, Cocoa Research Centre CEPEC/CEPLAC (Itabuna, BA, Brazil) and at EcoFoG in Kourou.

Although the mitochondrial gene encoding the cytochrome *c* oxidase subunit 1 (COI) has been accepted has the consensus marker (Kress et al. 2015), its sequence length, i.e. about 650 bp, turned out to be problematic when using High-Throughput sequencing technology. Indeed, Illumina technology, the most used and accurate sequencing technology, provides reads of 100 to 500 bp. As an alternative, an informative region of 16S rRNA gene of 135-276 bp has been shown as a suitable alternative to COI for DNA barcoding in insects (Elbrecht et al. 2016). Moreover, the variability in the COI primer binding sites result in amplification biases that impair its use in metabarcoding studies (Deagle et al. 2014). The 16S fragment provides promising results at least in insect metabarcoding studies (Elbrecht et al. 2016, Marquina et al. 2018), but reference libraries are still underdeveloped. Accordingly, this short 16S fragment has been sequenced here as described below.

DNA extraction was performed from single leg or whole specimen for the smallest, with at least three specimens per species. Each extract was amplified by PCR with the 16S rRNA primer Ins16S_1 (Clarke et al. 2014) (TRRGACGAGAAGACCCTATA / TCTTAATCCAACATCGAGGTC), using the "HotShot" protocol (Truett et al. 2000) with the following cycles: 15 min at 95°C, 38 cycles of 95°C for 20 s (denaturation), 49°C for 30 s (hybridation) and 72°C for 30 s, (elongation) and a final extension at 72°C for 5 min for the six first runs and with the cycle: 15 min at 95°C, 40 cycles of 95°C for 30 s (denaturation), 50°C for 30 s (hybridisation) and 72°C for 30 s (elongation) and a final extension at 72°C for 10 min for the two last runs. Samples were multiplexed with tagged primers to identify sequences from each specimen. Products were verified and visualised by electrophoresis on 0.8% agarose gels. Sequences shorter than 100 bp were removed by purification from PCR reaction with the GeneClean Turbo Kit (MP Biomedicals, LLC, Sante Ana, CA., USA). Finally, amplicon sequencing was performed using Illumina Miseq technology (2 × 250 bp) by Fasteris (Plan-les-Ouates, Switzerland) or at the Genotoul platform (www.genotoul.fr).

Data processing

Sequence data (Suppl. material 1, Suppl. material 2, Suppl. material 3) were analysed using Obitools, Obitools3 (Boyer et al. 2015) and dada2 (Callahan et al. 2016) packages in R (R Core Team 2020). The two approaches provided complementary results despite their different strategy of data processing and assignation. In Obitools and Obitools3 (Boyer et al. 2015), paired-end read assembly, read demultiplexing and read dereplication were first performed. Then, low-quality sequences (i.e. shorter than expected - under 80 bp), singletons and sequences not assigned to samples were discarded. Chimera sequences were also excluded using the uchime3_denovo algorithm from usearch tools (Edgar 2016). Remaining sequences were assigned using the EMBL invertebrate database (Baker 2000) with the Obitools assignation process. In dada2 (Callahan et al. 2016), sequences were trimmed and demultiplexed using cutadapt (Martin 2011) and deML tools (Renaud and Schmidt 2017), respectively. Then, low-quality sequences were discarded and remaining sequences dereplicated. An error model was generated from data themselves and used for creating amplicon single variants (ASVs). Finally, chimeras were deleted using the “removeBimeraDenovo” function from dada2 and remaining sequences were identified using the 16 rRNA sequences of the EMBL invertebrate database (Baker 2000) with the RDP classifier algorithm implemented directly in dada2 (Wang et al. 2007). Finally, results from the two workflows were assembled, the most abundant sequence was kept for each sampled specimen and the molecular identification was compared with the morphological one. Only groups of similar sequences corresponding to identical morphological taxonomic assignation were conserved.

Quality checking

The quality of the sequences (Suppl. material 1) was checked using a taxonomic congruence approach. For each species, multiple specimens were sequenced and the corresponding sequences were expected to form a monophyletic group. Sequences were aligned using Muscle (Edgar 2004) and a distance tree was performed using the BIoNJ (Gascuel 1997) algorithm in phyml (Guindon et al. 2010). For species for which only a single specimen was available, we considered the sequence to be correct if it was placed in the correct genus and significantly different from the remaining species.

Taxonomy

Temporal coverage

Notes: 2013-present

Taxonomic coverage

This dataset (Suppl. material 1, Suppl. material 2) complements the GenBank library with Ants from French Guiana sequences. Most of the sequences are from species that have not been sequences so far using this maker or even sequenced at all. A total of 3931 sequences have been deposited, representing 344 species distributed in 57 genera. Most of the sequences (n = 3920, 99.7%) have been assigned at the species level and the remaining (n = 11) were at the genus level (i.e. close enough to sequence groups belonging to the same genus, but not close enough to a sequence group forming a species). Amongst the sequences assigned at the species level, 69% (i.e. 2698 sequences) have been attributed to fully described species, while the remaining (31%, 1222 sequences) represent morphospecies. On average, intraspecific species variation was 4.5% (Suppl. material 4) when calculating with the identity matrix obtained through a multiple alignment with clustalw (Sievers et al. 2011). New sequences will be added periodically to the dataset when available.

Data Resources

This Targeted Locus Study project has been deposited at DDBJ/EMBL/GenBank under the accession number KFFS00000000. The version described in this paper is the first version, KFFS01000000.

Resource 1

Download URL

https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=KFFS01

Resource identifier

KFFS01000001-KFFS01003931

Data format

Usage Rights

Acknowledgements

Financial support for this study was provided by Investissement d'Avenir grants of the Agence Nationale de la Recherche (CEBA: ANR- 10-LABX-25-01; DRIIHM: ANR-11-

LABX-0010; TULIP: ANR-10-LABX-41), by the Programme Convergence 2007-2013, Région Guyane from the European community (BREGA, 757/2014/SGAR/DE/BSF) and by the PO-FEDER 2014-2020, Région Guyane (BiNG, GY0007194 and BUG, GY0024253). We would like to thank Sébastien Cally and Anna Grandchamp who participated with specimen sequencing and data analysis. We thank the national park and natural reserve managers for allowing our research programme in the protected areas. Specimens from Itoupé and Mitaraka were collected in the core area of the Parc Amazonien de Guyane. The Itoupé expedition was organised and conducted in collaboration with the Parc Amazonien de Guyane. The Mitaraka expedition was part of the “Our Planet Reviewed” French Guiana-2015 initiative organised by the Muséum National d’Histoire Naturelle (Paris) and the NGO Pro-Natura International and funded by the European Regional Development Fund (ERDF), the Conseil Régional de Guyane, the Conseil Général de Guyane, the Direction de l'Environnement, de l'Aménagement et du Logement (DEAL) and by the Ministère de l'Éducation nationale, de l'Enseignement Supérieur et de la Recherche. Specimens from the Trinité area were collected in the Réserve Naturelle Nationale de La Trinité managed by the Office National des Forêts. The expedition was funded by the Réserve Naturelle Nationale de La Trinité and the DEAL Guyane. Data have been collected from access to genetic resources in French Guiana, that has come through a declarative process with non-commercial uses at the competent administrative authority, in accordance with article L.421-7 of the environmental code (Authorization number TREL1820249A/51; APA-973-1 and ABSCH-IRCC-FR-253854-1).

Conflicts of interest

The authors declare no conflicts of interests.

References

- Agosti D, Alonso LE (2000) The ALL protocol: A standard protocol for the collection of ground-dwelling ants. In: Agosti D, Majer JD, Alonso LE, Schultz TR (Eds) Ants: Standard methods for measuring and monitoring biodiversity. Smithsonian Institution Press, Washington and London, 280 pp. URL: http://antbase.org/databases/publications_files/publications_20330.htm [ISBN 1560988851, 9781560988854].
- Baker W (2000) The EMBL nucleotide sequence database. Nucleic Acids Research 28 (1): 19-23. https://doi.org/10.1093/nar/28.1.19
- Bestelmeyer BT, Agosti D, Alonso LE, et al. (2000) Field techniques for the study of ground-dwelling ants: An overview, description and evaluation. In: Agosti D, Majer JD, Alonso LE, Schultz TR (Eds) Ants: Standard methods for measuring and monitoring biodiversity. Smithsonian Institution Press, Washington and London, 280 pp. URL: http://antbase.org/databases/publications_files/publications_20330.htm [ISBN 1560988851, 9781560988854].

- Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E (2015) Obitools Unix-inspired software package for DNA metabarcoding. Molecular Ecology Resources 16 (1): 176-182. https://doi.org/10.1111/1755-0998.12428
- California Academy of Science (2022) AntWeb. https://www.antweb.org. Accessed on: 2022-4-15.
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13 (7): 581-583. https://doi.org/10.1038/nmeth.3869
- Clarke L, Soubrier J, Weyrich L, Cooper A (2014) Environmental metabarcodes for insects: In sillico PCR reveals potential for taxonomics bias. Molecular Ecology Resources 14 (6): 1160-1170. https://doi.org/10.1111/1755-0998.12265
- Deagle B, Jarman S, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochromecoxidase subunit I marker: not a perfect match. Biology Letters 10 (9). https://doi.org/10.1098/rsbl.2014.0562
- Edgar R (2016) UCHIME2: Improved chimera prediction for amplicon sequencing. bioRxiv https://doi.org/10.1101/074252
- Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5 (1). https://doi.org/10.1186/1471-2105-5-113
- Elbrecht V, Taberlet P, Dejean T, Valentini A, Usseglio-polatera P, Beisel J, Coissac E, Boyer F, Leese F (2016) Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. PeerJ https://doi.org/10.7287/peerj.preprints.1855v1
- Engel MS, Ceríaco LMP, Daniel GM, Dellapé PM, Löbl I, Marinov M, Reis RE, et al. (2021) The taxonomic impediment: A shortage of taxonomists, not the lack of technical approaches. Zoological Journal of the Linnean Society 193 (2): 381-387. https://doi.org/10.1093/zoolinnean/zlab072
- Fontaine B, Perrard A, Bouchet P (2012) 21 years shelf life discovery and description of new species. Current Biology 22 (22). https://doi.org/10.1016/j.cub.2012.10.029
- Franco W, Ladino N, Delabie JHC, Dejean A, Orivel J, Fichaux M, Groc S, Leponce M, Feitosa RM (2019) First checklist of the ants (Hymenoptera: Formicidae) of French Guiana. Zootaxa 4674 (5): 509-543. https://doi.org/10.11646/zootaxa.4674.5.2
- Gascuel O (1997) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution 14 (7): 685-695. https://doi.org/10.1093/oxfordjournals.molbev.a025808
- Guindon S, Dufayard J, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenesis: Assessing the performance of PhyML 3.0. Systematic Biology 59 (3): 307-321. https://doi.org/10.1093/sysbio/syq010
- Guitet S, Brunaux O, Granville JJd, Gonzalez S, C R (2015a) Catalogue des habitats forestiers de Guyane. DEAL Guyane. URL: https://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers15-09/010065207.pdf
- Guitet S, Pélissier R, Brunaux O, Jaouen G, Sabatier D (2015b) Geomorphological lanscape features explain floristic patterns in French Guiana rainforest. Biodiversity and Conservation 24 (5): 1215-1237. https://doi.org/10.1007/s10531-014-0854-8
- Hölldobler B, Wilson E (1990) The Ants. Springer-Verlag, Berlin, 732 pp. https://doi.org/10.1046/j.1420-9101.1992.5010169.x
- Hubert N, Hanner R (2015) DNA Barcoding, species delineation and taxonomiy: A historical perspective. DNA Barcodes 3 (1). https://doi.org/10.1515/dna-2015-0006

- Kress WJ, García-Robledo C, Uriarte M, Erickson D (2015) DNA barcodes for ecology, evolution and conservation. Trends in Ecology & Evolution 30 (1): 25-35. https://doi.org/10.1016/j.tree.2014.10.008
- Lori L, Catherine LP, Kirsti LA (2010) Ant ecology. Oxford University Press, New York, 420 pp. [ISBN 9780199544639] https://doi.org/10.1093/acprof:oso/9780199544639.001.0001
- Marquina D, Andersson A, Ronquist F (2018) New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. Molecular Ecology Resources 19 (1): 90-104. https://doi.org/10.1111/1755-0998.12942
- Martin M (2011) Cutadapt removes adapter sequences from high-throughout sequencing reads. EMBnet.journal 17 (1). https://doi.org/10.14806/ej.17.1.200
- R Core Team (2020) R: A language and environment for statistiacal computing. R foundation for statistiacal computing, Vienna, Austria . URL: https://www.R-project.org/
- Renaud G, Schmidt J (2017) deML: Maximum likelihood demultiplexing for NGS data. 1.13. URL: https://github.com/grenaud/deML
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology 7: 539. https://doi.org/10.1038/msb.2011.75
- Truett GE, Heeger P, Mynatt RL, Truett AA, Walker JA, Warman ML (2000) Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide ad tris (HotSHOT). BioTechniques 29 (1): 52-54. https://doi.org/10.2144/00291bm09
- Wang Q, Garrity G, Tiedje J, Cole J (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73 (16): 5261-5267. https://doi.org/10.1128/aem.00062-07

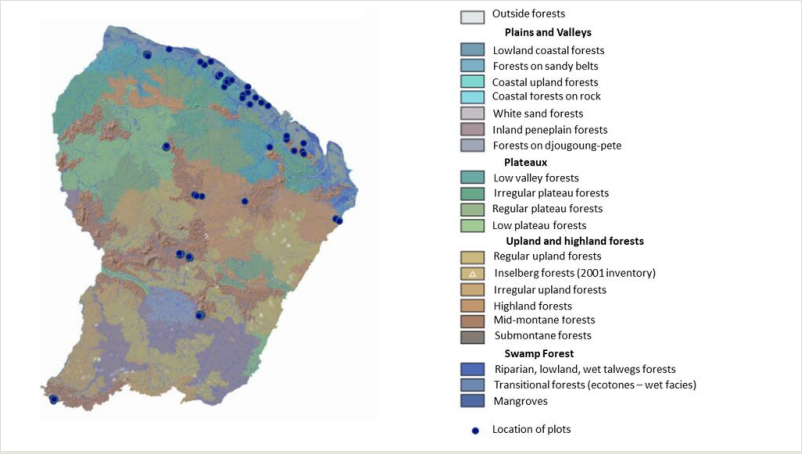

Figure 1.

Distribution of the sampling plots across French Guiana. Background colours represent the main forest habitats in the region and a topographic layer from a 30 m resolution SRTM radar image produced by NASA resolution *sensu* Guitet et al. (2015a).

Supplementary materials

Suppl. material 1: Non-aligned sequences dataset

Authors: Gaëtan Rongier
Data type: FASTA
Download file (1.02 MB)

Suppl. material 2: Aligned sequences dataset

Authors: G. Rongier, J. Orivel
Data type: FASTA
Download file (2.24 MB)

Suppl. material 3: Specimen-associated metadata

Authors: G. Rongier
Data type: Collection data
Download file (166.74 kb)

Suppl. material 4: Intraspecific variations in sequences

Authors: G. Rongier
Data type: % of variation in sequences at the intraspecific level
Download file (17.15 kb)