The emergence of DNA barcoding and metabarcoding opened new ways to study biological diversity, however, the completion of DNA barcode libraries is fundamental for such approaches to succeed. This dataset is a DNA barcode reference library (fragment of Cytochrome Oxidase I gene) for 2,190 specimens representing at least 540 species of shore fishes collected over 10 years at 154 sites across the four volcanic archipelagos of French Polynesia; the Austral, Gambier, Marquesas and Society Islands, a 5,000,000 km² area. At present, 65% of the known shore fish species of these archipelagoes possess a DNA barcode associated with preserved, photographed, tissue sampled and cataloged specimens, and extensive collection locality data. This dataset represents one of the most comprehensive DNA barcoding efforts for a vertebrate fauna to date. Considering the challenges associated with the conservation of coral reef fishes and the difficulties of accurately identifying species using morphological characters, this publicly available library is expected to be helpful for both authorities and academics in various fields.

Background & Summary
DNA barcoding aims to identify individuals to the species level by using a short and standardized portion of a gene as a species tag. This standardized procedure has revolutionized how biodiversity can be surveyed as the identification of a species then becomes independent of the level of taxonomic expertise of the collector, the life stage of the species or the state of conservation of the specimen. Due to its large spectrum of potential applications, DNA barcoding has been employed in a large array of scientific fields such as taxonomy, biogeography, biodiversity inventories and ecology; but see Hubert and Hanner for a review. In the genomic era, this approach has been successfully applied to the simultaneous identification of multiple samples (i.e.
metabarcoding approach), extending its applications to surveys of whole ecological communities11, but also monitoring species diet12,13, identifying the presence of specific species in a region14, or studying changes in the community through time by sampling environmental DNA15,16.

By design, DNA barcoding has proved to be fast and accurate, but its accuracy is highly dependent on the completeness of DNA barcode reference libraries. These libraries turn surveys of Operational Taxonomic Units (OTUs) into species surveys through the assignment of species names to OTUs17,18, hence giving meaning to data for ecologists, evolutionary biologists and stakeholders. Taxonomists increasingly provide DNA barcodes of new species they are describing; but thousands of species of shore fishes still lack this diagnostic molecular marker.

In the South Pacific, an early initiative led by the CRIOBE Laboratory was successfully carried out for French Polynesian coral reef fishes at the scale of one island, Moorea (Society Island)19. The fish fauna of Moorea’s waters is one of the best known of the region given the historical operation of research laboratories and long term surveys20,21, The Moorea project revealed a high level of cryptic diversity in Moorea’s fishes19 and motivated the CRIOBE Laboratory to extend this biodiversity survey of shore fishes to the remaining islands of French Polynesia. French Polynesia (FP) is a 5,000,000 km² region located between 7° and 27° South Latitude that constitutes a priority area for conducting a barcoding survey. This region is species rich due to its position at the junction of several biogeographic areas with varying levels of endemism. For example, the Marquesas Islands (northeastern FP) rank as the third highest region of endemism for coral reef fishes in the Indo-Pacific (13.7%)22. The Austral Islands (southwestern FP) and Gambier Islands (southeastern FP) host numerous southern subtropical endemic species23–25. Finally, the Society Islands (western FP) possess the highest species richness (877 species) and the highest number of widespread species in French Polynesia26.

Here, we present the result of a large-scale effort to DNA barcode the shore fishes in French Polynesia. Conducted between 2008 and 2014, a total of 154 sites were inventoried across these four archipelagoes. Islands of varying ages and topographies were visited ranging from low-lying atolls to high islands surrounded by a barrier reef, or solely fringing reefs. Furthermore, inventories were conducted across different habitats at each island (i.e. sand bank, coral reefs, rubble, rocky, etc.). In total, 2,190 specimens were identified, preserved, photographed, tissue sampled, DNA barcoded and cataloged with extensive metadata to build a library representing at least 540 species, 232 genera and 61 families of fishes (Fig. 1). Merged with previous sampling efforts at Moorea, a total of 3,131 specimens now possess a DNA barcode representing at least 645 nominal species for a coverage of approximately 65% of the known shore fish species diversity of these four archipelagoes. These biodiversity surveys have already resulted in the publication of updated species checklists22,26 and in the description of 17 new species27–34. This comprehensive library for French Polynesia shore fishes will certainly benefit a wide community of users with different interests, ranging from basic to applied science, and including fisheries management, functional ecology, taxonomy and conservation. Furthermore, many newly detected taxa for science are revealed here, along with complete collection data and DNA barcodes, which should facilitate their formal description as new species. While shedding new light on the species diversity of the Pacific region, this publicly available library is expected to fuel the development of DNA barcode libraries in the Pacific Ocean and to provide more accurate results for the growing number of studies using DNA metabarcoding in the Indo-West Pacific.
Methods

Sampling strategy. We explored a diversity of habitats across the four corners of French Polynesia with shallow and deep SCUBA dives (down to 50–55 m) for a total of 154 sampled sites (Fig. 2, Table 1). A total of 2,190 specimens, representing at least 540 species, 232 genera and 61 families (Fig. 3a) have been collected across four archipelagos representing the four corners of French Polynesia (FP), through six scientific expeditions: Marquesas Islands (1) in 2008 at Mohotani and (2) in 2011 at every island of the archipelago aboard the M.V. Braveheart (Clark Bank, Motu One, Hatutaa, Eiao, Motu Iti, Nuku-Hiva, Ua-Huka, Hiva-Oa, Tahuata, Fatu-Hiva; 52 sites), (3) in 2010 at Gambier Islands aboard the M.V. Claymore (Mangareva, Taravai, Akamaru, and all along the barrier reef; 53 sites), (4) at Austral Islands in 2013 aboard the Golden Shadow (Raivavae, Tubuai, Rurutu, Rimatara, Maria Islands; 25 sites), (5) at westernmost atolls of the Society Islands in 2014 aboard the M.V. Braveheart (Manuae and Maupiha’a; 20 sites). A sixth scientific expedition took place on Moorea’s deep reefs in 2008 (Society Islands) as a small scale scientific expedition that included the exploration and sampling of some of the deep reefs of Moorea (53 to 56 m depth; 4 sites) (Fig. 2).

Specimen collection. Specimens were captured using rotenone (powdered root of the Derris plant) and spear guns while SCUBA diving. These complementary sampling methods allowed us to sample both the cryptic and small fish fauna as well as the larger specimens of species not susceptible to rotenone collecting. Four individuals per species were collected on average. Fishes were sorted and identified onboard to the species level using identification keys and taxonomic references and representative specimens of all species collected were photographed in a fish photo tank to capture fresh color patterns, labeled and tissue sampled for genetic analyses (fin clip or muscle biopsies preserved in 96% ethanol). The photographed/sampled voucher specimens were preserved in 10% formalin (3.7% formaldehyde solution) and later transferred into 75% ethanol for permanent storage.

Table 1. Overview of the dataset. Number of specimens and species collected for each scientific expedition. Sampling effort expressed in number of sampling days and number of sites.

| BOLD project | Geographical location | No. of specimen collected | No. of species collected | Sampling effort (No. of sampling days/No. of sites) |
|--------------|-----------------------|---------------------------|--------------------------|----------------------------------------------------|
| AUSTR        | Austral Islands       | 560                       | 263                      | 12/25                                              |
| GAMBA        | Gambier Islands       | 705                       | 290                      | 18/53                                              |
| MARQ         | Marquesas Islands     | 386                       | 182                      | 18/41                                              |
| MOH          | Marquesas Islands     | 190                       | 107                      | 5/11                                               |
| MOOP         | Society Islands       | 42                        | 27                       | 4/4                                                |
| SCIL         | Society Islands       | 309                       | 213                      | 8/20                                               |

Fig. 2  Sampling localities across French Polynesia. The number of sampling sites and the number of specimens collected are displayed for each archipelago. Several sampling localities may be represented by a single dot due to the geographic scale of French Polynesia. Map data: Google, DigitalGlobe.
archival storage. Preserved voucher specimens and tissues were deposited and cataloged into the fish collection at the Museum Support Center, National Museum of Natural History, Smithsonian Institution, Suitland, Maryland, USA. Nomenclature follows Randall23 and we followed recent taxonomic changes using the California Academy of Sciences Online Eschmeyer’s Catalog of Fishes37.

DNA barcode sequencing. We extracted whole genomic DNA using QIAxtractor (QIAGEN, Crawley) and Autogen AutoGenPrep 965 according to manufacturer’s protocols. A 655 bp fragment of the cytochrome oxidase I gene (COI) was amplified using Fish COI primers FISHCOILBC (TCAACYAATCAYAAAGATATYGGCAC) and FISHCOIHBC (ACTTCYGGGTGRCCRAARAATCA) and Polymerase Chain Reaction (PCR) and Sanger sequencing protocols as in Weigt et al.38. PCR products were Sanger sequenced bidirectionally and run on an ABI3730XL in the Laboratories of Analytical Biology (National Museum of Natural History, Smithsonian Institution). Sequences were edited using Sequencer 5.4 (Gene Codes) and aligned with Clustal W as implemented in Barcode Of Life Datasystem (BOLD, http://www.boldsystems.org). Alignments were unambiguous with no indels or frameshift mutations. A total of 2,190 DNA barcodes have been generated.

Specimen identification. All morphological identifications were revised as needed after the specimens were deposited in the archival specimen collection to confirm initial identifications made in the field. Specimens of specific groups like Antennaridae, Bythitidae, Chlopsidae or Muraenidae were revised by additional taxonomist specialists (David Smith, John McCosker, Leslie W. Knapp, Werner Schwarzhans). After the morphological identification, we used the Taxon-ID Tree tool and Barcode Index Numbers (BIN) discordance tools as implemented in the Sequence Analysis module of BOLD to check every identification using the DNA barcodes generated. The Taxon-ID tool consists of the construction of a neighbor-joining (NJ) tree using K2P (Kimura 2 Parameter) distances by BOLD to provide a graphic representation of the species divergence39. The BIN discordance tool uses the Refined Single Linkage algorithm (RESL40) to provide a total number of OTUs.

Data Records
This library is composed of three main components: (1) voucher specimens archived in the national fish collection at the Smithsonian Institution (Washington, DC), which were photographed in the field, (2) complete collection data associated with each voucher specimen, and (3) DNA barcodes (Fig. 1).

All photographs, voucher collection numbers, DNA barcodes and collection data are publicly available in BOLD41 in the Container INDOF “Fish of French Polynesia” or by scientific expedition (“AUSTR”, “GAMBA”, “MARQ”, “MOH”, “MOOP” and “SCILL”) and in Figshare42. DNA barcodes have also been made available in GenBank, and have accessions KC56766143 to KC56766344, KC68499045, KC68499146, KU90570947 to KU90572748, KY57069849, KY57070350 to KY57070551, KY57070852, KY68354953, MH70784654 to MH70788155, MK56677456 to MK56715357, MK65696958 to MK65871359 and this database is accessible through the CRIOBE portal (http://fishbardb.criobe.pf).

The library fulfills the BARCODE data standard60,61 which requires: 1) Species name, 2) Voucher data, 3) Collection data, 4) Identifier of the specimen, 5) COI sequence of at least 500 bp, 6) PCR primers used to generate the amplicon, 7) Trace files. In BOLD, each record in a project represents a voucher specimen with its photographs, voucher collection numbers, associated sequences and extensive collection data related to (1) the Voucher: Sample ID, Field ID, Museum ID, Institution Storing; (2) the Taxonomy: Phylum, Class, Order,
Family, Subfamily, Genus, species, Identifier, Identifier E-mail, Taxonomy Notes; (3) Specimen Details: Sex, Reproduction, Life Stage, FAO Zone, Notes such as sizes of the specimens, Voucher Status, and (4) Collection Data: Collectors, Collection Date, Continent, Country/Ocean*, State/Province, Region, Sector, Exact Site, GPS Coordinates, Elevation, Depth, Depth Precision, GPS Source, and Collection Notes42.

Table 2. Potential cryptic species. Species with number of specimens collected displaying taxonomic paraphyly most likely representing undescribed cryptic species. Sample ID includes sampling location (AUST: Austra Islands, GAMB: Gambier Islands, MARQ and MOH: Marquesas Islands, SCIL and MOOP: Society Islands).
Acanthuridae | Acanthurus reversus | 0.08 | 0.15 | AUSTR453-13 | Acanthurus olivaceus | 0
Holocentridae | Myripristis carlsii | 0.28 | 0.62 | SCILL065-15 | Myripristis berndti | 0
Monacanthidae | Pervagor marginalis | 0.36 | 0.62 | SCILL083-15 | Pervagor aspersus | 0
Tetraodontidae | Canthigaster crusei | 0 | 0 | MOH030-16 | Canthigaster janthina | 0
Mullidae | Mulliodichthys microps | 0.52 | 0.52 | AUSTR089-13 | Mulliodichthys vanicolensis | 0.17
Pomacentridae | Chromis abrupta | 0 | 0 | SCILL209-15 | Chromis marginatigera | 0.31
Labridae | Coris marquesensis | 0 | 0 | SCILL040-15 | Coris gaimard | 0.46
Apogonidae | Ostorhinusca relativus | N/A | 0 | SCILL142-15 | Ostorhinusca angustatus | 0.93
Tetraodontidae | Canthigaster ruffiensis | 0.21 | 0.31 | MARQ456-12 | Canthigaster marquesensis | 1.1
Pomacentridae | Abudelfyus conformis | 0.15 | 0.15 | GAMBA444-12 | Abudelfyus sexiacutis | 1.24
Monacanthidae | Cantherhines nukuhwara | 0.15 | 0.31 | GAMBA711-12 | Cantherhines sandwichiensis | 1.4
Pomacentridae | Plectroglyphidodon sagmarius | 0.08 | 0.15 | AUSTR222-13 | Plectroglyphidodon imparipinnis | 1.56
Holocentridae | Sargocentron casimiliarum | 0.68 | 1.1 | SCILL104-15 | Sargocentron tere | 1.57
Acanthuridae | Zebrasoma rostratum | 0 | 0 | AUSTR376-13 | Zebrasoma scopas | 1.72
Apogonidae | Apogon marquesensis | 0.23 | 0.31 | GAMBA657-12 | Apogon susanae | 1.88
Chaetodontidae | Chaetodon flavirrostris | 0.08 | 0.15 | SCILL269-15 | Chaetodon lunula | 1.88
Chaetodontidae | Chaetodon banula | 0.1 | 0.15 | GAMBA555-12 | Chaetodon flavirrostris | 1.88

Table 3. Species displaying either incomplete lineage sorting or shallow inter-species divergence. Mean and Maximum intra-Species distances (Mean Intra-Sp and Max Intra-Sp), and Kimura 2 Parameter distances from Nearest Neighbour (NN).

and the very shallow genealogies involved (maximum K2P genetic distances lower than 2%), both incomplete lineage sorting and past introgressive hybridization might be responsible of the mixing of species genealogies in those 17 cases. In summary, 94% of the BINs match species identified using morphological characters, meaning that it was possible to successfully identify a species using DNA barcodes in 94% of the cases.

Usage Notes
This Barcode release dataset is freely available to use in barcoding or metabarcoding surveys for specimen identification. Several approaches can be considered:

1. directly downloading the sequences in fasta format, and working offline by merging this dataset with an ongoing barcoding project;
2. working online, through the BOLD website (registration is free), and merging the Container INDOF “Fish of French Polynesia” or parts of the scientific expeditions (Table 1) with an ongoing BOLD project;
3. through online identification tools, as data are indexed in both BOLD and Genbank databases. This library will be considered when any queries of molecular identification will be made through the identification engine of BOLD (http://www.boldsystems.org/index.php/IDS-OpenIdEngine) or the standard nucleotide Basic Local Alignment Search Tool (BLAST; https://blast.ncbi.nlm.nih.gov/). In the same manner, this dataset should also be indexed in the MIDORI database. Composed of both endemic and widespread species, this library is expected to benefit a large community from academics to authorities who use molecular data to monitor and survey biodiversity.

References
1. Hebert, P. D. N. & Gregory, T. R. The promise of DNA barcoding for taxonomy. Syst. Biol. 54, 852–859 (2005).
2. Garnett, S. T. & Christidis, L. Taxonomy anarchy hampers conservation. Nat. News 25 (2017).
3. Hubert, N., Delrieu-Trottin, E., Irisson, J. O., Meyer, C. & Planes, S. Identifying coral reef fish larvae through DNA barcoding: A test case with the families Acanthuridae and Holocentridae. Mol. Phylogenet. Evol. 55, 1195–1203 (2010).
4. Ko, H.-L. et al. Evaluating the accuracy of morphological identification of larval fishes by applying DNA barcoding. PLoS One 8, e53451 (2013).
5. Wong, E. H.-K. & Hanner, R. H. DNA barcoding detects market substitution in North American seafood. Food Res. Int. 41, 828–837 (2008).
6. Holmes, B. H., Steinkne, D. & Ward, R. D. Identification of shark and ray fins using DNA barcoding. Fish. Res 95, 280–288 (2009).
7. Riedel, A., Sagata, K., Suhardjono, Y. R., Tänzler, R. & Balke, M. Integrative taxonomy on the fast track: towards more sustainability in biodiversity research. Front. Zool. 10, 15 (2013).
8. Monaghan, M. T. et al. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Syst. Biol. 58, 298–311 (2009).
9. Tänzler, R., Sagata, K., Surbakti, S., Balke, M. & Riedel, A. DNA barcoding for community ecology–how to tackle a hyperdiverse, mostly undescribed Melanesian fauna. PLoS One 7, e28652 (2012).
10. Hubert, N. & Hanner, R. DNA barcoding, species delineation and taxonomy: a historical perspective. DNA barcodes 3, 44–58 (2015).
11. Leray, M. & Knowlton, N. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proc. Natl. Acad. Sci. 112, 2076–2081 (2015).
12. Leray, M., Boehm, J. T., Mills, S. C. & Meyer, C. P. Moorea BIOCODE barcode library as a tool for understanding predator-prey interactions: insights into the diet of common predatory coral reef fishes. Coral Reefs 31, 383–388 (2012).
13. Leray, M., Meyer, C. P. & Mills, S. C. Metabarcoding dietary analysis of coral dwelling predatory fish demonstrates the minor contribution of coral mutualists to their highly partitioned, generalist diet. PeerJ 3, e1047 (2015).
Acknowledgements

This work was financially supported by the French National Agency for Marine Protected Area in France (‘Pakaiai Te Moana’ expedition), the ANR IMODEL and Contrat de Projet Etat-Territoire in French Polynesia and the French Ministry for Environment, Sustainable Development and Transport (MEDDTL) (‘CORALSPOT’ expeditions), the Living Oceans Foundation (‘Australs’ expedition) and the LabEx CORAIL and the GOPS, (‘Scilly’ expedition) and the Gordon and Betty Moore Foundation (Mo’orea Biocode Project). Additional funding was provided by the IFRECOR in French Polynesia and the TOTAL Foundation. We are grateful to T. Froger, P. Mery and the Centre Plongée Marquises (Xavier (Pipapo) and Marie Curvat), for their field assistance in the Gambier, the Marquesas, the Australs and Scilly Islands along with the crew of the Claymore II, Braveheart and the Golden Shadow. We thank the Ministère de l’Environnement de Polynésie, the Délégation à la Recherche Polynésie, the Mairie of Nuku-Hiva, and the people of the Marquesas Islands for their kind and generous support of the project as we traveled throughout the islands. We thank Jerry Finan, Erika Wilbur, Shirleen Smith, Kris Murphy, David Smith and Sandra Raredon of the Division of Fishes (National Museum of Natural History) for assistance in preparations for the trip and processing specimens and Jeffrey Hunt and Kenneth Macdonald III and Meaghan Parker Forney of the Laboratories of Analytical Biology (Smithsonian Institution) for assistance in molecular analysis of samples. Finally, we thank the staff of the CRIOBE and particularly Yannick Chancerelle for logistical support in French Polynesia. Specimens were collected under the permit “Permanent agreement, Délégation à la Recherche, French Polynesia”. This publication has ISEM Number 2019-105 SUD.

Author Contributions

E.D.T. drafted the first manuscript, J.T.W., D.P., A.D. and N.H. provided extensive edits and comments. E.D.T., J.T.W., T.C., R.G., M.K., T.L.M., J.M., G.M.-T., VP., P.P., P.S., G.S., N.T., M.V. and S.P. collected fish specimens. E.D.T., J.T.W., D.P., A.D., N.H., J.V., B.E., C.M., L.W. and S.P. produced DNA barcodes and cleaned the database. R.G. and S.P. financed the scientific expeditions. C.M. and S.P. financed the sequencing. All authors read and approved the final manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2019